Computer vision and natural language processing for people with vision impairment

Daniela Massiceti

Pembroke College

University of Oxford

A dissertation submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Engineering Science

Trinity 2019

To my mom, who is not here
to see me reach the end of this journey

To my dad, who has never stopped believing in me

Abstract

Globally 253 million people live with severely impaired vision. They face extensive challenges in their day-to-day lives, from independently navigating and socialising, to fine-grained tasks like reading, and identifying and interacting with objects. Assistive tools have been developed to ease these challenges, ranging from the white cane to talking wearables, however, most remain simplistic, doing little in the way of parsing and understanding the dynamic, 3D world required for safe and easy interaction. Rapid advances in machine learning, computer vision and natural language processing, however, coupled with the miniaturisation of electronics and proliferation of mobile devices and wearables, are redefining the landscape of assistive technologies for people with vision impairment.

This thesis takes concrete steps toward the goal of data-driven assistive technologies by exploring methods for i) understanding visual scenes, ii) relaying this information to visually impaired (VI) users, and iii) evaluating models for relaying information through natural language. In the first direction, we develop a state-of-the-art weakly-supervised semantic segmentation method which segments objects using only classification labels. These labels can easily be collected from VI users as they interact with objects in the real-world. In the second, we develop two methods for relaying information about the environment. We do this by i) creating spatial audio soundscapes of 3D scenes, and ii) allowing users to directly ask questions about their visual environment, which the system then answers. We validate the first on human participants in virtual reality environments, and the second using quantitative metrics on experimental datasets. In the third direction we investigate a widely-used dataset for a sequential visual question-answering task and find that it contains exploitable biases which are exacerbated by poor evaluation metrics. We then propose a better method for evaluating, and quantifying performance on this task. In the future data-driven assistive technologies hold much promise for people with vision impairment. Efforts must, however, consider how well data-driven models port to real-world scenarios, where i) the incoming data will be considerably different from that in established visual perception datasets, and ii) the portability constraints of mobile devices are much more stringent.

Acknowledgements

Reaching the end of this journey has only been possible because of the support of so many.

To my supervisors, Professor Philip Torr and Dr Stephen Hicks, for taking me on as a student and guiding me through my doctoral research endeavours. To Phil, for his wisdom on whether ideas 'have legs', and his enthusiasm for new research problems. To Steve, for inspiring me to do research that matters. To Puneet Dokania and Siddharth Narayanaswamy, for their valuable guidance in each of our collaborations, and for teaching me not only how to define a problem, but how to investigate and write about it.

To my funders, without whom my graduate study would not have been possible. To the Skye Foundation and Dr Neville Passmore, for their support of South African students studying abroad, and to the University of Oxford's Clarendon Fund and Department of Engineering Science, both which have enabled me to pursue and share my research over the past years.

To Pembroke College and the Middle Common Room, which, especially at the start, were my home away from home in this beautiful and historic corner of the world. Through the people I have met, and conversations I have had, I am grateful for the ways in which each interaction has broadened my view of the world.

To all the committees and teams I have worked with over the years: Deep Learning Indaba, Pembroke MCR committee, Oxford Women in Computer Science, and Oxford Women in Engineering. Each one's dedication to a meaningful cause has been an inspiration to me.

To my dear Oxford friends who have been the brightness I needed through the many dark and rainy days. To Maria Laura Tognoli. From the many meals together, travels, lucky buttons, yoga in the lounge, cremini, and more, I am grateful to have shared our time here together. To Saumya Jetley, my fellow computer visionary. Thank you for lightening my days in the lab with chats, laughs, caring for our lab plants, and stories about family and friends. To Jonny Downing, for always checking-in, for caring, and for listening. To Cristiana Vagnoni, for our weekly breakfasts, and for being my fellow guilty feminist. To the Saffa crew: Yossi Singer, Margot Leger, and Matthew Davey. Our friendship holds such a special place in my heart, and I hope the future will bring us many more travel adventures. To my long-time friend, and fellow engineer and knitter, Bianca van Velze. Despite the distance, I am thankful that we can still share so much of our lives with each other. To Jessica Standish-White, from our long runs and banana bread pitstops at Zappi's, I could not be happier that Oxford brought

us together. To the many others who have been such special parts of my life in Oxford: Greg Dobrynin, William Fawcett, Juan Pablo Ruiz, Friederike Haberstroh, Sibyl Bertrand, and Miren Tamayo Elizalde. I cherish having met and gotten to know each of you.

To my family. To my dad, my biggest supporter, and always a voice of logic and reason in times of need. Thank you for supporting me in the many decisions I have made, including those which have led me further away from home. I could not be luckier to have a dad like you. To my mom, who showed me true dedication to a cause. From your perch up in heaven, I am sure you are doing your Egyptian dance moves in celebration. To my sister, Lara, whose love for family has been our family glue across the oceans. Thank you for putting life into perspective for me. To my brother, Marco, who is beginning his time as a student just as I am ending mine. You have so much to come in your bright future, and I hope that you can see it too. To Zia Giulie and Nonna Liliana. Thank you for taking care of Dad, Lara, and Marco during this time. I know it has not been easy for you with me being so far away.

Finally, to Samuel Wills, who, no matter the location or time zone, has never faltered in his belief in me throughout this journey. Thank you for being my sounding board and in-house consultant, for inspiring me with your passion for solving big problems, and for showing me the exciting parts of my own research when I couldn't see them myself. Of all the things that Oxford has given me, from the very first ice cream, you have by far been the best one.

Contents

List of Figures

List of Tables

Acronyms

1VD one-way visual dialogue.

2VD two-way visual dialogue.

AMD age-related macular degeneration.

AMT Amazon Mechanical Turk.

CCA canonical correlation analysis.

CE cross entropy.

CNN convolutional neural network.

CVAE conditional variational auto-encoder.

DB diabetic retinopathy.

DCG discounted cumulative gain.

DNN deep neural network.

ELBO evidence lower bound.

EM Expectation Maximisation.

ETA electronic travel aid.

FOV field of view.

GIS global information system.

GPS global positioning system.

HCI human-computer interaction.

IMU inertial measurement unit.

IOU intersection-over-union.

KL Kullback-Leibler.

ML machine learning.

NDCG normalised discounted cumulative gain.

NHS National Health Service.

NLP natural language processing.

OCR optical character recognition.

PDA portable digital assistant.

RNIB Royal National Institute of Blind People.

RNN recurrent neural network.

SLAM simultaneous localisation and mapping.

SOTA state-of-the-art.

SSD Sensory Substitution Device.

UK United Kingdom.

VAE variational auto-encoder.

VD visual dialogue.

VI visually impaired.

VQA visual question-answering.

VR virtual reality.

WHO World Health Organisation.

"*Blindness is a world. I've sought to show that that it's one of a number of human worlds.*"
John Hull, *Notes on Blindness*[1]

Introduction

Globally an estimated 1.3 billion people are living with some degree of visual impairment. Of this group, 36 million are considered blind, 217 million as moderately to severely, and 188.5 million as mildly visually impaired [314, 362]. In the United Kingdom (UK), just over 2 million people are living with sight loss, with the prediction that this number will double by 2050 [230, 239].

The consequences of sight loss are evident in the extensive challenges faced by visually impaired (VI) people in everyday life. These challenges range from difficulties with navigating, and socialising, to finer-grained tasks like reading, and finding, identifying, and interacting with objects, all of which are exacerbated in unfamiliar environments [130, 168, 183, 213, 313]. As evidence of the profound impact vision impairment can have on one's quality of life and productivity: in the UK, it is estimated that fewer than half of all registered blind people leave their house each day, and only 1 in 4 VI people of working age is in employment, compared with 4 in 5 non-disabled people [230, 307]. This, of course, introduces concomitant strains on the economic and welfare state of a country [230, 266, 283]. The National Health Service (NHS) estimates the total cost of eye health and sight in the UK to be £28 billion every year [266]. This accounts for the costs of preventing and treating eye conditions, as well as the indirect costs associated with lower employment and reduced well-being, including the increased risk of related health issues, like injuries due to falls [150, 184, 212, 266, 283, 330]. The main aim of this thesis is to ameliorate some of these profound individual and societal costs through the development of technology to assist VI people.

[1] www.notesonblindness.co.uk

While some causes of visual impairment can be prevented or treated, a large proportion of sight loss remains without a cure [94]. New treatments such as retinal prosthetics, optogenetics, and gene therapy offer hope for the future, but they are still at a research or early implementation stage and await evidence of real-life benefit to patients [93]. Assistive devices and aids have, therefore, come to be ubiquitously used by VI people to ease their day-to-day living. The Royal National Institute of Blind People (RNIB) estimates that nearly three quarters of the VI population in the UK make use of mobility aids and other accessibility tools for reading, writing, and computer usage [307]. By far the largest class of these aids employ either no or only very simple technology, and are limited in what help they can provide. The white cane, for example—a commonly used tool for orientation and mobility, and a global symbol for vision impairment—only conveys information about the user's proximal environment that makes direct contact with the cane.

The rapidly changing technological landscape, however, is opening the door to revolutionary changes in the way assistive devices may come to help VI people. Advances in machine learning, facilitated by greater compute power and data, have furthered machine performance across a range of perceptual vision and language tasks. In computer vision, impressive gains have been seen in image, video, and 3D scene understanding [61, 108, 131, 142, 240, 276, 281, 305, 320, 327], in natural language processing (NLP), in a host of language understanding tasks, including language modelling, question-answering, and speech recognition [17, 84, 136, 319, 337, 370], and in combinations of the two [21, 24, 342, 368, 374]. Alongside this, the proliferation of smart-phones and ubiquity of internet is providing a portable platform for deploying these data-driven technologies. There is thus a great opportunity for these developments to assist the VI community, and we are already seeing the beginnings of such [67, 74, 226, 249, 328].

This thesis, therefore, is positioned at the intersection of computer vision and NLP with the aim of developing assistive technologies for VI people. In what follows, Section 1.1 discusses vision impairment, and in broad strokes, the challenges facing VI people. Section 1.2 provides a brief treatment of the existing assistive solutions for the VI community (with Appendix A containing a more thorough coverage). Section 1.3 goes on to present the core research contributions of this thesis, contained in Chapters 2 to 5, which take steps toward developing data-driven assistive technologies for the VI community.

1.1 Vision impairment

The World Health Organisation (WHO) formally defines vision impairment as the loss of part or all of one's ability to see, with a categorisation into either distance or near vision impairment [251][2].

1.1.1 Causes

Globally the leading causes of vision impairment are cataracts, uncorrected refractive errors, trachoma, glaucoma, age-related macular degeneration (AMD), diabetic retinopathy (DB), and corneal opacity (Figure 1.1a). The prevalence of each of these conditions varies by region: cataracts and uncorrected refractive errors, for example, have higher incidence rates in low- to middle-income countries, while conditions like AMD, DB, and glaucoma, are more common in high-income countries (Figure 1.1b).

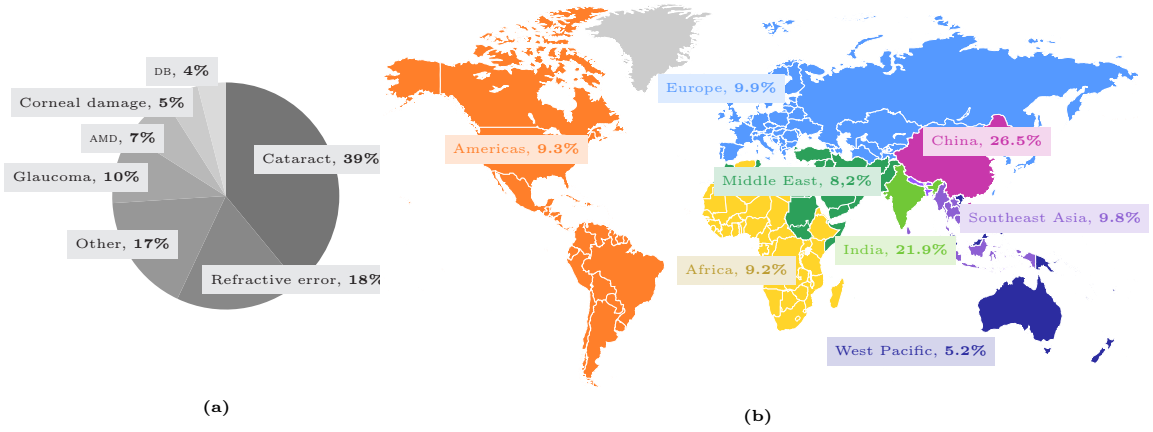


Figure 1.1 Global **(a)** causes and **(b)** prevalence of vision impairment. Source: World Health Organisation

Each eye condition is characterised by a distinctive field of view (FOV) profile. AMD, for example, causes damage to the macula, resulting in a blurring of the central vision, while glaucoma, caused by a build-up of fluid pressure in the eye, degrades the peripheral vision over time, resulting in a vignette-like visual effect.

[2]Distance vision impairment encompasses a visual acuity of worse than 6/12, considered mild vision impairment, all the way up to full blindness, characterised by a visual acuity of worse than 3/60. Near vision impairment accounts for a near visual acuity of worse than N6 or M.08 with existing correction.

1.1.2 Challenges of vision impairment

The challenges faced by VI people impact all facets of their life, from day-to-day activities and movement, employment prospects, the ability to socialise, and overall physical and mental well-being. Here, we discuss four areas that are the most challenging for VI people: orientation and mobility, object interaction, reading, and social integration. This list is by no means exhaustive and the severity of each challenge will additionally depend on eye condition, degree of functional sight loss, and age, amongst other factors.

Orientation & mobility

Independent navigation, composed of way-finding (finding and orienting oneself relative to landmarks) and locomotion (moving oneself from origin to destination efficiently and effectively), presents one of the biggest challenges for people with limited sight [104, 107, 197, 198]. This is because vision is able to deliver information at a high bandwidth which is required for navigating highly dynamic real-world environments [104, 262]. Without it, alternate senses and cues, which may not be as informative, must be used, and introduces heavy planning and cognitive loads on the VI user [104, 107, 200, 259]. Using public transport, for example, is cited as one of the biggest day-to-day challenges for VI people, with the RNIB estimating that 40% of VI people are unable to make all the journeys they want or need to make [307]. Inevitably, and because our living and public spaces are usually designed with the sighted population in mind [107], VI individuals are at an increased risk of mobility-related accidents, injuries, and falls [150, 184, 212, 283, 330]. Over a three-month period, the RNIB found that two-thirds of VI people were reported to collide with an obstacle on the pavement [307], including street furniture, wheelie bins, parked cars, advertising boards, temporary obstructions such as building works, trees, and bicycles. Even in familiar environments, the variable presence of obstacles, changes in walking surfaces, and drop-offs can present significant mobility hazards [308], and even experienced non-sighted navigators may still experience difficulties [122].

Orientation and mobility training can be undertaken to ensure safer navigation without vision [357]. This training typically involves learning to i) use mobility aids like the white cane, ii) be accompanied by a sighted human companion, and iii) use orienting techniques like

trailing, squaring-off, and landmark establishment [263]. It also involves learning to maximise information from other available senses, most commonly, sound and touch[3].

Navigating independently and safely from one point to another facilitates many other tasks, including gaining education, seeking employment and socialising, giving rise to many other knock-on difficulties for VI individuals.

Object interaction

A large part of our interaction with the world is with objects: in the morning, for breakfast, we find the necessary crockery and cutlery, select the cereal we like, and ensure the milk is not out-of-date. About our workday, we may read and respond to e-mails via a computer or mobile phone. In the evening, we may read a book to relax, or do the laundry.

Such day-to-day activities present challenges for VI people [130], requiring not only the ability to locate the objects in question, but also interact with them in a fine-grained manner—reading small text on a label, or determining if a device is powered on, for example. The navigational aids in common use—the white cane, or a guide dog—offer only very little of the detail needed to accomplish these tasks. Illustrating these difficulties, in a study conducted by the RNIB in 2015, nearly half of the VI people surveyed require help around the house, including support for tasks like preparing meals, personal care such as washing and dressing, and moving around the house [307]. Difficulties also arise in operating home appliances and heating controls, stemming from interfaces not being well-designed for VI users [38, 126]. Even something as simple as a flat-screen display, a trend in modern household appliances, introduces difficulties for VI people who otherwise rely on tactile feedback provided by buttons [126].

Reading & learning

The ability to read is central to enabling this fine-grained interaction, and indeed indispensable for acquiring information more generally, whether it is from carton labels, street signs, device screens, or books. There are a growing number of services to facilitate easier reading and information access, including magnification tools, Braille, optical character recognition (OCR) software, text-to-speech engines, and screen readers[4] [23, 26, 97, 98, 99, 246, 250, 273, 279,

[3] There exists a rich literature in sensory substitution—employing intact senses to obtain information about the world in place of a damaged sense—for VI persons [34, 318, 361].

[4] Platform-specific software which relays screen content, typically via speech.

287, 303, 315, 317]. Despite this, not all written information may be accessible to a VI person. In fact, the RNIB estimate that half of the VI community regularly require assistance with reading written information [307]. In accessing printed information, difficulties are reported with instructions on medication boxes and food packaging, information on voting ballots[5], and information from banks and healthcare providers [307]. In accessing digital information, while screen readers have been widely adopted by the VI community, information is now often presented as highly formatted or integrated with multimedia content which may go beyond screen reader capabilities. The heterogeneity of digital platforms too, whether website, mobile app, or other, may present further challenges for a VI person navigating the digital space. Without a focus on accessible design, which may be expensive to develop, both printed and digital information may not be accessible by a VI person.

Reading and, closely associated, writing, are crucial tools to help people learn. This can go on to adversely affect employment: only 1 in 4 registered VI people of working age in the UK are employed [307]. While worth considering for VI adults, it is for VI *children* where the impacts of impaired vision on development and learning are most severe if left unaided [167, 353]. In the UK, 25,000 children under the age 16 are registered as blind or partially sighted [235]. The majority of these children are in mainstream eduction [353], with an increasing number without access to the specialist support needed for optimal learning and development [167]. Furthermore, learning materials and exams are not consistently made available in suitable formats [95]. Programming, for example, an increasingly relevant skill for the future of work, has been largely omitted from the curricula of VI students because coding interfaces are inaccessible, amongst other factors [53, 123], with only recent efforts to address this [236].

Social integration

As well as interacting with objects, we also need to engage with other people, for example, by identifying the faces of friends and family, recognising gestures used to convey information, and socialising. Limited vision makes these activities difficult, and hence may affect a VI person's ability to engage in social events and physical activities, meet new people, and develop personal relationships. Supporting this, VI people often report feeling isolated or cut-off from other people and things around them [307], with symptoms of clinical depression

[5] 1 in 5 VI people report being unable to vote in secret—projecting to around 70,000 people across the UK!

and lower well-being more frequently observed in the VI population compared to the sighted population [65, 243, 270]. This is particularly the case for those who have recently lost their sight, or a part of it [325]. Feeling well is also influenced by a sense of being socially accepted [270]. Perceived negative attitudes from the general public because of their sight loss, as reported by over one-third of VI people in the UK [307], may therefore also contribute to poorer mental health in VI people.

1.2 Assistive technologies for vision impairment

A number of assistive tools have been developed to ease the above-discussed extensive challenges faced by the VI community. Beginning with very simple solutions like the white cane and the guide dog, a host of technology-based aids have subsequently been developed to extend the range of assistance possible. These solutions can broadly be grouped into three categories [105, 294, 307]: i) long-range assistive tools, for example, those that enable easier way-finding and orientation within a global environment, ii) medium-range assistive tools, for example, those that aid mobility, and avoiding and identifying obstacles in the user's proximity, and iii) short-range assistive tools, for example, those that facilitate fine-grained interaction with objects, and access to information. An extended review of aids within each of these categories is provided in Appendix A, with a brief sampling discussed below.

The white cane, a non-technological aid, is a simple mechanical device which allows the VI user to detect the presence of, and sometimes identify, obstacles in their path. The guide dog extends the navigational functions of the white cane by additionally assisting with coarse way-finding, judging height and width constraints along a route, and reducing veering, a challenge when traversing open areas and crossing streets. Guide dogs can also offer companionship to VI people. Both of these solutions, however, are limited by the *amount* of information about the user's environment they can convey [299, 357]. This, however, does not make them poor solutions. In fact, *because* it is simple, economical, and reliable, the white cane remains a popular choice for navigational assistance in the VI community [332, 357].

Completing day-to-day tasks, however, requires a wealth of higher-order information, most of which is unavailable to VI people. Furthermore, if we look beyond simply *completing* tasks,

day-to-day life involves, in parts, discovery and exploration—being able to read a poster advertising a music concert, or admire the architecture of a nearby building, for example. The challenge of developing assistive devices for the VI community, therefore, should not only be to facilitate independence, but also offer a wholesome sight-like experience. Many technology-based assistive tools, therefore, have been developed with these goals in mind, and we discuss these in the grouping of long, medium- and short-range aids below (Table 1.1).

Long-range	Medium-range	Short-range
▶*VictorReaderTrek*, £690 [316]	♦*Kay Sonic Torch* [R] [164]	★*ZoomText*, £540 [99]
▶*SeeingEye*, £54/year [119]	♦*Bat K Sonar Cane* [R] [204]	★*MAGic*, £461 [97]
▶*Talking Signs* [72]	♦*Sonic Guide* [R] [165]	★*IrisVision*, £2,495 [149]
▶*BlindSquare*, £39 [231]	♦*TriSensor* [R] [85]	★*NuEyes*, £5,995 [246]
▶*Nearby Explorer*, £61 [16]	♦*vOICe*, £0 [224]	★*eSight*, £8,995 [86]
♦*Autour*, £0 [182]	♦*EyeMusic*, £0 [1]	★*OXSIGHT Crystal*, £6,000 [252]
♦*MS Soundscape*, £0 [227]	♦*Sonic Pathfinder* [R] [134]	▶*VoiceOver*, £0 [26]
♦*SWAN* [R] [359]	♦*Capelle* [R] [54]	▶*TalkBack*, £0 [23]
◀*BeltMap* [42]	◀*UltraCane*, £635 [312]	▶*JAWS*, £0 [98]
◀*NavBelt* [R] [47]	◀*MiniGuide*, £300 [102]	▶*KNFB Reader*, £77 [287]
	◀*EyeCane* [R] [211]	▶*Giraffe Reader*, £32 [279]
	◀*GuideCane* [R] [48]	▶*Orcam Reader 2*,£2,450 [250]
	◀*CyArm* [R] [8]	▶**MS SeeingAI**, £0 [226]
	◀*TVSS* [R] [34]	▶**TapTapSee**, £0 [67]
	◀*BrainPort*, £4000 [35]	▶**iDentifi**, £0 [328]
	◀*Vest* [R] [244]	▶**Orcam MyEye 2**, £3,250 [249]
	◀*FingerSight* [R] [137]	▶**CyberEyez**, £1,418 [74]
		▶*Google Home*, £89 [111]
		▶*Amazon Echo*, £90 [11]
		▶*Be My Eyes*, £0 [39]
		▶*Aira*, £914/year [7]

Table 1.1 A sampling of long-, medium-, and short-range VI assistive technologies (and their prices). [R] indicates a non-commercialised research exploration. **Bold** indicates real-time scene understanding applied to image/video stream, excluding OCR. Colour indicates feedback type: ▶voice description, ♦spatial audio, ◀electro/vibro-tactle, ★visual.

Long-range tools are dominated by beacon-based navigational aids [16, 42, 47, 117, 119, 148, 182, 227, 231, 316, 359]. These aids employ global positioning system (GPS) tracking and landmark databases to guide the VI user toward or along a route of pre-established beacons. Navigational guidance is usually provided via audio cues, like voice descriptions [16, 117, 119, 148, 231, 316] and spatialised sound [182, 227, 359], or haptic cues, like vibrations [42, 47]. Because they are GPS-based, these aids are primarily limited to outdoor navigation, and are able to localise the user with reference to the beacons with only a coarse degree of accuracy. Furthermore, they rely on the availability of databases with (relevant) tagged landmarks.

Medium-range tools primarily focus on conveying information about the user's proximal environment for the purposes of mobility. This includes the white cane, the guide dog, and their virtual extensions [8, 48, 85, 102, 127, 164, 165, 204, 211, 312], so called electronic travel aids (ETAs), which employ sonar or infrared technologies to detect obstacles further afield. Since the primary function of these aids is to simply detect obstacles in the path of the user, they are only able to capture a slice of any environment. In addition, ETAs are often based on sonar or infrared, both line-of-sight technologies, and hence their pulse reflections are prone to being blocked or distorted, for example by a passing person, or an opening door. As a result, these aids are insufficient for safely navigating in all situations. Efforts have thus been made to widen the environmental slice by capturing full images of the environment and converting them to other sensory representations [1, 34, 35, 54, 224, 244]. However, uptake of these developments by the VI community has been low [105, 210, 289] due to the inherent challenges in extracting mobility-relevant information from an image, intuitively mapping that information to another sense, and ensuring real-time performance [105].

Short-range aids assist with a selection of close-range tasks, including reading, and interacting with objects and people [23, 26, 86, 97, 98, 99, 149, 246, 250, 252, 279, 287, 293]. Historically Braille has been used by VI people to access written information. However, the increase in digital content, and the effectiveness of screen readers and text-to-speech engines, has reduced dependence on the tactile alphabet [355]. Screen readers [23, 26, 98], perhaps the most commonly used of the short-range aids [307], have made computer usage more accessible by relaying screen content and layout through synthesised speech or refreshable Braille displays. Other tools magnify text, screens, and scenes [86, 97, 99, 149, 246], or read printed text aloud [149, 246, 250, 279, 287], with both hand-held and wearable options available.

Despite recent advances, few VI assistive devices parse and understand scenes in real-time, which is essential for handling the dynamic and visual nature of our world. The beacon-based navigational services [117, 119, 148, 316], for example, use GPS to localise and guide a user relative to static landmarks, doing no real-time scene understanding. This may often be the most crucial for safe navigation—for example, identifying a moving bus, a red traffic light, or a temporarily open man-hole. Commercially available devices which transform 2D images

to alternate representations[6] simply employ deterministic mapping functions based on pixel (x, y) locations and intensities, without identifying what is actually *in* the image [1, 54, 224]. Screen readers simply operate by attaching verbal tags to elements on the screen, which are then read out as a user hovers or clicks on them [23, 26, 98][7]. Of those devices which *do* perform some form of scene understanding, most employ only very rudimentary methods, for example, OCR to read out printed text [250, 287], or simple edge and contrast detection to enhance high frequency regions of the scene [86, 252].

The design choices made in these existing technologies have largely been governed by the available hardware and software. Meeting the computational requirements for capturing, processing, and relaying information to users in real-time was difficult, and better functionality came at the cost of size, portability, aesthetics, and battery life [105]. In addition, hand-crafted features and small data hindered the perceptual abilities of computer vision models, which reduced their functionality and robustness in real-world environments.

The exponential developments seen in machine learning, computer vision and NLP in recent years [61, 84, 281, 305, 319, 320, 327, 337, 342, 370, 374], along with the proliferation of data and compute power, is paving the way for a new generation of data-driven assistive technologies for VI people. If these technologies can capture and understand a user's visual environment, tailored to their individualised preferences and the current task, then it will be possible for them to offer around-the-clock, generalised assistance. The miniaturisation of electronics, improvements in on-board sensors, and expansion of cell coverage and internet speeds, complement these advances by allowing mobile devices like smart-phones to serve as platforms for deploying these technologies.

As an example of a possible tool, consider an app-based assistant which interprets a VI user's visual environment, streamed from a wearable or hand-held device, and delivers relevant information about the scene in a form that is natural for the user to understand. The visual representation can be obtained from an image classification, object detection, or semantic segmentation pipeline [60, 61, 131, 142, 238, 281, 305, 320], and integrated with a map of the local environment constructed using real-time simultaneous localisation and mapping

[6] Live demos can be found at `www.seeingwithsound.com` and `http://youtu.be/jVBp2nDmg7E?t=2m32s`

[7] Screen readers are nowadays default on most mobile devices and laptops—a simple experiment on my own computer ended in cacophony!

(SLAM) [189, 240, 241, 274, 352]. The representation could yet further be augmented by fusing information from other on-board sensors, like a GPS or inertial measurement unit (IMU), or even online information, like building layouts, shop operating hours, and live public transport timetables. When delivering this information, the assistant could interact with the user in a similar way to Apple's *Siri* [25] or Amazon's *Alexa* [10], through natural language descriptions, instructions, and answers to the user's questions about their environment. The assistant could equally deliver information via spatial sound or haptic cues, similar to existing assistive tools [1, 182, 224, 227], or, if the user is partially sighted, by projecting enhanced visual effects on a wearable's augmented reality display [86, 252].

Taking a step toward this ultimate ideal, assistive tools which are based on data-driven methods for scene understanding are already commercially available. Perhaps most well known is Microsoft's *SeeingAI* [226], a camera-based mobile app that is able to i) recognise people nearby including their emotions, ii) describe the user's surroundings, iii) describe perceived colours, and iv) identify currency bills, along with reading text aloud with standard OCR-based methods. Other mobile apps like *TapTapSee* [67] and *iDentifi* [328] perform object recognition on an image or video segment, delivering audible descriptions of visual content, including details like brand and colour (Figure 1.2). Wearables with similar capabilities are also available [74, 249], although typically not at the affordable cost of a mobile app. In addition, advances in NLP and speech processing, have spurred the development of home voice assistants likes Google's *Home* [111], and Amazon's *Echo* [11]. With the ability to control appliances around the house, purchase supplies, play the news, podcasts, and audio-books, and manage calendars on behalf of the user, these assistants, as a rather fortuitous by-product, may help with some of the in-house challenges faced by VI people [44]. While a step in the right direction, camera-based mobile apps offer only prescriptive sets of features, and home assistants are solely voice-based, doing no visual processing of the scene. This, therefore, restricts their usefulness to only specific types of tasks and scenarios faced by VI people.

To address these limitations and provide complete visual-based assistance across diverse scenarios requires the help of a human. Meeting this need, mobile video-calling apps *Be My Eyes* [39] and *Aira* [7] call on *virtual* humans and connect sighted call-centre helpers with VI users via a live video connection. Facilitated by the ubiquity of smart-phones and

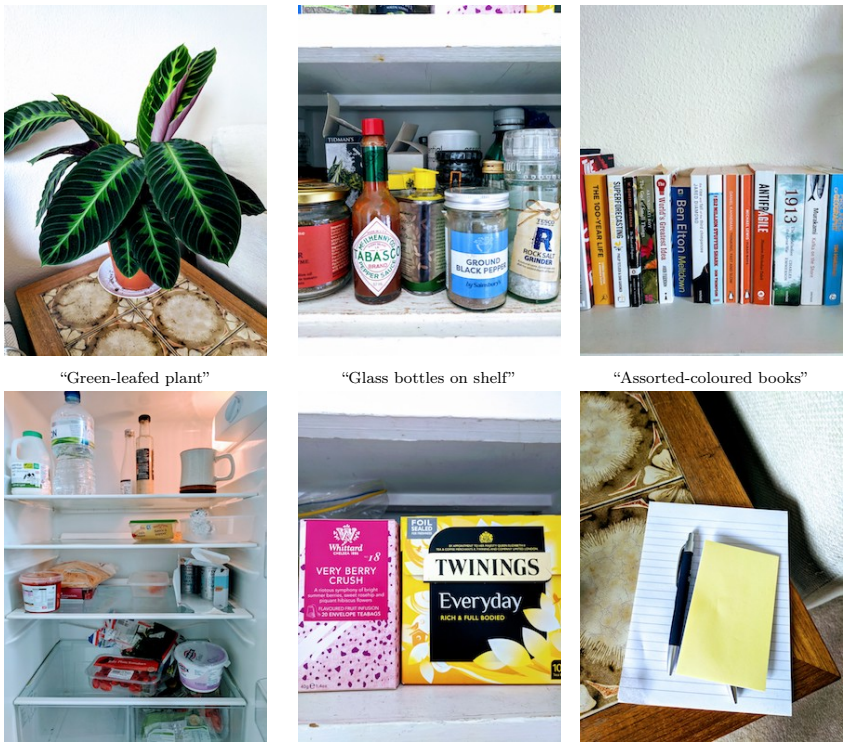


Figure 1.2 Verbal descriptions from the *TapTapSee* app [67]. When multiple objects are present, the descriptions fail to differentiate between, or provide specific detail about, each of the objects.

internet access, assistance is real-time and can be provided across a range of tasks, including navigation, sorting mail, identifying medication, reading labels, locating an empty seat or the elevator, and colour-matching clothing. These services, however, rely on human helpers who may not always be available, may not speak the VI user's language, and may greatly vary in the quality of assistance they deliver if not given proper training. Furthermore, using the service requires a permanent (and fast) internet connection, which may not be available in all scenarios, for example, on the underground. With rapid developments in machine learning facilitating perceptual advances in machine vision and language understanding, it is becoming increasingly possible that these human assistants may be replaced with data-driven models with similar capabilities, hence circumventing many of these issues.

These steps toward data-driven assistive technologies for VI people are promising, however, work remains to be done. Most of the current devices offer only basic scene and object recognition under specific settings, with human assistants required for all others. It is promising that advances in computer vision and natural language/speech processing will allow more complex scenes to be understood, and relevant information to be more efficiently relayed, thereby extending the range of assistance these devices can offer the VI community.

1.3 Research contributions

This thesis aims to enable fully autonomous, data-driven assistive tools which are able to understand visual and linguistic information, and deliver useful assistance to a VI user as a sighted human assistant would be able to do. With the problem statement and relevant background framed in Section 1.1 and 1.2, the primary contributions of this thesis are in:

1. Understanding visual scenes (Chapter 2),
2. Relaying information contained within scenes to VI users (Chapters 3 and 4), and
3. Evaluating the effectiveness of relay methods (Chapter 5).

1.3.1 Understanding visual scenes

The first challenge in developing VI assistive technologies is to understand the visual scene. In recent years, the computer vision community has made spectacular advances in scene understanding from images and videos [2, 61, 108, 131, 142, 240, 276, 281, 285, 305, 320, 327, 347, 374]. The most widely adopted approach has been through understanding objects within the scene, both in 2D and 3D [2, 60, 61, 108, 131, 142, 238, 281, 305, 320, 374], although others have also focussed on understanding scene structure [128, 189, 240, 241, 274, 341, 352, 372]. Object-centric representations are indeed natural because most of our interactions, including those of VI people, are with objects. Supporting this, Brady et al. [51] analyses the types of questions asked by VI people and finds a majority about objects: object identification (*'What is this?'*, *'Is this a Coke or Pepsi?'*, *'What bank note is this?'*), object description (*'What colour are these trousers?'*, *'Is the washing machine on?'*) and reading text on objects (*'What is the expiry date on these tomatoes?'*, *'What temperature is the thermostat set to?'*).

An object-centric understanding of the world is therefore important for assistive technologies and has largely been derived by asking *where is it?* and *what is it?* with reference to objects in a scene. Starting with the first question, much work has focussed on localising objects as 2D bounding boxes or masks in images and videos [131, 153, 281, 285, 347]. Although useful, 2D object localisations are limited in what they can convey about a 3D scene of 3D

objects. Consider a VI person navigating in a public space: it will be necessary to know where obstacles are relative to the user (e.g. radial distance and angle) and how they are oriented (e.g. facing away or toward the user). Much work, therefore, has focussed on estimating the 3D position and 3D pose of objects [50, 253, 366, 367]. Beyond navigating and avoiding obstacles, having objects' locations and poses may also help VI people to find objects (for example, their house keys amongst a clutter) and to interact with objects in a more fine-grained manner (for example, orienting their chair appropriately in a social group).

The second question of *what is it?* aims to identify the class or category of objects within a scene. This can be done at multiple levels of granularity: at a frame level, referred to as classification [2, 131, 142, 305, 320], a bounding box level [131, 281, 285], as part of object detection, or a pixel level [60, 61, 153, 238, 347], referred to as semantic segmentation. Each level of granularity can provide useful information to a VI person: a classification pipeline could identify specific items in a grocery store, while semantic segmentation masks could visually highlight objects/faces in an augmented reality wearable headset for partially-sighted people [86, 252]. At each of these resolutions, both class and instance labels [28, 133, 290] may be valuable for making inter- and intra-class distinctions, respectively. A class label could identify a T-shirt or food tin, while an instance label could isolate a *specific* T-shirt ("my yellow Nike T-shirt") or food tin ("Tesco baked beans")—common day-to-day tasks for VI people [51]. Ultimately, a scene representation which incorporates both positional and semantic information about objects will be essential for inferring higher-order spatial and functional relationships between objects and helping VI people complete more complex tasks.

Learning to recognise objects relies on the availability of labelled data—an object detector, for example, requires bounding box annotations, while a semantic segmentation pipeline needs class labels down to a pixel-level. In many cases, labelled datasets for these tasks already exist and can be used for learning. The time-varying nature of our environments, however, requires that models can integrate their existing knowledge for specific tasks with that learned from incoming (unlabelled) data streamed as the user goes about the real-world. To do this, labels for some[8] or all of the new data will be needed, collected either by remote annotators or the users themselves. A human-in-the-loop labeller is, in particular, attractive: human-centric

[8] Referred to as a semi-supervised setting in which only a portion of the full dataset is labelled.

labels may enable more personalised solutions [232] as well as facilitate faster (on-device) learning of new concepts. If the labeller is visually impaired, however, by virtue of their limited sight, the labels they provide may be noisy, and not exactly matched to the task at hand. The user may also be constrained to labelling mechanisms which are hands-free or "hands-light", for example verbal or short text inputs. As an example, consider a VI user purchases three food items and wishes to train an object localisation system that can guide them to each of the items in their pantry using audio cues. The simplest label the user could provide for each item is its classification label. With access to only these, the system must then learn to recognise and importantly *localise* the objects, a task for which it does not have explicit annotations. Building models that can learn to understand scenes and recognise objects with weak labels[9] that can be collected in an easy and user-friendly manner by VI users, is therefore important for the development of assistive devices that work in the real-world.

Toward this goal, in Chapter 2, we develop a weakly-supervised method for semantic segmentation which uses only classification labels. This differs from the traditional fully-supervised setting in which class labels for all pixels are available. Treating the segmentation as a latent variable, we propose an Expectation Maximisation (EM)- and curriculum learning-based algorithm to infer a per-pixel class distribution for all pixels in a given image. First, using a combination of attention and saliency maps as coarse ground-truth segmentations for simple, single-object images, we train an initial segmentation model. This serves to initialise a segmentation model which we then train on complex, multi-object images using an EM approach. Since only classification labels are available, within each EM iteration, we make the latent per-pixel posterior distribution (computed in the E-step) more robust by i) constraining the model to only place probability mass on the object classes *present* in the image, as specified by its classification label, and ii) linearly combining this constrained posterior with a delta distribution on the current iteration's most-likely label for the pixel. These steps help to guard against incorrect predictions in the current EM iteration derailing subsequent iterations. We also incorporate an approximate intersection-over-union (IOU) loss term [6, 68, 245] which optimises the model for the commonly used IOU evaluation metric. With this, our algorithm achieves state-of-the-art (SOTA) results on *PASCAL VOC 2012* [89], improving on methods which use classification label supervision by 4%.

[9] Labels for a necessarily lighter-weight task than the task of interest.

1.3.2 Relaying scene information

Learning meaningful visual representations of scenes, however, is only half the task of developing an assistive device. The second step is to *relay* the information contained within a given representation in a concise, correct, useful, and intuitive way. Existing assistive devices (see Appendix A) have employed a variety of creative methods for delivering information, ranging from an electro-tactile pin-array placed on the tongue [35], to producing natural language descriptions of an image [67, 226, 328], to employing spatial audio cues to highlight nearby objects in the user's environment [165, 182, 227].

This thesis, as part of its second contribution, explores ways to relay information through:

1. Spatial audio soundscapes encoding the presence and location of objects in novel virtual reality (VR) environments (Chapter 3), and
2. Natural language interactions between the system and the user, in the form of question-answer dialogues (Chapter 4).

Spatial audio soundscapes

Relaying information to VI people often requires appealing to other intact senses. Motivated by the increasingly better understood cross-modal plasticity of the brain [12, 13, 14, 113, 225, 343] and the extensive literature in Sensory Substitution Devices (SSDs) for VI people [1, 33, 34, 35, 47, 54, 137, 161, 224, 244, 284], common feedback forms include tactile cues, for example heat or vibrations, and audio cues, for example spatial sound or voice descriptions.

Most existing SSDs, however, either encode 2D images of environments [1, 54, 224], or single slices of 3D environments [48, 102, 164, 165, 204, 211, 312]. This is limiting for independent navigation and interaction within real-world 3D environments. Promisingly, however, there has been a recent proliferation in portable devices that can rapidly scan and reconstruct real-world scenes [189, 240, 341, 372]. Together these hint at the possibility of relaying information about the structure and semantics of 3D environments in the real-world.

Chapter 3 develops two novel methods for conveying visual information through spatial audio representations of 3D virtual reality environments. Our choice of audio over other

forms of feedback is motivated by the fact that hearing, like seeing, is able to deliver spatial information [202] which is required for our task of interest: navigation. The first method converts distances between the user and objects in the environment into audible echoes, similar to echolocation which is a navigational technique used by some VI people [173, 175, 322]. The second attaches continuous humming sounds to objects, and modulates their volume based on the user's distance to them. Importantly, in both methods, objects' positions on the azimuthal plane are spatially mapped using stereo sound sources, and distance to objects can be estimated by using the delay of the echoes or volume of the hums. To assess these representations, we task human participants to use audio to navigate and avoid obstacles in the VR environments using an untethered VR headset. Based on a bank of quantitative metrics to assess navigational performance and strategy, we find that participants are able to successfully navigate after only a brief training period (≤ 6 trials). This suggests that the audio soundscapes are useful and easy to learn for navigational tasks in 3D environments.

In addition to the soundscapes, the work makes a novel contribution in the way it evaluates human navigation. Existing methods rely on keyboard- and joystick-based set-ups on a computer. Instead, we test participants physically walking through VR environments, allowing for proprioceptive feedback to be incorporated. In addition to studies of mobility and sensory substitution, this has potential applications in spatial cognition studies where proprioceptive information is relevant; for instance, spatial learning and memory [221, 298] and how they may be affected by clinical conditions such as Alzheimer's Disease [73, 101] or depression [114].

Natural language interactions

Along with passive cues conveying information about the structure of an environment, people also actively probe and query their environments to extract the information they require. Language is the natural choice for doing this because of its expressiveness, efficiency, and natural intuitions which guide our use of it [116]. Services like *Be My Eyes* [39] and *Aira* [7] already use human assistants to help VI people through natural language interactions. It is the hope that one day these services may be fully autonomous. It is, therefore, important to build models which can exchange questions and answers with a human about a visual input.

There has been extensive work in visual question-answering (VQA) in recent years in both curating data [4, 5, 24, 77, 79, 80, 115, 121, 160, 278, 385] and modelling [5, 20, 22, 143, 158, 205, 373]. Most of these approaches, however, focus on answering single questions. While useful, autonomous assistants will likely need to engage in *multi-step* exchanges of questions and answers (and even general conversation) with VI users. In addition, most of the current approaches to both single- and multi-step VQA, frame the problem discriminatively: given the image, question, and a set of answers, simply select the correct one from the set [22, 24, 77, 100, 156, 205, 242, 373, 375]. These models, however, are not able to *generate* answers—the very goal of a VQA system—and additionally place the constraint that a set of answers be available at test time. A second class of model, therefore, employs a generative approach, with the common method being to train a recurrent neural network (RNN) and token-wise sample its learned likelihood to generate an answer [77, 78, 206, 319, 364]. The deterministic encoding of a RNN's inputs, however, results in it yielding homogeneous answers without a computationally intensive post-processing beam search [280, 319]. This differs from the way humans easily phrase sentences with the same meaning in a multitude of ways.

Chapter 4, therefore, proposes a fully generative approach to visual dialogue (VD)—answering a *sequence* of questions based on an image. We employ a variational auto-encoder (VAE) [171, 310] to learn distributions over answer embeddings, conditioned on the image, its caption (a short textual description), and the previous dialogue history. This enables us to sample *sets* of answers which are diverse, without the need for beam search, and relevant, given the distribution's conditional nature. Motivated by the time and compute cost of training recurrent models, and the recent successes of convolutional neural networks (CNNs) in language generation and prediction tasks [141, 162, 267], our approach is additionally fully convolutional, learning sentence-wise rather than token-wise conditional distributions.

The second contribution of Chapter 4 incorporates the intuition that an assistant should not just answer questions asked about an image, but should also be able to *ask* questions itself, to clarify or request pieces of information. As an example, if a VI user asks *'What is in front of me?'*, then an assistant may ask *'Near or far?'* in order to provide a more informative answer. With the exception of a few works [152, 233, 237], the larger focus has been on one-way question-answering models that solely *answer* questions about images [5, 20, 22, 24,

77, 100, 143, 158, 205, 206, 364, 373]. We, therefore, formulate a two-way question-answering task, and propose a generalised version of our conditional VAE which is able to generate a sequence of both questions *and* answers about a given image.

On the one-way question-answering task, under the established evaluation protocol [77], our method improves the performance of SOTA methods, and also has the ability to generate multiple correct answers without the need for complex post-processing. On the two-way question-answering task, to our knowledge, we establish the first baseline, and introduce two novel metrics for evaluating the question generation performance of distributional models.

1.3.3 Evaluating the effectiveness of relay methods

Correctly answering a visually-grounded question requires a clear understanding of the visual stimulus, for example, the locations of objects, their categories, attributes, functional affordances, and spatial relationships with other objects. It may also require higher-order capabilities like being able to make logical inferences, count, compare, and draw on common-sense knowledge from an external source. Progress toward autonomous systems with these abilities has, however, been hampered by strong biases in many widely used and community-accepted VQA and VD datasets [24, 77, 79]. Specifically, datasets have highly skewed language distributions, with a specific few words, questions and answers occurring frequently, and these only being loosely grounded in the visual component [3, 5, 18, 115, 217, 378]. The consequences of this are VQA/VD models which i) answer not by reasoning, but by exploiting linguistic correlations, and ii) ignore the visual stimulus [24, 151, 383]. Solutions have been proposed to mitigate these issues, primarily by rebalancing the datasets [4, 5, 115, 378]. Still, the limited role of the visual component across a number of VQA and VD studies is concerning for the development of tools to help VI users understand their environment.

A second factor delaying progress in visual question-answering is the challenge inherent in evaluating language [159]. Image classification and segmentation tasks are evaluated with simple binary decisions of correctness. In contrast, the scope and subjectivity in how a sentence can be expressed blurs the notion of correctness, and its assessment. Because of the time and cost of turning to human evaluators, especially at scale, it has thus been the focus of many to develop robust *automatic* evaluation metrics for language tasks [19, 186, 191, 255, 302].

Following this direction, Chapter 5 investigates the extent to which *VisDial* [77], a ubiquitously used VD dataset, including in Chapter 4, is subject to the above concerns. We establish an embarrassingly simple baseline employing canonical correlation analysis (CCA) between just questions and answers, and show that it is able to achieve near SOTA performance on one of the dataset's established evaluation measures. We delve into the reasons for this surprising finding, and reveal the limitations of the dataset's current evaluation. In this paradigm, models trained for the VD task are evaluated by ranking a given set of candidate answers, of which the ground-truth is one, for a given question and image. This set-up ignores the answer actually generated by the model, and additionally dismisses the fact that multiple answers in the candidate set are, by construction, equally feasible, making rank-based measures uninformative (up to a limit). These factors together lead to perverse scenarios, as we show, in which the performance of a model on the VD task is not truly reflected by its score.

In light of these findings, this thesis's final contribution is to develop a revised evaluation paradigm for the *VisDial* dataset to mitigate the issues exposed by our analysis. Our proposed paradigm forgoes the assumption that only a single answer is considered correct, and instead draws on existing metrics in the NLP literature [186, 255, 302] to measure the consensus between an answer generated by the model, and a reference set of equally-feasible answers. To construct these reference sets at scale, we apply a semi-supervised method based on correlation to a small subset of human-annotated examples, which we then scale across the entire dataset. We verify that the reference sets obtained via weak supervision are sufficiently high quality, by measuring i) their overlap with those annotated by humans, and ii) their ability to improve performance on the downstream task of VD. In the face of the limitations imposed by the dataset itself, and the experimental design choices made in its curation, we hope for this revised paradigm to be adopted by the community when exploring the *VisDial* dataset in the future. The expanded version of the dataset, including reference set annotations, will be made publicly available as a baseline for future evaluation and model development.

1.3.4 Summary

The main contributions of this thesis are in developing methods for i) learning visual scene representations with low-cost, easy-to-acquire labels; ii) relaying information contained in these representations in forms that are interpretable and accessible to VI people; and iii) more robustly evaluating relay methods, specifically via natural language interactions. Each is motivated by the development of data-driven, autonomous assistive technologies that will enable VI people to more easily navigate and interact with the world.

Of course, for these technologies to be used, they must be deployed on real-world devices. It is, therefore, essential to consider the practical implications of this alongside model development. Beyond model performance, important factors will include model size, speed, and robustness to real-world inputs which may differ from those in existing datasets. Models should additionally be able to learn flexibly, leveraging existing knowledge to understand new concepts which the user encounters in varied and dynamic environments.

1.4 Structure of thesis

Chapters 2 to 5 each contain a paper which has been peer-reviewed and accepted for publication in a conference or journal. The papers have been left unmodified from their published forms, with the exception of formatting changes. The structure of the thesis is as follows:

Chapter 1 Introduction

Chapter 2: Bottom-Up Top-Down Cues for Weakly-Supervised Semantic Segmentation

Q. Hou*, **D. Massiceti***, P.K. Dokania, Y. Wei, M-M. Cheng, and P.H.S. Torr

In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, 2017

Chapter 3: Stereosonic Vision: Exploring Visual-to-Auditory Sensory Substitution Mappings in an Immersive Virtual Reality Navigation Paradigm

D. Massiceti, S.L. Hicks, and J.J. van Rheede

Public Library of Science (PLOS ONE), 13(7):e0199389, 2018

Chapter 4: FlipDial: A Generative Model for Two-Way Visual Dialogue

D. Massiceti, N. Siddharth, P.K. Dokania, and P.H.S. Torr

In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018 [oral]

Chapter 5: On the Evaluation of Visual Dialogue (under review)

D. Massiceti, V. Kulharia, P.K. Dokania, N. Siddharth, and P.H.S. Torr

Chapter 6 Discussion

Publications not included in this thesis

- Visual Dialogue without Vision or Dialogue[10]

  D. Massiceti*, P.K. Dokania*, N. Siddharth* and P.H.S. Torr

  In *Advances in Neural Information Processing Systems (NeurIPS) Workshop on Correcting and Critiquing Trends in Machine Learning*, 2018

- Random Forests versus Neural Networks—What's Best for Camera Localization?[11]

  D. Massiceti, A. Krull, E. Brachmann, C. Rother, and P.H.S. Torr

  In *International Conference on Robotics and Automation (ICRA)*, 2017

[10] Chapter 5 extends and therefore supersedes this publication
[11] Camera localisation is only loosely related to the topic of this thesis, and hence not included

Chapter 2

Bottom-Up Top-Down Cues for Weakly-Supervised Semantic Segmentation

Qibin Hou[1*] Daniela Massiceti[2*] Puneet K. Dokania[2] Yunchao Wei[3]
Ming-Ming Cheng[1] Philip H. S. Torr[2]
[1]Nankai University, China [2]University of Oxford, UK [3]National University of Singapore
([*]contributed equally)

Abstract

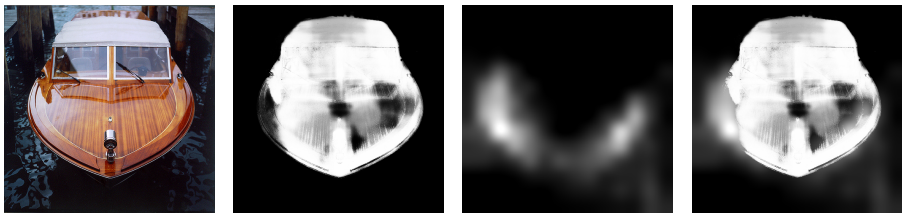
We consider the task of learning a classifier for semantic segmentation using weak supervision in the form of image labels which specify the objects present in the image. Our method uses deep convolutional neural networks (CNNs) and adopts an Expectation Maximisation (EM)-based approach. We focus on the following three aspects of the EM algorithm: (i) the initialisation; (ii) the latent posterior estimation (E-step) and (iii) the parameter update (M-step). We show that saliency and attention maps, bottom-up and top-down cues respectively, of simple images provide highly reliable cues for learning an initialisation of the EM algorithm. Given the weak labels and this initialisation, we learn to segment these simple images, before progressing onto more complex ones. In the E-step, we estimate the latent posterior class-wise distribution per pixel, and in the subsequent M-step, we minimise the combination of the standard *softmax* loss and the Kullback-Leibler (KL) divergence between this distribution and the likelihood given by the CNN. This combination is more robust to wrong predictions made in the E-step, which is likely given the use of only image-level labels. Extensive experiments and discussions show that our method is simple and outperforms the state-of-the-art method with a margin of 3.7% and 3.9% on the *PASCAL VOC 2012* [89] validation and test sets, respectively, thus setting new state-of-the-art results.

Published in the *Proceedings of the 2017 International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*[1]

[1]https://bit.ly/2ECeUab

2.1 Introduction

Semantic segmentation performance has rapidly advanced with the use of convolutional neural networks (CNNs) [56, 60, 196, 382]. The performance of CNNs, however, is largely dependent on the availability of a large corpus of annotated training data, which is both cost- and time-intensive to acquire. The pixel-level annotation of an image takes on average 4 minutes [40]. This is likely a conservative estimate given that it is based on the COCO dataset [193] in which ground-truths are obtained by annotating polygon corners rather than pixels directly. In response, recent work has focussed on weakly-supervised semantic segmentation [40, 176, 254, 261, 269, 277, 350]. These works differ from the fully-supervised case in that rather than having pixel-level ground-truth segmentations, a lower degree of supervision is provided. For example, image-level labels [176, 254, 261, 269], bounding boxes [254], and points and scribbles [40, 192, 369].



(a) Input **(b)** Saliency **(c)** Attention **(d)** Combined

Figure 2.1 Combining bottom-up (*saliency*) and top-down (*attention*) cues for simple images. Both cues complement each other by putting high probability mass on regions missed by the other.

In our work, we address the semantic segmentation task using only image-level labels. These labels specify the object categories present in the image. Our motivation for this is two-fold: i) the annotation of an image with the 20 PASCAL VOC object classes is estimated to take 20 seconds, which is at least 12 times faster than a pixel-level annotation and is also scalable; ii) images can easily be downloaded from the Internet with their image labels or tags, thus providing a rich and virtually infinite source of training data. The method we adopt, similar to [254], employs the Expectation Maximisation (EM) algorithm [82, 223]. We focus on the three key steps of an EM-based approach: i) initialisation; ii) latent posterior estimation (E-step); and iii) parameter update (M-step). The following addresses each of these points.

We provide an informed initialisation to the EM algorithm as follows. We first train a network to segment *simple* images with one object category (from a subset of *ImageNet* [83]) using an *approximate* per-pixel class distribution obtained using the combination of class-agnostic *saliency* maps [139] and class-specific *attention* maps [377]. Note that obtaining saliency and attention maps does not require pixel-level semantic segmentation. We use this trained model to initialise our EM algorithm in order to learn to segment complex images. Intuitively, we first learn to segment simple images and then move towards the complex ones, similar to the work of [350]. In more detail, given a simple image, the saliency map *finds* the object (Figure 2.1b)—this is a class-agnostic "bottom-up" cue. To complement this, once provided with the class present in the image, the attention map (Figure 2.1c) gives the "top-down" class-specific regions in the image. Since both saliency and attention maps are tasked to find the same object, their combination is more powerful than if either one is used in isolation (Figure 2.1d). The combined probability map is then used as the per-pixel class distribution for training an initialisation model for the semantic segmentation task, and provides the initialisation parameters for the follow-up E- and M-steps of the EM algorithm. Note that this process is in contrast to [254] where the initialisation model is trained for the image classification task on the same *ImageNet* dataset. To our surprise, experimentally we found that this initialisation model, which is trained on just *ImageNet* images under a weakly-supervised setting (with *no* images from *PASCAL VOC 2012*) outperforms all of the current state-of-the-art (SOTA) algorithms for the weakly-supervised semantic segmentation task on the *PASCAL VOC 2012* dataset. Note that the existing algorithms are significantly more complex and most of them rely on higher degrees of supervision such as bounding boxes, points/squiggles and superpixels. This indicates the importance of learning from simple images before delving into more complex ones. With the trained initialisation model, we then incorporate *PASCAL VOC 2012* images (with multiple objects) for the E- and M-steps of our EM-based algorithm.

In the E-step, we obtain the latent posterior probability distribution by constraining (or regularising) the CNN likelihood using a prior based on the weak image labels. This reduces false positives by redistributing the probability masses (which are initially over the 20 object categories) over only the categories present in the image and the background. In the M-step, the parameter update, we then minimise a combination of the standard *softmax* loss (where the ground-truth is assumed to be a Dirac delta distribution) and the KL divergence [180]

between the latent posterior distribution (obtained using the E-step) and the likelihood given by the CNN. This makes the approach more robust to difficult classes and incorrect labels, which are likely under the weakly-supervised setting. In addition to this, to obtain better CNN parameters, we add a probabilistic approximation of the Intersection-over-Union (IoU) [6, 68, 245] to the above loss function. This optimises the model specifically for the metric on which, at test time, it is evaluated. With this approach we obtain SOTA results in the weakly-supervised semantic segmentation task on the *PASCAL VOC 2012* dataset [89].

2.2 Related work

Work in weakly-supervised semantic segmentation has explored varying levels of supervision including combinations of image labels [176, 254, 261, 350], annotated points [40], squiggles [192, 369] and bounding boxes [254]. Papandreou et al. [254] employ an EM-based approach with supervision from image labels and bounding boxes. Their method iterates between inferring a latent segmentation (E-step) and optimising the parameters of a segmentation network (M-step) by treating the inferred latents as the ground-truth segmentation. Similarly, [350] train an initial network using saliency maps, following which a more powerful network is trained using the output of the initial network. The MIL frameworks of [269] and [260] use fully convolutional networks to learn pixel-level semantic segmentations from only image labels. The image labels, however, provide no information about the position of the objects in an image. To address this, localisation cues can be incorporated [269, 277]. These can be obtained from bottom-up proposal generation methods (for example, multi-scale combinatorial grouping (MCG) [27]), or saliency [350] and attention [377] mechanisms. The work of [349] uses saliency maps and iteratively erases areas of the image, from most to least salient, thereby forcing the network to learn increasingly discriminative features for segmentation from image labels. Localisation cues can also be obtained directly through point/squiggle annotations [40, 192, 369].

Our method is most similar to the EM-based approach of [254]. We use saliency and attention maps to learn a network for a simplified semantic segmentation task which provides a better initialisation for the EM algorithm. This is in contrast to [254] where a network trained for a classification task is used as initialisation. Also different from [254] where the latent posterior

is approximated by a Dirac delta function (which we argue is too harsh of a constraint in a weakly-supervised setting), we instead propose to use the combination of the true posterior distribution and the Dirac delta function to learn the parameters.

2.3 Semantic segmentation

Consider an image I consisting of a set of pixels $\{y_1, \cdots, y_n\}$ where each pixel represents a random variable taking on a value from a discrete semantic label set $\mathcal{L} = \{l_0, l_1, \cdots, l_c\}$, where c is the number of classes (l_0 for the background). Under this setting, a semantic segmentation is defined as the assignment of all pixels to their corresponding semantic labels, denoted as $\mathbf{y}$.

CNNs are extensively used to model the class-conditional likelihood for this task. Specifically, assuming each random variable to be independent, a CNN models the likelihood function as $P(\mathbf{y}|I;\theta) = \prod_{m=1}^{n} p_\theta(y_m|I;\theta)$ where $p_\theta(y_m = l|I;\theta)$ is the *softmax* probability (or the marginal) of assigning label l to the m^{th} pixel. The *softmax* probability is obtained by applying the *softmax*[2] function to the CNN outputs $f(y_m|I;\theta)$ such that $p_\theta(y_m = l|I;\theta) \propto \exp(f(y_m = l|I;\theta))$. Given a training dataset $\mathcal{S} = \{I_i, \mathbf{y}_i\}_{i=1}^{N}$, where I_i and $\mathbf{y}_i$ represent the i^{th} image and its corresponding ground-truth semantic segmentation, the log-likelihood is maximised by minimising the cross-entropy loss function using the back-propagation algorithm to obtain the optimal θ. At test time, for a given image, the learned θ is used to obtain the *softmax* probabilities for each pixel. These probabilities are either post-processed or used directly to assign semantic labels to each pixel.

2.4 Weakly-supervised semantic segmentation

To find the optimal θ for the semantic segmentation task, we need a dataset with ground-truth pixel-level semantic labels. Obtaining this, however, is highly time-consuming and expensive: for a given image, annotating its pixel-wise segmentation (for 20 object classes) takes nearly 240 seconds, or 4 minutes [40]. This is highly non-scalable to higher numbers of images and classes. Motivated by this, we use an EM [82, 223] approach for weakly-supervised semantic

[2]The *softmax* function is defined as $\sigma(f_k) = \frac{e^{f_k}}{\sum_{j=0}^{c} e^{f_j}}$

segmentation using only image-level labels. Image-level labels tag the object classes present in an image and are over 10 times faster to obtain that pixel-level annotations. Let us denote a semantic label set as $Z = \mathcal{L} \setminus l_0$, and a weak dataset as $\mathcal{D} = \{I_i, \mathbf{z}_i\}_{i=1}^{N}$ where I_i is the i^{th} image and $\mathbf{z}_i \subseteq Z$ is the image labels corresponding to the objects present in the i^{th} image. The task is to learn an optimal θ using $\mathcal{D}$.

2.4.1 Expectation Maximisation algorithm

Similar to [254], we treat the unknown semantic segmentation $\mathbf{y}$ as the latent variable. Our probabilistic graphical model is of the following form (Figure 2.2):

$$P(I, \mathbf{y}, \mathbf{z}; \theta) = P(I)P(\mathbf{y}|I, \mathbf{z}; \theta)P(\mathbf{z}), \tag{2.4.1}$$

Briefly, to learn θ while maximising the above joint probability distribution, the three major steps of an EM algorithm are: i) initialise the parameters θ_t; ii) E-step: compute the expected complete-data log-likelihood $F(\theta; \theta_t)$; and iii) M-step: update θ by maximising $F(\theta; \theta_t)$. In what follows, we first address how to obtain a good initialisation θ_t in order to avoid poor local maxima and then describe the method for optimising parameters (E- and M-steps) for a given θ_t.

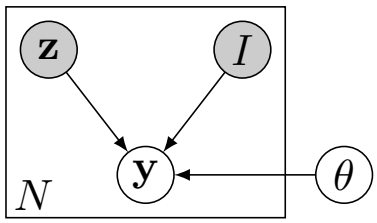


Figure 2.2 The graphical model. I is the image. $\mathbf{z}$ is the set of objects present in the image. $\mathbf{y}$ is the latent variable (semantic segmentation). θ is the set of parameters.

2.4.2 Initialisation using bottom-up top-down cues

It is well known that if the log-likelihood has several maxima or saddle points, an EM-based approach is highly susceptible to finding local maxima. In such cases, a good initialisation is crucial [363]. We argue that instead of initialising the algorithm with parameters learned for the classification task using the *ImageNet* dataset, as is done by most SOTA methods irrespective of their nature, it is much more effective to initialise with parameters learned

for solving an easier version of the task at hand—semantic segmentation in our case. In the following we show how to use *ImageNet* images with image-level labels to learn parameters for the weakly-supervised semantic segmentation task. These learned parameters will be used to initialise the EM algorithm.

Let us denote $\mathcal{D}(I)$ as the subset of images from the *ImageNet* dataset containing objects of the categories we are interested in (details in Section 2.5). Dataset $\mathcal{D}(I)$ contains simple images, which have mainly centred and clutter-free single objects. This is unlike the challenging *PASCAL VOC 2012* dataset [89]. To train a model for the semantic segmentation task using $\mathcal{D}(I)$, and hence obtain θ_t, we need pixel-level semantic labels which are not available in the weakly-supervised setting. To circumvent this, for each simple image in $\mathcal{D}(I)$ we generate a class-agnostic saliency map [64, 139] (bottom-up cue) and a class-specific attention map [377] (top-down cue) to *construct* the probability distribution over labels for each pixel. Intuitively, a saliency map gives the probability of each pixel belonging to *any* foreground class, and an attention map gives the probability of it belonging to a particular object class. Combining these two maps allow us to estimate a reasonably accurate probability distribution over object classes for each pixel in the image (see Figure 2.1).

We formalise this in Algorithm 1: for a given simple image $I \in \mathcal{D}(I)$ and its corresponding image label $z \in Z$, we combine the attention and saliency values per pixel to obtain M. $M(m) \in [0, 1]$ denotes the probability of the m^{th} pixel being the z^{th} object category. Similarly, $1 - M(m)$ denotes the probability of it being the background. The combination function $h(\cdot, \cdot)$ in Algorithm 1 is a user-defined function which combines the saliency and the attention maps. In this work we employ the *max* function which takes the union of the two maps (Figure 2.1).

To construct the per-pixel class distribution, let us define the distribution for the m^{th} pixel in image I as δ_m^I. Thus, $\delta_m^I \in [0, 1]^{|\mathcal{L}|}$ is the class distribution, where $\delta_m^I(z) = M(m)$ is the probability of the pixel belonging to object category z, $\delta_m^I(0) = 1 - M(m)$ is the probability of the pixel being the background, and all other entries are zero. Given δ_m^I for each pixel, we find θ_t by optimising a CNN with the per-pixel cross-entropy loss between δ_m^I and $p(y_m|I, \theta)$—$\sum_{k \in \mathcal{L}} \delta_m^I(k) \log p(k|I; \theta)$—where $p(y_m|I, \theta)$ is the CNN likelihood.

Algorithm 1 Approximate ground-truth distribution

Input: Image I with one object category; Image-level label z

1: $M = zeros(n)$, n is the number of pixels.
2: $\mathbf{s} \leftarrow SaliencyMap(I)$ [139]
3: $\mathbf{a} \leftarrow AttentionMap(I, z)$ [377]
4: **for** each pixel $m \in I$ **do**
5: $\quad M(m) = h(\mathbf{s}(m), \mathbf{a}(m))$
6: **end for**

Output: M

Using the probability value $M(m)$ directly rather than a Dirac delta distribution makes our method more robust to noisy attention/saliency maps. This can also be viewed as a way to mine class-specific noise-free pixels, and is motivated by [40] where humans annotate points and squiggles in complex images. Their work showed that the optimisation can be sufficiently guided using only a few supervised pixels, which are easily obtained. We improve on this by completely removing the need for human annotators: the per-pixel label distribution can be obtained using *only* image-level labels, making our approach highly scalable.

2.4.3 E- & M-steps: optimising parameters

E-step (approximate complete-data log-likelihood)

We now describe how to define and optimise the expected complete-data log-likelihood, $F(\theta; \theta_t)$. By definition, $F(\theta; \theta_t) = \sum_{\mathbf{y}} P(\mathbf{y}|I, \mathbf{z}; \theta_t) \log P(I, \mathbf{y}, \mathbf{z}; \theta)$, where the expectation is taken over the posterior over the latent variables at a given set of parameters θ_t. The expectation is denoted as $P(\mathbf{y}|I, \mathbf{z}; \theta_t)$. In the case of semantic segmentation, the latent space is exponentially large $|\mathcal{L}|^n$, and computing $F(\theta; \theta_t)$ is therefore infeasible. However, as will be shown, the independence assumption over the random variables, namely $P(\mathbf{y}|I; \theta) = \prod_{m=1}^{n} p(y_m|I; \theta)$, allows us to maximise $F(\theta; \theta_t)$ efficiently by decomposition. By using Eq. 2.4.1, the independence assumption, the identity $\sum_{\mathbf{y}} P(\mathbf{y}|I, \mathbf{z}; \theta_t) = 1$, and ignoring the terms independent of θ, $F(\theta; \theta_t)$ can be written in a simplified form as:

$$\bar{F}(\theta; \theta_t) = \sum_{m=1}^{n} \sum_{\mathbf{y}} P(\mathbf{y}|I, \mathbf{z}; \theta_t) \log p(y_m|I; \theta) \tag{2.4.2}$$

Without loss of generality, we can write $P(\mathbf{y}|I,\mathbf{z};\theta_t) = P(\mathbf{y} \setminus y_m|I,\mathbf{z},y_m;\theta_t)p(y_m|I,\mathbf{z};\theta_t)$, and using the identity $\sum_{\mathbf{y}\setminus y_m} P(\mathbf{y} \setminus y_m|I,\mathbf{z},y_m;\theta_t) = 1$, we obtain:

$$\bar{F}(\theta;\theta_t) = \sum_{m=1}^{n} \sum_{y_m \in \mathcal{L}} p(y_m|I,\mathbf{z};\theta_t) \log p(y_m|I;\theta) \tag{2.4.3}$$

M-step (parameter update)

The M-step parameter update, which maximises $\bar{F}(\theta;\theta_t)$ with respect to θ, can be written as:

$$\theta_{t+1} = \arg\max_{\theta} \sum_{m=1}^{n} \sum_{y_m \in \mathcal{L}} p_\theta(y_m|I,\mathbf{z};\theta_t) \log p(y_m|I;\theta) \tag{2.4.4}$$

We make the assumption that the posterior $p_\theta(y_m|I,\mathbf{z};\theta_t)$ belongs to the exponential family distribution such that $p_\theta(y_m|I,\mathbf{z};\theta_t) \propto \exp(f(y_m|I;\theta_t) + g(y_m,\mathbf{z}))$. Here, $f(y_m|I;\theta_t)$ is the likelihood obtained for pixel m using the CNN at a given θ_t, and $g(y_m,\mathbf{z})$ is a user-defined function which we use to regularise the CNN likelihood.

More specifically, we use $g(y_m,\mathbf{z})$ to explicitly impose constraints based on the image label information, namely the network should suppress the probability of objects not present in the image. For example, if we know that there are only two classes in a given training image, for example "cat" and "person", then we would like to push the latent posterior probability $P(\mathbf{y}|I,\mathbf{z};\theta_t)$ of absent classes to zero and increase the probability of the present classes. In order to impose the above mentioned constraints, we use $g(\cdot,\cdot)$ as:

$$g(y_m,\mathbf{z}) = \begin{cases} -\infty, & if\ y_m \notin \mathbf{z} \cup l_0 \\ 0, & otherwise \end{cases} \tag{2.4.5}$$

Imposing the above constraint is equivalent to retaining the *softmax* probabilities for only those classes (including background l_0) present in the image, and assigning a probability of zero to all other classes. In other words, the above definition of $g(\cdot,\cdot)$ inherently defines a uniform distribution over the object classes present in the image including the background and zero for the remaining ones. Other forms of $g(\cdot,\cdot)$ can also be used to impose different task-specific label-dependent constraints.

Optimising Eq. 2.4.4 is equivalent to minimising the cross-entropy or the KL divergence between the latent posterior distribution $p(y_m|I, \mathbf{z}; \theta_t)$ and the CNN likelihood $p(y_m|I; \theta)$. Papandreou et al. [254] use a Dirac delta approximation $\hat{p}$ of the posterior distribution, where $\hat{p}(\hat{l}_m)$ is 1 at $\hat{l}_m = \arg\max_{l \in \mathcal{L}} p(y_m = l|I, \mathbf{z}; \theta_t)$, and is otherwise zero. We instead propose to use the combination of the Dirac delta approximation and the actual latent posterior distribution (or the regularised likelihood) in Eq. 2.4.4 as follows:

$$J_m(I, \mathbf{z}, \theta_t; \theta) = \sum_{y_m \in \mathcal{L}} \bar{p}(y_m|I, \mathbf{z}; \theta_t) \log p(y_m|I; \theta) \tag{2.4.6}$$

where, $\bar{p}(y_m|I, \mathbf{z}; \theta_t) = (1 - \epsilon)p(y_m|I, \mathbf{z}; \theta_t) + \epsilon\hat{p}(y_m)$. We argue that using a Dirac delta distribution alone imposes a hard constraint that is suitable only when we are very confident about the true label assignment (for example, in the fully-supervised setting). In the weakly-supervised setting where the latent posterior, which decides the label, can be noisy (mostly seen in the case of difficult classes), it is more suitable to use the full posterior distribution. Eq. 2.4.6 therefore provides the best of both worlds. In defining the weighting factor ϵ, we explore two heuristics. The first, which we term the *max* heuristic, defines ϵ as:

$$\epsilon = \begin{cases} 1, & if\ p_{max} \geq \eta \\ p_{max}, & otherwise \end{cases} \tag{2.4.7}$$

where $p_{max} = \max_{l \in \mathcal{L}} p(y_m = l|I, \mathbf{z}; \theta)$ is the maximum probability value in the latent posterior distribution, and $\eta \in [0, 1]$ is a hyper-parameter. Intuitively, this implies that if the posterior is confident about a particular class label (relative to η) then the traditional Dirac delta posterior is used. Otherwise, a weighted combination of the true posterior- and Dirac delta-based cross-entropy is used, where the weight is decided by p_{max}. The second heuristic, which we term the *relative* heuristic, defines ϵ as:

$$\epsilon = \begin{cases} 1, & if\ r \geq \eta \\ r, & otherwise \end{cases} \tag{2.4.8}$$

where $r = (p_1 - p_2)/p_1$. Values p_1 and p_2 are the highest and second highest probabilities in the latent posterior distribution. Intuitively, $\eta = 0.05$ implies that the most probable score should be at least 5% better than the second most probable score in order to use the Dirac delta posterior alone, otherwise, the weighted combination should be used.

IOU gain function Along with minimising the cross-entropy loss as shown above, in order to obtain a better parameter estimate, we also maximise the probabilistic approximation of the intersection-over-union (IOU) between the posterior distribution and the likelihood [6, 68, 245]:

$$\mathcal{J}_{IOU}(P(\mathbf{y}|I,\mathbf{z};\theta_t),P(\mathbf{y}|I;\theta)) \approx \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}}\frac{\sum_{m=1}^{n}p_m^t(l)p_m^\theta(l)}{\sum_{m=1}^{n}\{p_m^t(l)+p_m^\theta(l)-p_m^t(l)p_m^\theta(l)\}} \tag{2.4.9}$$

where, $p_m^t(l) = p(y_m = l|I,\mathbf{z};\theta_t)$ and $p_m^\theta(l) = p(y_m = l|I;\theta)$. We refer the reader to [68] for further details about Eq. 2.4.9.

Overall objective function and learning algorithm

Combining the cross-entropy loss function Eq. 2.4.6 and the IOU gain function Eq. 2.4.9, the M-step parameter update is:

$$\theta_{t+1} = \arg\max_{\theta}\sum_{m=1}^{n}J_m(I,\mathbf{z},\theta_t;\theta) + \mathcal{J}_{IOU} \tag{2.4.10}$$

We use a CNN model along with the back-propagation algorithm to optimise the above objective function. Recall that our evaluation is based on the *PASCAL VOC 2012* dataset, therefore, during the M-step of the algorithm we use both the *ImageNet* $\mathcal{D}(I)$ and the *PASCAL trainval* $\mathcal{D}(P)$ datasets (see Section 2.5 for details). Our overall approach is summarised in Algorithm 2.

Algorithm 2 Final algorithm

Input: Datasets $\mathcal{D}(P)$ and $\mathcal{D}(I)$; θ_0; η; K

1: Use $\mathcal{D}(I)$ to obtain initialisation parameters θ_t (see Section 2.4.2)
2: **for** $k = 1 : K$ **do**
3: $\quad\theta \leftarrow \theta_t$
4: $\quad$**for** each pixel m in $\mathcal{D}(P) \cup \mathcal{D}(I)$ **do**
5: $\quad\quad$Obtain latent posterior: $p_m(y_m|I,\mathbf{z};\theta) \propto \exp(f_m(y_m|I;\theta) + g(y_m,\mathbf{z}))$
6: $\quad$**end for**
7: $\quad$Optimise Eq. 2.4.10 using CNN to update θ_t
8: **end for**

2.5 Experimental set-up & analysis

We show the efficacy of our method on the challenging *PASCAL VOC 2012* benchmark and outperform all existing SOTA methods by a large margin. Specifically, we improve on the current SOTA method [349] by 3.7% and 3.9% on the validation and test sets, respectively.

2.5.1 Experimental set-up

Dataset $\mathcal{D}(I)$ for obtaining initialisation To train our initialisation model (Section 2.4.2), we download 80,000 images from the *ImageNet* dataset [83]. These images contain objects in the 20 foreground object categories of the *PASCAL VOC 2012* segmentation task. We then filter this dataset using simple heuristics. First, we discard images with width or height less than 200 or greater than 500 pixels. Using the attention model of [377], we generate a per-class attention map for each image and record the most probable class label with its corresponding probability. We discard images for which the most probable class label does not match the given image label and has a probability of less than 0.2. We also generate saliency maps using the saliency model of [139] which is trained with class-agnostic saliency masks.

We then combine attention and saliency in the following way: we first generate an attention binary mask from each attention map by setting a mask pixel to 1 if its corresponding attention probability is greater than 0.5. We do the same to the saliency maps to obtain saliency binary masks. We then find the pixel-wise intersection between the saliency and the attention masks. For each object category, the images are sorted by this intersection area (i.e. the number of overlapping pixels between the two masks) with the intuition that larger intersections correspond to higher quality saliency and attention maps. The top 1500 images are then selected for each category. The only exceptions are the "person" category in which the top 2500 images are kept, and categories with fewer than 1500 images, in which case all images are kept. This filtering process leaves us with 24,000 simple images of uncluttered and mainly-centred single objects. We denote this dataset as $\mathcal{D}(I)$ and highlight that it does not contain any additional images relative to those used by other weakly supervised-works (see Dataset column in Table 2.1).

Datasets $\mathcal{D}(P)$ and $\mathcal{D}(I)$ for M-step. For the M-step, we use a filtered subset of *PASCAL VOC 2012* images, denoted $\mathcal{D}(P)$, and a subset of $\mathcal{D}(I)$. To obtain $\mathcal{D}(P)$, we take complex *PASCAL VOC 2012* images ($10,582$ in total, made up of $1,464$ training images [89] and the extra images provided by [125]), and use the trained initialisation model to generate a (hard) ground-truth segmentation for each. The hard segmentations are obtained by assigning each pixel with the class label for which the initialisation model predicts the highest probability. The ratio of the foreground area to the whole image area (where area is the sum of the number of pixels) is computed. If the ratio is below 0.05, the image is discarded. This leaves $10,000$ images. We also further filter $\mathcal{D}(I)$: using the trained initialisation model, we generate (hard) segmentations for all simple *ImageNet* images in $\mathcal{D}(I)$. We compute the intersection area (as above) between the attention binary mask and the predicted segmentation (rather than the saliency mask as before). We then select the top $10,000$ of $24,000$ images based on this metric. Together $\mathcal{D}(P)$ and this subset of $\mathcal{D}(I)$ make up $20,000$ images which are used for the M-step.

CNN architecture & parameter settings Similar to [176, 254, 350] our initialisation model and our EM model are based on the *DeepLab* architecture [60]. We use simple bilinear interpolation to map the down-sampled feature maps to the original image size as suggested in [196]. We use the publicly available Caffe toolbox [154] for our implementation. We use weight decay (0.0005), momentum (0.9), and iteration size (10) for gradient accumulation. The learning rate is 0.001 at the beginning and is divided by 10 every 10 epochs. We use a batch size of 1 and randomly crop the input image to 321×321. Images with width or height less than 321 are padded with the mean pixel values and the corresponding places in the ground-truth are padded with ignore labels to nullify the effect of padding. We flip the images horizontally, resulting in an augmented set twice the size of the original one. We train our networks for 30K iterations by optimising Eq. 2.4.10 as per Algorithm 2 with $\eta = 0.05$ and $K = 2$. Performance gains beyond two EM iterations were not significant compared to the computational cost.

Table 2.1 Comparison table. All dependencies, datasets, and degrees of supervision used by the current SOTA methods for weakly-supervised semantic segmentation. $\mathcal{D}(I)$: *ImageNet* dataset; $\mathcal{D}(P)$: *PASCAL VOC 2012* dataset (see Section 2.5); and $\mathcal{D}(F)$: 41K images from *Flickr* [350]. Note that the cross validation of CRF hyper-parameters and the training of MCG are performed using a fully-supervised pixel-level semantic segmentation dataset. Methods with equivalent supervision are underlined for fair comparison to our own method.

Method	Dataset	Dependencies	Supervision	CRF [178]	mIoU (Val)	mIoU (Test)
EM Adapt [254]	$\mathcal{D}(I), \mathcal{D}(P)$	No	Image labels	✗	–	–
				✓	38.2%	39.6%
CCNN [261]	$\mathcal{D}(I), \mathcal{D}(P)$	No	Image labels	✗	33.3%	35.6%
				✓	35.3%	–
		Class size		✗	40.5%	43.3%
				✓	42.4%	45.1%
SEC [176]	$\mathcal{D}(I), \mathcal{D}(P)$	Saliency [306] & Localization [384]	Image labels	✗	44.3%	–
				✓	50.7%	51.7%
MIL [269]	$\mathcal{D}(I)$	Superpixels [92]	Image labels	✗	36.6%	35.8%
		BBox BING [63]			37.8%	37.0%
		MCG [27]			42.0%	40.6%
WTP [40]	$\mathcal{D}(I), \mathcal{D}(P)$	Objectness [9]	Image labels	–	32.2%	–
			Image labels + 1 Point/Class		42.7%	–
			Image labels + 1 Squiggle/Class		49.1%	–
STC [350]	$\mathcal{D}(I), \mathcal{D}(P), \mathcal{D}(F)$	Saliency [155]	Image labels	✓	49.8%	51.2%
AugFeed [277]	$\mathcal{D}(I), \mathcal{D}(P)$	SS [331]	Image labels	✗	46.98%	47.8%
				✓	52.62%	52.7%
		MCG [27]		✗	50.41%	50.6%
				✓	54.34%	55.5%
AE-PSL [349]	$\mathcal{D}(P)$	Saliency [155]	Image labels	✓	55.0%	55.7%
Ours Initial Model	$\mathcal{D}(I)$	Saliency [139] & Attention [377]	Image labels	✗	53.53%	54.34%
				✓	55.19%	56.24%
Final	$\mathcal{D}(I), \mathcal{D}(P)$			✗	56.91%	57.74%
				✓	**58.71**%	**59.58**%

2.5.2 Experimental analysis

Table 2.1 provides an extensive comparison between our method and other SOTA methods, in terms of performance, dependencies, and degrees of supervision. Regarding the dependencies of our method, our saliency network [139] is trained using salient region masks. These masks are class-agnostic, and, therefore, once trained the network can be used for any salient semantic object category, without the need to retrain the network for new object categories. Our second dependency, the attention network [377] is trained using solely image labels.

Comparison to SOTA We outperform all existing SOTA methods. The most directly comparable method in terms of supervision and dependencies is AE-PSL [349] which uses image-level labels and a saliency network trained on bounding boxes [155] (whereas our saliency network uses class-agnostic saliency masks [139]). Our method obtains almost 4% better mean IOU than AE-PSL on both the validation and test sets. Even if we ignore equivalent supervision and dependencies, our method still outperforms all existing methods.

Importance of good initialisation The initialisation model is essential to the success of our method. We train this model in a very simple way by learning the semantic segmentation task using a filtered subset of simple *ImageNet* images. Importantly, this process uses only image labels and is fully automatic, requiring no human intervention. The learned θ_t provides a very good initialisation for the EM algorithm, enabling it to avoid poor local maxima. This is shown visually in Figure 2.3: the initialisation model (3^{rd} column) is already a good prediction, and the 1^{st} and 2^{nd} EM iterations (4^{th} and 5^{th} columns) improve the semantic segmentation even further. We highlight that with this simple approach, surprisingly, our initialisation model beats *all* current SOTA methods, which are more complex and often use higher degrees of supervision. By implementing this intuitive modification, we believe that many methods could easily boost their performance.

Post-processing with a CRF Even though Table 2.1 includes the performance of our method with a conditional random field (CRF) [178] applied as a post-processing step, we believe that CRFs do not satisfy the requirements of a truly weakly-supervised setting. Firstly, CRF hyper-parameters are normally cross-validated over a fully-supervised pixel-wise segmentation dataset. This is likewise the case for MCG [27] which is trained on a pixel-level semantic segmentation dataset. Secondly, the CRF hyper-parameters are incredibly sensitive: incorporate new object categories requires a pixel-level annotated dataset of the new categories along with the old ones for the cross-validation of the CRF hyper-parameters. This is highly non-scalable. For completeness, however, we show that our method's performance with a CRF, which is boosted by 1.8%. We note that even without a CRF, our approach exceeds the SOTA method [349], which uses a CRF, by almost 2% on the validation and test sets.

(1) Source (2) Ground truth (3) Initialisation (4) 1st EM iteration (5) 2nd EM iteration

Figure 2.3 Qualitative results for weakly-supervised semantic segmentation using our proposed EM-based method. From the initialisation model (3rd column) to the second iteration of the EM algorithm (5th column) the segmentation quality improves incrementally. The bottom two rows show failure cases.

2.6 Conclusion

In this work we address weakly-supervised semantic segmentation using only image-level labels. We propose an EM-based approach and focus on the three key components of the algorithm: i) initialisation, ii) E-step and iii) M-step. Using only the image labels of a filtered subset of simple *ImageNet* images, we learn a set of initialisation parameters for the semantic segmentation task. Following this, with each EM iteration, we empirically and qualitatively verify that our method improves the segmentation accuracy on the challenging *PASCAL VOC 2012* benchmark. Furthermore, we show that our method outperforms all SOTA methods. Future directions include making our method more robust to noisy labels, for example, when images downloaded from the Internet have incorrect labels, as well as improving the performance on images with multiple object classes.

Acknowledgements DM, PKD and PHST were supported by *grant ERC-2012-AdG 321162-HELIOS*. DM was supported by the *Skye Foundation*. QH, YW and MMC were supported by *NSFC (61620106008, 61572264)*, *CAST (YESS20150117)*, *Huawei Innovation Research Program (HIRP)*, and *IBM Global SUR award*.

2.7 Supplementary

2.7.1 Extended experimental analysis

To further explore our method and better understand its successes and failures, here we include additional empirical results on the *PASCAL VOC 2012* validation dataset.

Class-wise results Tables 2.3 to 2.5 show the class-wise mean IOU for different values of η in the *max* and *relative* heuristic settings defined in Section 2.4.2. Our method supersedes the results of all other current methods with equivalent supervision [176, 260, 261, 277] on the majority of object classes. We observe, however, that we consistently under-perform on the "bike", "plant" and "sofa" classes. We attribute this to failures in the attention and saliency networks, which are not able to detect and attend to these classes well. This may be because these classes often occur simultaneously with more salient-worthy object classes like people, and are hence ignored by saliency and attention mechanisms. These failures result in our mining method not extracting many nor high quality training pixels for these classes, which goes on to adversely affect our initialisation model's performance on these classes.

Thresholds As detailed in Section 2.4.2, threshold η is a user-defined hyper-parameter controlling the weighted combination of a traditional Dirac delta ("hard") cross-entropy loss and a true posterior ("soft") cross-entropy loss in the overall objective function Eq. 2.4.6. For the *max* heuristic, $\eta = 0.1$ for example, enforces that the maximum probability (p_{max}) must be at least 0.1 (i.e. 10% confident in the label) in order to use the "hard" cross-entropy. Otherwise, for a probability of less than 10%, a weighted combination of both cross-entropy terms should instead be employed. For the *relative* heuristic, a variable r is instead defined as the relative difference between the highest and second-highest probability for a pixel. A threshold of $\eta = 0.05$ enforces that this relative difference should be at least 5% in order to employ the "hard" cross-entropy, otherwise the weighted combination should be used. A threshold of $\eta = 0$ results in only the "hard" cross-entropy term being used.

Table 2.3 shows the class-wise results for different values of η, within suitable ranges, for each heuristic. The results shown in Table 2.1 in the main paper use the threshold values highlighted in blue. Table 2.4 and Table 2.5 show the class-wise mean IOU from the initialisation model to the final EM iteration ($K = 2$) with the best-performing thresholds for the *max* and *relative* heuristic, respectively. Interestingly, while final performance (at $K = 2$) is similar between the "hard" cross-entropy ($\eta = 0$) and the weighted loss ($\eta \neq 0$), the weighted loss achieves this after only 1 EM iteration, suggesting it facilitates faster convergence of the EM algorithm.

2.7.2 Dependency comparison

Table 2.2 summarises the dependencies of the methods in Table 2.1, including datasets and levels of supervision. The methods which use image labels and other equivalent forms of supervision are comparable to our weakly-supervised method. Of the dependencies listed, we bring particular focus to the CRF. A CRF is commonly used in fully- and weakly-supervised semantic segmentation pipelines. In this work, however, we do not employ a CRF for several reasons. The first is that CRF hyper-parameters are typically fine-tuned on pixel-level annotations. This deviates from a weakly-supervised setting, and so we do not compare against methods which employ a CRF. Secondly, our EM algorithm predicts a latent segmentation (in the E-step) which we fix as the ground-truth for the subsequent M-step. Employing a CRF at this stage may enforce smoothness in the segmentation which may lead to noisy predictions. Finally, the hyper-parameters of the CRF are highly sensitive, and if, for example, the framework is extended to incorporate new object categories, a fully-supervised dataset of the new categories would be required to cross-validate the CRF hyper-parameters. This is does not scale well nor align with the overall motivation for doing weakly-supervised learning.

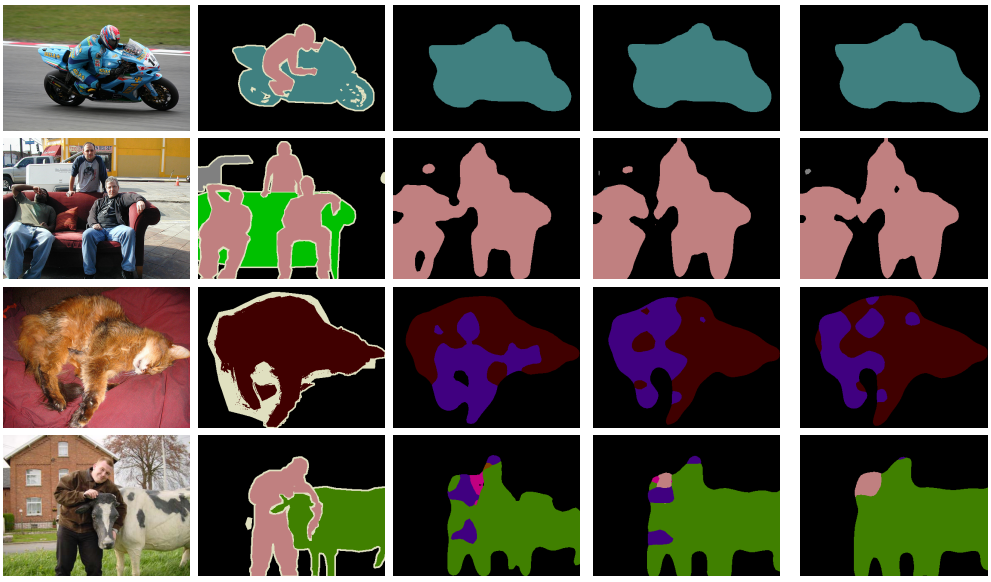
The two dependencies used by our method are the saliency network of [194] and the attention network of [377]. Training the saliency network requires supervision in the form of saliency masks which are class-agnostic foreground masks of the salient object/s in an image. We consider this a suitably weak form of supervision since once trained, the network can generalise to any semantic object category. Training the attention network requires supervision in the form of image labels, which mirrors our weakly-supervised semantic segmentation setting.

Table 2.2 Dependency table. Table summarising the dependencies of existing weakly-supervised methods for the semantic segmentation task. $\mathcal{D}(P)$ and $\mathcal{D}^{07}(P)$ represent the *PASCAL VOC 2012* and *2007* datasets [89], respectively, $\mathcal{D}(I)$ represents the *ImageNet* dataset [83], and $\mathcal{D}(C)$ represents the Microsoft COCO dataset [193].

Dependency	Dataset	Supervision
Class size	$\mathcal{D}(P)$	Image labels + Bboxes
Saliency	$\mathcal{D}(I)$, $\mathcal{D}(P)$	[306] Image labels
	MSRA-B [195]	[155] Bboxes
	MSRA10K [64] and DUT-OMRON [371]	[194] Saliency masks
Attention [377]	$\mathcal{D}(I)$, $\mathcal{D}^{07}(P)$, $\mathcal{D}(C)$	Image labels
Objectness [9]	$\mathcal{D}^{07}(P)$	Image labels + Bboxes
Localisation [384]	$\mathcal{D}(I)$, $\mathcal{D}(P)$	Image labels
Superpixels [92]	NA	None
BBox BING [63]	$\mathcal{D}^{07}(P)$	Bboxes
MCG [27]	$\mathcal{D}(P)$, BSDS [214]	Pixel labels
Selective Search [331]	$\mathcal{D}^{07}(P)$	Bboxes
CRF [178]	–	Pixel labels (parameter cross-val)

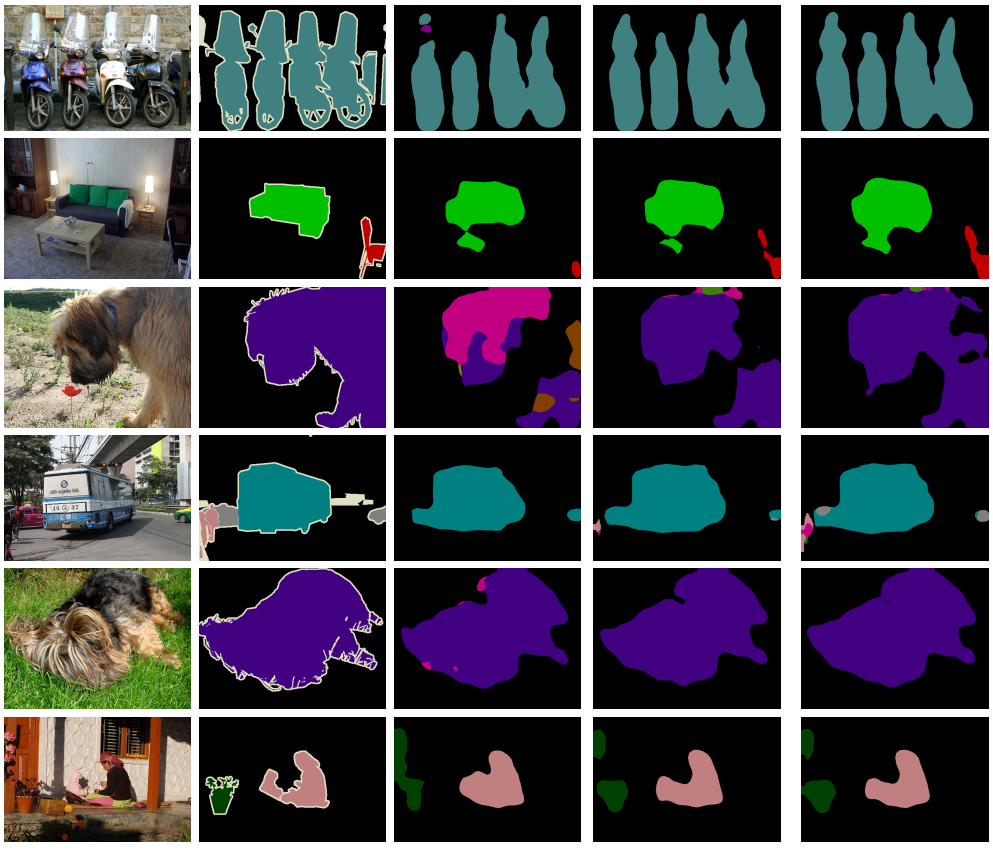
2.7.3 Success and failure cases

Figure 2.4 and Figure 2.5 show qualitative examples of failure and success cases of our weakly-supervised semantic segmentation method. In terms of failure cases, our method struggles to segment thin or fine structures, for example, the wheels of bicycles and the legs of chairs and tables. Conventionally, the application of a CRF either in training or as a post-processing step helps to better delineate such structures, however, for the above-mentioned reasons, we do not employ a CRF in our pipeline. Another common failure arises in the case of overlapping objects. Again, this can largely be attributed to a failure of the saliency and attention network, where the background object is likely overlooked or considered non-salient. In terms of success cases, Figure 2.5 shows the segmentation predictions from the initialisation model to the final iteration of the EM algorithm. At each step we observe an improvement in the segmentations suggesting that the iterative nature of the EM algorithm and our learning objective are well suited to the segmentation task.



(1) Source (2) Ground truth (3) Initialisation (4) 1st EM iteration (5) 2nd EM iteration

Figure 2.4 Failure cases. Visual results for the weakly-supervised semantic segmentation using our proposed EM-based method, from the initialisation model (3rd column) to the final iteration of the EM algorithm (5th column).



(1) Source (2) Ground truth (3) Initialisation (4) 1st EM iteration (5) 2nd EM iteration

Figure 2.5 Success cases. Visual results for the weakly-supervised semantic segmentation using our proposed EM-based method. As we move from the initialisation model (3rd column) to the second iteration of the EM algorithm (5th column), the segmentation quality improves.

Table 2.3 Experimental results (in %) on *PASCAL VOC 2012* validation set using different values of threshold η for the *max* and *relative* heuristic defined in Section 2.4.2. *Threshold* shows η values for EM *Iteration 1* and EM *Iteration 2*, respectively, with blue values showing best-performing thresholds. Results are shown for the final iteration of the EM algorithm.

Heur	Threshold	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
Max	(0, 0)	87.9	73.0	29.6	66.7	58.8	57.9	77.4	69.0	70.6	16.7	64.7	23.2	64.4	59.3	62.4	58.4	37.5	68.3	31.2	71.0	44.5	56.78
	(0.1, 0)	87.9	72.9	29.2	66.9	58.7	57.4	77.2	69.1	70.8	17.3	64.1	24.4	63.4	60.4	62.4	58.7	37.7	68.5	31.6	71.5	44.0	56.86
	(0.1, 0.1)	87.8	73.4	29.5	67.3	60.0	56.4	77.5	68.4	70.0	16.9	64.1	23.2	64.8	60.7	61.8	57.4	39.5	69.5	29.1	70.6	44.5	56.78
Rel	(0.05, 0)	87.9	72.7	28.5	67.1	59.4	55.5	77.8	69.3	70.6	18.4	62.4	23.9	65.3	58.9	61.3	58.0	39.7	69.8	29.9	71.7	44.7	56.80
	(0.05, 0.05)	87.8	72.4	28.7	67.9	58.8	55.8	78.0	69.7	70.2	17.8	63.3	23.2	65.7	60.5	63.1	58.7	40.0	68.2	28.9	70.9	45.5	56.91

Table 2.4 Experimental results on *PASCAL VOC 2012* validation and test set using the *max* heuristic, with a threshold η of (0.1, 0) for EM *Iteration 1* and EM *Iteration 2*, respectively.

Data	Stage	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
val	*Initial*	87.1	74.7	29.0	69.8	55.8	55.6	73.3	65.2	63.4	15.8	61.5	15.9	60.0	56.4	57.5	53.7	32.9	65.6	23.9	64.6	42.2	53.53
	EM-*Iter1*	87.6	74.3	30.5	69.4	58.3	58.7	76.5	66.9	69.6	16.3	64.9	19.4	63.8	61.0	61.3	56.4	36.8	68.4	28.3	69.6	44.1	56.27
	EM-*Iter2*	87.9	72.9	29.2	66.9	58.7	57.4	77.2	69.1	70.8	17.3	64.1	24.4	63.4	60.4	62.4	58.7	37.7	68.5	31.6	71.5	44.0	56.86
test	*Initial*	87.9	69.2	29.2	74.9	41.7	53.4	70.6	69.6	59.9	18.3	66.1	24.9	62.5	63.3	68.8	55.4	33.7	63.8	18.6	64.3	44.9	54.35
	EM-*Iter1*	88.2	70.3	30.0	75.3	45.2	58.3	72.0	72.0	66.8	19.0	67.4	29.8	67.6	65.1	72.2	57.9	43.2	69.2	23.9	65.4	44.5	57.31
	EM-*Iter2*	88.2	70.8	29.5	72.3	45.3	57.3	72.8	72.1	69.0	19.8	64.4	34.6	67.4	63.5	72.4	59.1	45.8	69.3	28.1	64.3	45.2	57.67

Table 2.5 Experimental results on *PASCAL VOC 2012* validation and test using *relative* heuristic, with a threshold η of (0.05, 0.05) for EM *Iteration 1* and EM *Iteration 2*, respectively.

Data	Stage	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
val	*Initial*	87.1	74.7	29.0	69.8	55.8	55.6	73.3	65.2	63.4	15.8	61.5	15.9	60.0	56.4	57.5	53.7	32.9	65.6	23.9	64.6	42.2	53.53
	EM-*Iter1*	87.6	74.3	29.7	69.9	58.3	57.1	76.3	68.0	68.7	17.7	62.9	19.8	64.1	59.6	61.2	57.2	37.3	67.3	27.6	69.8	43.5	56.09
	EM-*Iter2*	87.8	72.4	28.7	67.9	58.8	55.8	78.0	69.7	70.2	17.8	63.3	23.2	65.7	60.5	63.1	58.7	40.0	68.2	28.9	70.9	45.5	56.91
test	*Initial*	87.9	69.2	29.2	74.9	41.7	53.4	70.6	69.6	59.9	18.3	66.1	24.9	62.5	63.3	68.8	55.4	33.7	63.8	18.6	64.3	44.9	54.35
	EM-*Iter1*	88.1	70.4	30.2	75.4	46.4	56.3	73.3	71.7	67.3	19.6	66.4	29.9	67.3	63.4	72.4	58.2	42.9	70.8	22.6	64.7	46.1	57.31
	EM-*Iter2*	88.2	69.5	29.7	72.2	45.1	57.3	73.2	72.7	69.3	20.5	65.4	33.5	67.8	64.0	72.3	58.9	45.5	69.8	26.8	63.8	46.8	57.74

Chapter 3

Stereosonic Vision: Exploring Visual-to-Auditory Sensory Substitution Mappings in an Immersive Virtual Reality Navigation Paradigm

Daniela Massiceti[1] Stephen L. Hicks[2,3] Joram J. van Rheede[2]

[1]Engineering Science, University of Oxford

[2]Nuffield Department of Clinical Neurosciences, University of Oxford [3]OXSIGHT

Abstract

Sighted people predominantly use vision to navigate spaces, and sight loss has negative consequences for independent navigation and mobility. The recent proliferation of devices that can extract 3D spatial information from visual scenes opens up the possibility of using such mobility-relevant information to assist visually impaired people by presenting this information through modalities other than vision. In this work, we present two new methods for encoding visual scenes using spatial audio: simulated echolocation and distance-dependent hum volume modulation. We implement both methods in a virtual reality (VR) environment and test them using a 3D motion-tracking device. This allows participants to physically walk through virtual mobility scenarios, thus simulating real locomotion behaviour. Blindfolded sighted participants completed two tasks: maze navigation and obstacle avoidance. Results were measured against a visual baseline in which participants performed the same two tasks without blindfolds. Task completion time, speed and number of collisions were used as indicators of successful navigation, with additional metrics defined to investigate other dynamics of their navigation. In both tasks, participants were able to navigate using only audio information after minimal instruction. While participants were 65% slower using audio compared to the visual baseline, they reduced their audio navigation time by an average 21% over just 6 trials. Hum volume modulation allowed over 20% faster navigation, and quicker improvement over the trials, compared to simulated echolocation in both mobility scenarios. Despite this, we speculate that simulated echolocation remains worth exploring as it provides more spatial detail and could therefore be more useful in more complex environments. The fact that participants were intuitively able to successfully navigate space with two new visual-to-audio mappings for conveying spatial information motivates further exploration of these and other mappings with the goal of assisting visually impaired individuals with independent mobility.

Published in the *Public Library of Science (PLOS ONE)*[1]

[1] https://bit.ly/2ETYueJ

3.1 Introduction

Globally more than 250 million people are visually impaired (VI), with over 36 million of this group classified as blind [314, 362]. While certain causes of visual impairment can be prevented or treated, a large proportion of sight loss remains without a cure [94]. New treatment approaches such as retinal prosthetics, optogenetics, and gene therapy offer hope for the future, but at present are at a research or early implementation stage and await evidence of real-life benefit to patients [93].

Vision loss affects the ability to independently carry out activities of daily living [130, 168, 183, 313], in part due to its negative impact on mobility and navigation [43, 52, 103, 129, 181, 213, 329]. While blind or visually impaired individuals are often able to learn to successfully navigate without vision through orientation and mobility training [357], they face significant challenges not faced by the sighted population [104, 107, 197, 198]. Non-sighted navigation requires more planning and cognitive resources [104, 107, 200, 259], and blind and visually impaired individuals are at an increased risk of mobility-related accidents, injuries, and falls [150, 184, 212, 283, 330]. Even when walking around a familiar environment, the variable presence of obstacles, changes in walking surface or drop-offs can be significant mobility hazards [308], and even very experienced non-sighted navigators still regularly veer off their intended course [122]. It also remains the case that human living spaces are usually designed with sighted navigation in mind [107].

While academic debate exists around the representation of spatial information in blind individuals, in particular people who are congenitally blind, it is clear that the cognitive ability for representing spatial information is not the main limiting factor in navigation and mobility [104, 300]. Rather, the limitation lies in the rate at which non-sighted navigators are able to *acquire* spatial information about their current environment, whether that is in order to build an initial cognitive map of the space, or to update their current position in a cognitive map from memory and scan for the presence of any obstacles to safe mobility [104]. While vision is uniquely well-placed to deliver mobility-relevant spatial information rapidly [262], it is not the only source of such information. Already, many blind individuals will spontaneously learn to use non-visual cues to their advantage in sensing obstacles in their environment [318, 361].

Sensory Substitution Devices (SSDs) go one step further by converting information normally acquired through one sense into a representation that is compatible with another intact sense. They aim to exploit the increasingly well-understood cross-modal capacities of the brain [15, 284]. Approaches to substituting vision have generally focussed on the other spatially informative senses, namely hearing and touch [202]. The first SSDs were pioneered by Bach-y-Rita in the 1960s with his development of the Tactile Vision Sensory Substitution (TVSS) device [34]: subjects received vibrating patterns via an array of pins mounted on their backs and were able to differentiate between oriented parallel lines, simple geometric shapes and capital block letters. Extending this initial work, a host of studies have investigated "seeing" with vibro-tactile and electro-tactile stimulation applied to a number of body surfaces [33, 35, 127, 244]. Other SSDs have attempted to present an auditory representation of the visual scene. The *vOICe*, perhaps the most widely investigated vision-to-audio SSD, scans the visual environment from left to right, and converts the 2D grayscale image into a frequency spectrum or "soundscape" [224]. These efforts and others [1, 54] have largely focused on converting 2D camera images into corresponding sounds for the purpose of understanding the visual scene.

Mobility, however, is a 3D task, and requires knowledge about the distance of objects and their radial position to the user. An effective SSD for navigation must therefore provide this information as explicitly as possible. This moves toward viewing an SSD as a mobility tool which can provide the user with an improved spatial awareness of their surroundings. This task-specific approach has the additional advantage of reducing the bandwidth of the information needed to be encoded cross-modally—an important consideration [105, 202]. Several SSDs have been developed specifically with this in mind (for detailed surveys of such devices, see [105, 202, 289]). Work in this field has initially been led by Leslie Kay with the Kay Sonic Torch [164] followed by several others [102, 165, 204, 211, 312]. While the first approaches only provided users with a "virtually extended" cane that reported the distances of objects further afield in their path, the more ambitious devices use a frequency-modulated ultrasound sweep of the environment to provide a fuller "soundscape".

Despite the active research in this area and some encouraging results in academic settings, the uptake of SSDs for mobility by the VI community has been low [105, 210, 289]. Giudice and Legge [105] identify four factors where SSDs can fail: 1) the sensory translation rules—how

effectively does the sensory translation convey the relevant information in another modality?, 2) selection of the information to be transcoded between senses—is the information that is being translated actually the most adequate for carrying out the task at hand?, 3) is the device practical to use?, and finally, 4) does the device have an acceptable form factor? We believe that the availability of new, affordable, and portable devices that can sense and reconstruct the 3D environment and extract semantic information from it (discussed below) allows for novel approaches to sensory translation rules and the selection of task-relevant information that could be implemented on devices that are both practical and unobtrusive. In this work, we focus on sensory translation rules and information selection.

For a dynamic and time-constrained task such as navigation, there is a great need for sensory translation rules to be intuitive. In this work, we have therefore focused on visuospatial-to-audio sensory substitution as this has the advantage that we already use auditory information to inform our representations of near space. We are able to rapidly and accurately localise sound sources based on level differences, temporal delays, and their spectral envelope [247], and use this on a daily basis to direct our attention, head and gaze towards sound sources. Inferring spatial information from such cues should thus come naturally. In contrast, an array of stimulators on a patch of skin may have a natural correspondence with, for instance, the type of 2D image produced by a camera, but there is no such natural correspondence between a 2D patch of the skin and near space (beyond the immediate body surface).

This leaves the question of how best to convey such visuospatial information to the wearer via audio—a process we will refer to as sonification. Here, research thus far has taken inspiration from blind individuals such as Daniel Kish [174] who have learned to use echolocation for navigation and mobility (for recent reviews on the psychophysics of human echolocation see [175, 322]). Echolocation refers to the general process of using sound reflections to infer the spatial properties of the environment. Typically, human echolocators will use a sharp, self-generated sound such as a mouth click to sample their surroundings, and use properties of the sound reflections such as their volume, delay, spectral properties, and stereo components (left versus right ear) to infer the spatial structure of the environment [175, 322]. It is a skill that takes a long time to develop and is best learned at a young age [321]. The signal, however, contains the right sort of spatial information needed for navigation. Previous SSDs have tried

to make echolocation less disruptive to the auditory environment by emitting ultrasonic audio pulses, and making it easier to learn by playing the user a slowed-down recording of the echoes in the audible frequency range [144, 146, 228, 309, 348].

These SSDs work by presenting a spatially informative audio signal to the user, but they have no access to a 3D model of the user's surroundings. This limitation can be addressed, however, by the recent proliferation of portable devices that rapidly scan and reconstruct 3D environments [189, 240, 341, 372]. Explicit access to information about the spatial structure and semantic context of the environment opens up the possibility of conveying this information via a non-visual sense in a manner that is independent of how such information was acquired. It follows, therefore, to consider whether there are sonification strategies that would be easier to learn or more informative than those directly derived from sampling an environment with human-generated echolocation. Additionally, having access to a 3D reconstruction of the environment means it is now possible for devices to represent objects that are currently outside the field of view (FOV) of the device, enabling the creation of persistent audio beacons representing targets or obstacles around the wearer.

In this work, we have explored two novel, simple and sparse spatial audio representations of 3D environments: 1) simulated echolocation with discrete sound particles, and 2) distance-dependent hum volume modulation of beacon sounds attached to objects. Our simulated echolocation acknowledges the previous work that has been carried out in this area, but aims to present a less complex sampling of the environment by having the user virtually emit a fixed number of particles in a 90×90 degree FOV determined by their head direction. The emitted particles "pop" (i.e. are sonified) as they make contact with virtual obstacles in the user's nearby vicinity, and in this way are analogous to the echoes of sound that would bounce of real-world objects in traditional echolocation. The particles' time-of-flight delays (captured as the temporal delay between an initiating click sound and a particle's pop as it hits an obstructing object) and the volume and stereo components of the pop sound itself, convey 3D spatial information about the environment to the user. The object-centric distance-dependent hum volume modulation, on the other hand, departs from the principles of echolocation, and instead transforms features of the environment itself—in our case, virtual obstacles and walls—into sources of sound. Each type of object is assigned a humming sound

of characteristic frequency, and the volume of the hum is modulated by the user's distance to the object, with a higher volume indicating a shorter distance.

To test these mappings for the task of spatial navigation, we used a novel auditory virtual reality (VR) paradigm. The potential of VR for rapid prototyping and testing of audio-based navigation has been recognised previously [106, 199, 296, 345]. It allows researchers to generate any number of arbitrary and randomised environments for participants to navigate, and enables rapid and precise extraction of the dynamics of navigation and mobility behaviour [345]. Moreover, using virtual obstacles avoids exposing participants to real mobility hazards. Some have focused on understanding the neural mechanisms of navigation and developed paradigms that translate well to neuroimaging settings [296], capitalising on the ability of the subject to remain still while navigating in virtual space. This, however, does not take into account the important contributions of proprioceptive feedback of physically walking [55, 333]. For the development of a navigational aid that ultimately aims to have a real-world implementation, it is important to incorporate such proprioceptive cues, particularly because it is expected that the importance of these inputs is magnified in the absence of vision.

Until recently, accurate locational tracking in VR required a substantial amount of dedicated infrastructure (e.g. as in the SWAN system used by Walker and colleagues [344, 345, 346, 359]). Here, however, we present a portable implementation of an immersive VR with the ability to accurately track the movements of users wearing a cord-free VR headset. More specifically, we employ a tablet computer called the *Google Tango* which accurately tracks its 3D position and rotation. This tablet was worn as a head-mounted device by blindfolded sighted participants. We created two types of VR environment, viewed live through the VR tablet headset. Participants could thus navigate through them by physically walking, ensuring realistic proprioceptive mobility feedback. In each of the two environments, the blindfolded participants were tasked with navigating to an end point (unknown a priori), but crucially were presented *only* the audio cues of the sonification being tested via stereo headphones. For each of the two environments, navigation using the vision-to-auditory mappings was compared against a visual baseline condition: participants, without blindfolds, performed the navigation tasks under similar settings, with the end points randomised and thereby still unknown a priori. Central to our quantitative assessment of participants' navigational efficiency was the

3D tracking capability of the *Google Tango* tablet. The device captured the real-time 3D dynamics of participants' walking behaviour in the VR environments, thus enabling us to develop mobility-relevant metrics and provide an in-depth analysis of participants' movements in both auditory conditions and the visual baseline condition.

In summary, our three main research aims with this work were:

1. Develop new visual-to-audio mappings that simultaneously provide information about obstacle distance and radial position relative to the wearer;
2. Develop a locomotion-controlled, flexible navigation paradigm to test whether these new mappings are in principle sufficient for navigating an environment using sound alone;
3. Investigate whether there are differences in participants' navigation performance, strategies and speed of learning between the two sonification approaches.

Ultimately, the aim of exploring these new sonification strategies is to establish their suitability for stand-alone devices or depth-based electronic travel aids [135, 336] to assist independent navigation and mobility in VI individuals.

3.2 Methodology

3.2.1 Participants

18 participants were recruited locally in Oxford, United Kingdom. Participants were healthy volunteers with full sight and full stereo hearing. The choice to test a fully-sighted participant group had a twofold motivation: firstly, the test group was able to perform the navigation task with full sight in each virtual environment, thereby providing a visual control condition for our experiments. Secondly, the group was seen as a necessary first step given that this is a proof-of-principle study exploring novel sonification methods for 3D spatial representation and navigation within a new experimental paradigm.

The mean age of participants was 28.78 ± 8.00 with a male/female distribution of 11/7. All participants had normal or corrected-to-normal vision. Due to the physical limitations imposed by the testing equipment, it was necessary that participants with corrected vision were able to

wear contact lenses for the testing rather than glasses. Participants were rated on their prior experience with the developed SSD, and their experience with first-person-controller computer games and VR devices (Table 3.1).

Table 3.1 Participant demographics. SSD naivety rating: 1(>9 hours), 2(>6 hours), 3(>3 hours), 4(>0 hours), 5(=0 hours). VR/gaming frequency rating: 1(no experience), 2(rarely), 3(several times a year), 4(monthly), 5(regularly)

	Sex	Naivety	Age	Corrected vision	Glasses/contacts	VR experience
1	M	4	40	Y	Y	5
2	M	5	28	N	-	2
3	F	3	23	Y	Y	2
4	M	5	24	N	-	4
5	F	5	24	N	-	1
6	F	5	27	Y	Y	2
7	M	5	19	Y	Y	3
8	M	5	32	Y	Y	2
9	M	5	28	N	-	5
10	F	5	30	Y	Y	1
11	F	5	24	N	-	2
12	F	5	26	N	-	2
13	F	5	31	Y	Y	2
14	M	5	25	Y	Y	3
15	M	5	29	N	-	2
16	M	5	23	N	-	3
17	M	5	30	N	-	2
18	M	5	55	Y	Y	1

On the day of testing, participants were verbally instructed on the types of environments that they would be required to navigate (Section 3.2.2), the two sonification methods to be tested (Section 3.2.3), the equipment to be used (Section 3.2.4) as well as the experimental protocol (Section 3.2.5). Testing was conducted over two 1.5 hour sessions per participant. A brief re-instruction was conducted at the start of the second session. Verbal feedback was collected during and after each session. In addition to this, a voluntary and anonymous follow-up survey was sent to participants after their completion of both sessions.

Ethics This work received ethical clearance through the University of Oxford Central University Research Ethics Committee. All individuals involved in the study provided written informed consent to publish these case details.

3.2.2 Virtual reality environments

Participants were tasked with spatially navigating to randomised end points in VR environments, using each of the developed visual-to-audio mappings (sonification conditions) or using visual information (visual-only baseline condition). Two types of VR environments were

constructed: a maze and an obstacle corridor. Our motivation for using VR environments to test our methods was two-fold:

1. VR environments offer the ability to build randomised environments and obstacles on the fly which is an advantage over real-world testing.
2. Using a VR environment bypasses the problem of understanding the 3D structure of the environment, allowing us to focus on methods for conveying spatial information.

Maze. The maze environments were generated within a 5×7 grid of virtual cubes, each cube was $3 \times 3 \times 3m$ in size making a real-world sized arena of $15 \times 21 \times 3m$. Each maze was constructed such that there existed only one constant-length (7 cubes) path to a goal (Figures 3.1a and 3.1c). For each maze trial, the path was randomly selected from a pool of 20 pre-generated maze paths saved on the *Tango* tablet (see Section 3.2.4). Upon selection, the participant was placed at the starting point of the selected virtual maze.

Obstacle corridor. The maze environment simplifies the task of navigation to making a series of right-angled turns. To mimic a scenario closer to the real-world challenge of detecting and avoiding randomly-placed obstacles, based on the work of [336], we constructed an environment with obstacles. This VR environment was a 6m-wide corridor bounded by a left and right wall with the length of the corridor segmented as follows: an initial 3m of empty space, a 7m segment of obstacles, a second 3m of empty space, and finally the goal. For each trial, the obstacle segment was populated by 5 randomly-positioned columnar objects of 0.8m diameter and 1m height. Additionally, for each trial, the goal was placed at 13m from the starting line at a random point along the corridor's 6m width (Figures 3.1b and 3.1e).

The placement of obstacles was randomised such that there were always multiple possible paths through the obstacles to the goal, and these paths were sufficiently diverse and also non-trivial (for example, participants could not simply navigate around the edges of the obstacle segment where they would encounter no obstacles). As with the mazes, for each trial, the obstacle corridor arrangement was randomly selected from a pool of 20 pre-generated arrangements saved on the *Tango*. Upon selection, the corridor was dynamically constructed and the participant placed at the starting point in virtual space.

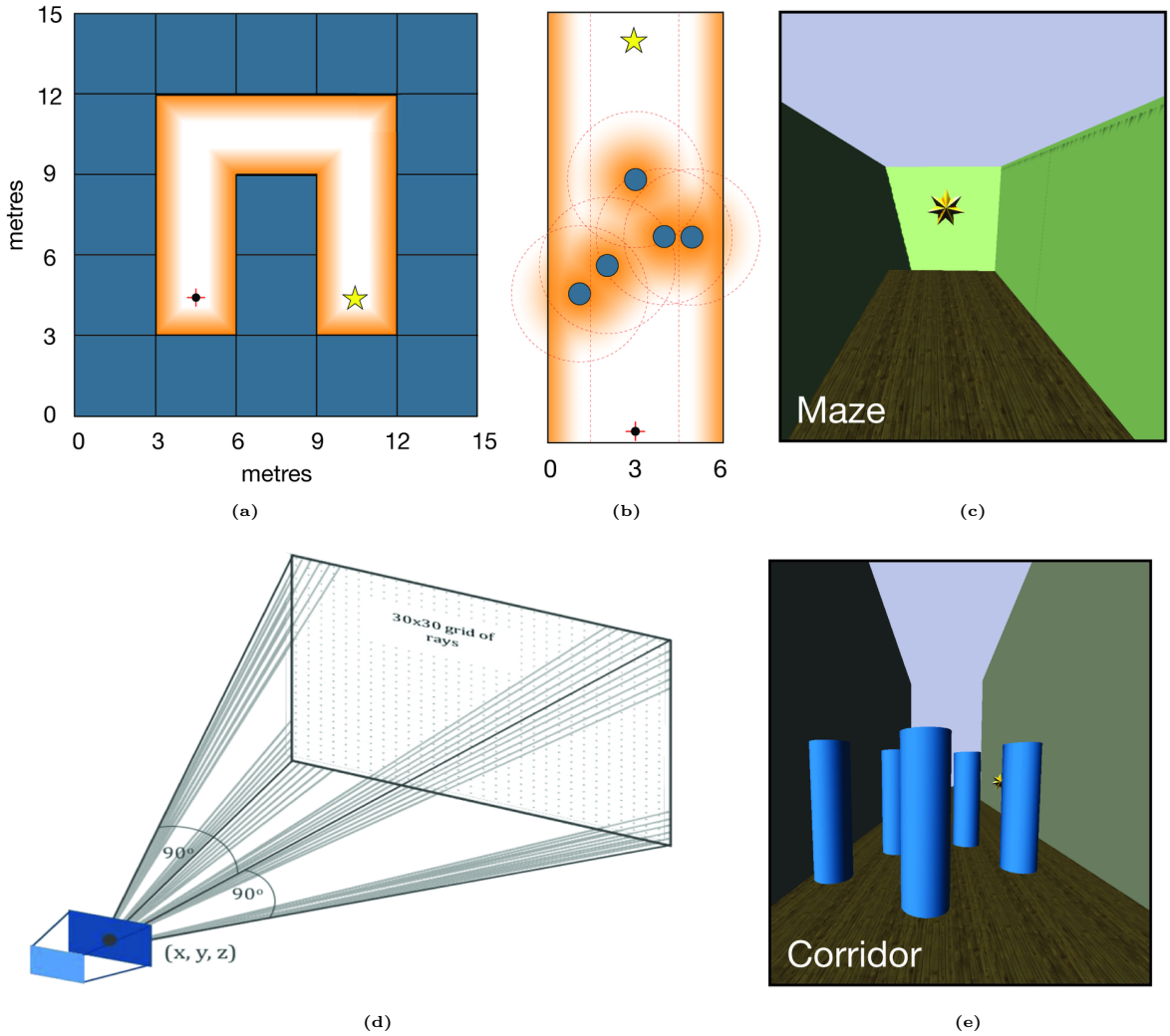


Figure 3.1 VR visualisation and implementation of simulated echolocation and hum volume modulation sonifications. **(a)** Example of a randomly-generated maze with a single path to goal of 7 cube steps. Orange glow indicates humming boundary in the humming sonification. Golden star indicates goal. Red cross and black dot indicate starting position. **(b)** Example of a corridor with randomly positioned obstacles. Orange glows and red dotted lines indicate humming boundaries in the humming sonification. Golden star indicates goal. Red cross and black dot indicate starting position. **(c)** VR visualisation of the maze. **(d)** Implementation of simulated echolocation. A pool of 900 particles is projected along a 30×30 grid array of rays from the (x, y, z) position of the *Tango*. **(e)** VR visualisation of the obstacle corridor.

3.2.3 Stereosonic vision

Our aim was to convert the 3D structure of the visual VR environment into stereo soundscapes providing spatial information to our participants for the task of navigation. We call this stereosonic vision. This section presents the two visual-to-audio mappings, or sonification methods, we implemented and explored in the VR environments: the first, simulated echolocation and the second, distance-dependent volume modulation of hums attached to objects. Resulting sounds were presented on stereo headphones worn by the participant.

Simulated echolocation

There is a large body of work supporting echolocation as a useful mobility technique for blind people [173, 175, 322]. Using self-generated sounds such as mouth clicks, echolocators use sonic reflections to infer the spatial properties of their surroundings. Learning to navigate using echolocation in the real world, however, is difficult and takes much training [175, 321, 322]. In this work, we simulated echolocation in the two VR environments. This alleviated the real-world difficulties of echolocation since we were able to slow down the speed of echolocation feedback, and also select the types of echo sounds. Our motivation was to make simulated echolocation easier to interpret and acoustically more pleasant than real world echolocation. The following describes our VR implementation of simulated echolocation, which from here on we will refer to as simply echolocation.

Much like echolocation in the real world, we implemented simulated echolocation in the virtual world such that it was dependent on the user's body and head direction. The *Tango* emitted a virtual "chirp" in the form of a sharp click (audible via the user's headphones) at a rate of $0.57Hz$, or every 1.75 seconds. With each click, a pool of 900 virtual particles was projected radially outwards from the (x,y,z) position of the *Tango* (where x and y were the lateral coordinates of the *Tango* in space, and the z coordinate was its height above the ground). Each of the 900 particles was projected along a unique radial line, following a 30×30 grid arrangement (Figure 3.1d). The radial projection of the particles had a FOV of $90°$ vertically and horizontally (Figures 3.2c and 3.2d) with the rays spaced $3°$ apart in the horizontal and vertical direction, respectively. As the *Tango* was moved through space, the pool of particles was recast from the *Tango*'s updated (x, y, z) position. Since the *Tango* was head-mounted on the participant (see Section 3.2.4 and Figure 3.2a), a head/body rotation caused a corresponding *Tango* rotation. In this way, the virtual particle projection was always outwards and forwards relative to the participant's head direction.

Following a click, each particle within the pool was projected along its ray but only generated a sound (played via the user's headphones) if it intersected with a virtual object in the VR environment. This was intended to be analogous to the echoes that would be reflected off real-world objects in traditional echolocation. In terms of implementation in our sound engine, this simply meant that a particle was "silent" when travelling along its ray trajectory and

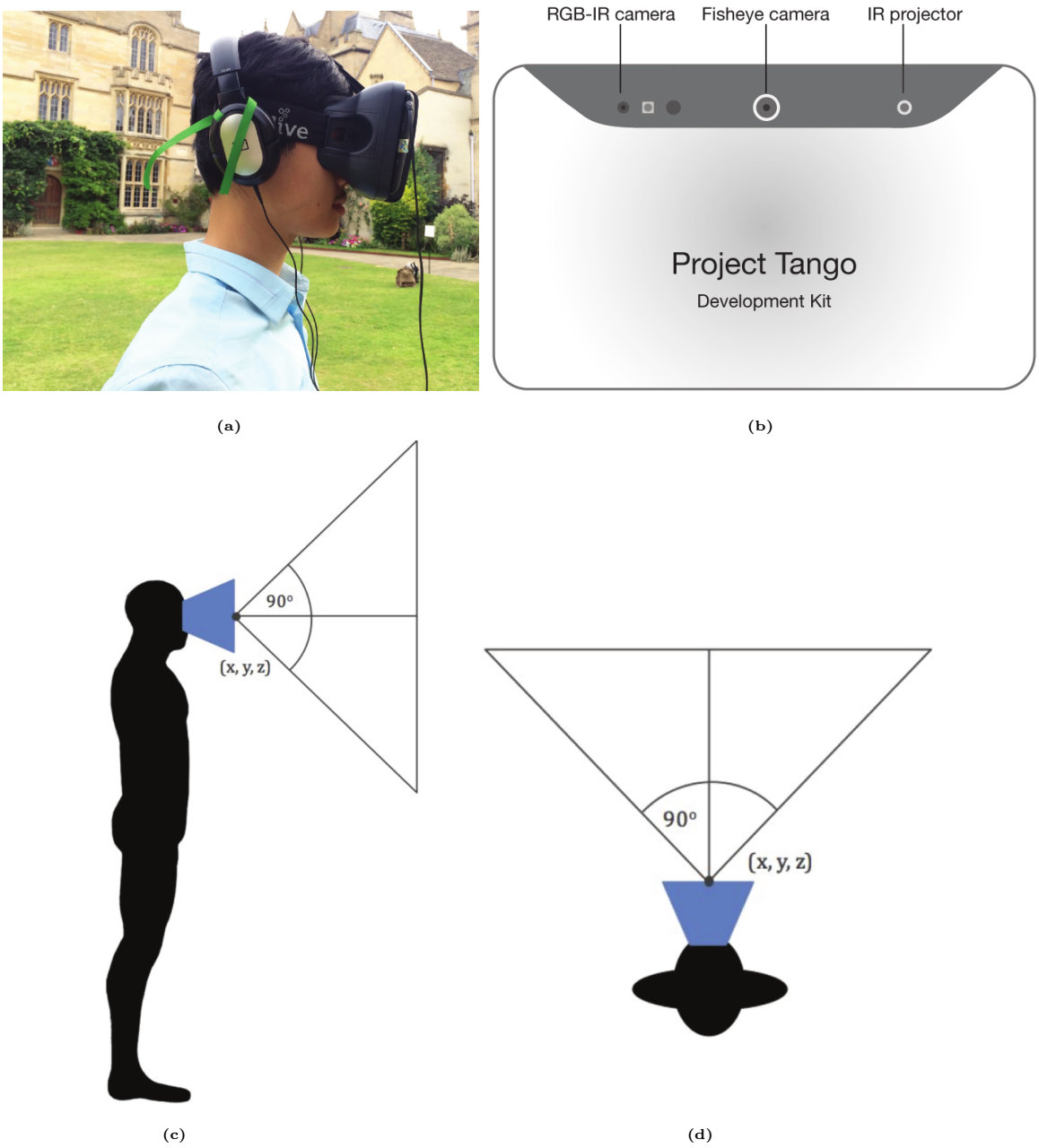


Figure 3.2 Hardware set-up with the *Project Tango* and *Durovis Dive 7* head mount
(a) Participant wearing the *Tango* in the *Durovis Dive 7* headset with stereo headphones. **(b)** Rear-facing camera hardware of the *Tango*. **(c,d)** 90° field of view vertically and horizontally from the position of the *Tango*.

emitted its sound when a virtual object in the VR world obstructed that trajectory. This was thresholded to virtual objects within a 3m radius of the *Tango*'s (x, y, z) position. Particles were sonified as "pops", intended to be harmonious and easy on the ear. The pairing of clicks and pops aimed to convey three pieces of information about the 3D spatial features of the environment: firstly, the presence of an object in a particular radial direction, secondly, an

estimate of the distance to that object, and thirdly, some information about the object's 3D form. This was done by employing the stereo capability of the headphones (i.e. a pop sonified on the left or right corresponded to an obstructing object on the left or right of the participant), the time delay between the initiating click and the corresponding pop as with traditional echolocation (a long delay indicated that the object was further away from the participant than a shorter delay), and the volume of the pop sound itself. Importantly, participants were required to attend to not just a single particle's pop, but the pops of all particles (of 900) that have landed on virtual objects in front of the participant. The speed of the click and particle projection was chosen such that there was no temporal overlap between the pops corresponding to different clicks. In this way, a participant had to use the click and the collection of popping sounds to construct and continually update a distance-based map of their immediate (< 3m) surroundings in the VR environment. A video example of echolocation is available online[2].

Distance-dependent hum volume modulation

Differing from the "snapshot" nature of the point-and-project simulated echolocation, the second sonification method we explored aimed to encode the spatial layout of objects using continuous audio beacons. Here, hums of different pitches were attached to different objects, and the volume of these hums was modulated based on a participant's distance to the objects: shorter distances to objects corresponded to their hums being louder. Furthermore, an object's hum was only triggered when the participant entered a defined "humming zone" around the object, and as they moved closer to the object, the volume of the hum was linearly increased. A linear volume roll-off curve was selected over a log-based roll-off since it allowed changes in distance to an obstacle (given the bounds of its humming zone) to be more easily discerned. In the maze, the humming zone extended 1.2m from the wall (Figure 3.1a). In the obstacle corridor, the humming zone extended 3m radially from each obstacle, and 1.5m from the walls (Figure 3.1b). Using the stereo headphones, the panoramic position of the hum enabled participants to determine its direction and hence the spatial position of the obstacle.

The types of humming sounds were chosen based on the objects emitting them. The walls in both the maze and the corridor were assigned a deep, resonant hum, while the obstacles in

[2] https://www.youtube.com/watch?v=WFHEJ8pOego

the obstacle corridor were each assigned a hum of unique pitch (to all the other obstacles and the walls). We opted for this approach because it was crucial for participants to be able to differentiate obstacles, especially when the humming zones of obstacles overlapped.

Our rationale for choosing humming sounds over other types of sounds was two-fold: firstly, we wanted to create a *continuous* or smooth soundscape capable of capturing the presence and spatial location of multiple objects in an environment. The humming sounds were ideal for this since they could be looped seamlessly (compared to single beeps or pulses). Their continuous nature also allowed us to fuse audio representations across multiple objects in the environments. Secondly, we wanted to create an acoustically pleasant soundscape with the motivation that our sonification mapping should be ambient and non-intrusive to users—essential in a real-world implementation of a SSD. For this reason, we selected hums since they have little temporal structure and are almost absorbed into the audio background unless specifically attended to. The hums themselves were additive blends of sinusoids, each of characteristic frequency, rather than pure tones which are harsher on the ear. The blends of frequencies were chosen to ensure perceptibly pleasant sounds. Figure 3.3 summarises the core frequency components of the hums used.

For brevity, we refer to this method of distance-dependent hum volume modulation as simply the humming sonification. A video example of humming is available online[3].

Miscellaneous environmental sounds

In addition to the audio cues described above, the target location (represented by a golden star) emitted a continuous pinging sound in order to help participants localise the goal. The pinging had its volume modulated by a participant's distance to it. The stereo headphones enabled left/right differentiation of the star's location. Additionally, at the end of each trial, to indicate that the goal had been reached, a completion sound (a jingle of chimes) was played.

[3] https://www.youtube.com/watch?v=aR5r10daK7Y

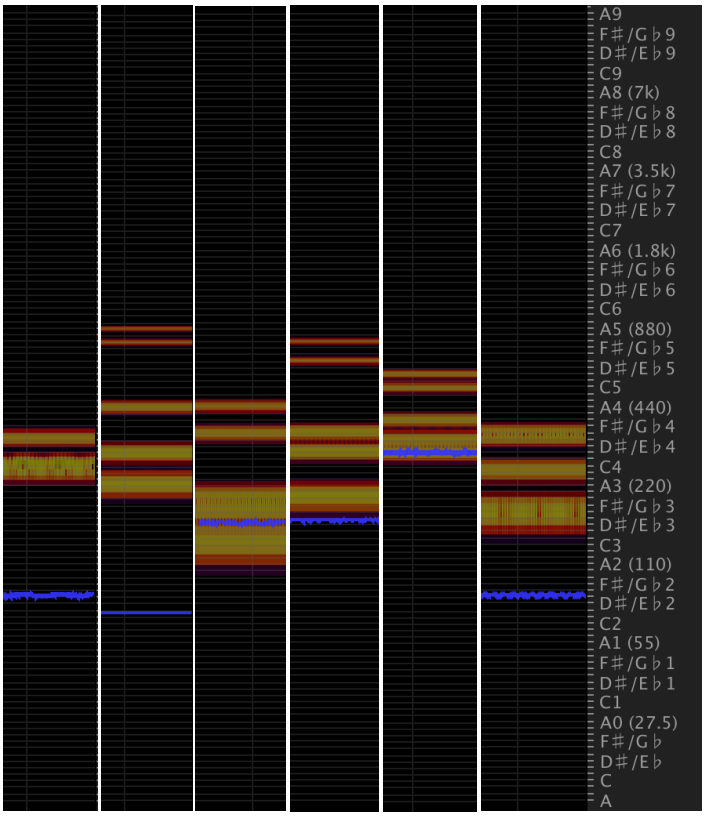


Figure 3.3 Spectral frequency displays of the additive sinusoidal hums used in the distance-dependent hum volume modulation sonification method. Each column (horizontal axis) corresponds to a hum, with its frequency components/notes shown on the vertical axis. The blue band indicates the hum's core frequency, while the red and orange bands indicate other prominent frequencies.

3.2.4 Hardware & software

Unity (Unity Technologies, San Francisco, CA, USA), a 3D game engine, was used to create the VR environments, and features in its built-in sound engine were used to encode the environments using the two sonification methods. The spatial audio features allowed for the localisation of sound sources on the horizontal plane by regulating the gains of a sound on the left and right side (relayed via the headphones) based on the distance and angle between it and the participant in the VR environment. The sonified environments were then deployed on a *Google Tango* tablet (Google, Mountain View, CA, USA). The tablet employs visual-inertial odometry using a 180° FOV fish-eye camera and its inertial measurement unit (Figure 3.2b) to map its 3D position and rotation in real-time.

The software packages used for development included the *Tango Tablet Development Kit* (kernel version: 3.10.24-g36a1dd3; April 14, 2015), *Unity 5.0.1 Personal Edition* with C#, *Android SDK* tools (release 23.0.2 for Mac), and the *Tango Unity SDK* (Ramanujan, version 1.17, July 2015). The *Tango* software was *Android KitKat* (version 4.4.2, API level 19).

Rather than using a keyboard or joystick-based spatial navigation paradigm, we aimed to create a fully immersive VR experience in which participants' movements in the real world corresponded to their movements in the virtual world. To do this, the *Tango* was mounted in a *Durovis Dive 7* VR headset (Figure 3.2a). The headset allowed the device to be fixed to participants' heads such that their 3D positional (translational) movement and 3D head rotation were tracked. The headset included a pair of lenses which projected the image on the *Tango* screen at a comfortable viewing distance for participants. The headphones (*Sennheiser HD 206 Stereo Headphones*) were wire-connected to the *Tango*. Importantly, the headphones were over-ear with a noise attenuation feature, thus ensuring that external sounds from the environment were suppressed.

3.2.5 Experimental protocol

Testing was carried out in a large indoor hall or a large outdoor space, thus comfortably fitting the VR environments and reducing the risk of participants colliding with physical objects. The physical size of the hall was 20×25 metres. The outdoor space was a flat manicured lawn, the weather was mild, and no other people, besides the experimenter and the participant, were present. The testing location was selected based on the availability of the hall, but also allowed the robustness of our methods to be tested in two different environments. Of the 18 participants, 5 were tested outdoors, and no difference in navigational performance was noted.

Six experimental conditions were undertaken by each participant. Each condition included a minimum of 6 repeated trials. Participants were instructed to reach the star goal as quickly as possible while avoiding walls and/or obstacles. Collisions with either were conveyed to the participant by the *Tango* vibrating gently. This vibration was intended to notify participants of a collision and, without being too noxious, to motivate them to avoid collisions as much as possible. Trials were discarded from analysis if 1) the *Tango* tracking was lost due to a device or initialisation failure, or 2) the participant failed to reach the goal within 150 seconds. Out of the total 641 valid trials across the 18 participants and 6 conditions (visual, humming and echolocation in both environments), 39 of these trials were discarded. Of these 39 trials, 28 were discarded due to device/initialisation failure, while the remaining 11 were discarded due to a participant taking over 150 seconds to reach the goal. The distribution of the 11 no-goal-reached trials was as follows: [5,1,1,2,2,0] from the first to sixth trial, respectively.

Stage 1: Visual-only condition A purely visual condition was conducted to familiarise participants with both of the VR environments and to determine a visual baseline for task performance. The VR environments were visible on the *Tango* screen mounted in the *Durovis* headset and participants were instructed to walk toward the star at a natural walking pace while avoiding walls and/or obstacles. The star emitted a ping that served as an audio beacon, and played the completion chime when reached. No other sounds were present. Six trials of the visual-only condition in each environment (maze and obstacle corridor) were conducted with the order of environments randomised across participants.

Stage 2: Spatial audio training Following the visual-only condition, the screen of the *Tango* was blacked out for the remainder of the experimental conditions. Henceforth, only audio information was available to participants via the headphones. In order to acquaint participants with the concept of spatial sound, a training stage was conducted: participants were instructed to walk toward a pinging goal positioned 13 metres directly ahead of them in an obstacle-free environment. Training was continued until participants were able to comfortably localise the goal using the stereo and distance-based volume changes of the pinging sound. This did not prove to be too difficult, with all 18 participants doing one training run, 10 doing a second, and 3 doing a third.

Stage 3: Sonification conditions Six trials of each of the two sonification conditions were then conducted, with a familiarisation period for each done prior to the test trials. The testing protocol was laid out as follows:

- *Echolocation*: a familiarisation period followed by 6 trials in the maze, and another 6 trials in the obstacle corridor
- *Humming*: a familiarisation period followed by 6 trials in the maze, and another 6 trials in the obstacle corridor

Since navigational improvement over trials using audio-only cues was of primary interest, it was assumed that any bias introduced by *seeing* an environment (visual-only condition) before navigating through it audially (sonification conditions) could be ignored. To mitigate other biases, however, the order of the echolocation and humming sets was randomised across participants, as well as the order of the environments tested within each set.

The familiarisation period for each sonification condition was conducted in a simplified version of the obstacle corridor set-up: a single obstacle was located centrally at 4m away from the starting point. In the echolocation training period, this obstacle and the walls received and projected the echolocation click and popping sounds as described in Section 3.2.3. In the humming training period, the object and walls hummed depending on the participants' distance to them as described previously. Familiarisation was considered complete when both the participant and the experimenter agreed that the participant understood the task aims and sonification rules. This proved to be a relatively short period: for echolocation training, all 18 participants did one training run, 13 did a second and 3 did a third, while for humming training, all 18 participants did one training run, 13 did a second, 2 did a third and 1 did a fourth. A two-sample *Kolmogorov-Smirnov* test did not reject the null hypothesis ($p = 1$) that the number of training runs for each sonification came from the same distribution. This suggests no significant difference in training duration between the two sonifications.

3.2.6 Experimental metrics

The 3D location (x, y, z) and 3D head rotation (yaw, pitch, roll) of the participant was recorded from the *Tango* at a rate of 2Hz. From these data, participants' 6D path trajectories through each of the environments in each of the conditions were reconstructed. Data were written to a text file on the *Tango* at the completion of each trial and later processed with *MATLAB* (MathWorks, Natick, MA, USA) using custom scripts. Analysis included looking at the "bird's eye view" path of participants, the length of this path, and participants' instantaneous, mean and peak velocities. Measures designed to probe navigational strategies were additionally investigated through the derivation of i) instantaneous and mean head rotation angle, and ii) deviation distances from obstacles and walls, both of which are described below.

Head rotation angle. Since sound in the VR environments was fully stereo, an indicator of exploration was taken to be the amount of lateral head rotation that participants made whilst navigating the VR environments. Given that the *Tango* was head-mounted, the *Tango*'s rotation corresponded to participants' head rotation. The Euler angle for rotation about the vertical axis (yaw) was extracted at each time point and compared to the instantaneous path angle (that is, the direction of walking calculated using the positional data). The difference

in angle was then transformed such that a left head rotation fell in the range from 0° to −90°, and a right head rotation fell in the range from 0° to +90°. The mean left and right head rotation was calculated per trial, and their absolute values summed to give a total head rotation value between 0° and +180°.

Deviation distances. Detection distance may be defined as the distance at which a participant explicitly identifies an obstacle ahead of them [110]. One proxy for detection distance is a subject's "deviation distance", the distance of a participant to an obstacle when they begin to adjust their trajectory to avoid it. This value was calculated by extrapolating participants' path trajectories, determining if their trajectories intersected with the area occupied by obstacles at each time point, and if so calculating the distance at which participants deviated from this "collision course" [336]. In the corridor environment, this calculation was applied to the five columnar obstacles. In the maze, a midline deviation was used instead. The midline path, considered the optimal path, is the path equidistant to all walls at any given time point. At each time step, the perpendicular distance of a participant's path trajectory to this midline was calculated, and the means of the deviations left and right of the midline computed. In the visual maze condition, participants often "clipped" the corners, and for this reason the corners of the maze were not included in the midline deviation calculation.

3.2.7 Statistical analysis

ANOVA analyses were used to investigate the relationship between experimental conditions across participants. To account for across-trial variability of each participant within a given condition, the mean over their six trials for each measure was calculated, and these values averaged across participants. Additionally, because the maze and obstacle corridor are fundamentally different, the analysis was split accordingly (see Section 3.3.1 and 3.3.2). Results are reported as mean ± standard error, with statistical significance indicated as: $p < .05$ (*), $p < .01$ (**) and $p < .001$ (***). The p-values in pairwise comparisons for multiple comparisons were corrected using the Sidak correction. Greenhouse-Geisser estimates were used to correct for violations in the assumption of sphericity.

3.3 Results

3.3.1 Maze

Participants were able to navigate the maze environment in both visual and sonification conditions. At the start of each trial, participants were placed equidistant from the walls and forward-facing at the starting point. By nature of the maze and sonification set-up, this meant that participants received no audio information if they did not move from this point and gained no benefit from remaining there. We, therefore, formally began each trial when the participant began walking (we took this to be when their walking speed increased above 20% of their peak-to-peak velocity for that trial). Across participants, this start-delay was 2.42s with echolocation and 2.28s with humming, a statistically insignificant difference between the sonifications. Figures 3.4a to 3.4c show the path trajectories of participants for the visual, echolocation and humming conditions, respectively.

Efficiency of maze navigation

We first investigated participants' ability to navigate without making mobility errors (i.e. collisions). Total collisions were tallied over the 6 trials per participant. A significant effect is seen between conditions ($F(2, 34) = 8.906, p = .001$): as expected, the visual condition (0.06 ± 0.06 collisions) outperforms the sonification conditions (echolocation: 4.44 ± 1.12 collisions; humming: 4.78 ± 1.22 collisions).

Beyond number of collisions, the time to goal, path length and mean velocity reveal more about the ease with which participants navigated through the maze (Figures 3.4d to 3.4f). In the visual condition, time to complete (16.24 ± 0.84s), path length (14.91 ± 0.46 m) and mean velocity (0.96 ± 0.05 m/s) were all significantly different from the same metrics in each sonification method: echolocation with 73.81 ± 6.39s ($p < .001$), 22.78 ± 0.90m ($p < .001$) and 0.34 ± 0.02 m/s ($p < .001$), and humming with 60.77 ± 6.15s ($p < .001$), 23.47 ± 0.97m ($p < .001$) and 0.44 ± 0.03 m/s ($p < .001$), respectively. Pairwise comparisons show no significant difference between sonification methods for path length ($p = .833$), however, a significant difference is noted in the time to complete ($p = .049$) and mean velocity ($p = .001$) achieved with echolocation versus humming. The higher humming mean velocities

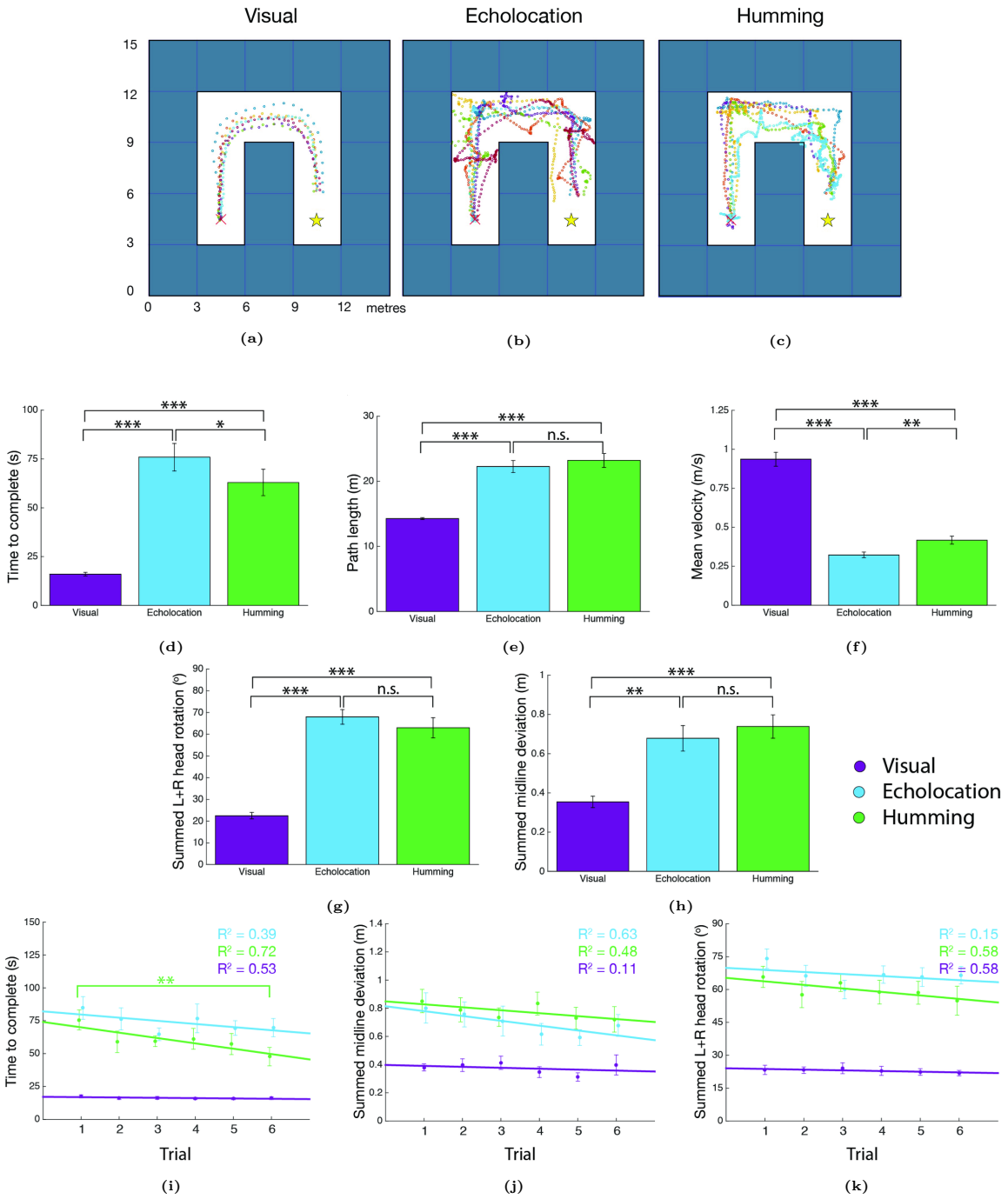


Figure 3.4 Maze navigational behaviours and results. Error bars indicate standard error. Statistical significance indicated as follows: not significant (n.s.), $p < .05$(*), $p < .01$(**) and $p < .001$(***). Colours indicate condition: visual (purple), echolocation (blue) and humming (green).
(a,b,c) Examples of participant trajectories through a maze, using visual-only cues (a), echolocation audio-only cues (b), and humming audio-only cues (c). **(d,e,f,h)** Basic efficiency and strategy of navigation through the maze in the visual, echolocation and humming conditions: time to reach goal in seconds (d), path length in metres (e), mean velocity in metres per second (f), and summed left and right midline deviation in metres (h). **(g)** Summed left and right head rotation in degrees. Head rotation from head straight (0°) direction, comparing visual, echolocation and humming conditions in the maze. **(i,j,k)** Learning curves over 6 trials comparing visual, echolocation and humming conditions in the maze: time to complete in seconds (i), summed left and right midline deviation in metres (j) and summed left and right head rotation in degrees (k). Significance symbols mark significance of effect between first and sixth trial. Linear regression line with R^2-statistic shown for each condition.

suggest that participants moved faster through the humming maze. Despite this, the paths taken in each sonification condition were still approximately 60% longer than in the visual

condition, indicating that participants were not taking the most efficient path and were more exploratory/cautious in their maze navigation. This is supported by the longer times to reach the goal and the lower mean speeds compared to the visual condition (Figures 3.4d to 3.4f).

A further point to draw from Figures 3.4d to 3.4f is the robustness of the visual control: the *Tango* was able to accurately register real-world distances. Given that the blocks of the virtual maze were 3×3m in real-world size and that every path to goal was standardised to 7 block steps in length (with the end of the trial triggered when the participant came within 1.75m of the goal), the midline path length is 16.25m. The mean path length of participants was just under 16.25m, explained by the fact that often participants clipped the corners in the visual condition. Also bolstering the visual control is participants' mean velocity of $\sim$1m/s which is close to the average human walking speed of 1.4m/s [234]. This suggests that participant performance in the visual control approaches normal (non-virtual) visual performance.

Beyond these basic metrics, we also explored more nuanced navigational behaviours of participants. Specifically, we hypothesised that participants adopted a "follow the wall" strategy in the maze, where by maintaining a constant distance to either wall (detected by a constancy in audio cues, whether echolocation or humming) participants would be able to eventually reach the goal. To investigate this, we computed the mean left and right deviations from the central maze midline per trial. Figure 3.4h shows the sum of the absolute left and right midline deviations for the visual (0.37 ± 0.03m), echolocation (0.70 ± 0.06m) and humming (0.78 ± 0.06m) conditions. A significant main effect of condition on the midline deviation is noted ($F(2, 34) = 20.519, p < .001$) with pairwise comparisons between the visual and echolocation condition ($p = .001$), and the visual and humming condition ($p < .001$) suggesting that without sight, participants deviate further from the midline than with sight.

Head rotation

The usefulness of a head rotation was hypothesised to differ for the two sonification conditions. With echolocation, a turn of the head changed the direction of the particle projection. This offered participants new information about the environment within their new FOV. Humming, on the other hand, benefited differently: a head rotation offered a change in stereo of the hum, but only when a participant was within the humming zone of an object. A head rotation

toward an object when outside its humming zone, for example, did not trigger its hum.

The mean left and right head rotation was calculated per trial (0° straight ahead, −90° maximum left rotation, +90° maximum right rotation), and their absolute values summed to find a total head rotation per trial between 0° and 180° (Figure 3.4g). We expected that in the visual condition simple shifts in gaze rather than full head turns would be sufficient for navigating the maze. In line with this hypothesis, the visual condition saw only $22.93 \pm 1.39°$ of total head rotation and this was significantly different ($p < .001$) from that in the echolocation ($66.64 \pm 3.14°$) and humming ($60.12 \pm 4.54°$) conditions. For the reasons described above, we also hypothesised that in the humming condition, a head rotation would be less useful than in the echolocation condition, however, pairwise comparisons showed no significant difference in total head rotation between the two sonification methods ($p = .303$).

Maze learning rates

The efficiency of participants' maze navigation improved over trials and since with each trial the maze path was randomised, the improved proficiency suggests that participants were increasing their understanding of each sonification method. To show this, a one-way repeated-measures ANOVA was performed on trial within each condition. Figures 3.4i to 3.4k show the learning curves over 6 trials for time to complete, summed midline deviation, and summed head rotation. Improvements were not expected in the visual condition and this was indeed the case, with no significant effect of trial observed for any of the three metrics in the visual condition. In the echolocation condition, no significant ($p = .145$) improvement was seen in time to complete (trial 1: 84.69 ± 8.65s to trial 6: 69.69 ± 6.93s) nor for summed midline deviation ($p = .386$) or summed head rotation ($p = .051$). However, in the humming condition there was a significant ($p = .008$) improvement over the six trials in time to complete (trial 1: 75.56 ± 7.61s to trial 6: 47.7 ± 7.00s). Summed midline deviation and summed head rotation did not significantly change in the humming condition over trials. Overall, even if statistical significance was not always achieved, all metrics showed improving trends for the sonification conditions.

3.3.2 Obstacle corridor

Compared to the maze, the obstacle corridor offered an environment much closer to a possible real-world scenario. The obstacles, by way of the randomness in their positioning, made for a far less predictable path to the goal compared to the mazes, with multiple possible paths available. As with the maze, each trial was formally started when participants' walking speed increased above 20% of their peak-to-peak velocity for that trial. Across participants, this start-delay was 2.07s with echolocation and 2.18s with humming, again a statistically insignificant difference. For illustration purposes, Figures 3.5a to 3.5c show the path trajectories of participants in selected corridor arrangements.

Efficiency of obstacle corridor navigation

As with the maze, the number of collisions with both walls and obstacles in the corridor was used as an indicator of navigational efficiency (Figure 3.5e). Like the maze, condition had a significant main effect on total collisions ($F(1.622, 27.566) = 21.36, p < .001$; sphericity $\chi^2(2) = 6.342, p = .042$ corrected with Huynh-Feldt estimates) with a significant difference observed between the visual condition (0.17 ± 0.09 collisions) and each sonification method (echolocation: $8.33 \pm 1.64, p < .001$; humming: $15.50 \pm 2.37, p < .001$). No statistically significant difference was noted between sonification methods ($p = .067$) however, by inspection it can be seen that a mean total of 7 more total collisions over the six trials occurred in the humming obstacle corridor compared with the echolocating obstacle corridor (Figure 3.5e). This suggests that humming was not as effective at signalling an imminent collision.

Time to complete, path length and mean velocity revealed further dynamics of the obstacle corridor navigation (Figures 3.5f to 3.5h). Condition had a significant effect on all three metrics due to the strong performance of participants in the visual condition. For time to complete (Figure 3.5f), the visual condition, taking 13.31 ± 0.82s, was significantly faster than the echolocation (101.89 ± 12.68s, $p < .001$) and humming (84.74 ± 9.09s, $p < .001$) conditions. A pairwise comparison between echolocation and humming showed no significant difference ($p = .537$). Path length yielded no interesting differences other than a shorter path length in the visual condition (12.12 ± 0.12m). Mean velocity was significantly affected by condition ($F(1.308, 22.239) = 207.73, p < .001$): there were significant pairwise differences

between the visual condition (0.98 ± 0.05m/s) and each sonification condition (echolocation: 0.25 ± 0.02m/s, $p < .001$; humming: 0.32 ± 0.03m/s, $p < .001$), respectively. Furthermore, there was a significant difference between the mean velocities of the two sonification conditions ($p = .008$), indicating that, like the maze, participants moved at higher speeds with humming.

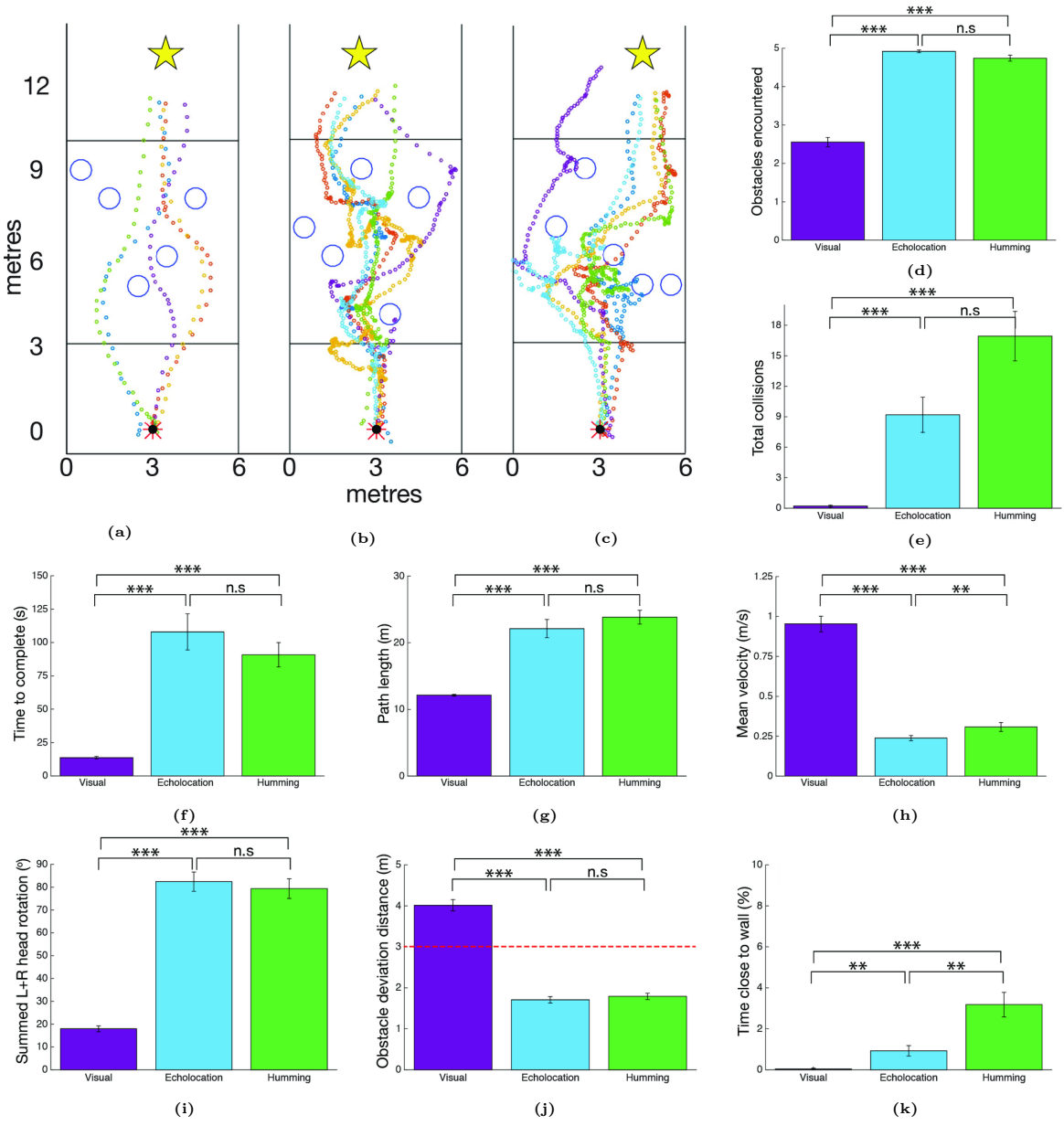


Figure 3.5 Obstacle corridor navigational behaviours and results. Error bars indicate standard error. Statistical significance indicated as follows: not significant (n.s.), $p < .05$(*), $p < .01$(**) and $p < .001$(***). Colours indicate condition: visual (purple), echolocation (blue) and humming (green).
(a,b,c) Examples of participant trajectories through particular obstacle corridor arrangements using visual-only cues (a), echolocation audio-only cues (b), and humming audio-only cues (c). **(d,j,k)** Active interaction with obstacles in the corridor in the visual, echolocation and humming conditions: number of obstacles encountered at least once per trial (d), obstacle deviation distance (j), and percentage of obstacle patch time per trial spent < 0.4m from left or right wall (k). Red dashed line shows 3m echolocation and humming threshold. **(e,f,g,h)** Basic efficiency of navigation through the obstacle corridor in the visual, echolocation and humming conditions: total number of collisions over 6 trials per participant (e), time to reach goal in seconds (f), path length in metres (g), mean velocity in metres per second (h). **(i)** Summed left and right head rotation from head straight (0°) direction in the corridor in the visual, echolocation and humming conditions.

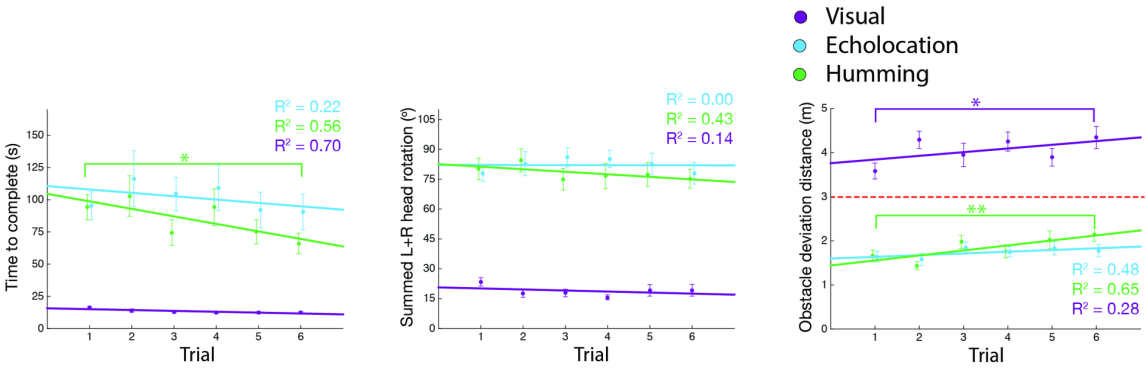


Figure 3.5 Obstacle corridor navigational behaviours and results (continued). Error bars indicate standard error. Statistical significance indicated as follows: not significant (n.s.), $p < .05$(*), $p < .01$(**) and $p < .001$(***). Colours indicate condition: visual (purple), echolocation (blue) and humming (green).
(l,m,n) Learning curves over 6 trials comparing visual, echolocation and humming conditions in the obstacle corridor: time to complete in seconds (l), summed left and right head rotation in degrees (m), obstacle deviation distance in metres (n). Red dashed line shows 3m echolocation and humming threshold. Significance symbols mark significance of effect between first and sixth trial. Linear regression line with R^2-statistic shown for each condition.

The deviation distance was computed for each columnar obstacle encountered in the obstacle corridor (see Section 3.2.6). Condition had a significant main effect on the obstacle deviation distance ($F(2, 34) = 172.712, p < .001$) with participants deviating from obstacles at 4.06 ± 0.13m in the visual condition, but only at 1.73 ± 0.08m and 1.82 ± 0.07m in the echolocation and humming conditions, respectively (Figure 3.5j). Pairwise comparisons show the difference between the visual and each sonification condition is significant ($p < .001$) but not between echolocation and humming directly ($p = .753$). As expected, with sight, participants gave obstacles a wide berth, whereas with only audio, participants moved closer to obstacles, which also resulted in a higher number of collisions. It is worth mentioning that obstacles only began to echolocate/hum at 3m which placed an upper limit on the achievable deviation distances in the sonification conditions (shown by the dotted red line in Figure 3.5j).

The obstacle deviation metric does not account for the portion of time per trial spent actively encountering the five obstacles. Given their random positioning, it was possible that a participant spent more time encountering obstacles in one trial than in another. In addition to this, the identifiable deep hum of the walls could lead participants to avoid the majority of the obstacle segment by simply walking against either one of the walls. To quantify this Figure 3.5d shows the number of obstacles (of the total 5) actively encountered per trial. An obstacle was considered to be actively encountered if a participant walked within the audio zone/threshold of that object. Repeat encounters were not factored in, although were

likely common. In the visual condition, only 2.54 ± 0.13 obstacles were encountered, whereas 4.88 ± 0.42 and 4.72 ± 0.07 obstacles were encountered in the echolocation and humming conditions, respectively. Additionally, since, in the corridor, all obstacles were positioned within a 6×7m obstacle patch (see Figures 3.1b and 3.1e), we could compute the percentage of the time per trial spent in the obstacle patch in which the participant was less than 0.4m from the left or right wall (Figure 3.5k). All conditions show values of less than 3% (visual: $0.03 \pm 0.03\%$; echolocation: $0.92 \pm 0.24\%$; humming: $2.95 \pm 0.56\%$). These results together suggest that in each sonification condition participants spent a large portion of the time per trial in the obstacle segment actively encountering and avoiding obstacles.

Head rotation

As with the maze, it was also relevant to look at the use of head rotation in the obstacle corridor. Condition had an effect on summed left and right head rotation ($F(2, 34) = 128.32, p < .001$) with the visual condition using a significantly smaller amount of head rotation ($18.82 \pm 1.63°$) compared to the echolocation ($82.19 \pm 3.97°$) and humming ($77.70 \pm 4.33°$) conditions. The difference between the visual and each sonification condition was significant ($p < .001$) but was not significant between sonification conditions ($p = .754$) (Figure 3.5i).

Obstacle corridor learning rates

Given the corridor's similarity to navigating a cluttered indoor environment, it was particularly interesting to analyse learning rates in this scenario. Just as with the maze, a one-way repeated-measures ANOVA was performed on trial within each condition.

Figure 3.5l shows that while the completion time remained largely consistent from trial 1 to 6 in the visual (16.11 ± 1.24s to 12.44 ± 0.75s) and echolocation (95.25 ± 10.91s to 90.53 ± 13.91s) condition, it dropped some 30 seconds for the humming condition (94.14 ± 9.76 to 66.00 ± 7.95s; $p = .029$). Faster navigation after 6 test trials of humming suggests learning and improved efficiency. Trial had no significant main effect on summed left and right head rotation within each condition, with only a slightly reduced use of humming head rotation observed over the six trials. This suggests that within 6 trials, participants were not varying their use of head rotation, for example as a strategy for navigation (Figure 3.5m). Obstacle deviation distance (Figure 3.5n) also showed significant improvement with trial in the humming

condition ($F(5, 75) = 4.021, p = .003$), but was not significant in the echolocation condition ($F(5, 80) = 0.731, p = .602$). Both, however, showed overall increasing trends in deviation distance from trial 1 to 6 ($R^2 = 0.652$ and $R^2 = 0.482$ for humming and echolocation, respectively) indicating that with more practice, participants were able to more efficiently "weave" within safer distances between the obstacles in the corridor.

3.3.3 Qualitative experience

A final component enhancing the quantitative results presented above is the qualitative feedback received from participants following their two testing sessions. Since the distribution of SSD naivety and gaming experience across participants was largely uniform, it was not expected to be correlated with participant performance. Beyond this, it was informative to look at participants' overall experience with each sonification method, and its ease or intuitiveness in each environment. Figure 3.6 show the ratings received from 14 out of 18 participants when asked about the intuitiveness of each method. The maze had a largely even spread across ratings from "easy" to "challenging even with training", with 36% and 43% of participants finding it easy for echolocation and humming, respectively. A common comment was that the mazes were more basic since one simply needed to walk in straight lines. One particular participant whose performance was notably poor in the mazes across both sonification conditions commented that her spatial memory and internal orientation were poor, and this made it difficult for her to spatially connect the corridors she was exploring. As a result, after periods of hesitation, she often began to navigate down a corridor from which she had just come. The obstacle corridor, on the other hand, had a more distinct qualitative split. Echolocation was perceived as manageable with training by 79% of participants, while humming was perceived as such by only 36% of participants, with 50% rating it as challenging even with training. The *perceived* ease/intuitiveness of the sonification methods, therefore, contradicts participants' actual navigational performance in the corridor environment. These results also suggest that of the two environments, the maze was easier to navigate, likely because of its limited layouts. It should be noted that no participant rated either sonification method as "impossible".

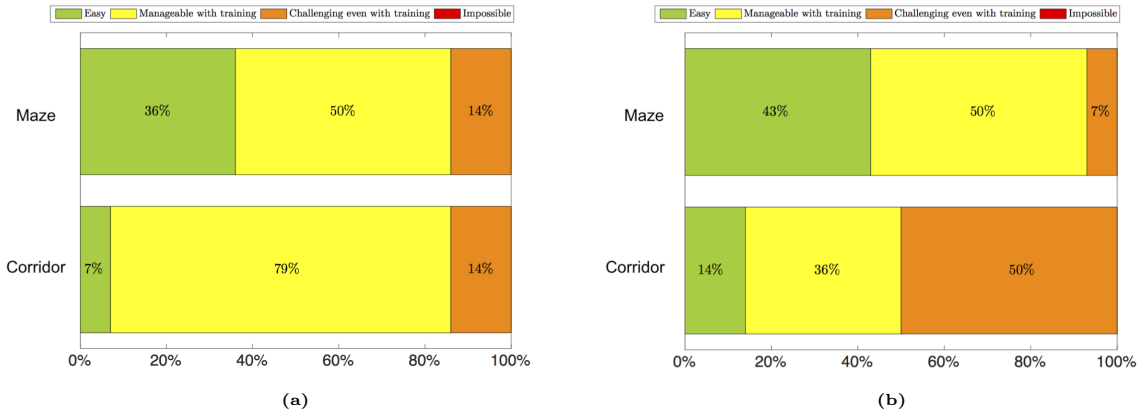


Figure 3.6 Qualitative participant feedback
(a) *'How easy/intuitive did you find the echolocation sonification in each environment?"*
(b) *'How easy/intuitive did you find the humming sonification in each environment?'*
Green: easy. Yellow: manageable with training. Orange: challenging even with training. Red: impossible.

3.4 Discussion

The principal outcome of this work is that participants without access to sight were able to navigate a virtual space and detect and avoid obstacles using our two novel sensory substitution sonification approaches. Furthermore, they learned to do so with a small amount of training (less than 3 hours) and minimal instruction. Our results are promising in the view of using these sonification strategies to help people with visual impairments to navigate real world environments. Moreover, the self-locomotion-guided VR environment we developed to test the utility of our stereosonic vision mappings provides a realistic yet safe, controlled, and flexible paradigm for testing navigation and mobility skills, and allows for the automatic extraction of many useful performance metrics.

3.4.1 Development of experimental set-up to test sonification mappings

In this work, we successfully built a visual-to-audio SSD using a head-mounted *Tango* tablet and a pair of stereo headphones. We established an experimental testbed for the device by using standardised metrics to analyse participants' detailed navigational behaviour through randomly generated VR environments. Participants' navigational data was obtained by extracting the 3D positional and 3D rotational tracks of a head-mounted *Tango* tablet. These tracks corresponded to participants' body position and head rotation at fine-grained time

steps over the length of each trial. This testbed allowed us to successfully measure the utility of two sonification methods for the task of spatial navigation.

Our VR experimental paradigm provides important improvements over keyboard- and joystick-based paradigms which to date have been ubiquitous in studies on spatial navigation [185, 295], capturing the important contribution of proprioceptive feedback of walking in VR navigation and mobility [55, 333]. Importantly, this was achieved using a portable, inexpensive tablet which can be used in stand-alone fashion in any sufficiently large open space, and did not require the dedicated infrastructure of previously published locomotion-controlled auditory VR [344, 345, 346, 359]. This makes locomotion-controlled, multi-sensory immersive VR accessible to smaller, independent research groups, or those who do not use immersive VR as their principal methodology.

Like other VR paradigms, our paradigm lends itself well to the automatic and randomised generation of simulated environments, and avoids real-world mobility hazards to participants. Moreover, the continuous tracking of real walking behaviour in the virtual world allows for the extraction of measures that capture the dynamics of walking and navigating, such as path length, speed, head rotation, and deviation distances from objects. In the current study, we generated simplified models of real-world scenarios as a necessary first step. The paradigm's flexibility will enable the testing of audio-only navigation in more complex virtual environments in the future, for example multi-room scenarios with multiple object types.

While we have used this navigation testing paradigm to investigate the utility of our sonification mappings, the paradigm would be equally suited to testing navigation using existing SSDs. Directional spatial-to-audio SSDs have already been tested in VR environments [59, 209, 211], and it would only require the use of a tracking device such as the *Tango* to convert this into a locomotion-based VR paradigm. To test realistic echolocation-type SSDs [144, 146, 228, 309], the most important additional requirement would be the accurate modelling of acoustics in the VR environment to accurately render the temporal and spectral information present in real-world echoes [41], something that has already been achieved [268, 348]. Finally, to test navigation in SSDs that take 2D images as input, whether auditory such as the *vOICe* or EyeMusic [1, 224], or tactile [33, 34, 35, 127, 244], the image of the VR environment as rendered on screen can simply be used as a substitute to the regular video input of such SSDs.

Additionally, the wealth of data on navigational dynamics which can be automatically extracted from the tracking device, and its lightweight, low-cost control via real locomotion opens up other exciting use cases. We believe this approach will not only prove useful for studies of mobility and sensory substitution, but also for studies of navigation and spatial cognition where a contribution of proprioceptive information is relevant; for instance, in immersive studies of spatial learning and memory [221, 298] and how they may be affected by clinical conditions such as Alzheimer's Disease [73, 101] or depression [114].

3.4.2 Comparison of navigational behaviour between sonifications

Visual/sighted control Participants' walking velocity in the visual control approached 1m/s, close to the average human walking speed of 1.4m/s [234]. These measurements verify the robustness of the visual control against which sonification performance is compared.

Simulated echolocation The emitted click and distance-dependent "pops" of simulated echolocation necessitated the following: firstly, participants needed to sample their surroundings by orienting their head, and secondly, using the returning waves of pops, participants needed to stitch their audio-based representations together in order to construct a representation of the full scene. For these reasons, echolocation differs from the passive humming sonification method, and is more akin to looking and "seeing". By comparing participant performance, however, the simulated echolocation sonification appeared to be less intuitive than the humming sonification. This was supported by overall slower echolocation navigation speeds in both the maze and the obstacle corridor, indicative of participants being more hesitant. This could have been the result of a number of factors. Firstly, compared to humming, there is a higher cognitive load in processing the click-pops of echolocation. Participants reported that the click-pop delays were difficult to interpret, however, with more training, they may have been able to draw more information from them. A second reason for higher hesitancy with simulated echolocation is the limitation imposed by its update speed. The click rate (every 1.75s) placed an upper bound on the speed at which an up-to-date acoustic snapshot of the environment is received. It is possible that with a higher click frequency, participants would have been able to move more quickly, however, this comes at the risk of losing the ability to temporally discriminate pops if they scatter too quickly.

Across both environments a unifying strategy for simulated echolocation was hypothesised to be one in which the "quietest" path or the path of least resistance was followed. In other words, if no obstacles were directly in front of a participant, then the returning pop sounds were few and so the soundscape relatively quiet. In this way, a path could be weaved between obstacles. Consistent with this, there was a trend towards fewer collisions in the echolocation condition compared to humming (by 46%, p=0.067).

Distance-dependent hum volume modulation Participant task performance indicated that encoding objects' spatial distance using volume-modulated humming was an intuitive sonification method, more so than simulated echolocation. In both the maze and obstacle corridor, navigation using the humming sonification was faster compared to simulated echolocation along equivalent path lengths. This suggests higher confidence levels with the humming technique. Looking at the more challenging obstacle avoidance in the corridor environment, with humming, walking speeds were on average 27% slower than those in the maze. The speed *difference* between the two environments using humming was greater than the difference using echolocation. Despite this, navigation with humming was faster than with echolocation by 0.7m/s in the obstacle corridor.

These results together suggest that of the two sonification methods, humming was quicker and easier to learn than echolocation. Humming, however, is inherently limited by the amount of information it can represent. A hum, based on its volume level conveys distance to an obstacle, and based on its stereo components conveys information about its spatial position. On the other hand, a hum offers no information about object shape. Furthermore, absolute hum volume, necessary for participants to calibrate their volume-distance correspondence, may be difficult to immediately discern. Critically too, the presence of multiple obstacles, resulting in the overlapping of humming zones, was reported to make it challenging to distinguish space between obstacles or clear paths ahead. We believe, therefore, that simulated echolocation has greater potential for representing more complex and detailed environments for the task of navigation. Pointing toward this, participants rated echolocation as easier than humming in the more complex corridor environment in their follow-up qualitative feedback, despite their poorer navigational performance here. One possible explanation for this difference could

be that participants are positively projecting their capacity to learn a richer sonification for more complex environments in the future (i.e. perceiving the method as easier in the present because of their predicted ability to improve or master the method in the future).

To enable the estimation of distance with a humming approach, future work could include modulating the hums in ways that would make judging absolute distance easier for the human auditory system. Hums could be pulsed with a frequency proportional to obstacle distance, or modulated in pitch or with a filter envelope as a participant approaches an obstacle. The presence of multiple objects as overlapping sources of hums also poses a challenge—it is easy to overwhelm a participant with a cacophony of sound. Here, an intelligent way of identifying the most mobility-relevant objects in the scene might be useful so that sounds from other less relevant objects or objects further away may be suppressed.

3.4.3 Observed learning effects in stereosonic navigation

For each sonification condition, participants received minimal instruction prior to the trials and no feedback during the trials. As part of the minimal instruction, participants underwent a training period with the device—this took place in a drastically simplified version of the obstacle corridor (with only one obstacle) and comprised only a small number of trials. The training period was primarily intended to familiarise participants with the headset and experimental procedure rather than the sonification methods themselves. The brevity of the instruction and training period was also intended to provide a fair baseline for comparison to other visual-to-audio SSDs in the task of navigation.

With this in mind, it is therefore impressive that learning effects were observed for the humming sonification over only 6 trials. The clearest learning effect was seen in the decrease in time taken to reach the goal: improvements of 24.2s and 28.1s were achieved in the maze and the obstacle corridor, respectively. Furthermore, in the obstacle corridor environment, obstacle deviation distances improved by $\sim$0.5m using the humming sonification. No significant effect was seen for echolocation, though the direction of the effects was consistent with improvement.

While the sonification performance does not reach performance when sight is available, the results speak to the possibility that with further training this performance difference could

be reduced. More experience with each sonification method as well as a more detailed understanding of the sonifications' formulations (in particular simulated echolocation which shows promise in representing more complex environments) may allow participants to narrow this performance gap with practice.

3.4.4 Considerations for real-life application

From the current study it is too bold to claim that the two sonification methods explored here encode visual spaces in sufficient detail for seamless spatial navigation. Our results, however, do suggest that participants were able to obtain a 3D spatial awareness of their virtual surroundings. The fact that participants were able to spatially place themselves and manoeuvre through two distinct VR environments, avoiding walls and simple obstacles, and also improve their navigational performance over time is indicative of this. We believe that it is, therefore, interesting to consider the potential for these strategies as a real-life sensory substitution device for visually impaired people.

To this effect, it will be necessary to investigate how VI participants specifically adapt to these different sensory substitution approaches, and how the approaches complement or interact with their existing navigation strategies. On one hand, VI participants may outperform our normally-sighted participants, as it has been shown that early-blind individuals in particular may display superior performance in auditory tasks like pitch discrimination [112] and spatial sound localisation [113, 187, 288, 343]. In addition, they may be more experienced in using such spatial properties of audio to navigate and detect obstacles in their environment [318, 361]. With regard to our hypothesised audio-based navigational strategies, VI individuals might in fact be more inclined than our sighted participants to use head rotation for echolocation as they will know from experience that this helps to sample the auditory environment.

Conversely, in particular those VI participants who have been blind from birth or a very young age may have altered representations of locomotor space that could interfere with performance [300, 339]. This may reflect a greater tendency to rely on an egocentric rather than allocentric spatial reference frames [258], which is thought to be particularly pronounced for larger-scale spatial representations such as those relevant to locomotor activity [145]. There is evidence that such altered spatial representations may negate the advantages of superior

auditory discrimination in sound localisation in larger spaces [340]. How well the performance of our sighted participants correlates to the performance of VI individuals is therefore an important question to address in future work.

Moreover, in order to translate our sonification mappings to the real world, it will be necessary to implement the current virtual world sonification in real 3D space, and ultimately transfer the mappings onto a stand-alone wearable SSD. By pairing an RGB camera with a depth sensor and applying computer vision algorithms to the incoming video streams, it is possible to extract 3D spatial and semantic information of the real-world environment. Object detection methods [281, 282, 347] can localise the presence/absence of objects, and object recognition methods [28, 153] can identify the *types* or classes of these objects (for example, desks, chairs, people, but also more abstract categories like walls, floors and ceilings). The incoming depth maps (from the depth sensor) will provide real-world distance estimates from the camera to densely scattered points in the environment. If the camera were mounted on the user's head, these two streams of information would allow for the building and updating of a 3D egocentric representation of the user's environment which could subsequently be converted into its corresponding soundscape. For simulated echolocation, the distances provided by the depth maps would be synonymous with the distances the projected particles travel. In a similar way, these distances could be used to modulate the volume/pulse frequency of the humming sounds. Central to this process will be the *selection* of objects which are relevant to mobility so as not to overwhelm the user with irrelevant audio cues.

Extending this, recent years have seen the proliferation of methods for the fusion of global 3D maps of environments [240, 274, 341, 352, 380]. These methods allow 3D maps of whole environments to be built and updated on-the-fly as the user moves within them. Access to such a map would offer the ability to integrate global-level information of an environment (which may not be available from the user's local camera view) into the sonification methods, for example, sonifying objects behind the user, objects that dynamically move in and out of the user's FOV, or objects at far distances or in occluded positions.

Implementing these real-world sonification methods into a wearable SSD presents several challenges. The system must be able to capture, build a representation of the user's environment, and sonify the relevant parts of the scene all in near real-time. Furthermore, the device

must be light in both weight and power consumption, placing significant restrictions on the on-board processing of the device. Real-time object detection and recognition methods already exist [153, 281, 297, 379], however, and we can expect increasingly reliable and light-weight solutions to follow in the near future.

A crucial consideration in building a wearable audio-based SSD for VI people is the fact that the resultant soundscapes must be delivered in a way that does not mask or impede the ambient sounds on which the user may already heavily rely to perceive their surroundings and remain oriented. Bone-conduction headphones, which transmit sounds through vibrating pads placed on the jaw bones, have been shown to allow equivalent performance in the spatial localisation of audio cues when compared with standard stereo headphones [208], and have been successfully used for auditory spatial navigation through VR environments [344]. This suggests that they do have the potential to accurately represent 3D spatial audio for navigation within a real-world visual-to-audio SSD.

Finally, it is worth noting that another successful real-life implementation of the sonification mappings presented in this study may yet be entirely virtual. Since the advent of VR environments, an important question has been how to make such environments accessible to the VI community, for example for gaming, and other experiences [354]. The focus of VR system development has largely been geared toward a better *visual* experience, but there have been notable exceptions focusing on the non-sighted population [69, 147, 209]. Our auditory VR navigation paradigm, therefore, could add to the available methods for making virtual worlds more accessible to those without sight.

3.5 Conclusion

The current study has explored the feasibility of two novel visual-to-audio mappings for the task of spatial navigation: simulated echolocation and distance-dependent volume modulation of hums. Both sonification methods were implemented and tested in two virtual reality environments using a head-mounted 3D motion-tracking device. The device created an immersive virtual world in which participants were able to physically walk around virtual scenes. To our knowledge, this is the first work to make use of such an experimental paradigm for the task of spatial navigation, and we believe this approach will be of interest to others working in the dynamics of mobility and spatial navigation. Our key findings showed that participants were able to navigate using both of the proposed sonification methods, with task completion time, velocity, number of collisions and other more nuanced navigational behaviours improving over the course of only six trials. Importantly, these improvements were achieved with participants receiving minimal instruction and a very short training period. This bodes well for future scenarios in which participants have more experience with the sonification methods. Although audio-only navigational performance clearly remains below that of visual navigation, our findings suggest that the sonification methods generated an awareness of nearby 3D spatial surroundings. This is a promising step in the direction of enabling independent mobility for visually impaired individuals through sensory substitution.

Acknowledgements DM was supported by the Clarendon Scholarship. SLH and JJR were supported by a private donor via the University of Oxford (grant *EVE: Expanded Visual Enhancement*).

Chapter 4

FlipDial: A Generative Model for Two-Way Visual Dialogue

Daniela Massiceti N. Siddharth Puneet K. Dokania Philip H.S. Torr

Engineering Science, University of Oxford

Abstract

We present FlipDial, a generative model for visual dialogue that simultaneously plays the role of both participants in a visually-grounded dialogue. Given context in the form of an image and an associated caption summarising the contents of the image, FlipDial learns *both* to answer questions and put forward questions, capable of generating entire sequences of dialogue (question-answer pairs) which are diverse and relevant to the image. To do this, FlipDial relies on a simple but surprisingly powerful idea: it uses convolutional neural networks (CNNs) to encode entire dialogues directly, implicitly capturing dialogue context, and conditional VAEs to learn the generative model. FlipDial outperforms the state-of-the-art model in the sequential answering task (one-way visual dialogue (1VD)) on the *VisDial* dataset by 5 points in Mean Rank using the generated answers. We are the first to extend this paradigm to full two-way visual dialogue (2VD), where our model is capable of generating both questions *and* answers in sequence based on a visual input, for which we propose a set of novel evaluation measures and metrics.

Published in the *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*[1] [presented as an oral].

[1] https://bit.ly/2XHNuIA

4.1 Introduction

A fundamental characteristic of a good human-computer interaction (HCI) system is its ability to effectively acquire *and* disseminate knowledge about the tasks and environments in which it is involved. A particular subclass of such systems, natural-language-driven conversational agents such as *Alexa* [10] and *Siri* [25], have seen great success in a number of well-defined language-driven tasks. Even such widely adopted systems suffer, however, when exposed to less circumscribed, more free-form situations. Ultimately, an implicit requirement for the wide-scale success of such systems is the effective understanding of the environments and goals of the user—a difficult problem as it involves getting to grips with a variety of sub-problems (semantics, grounding, long-range dependencies) each of which are extremely difficult problems in themselves. One avenue to ameliorate such issues is the incorporation of *visual* context to help explicitly ground the language used—providing a domain in which knowledge can be anchored and extracted from. Conversely, this also provides a way in which language can be used to characterise visual information in richer terms, for example with sentences describing salient features in the image (referred to as "captioning") [157, 163].

In recent years, there has been considerable interest in visually-guided language generation in the form of VQA [24] and subsequently visual dialogue [77], both involving the task of *answering* questions in the context of an image. In the particular case of visual dialogue, along with the image, previously seen questions and answers (i.e. the dialogue history) are also accepted, and a relevant answer at the current time produced. We refer to this one-sided or answer-only form of visual dialogue as one-way visual dialogue (1VD). Inspired by these models and aiming to extend their capabilities, we establish the task of two-way visual dialogue (2VD) whereby an agent must be capable of acting as both the questioner and the answerer.

Our motivation for this is simple—autonomous agents need to be able to both ask questions *and* answer them, often interchangeably, rather do either one exclusively. For example, a vision-based home-assistant like *Alexa* may need to ask questions based on her visual input (*'There is no toilet paper left. Would you like me to order more?'*) but may also need to answer questions asked by humans (*'Did you order the two-ply toilet paper?'*). The same question-answer capability is true for other applications. For example, with aids for visually

impaired people, a user may need the answer to *'Where is the tea and kettle?'*, but the system may equally need to query *'Are you looking for an Earl Grey or Rooibos teabag?'* to resolve potential ambiguities.

We take one step toward this broad research goal with FlipDial, a generative model capable of both 1vd and 2vd. The generative aspect of our model is served by using a conditional variational auto-encoder (cvae), a framework for learning deep conditional generative models while simultaneously amortising the cost of inference in such models over the dataset [171, 310]. Furthermore, inspired by the recent success of convolutional neural networks (cnns) in language generation and prediction tasks [141, 162, 267], we explore the use of cnns on sequences of sequences (i.e. a dialogue) to *implicitly* capture all sequential dependences through the model. Demonstrating the surprising effectiveness of this approach, we show sets of sensible and diverse answer generations for the 1vd task in Figure 4.1.

Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5	z_6
How old is the man?	I can't see his face, but maybe mid or late twenties	He looks thirty	Twenties	I can't tell	Thirtys	I say teenager	Twentys
What race is the man?	White	White	White	In five-thirty away	I can't see	Caucasian	He looks white
Is he wearing a hat?	I don't see his head	No	Yes	Yes	Yes	No	No
Is he wearing a shirt?	Yes	Yes	Yes	No	No	Yes	It
What color?	Dark grey	Grey	Grey	Black	White	Black	Black
What gender is the baby?	I can't really tell, maybe a girl	I	Boy	I	Male	I	I
What is the baby wearing?	A bib	Shirt	T-shirt has pants	Shirt and shirt	Shirt and shirt	Looks like white	Shirt and shirt
What color is the remote?	White and black	White	Silver	Silver	It is white and black	White	Black
Is the chair wood?	No, it's leather	No	No	Yes	No it's a chair	No	Yes
What color is the chair?	Like a light burgundy	Brown	White has white checkered	A light brown	Gray	Brown with white texture	Gray

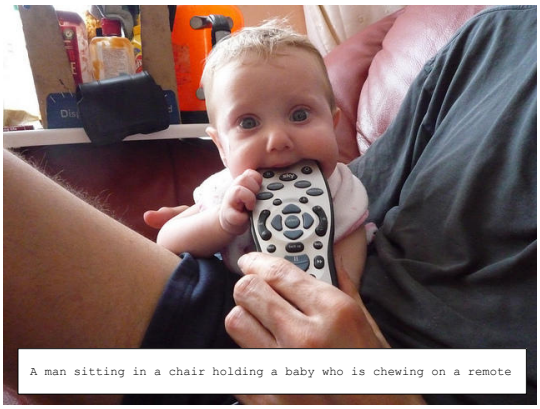


Figure 4.1 Diverse answers generated by FlipDial in the one-way visual dialogue (1vd) task. For a given time step (row), each column shows a *generated* answer to the current question. Answers are obtained by decoding a latent $\boldsymbol{z}_i$ sampled from a prior conditioned on the image, caption and dialogue history up until that time step.

We here provide a brief treatment of works related to visual dialogue. We reserve a thorough comparison to Das et al. [77] for Section 4.4.3, noting that our fully-generative convolutional extension of their model outperforms their state-of-the-art (sota) results on the answering of sequential visual-based questions (1vd). In another work, Das et al. [78] present a reinforcement learning-based model to do 1vd, where they instantiate two separate agents, one each for questioning and answering. Crucially, the two agents are given *different* information—with one (QBot) given the caption, and the other (ABot) given the image. While they formulate the interesting task of performing image retrieval from natural language descriptions, their

set-up is fundamentally different from having a single agent perform both roles. Jain et al. [152] explore a complementary task to VQA where the goal is instead to generate a (diverse) set of relevant *questions* given an image. In their case, however, there is no dependence on a history of questions and answers. Finally, we note that Zhao et al. [381] employ a similar model structure to ours, using a CVAE to model dialogue, but condition their model on discourse-based constraints for a purely linguistic (rather than visuo-linguistic) task. The tasks we target, our architecture (CNNs), and the dataset and metrics we employ are distinct.

Our primary contributions in this work are therefore:

1. A fully-generative, convolutional framework for visual dialogue that outperforms SOTA models on sequential question answering (1VD) using the generated answers, and establishes a baseline in the challenging two-way visual dialogue task (2VD).
2. Evaluation using the *predicted* (not ground-truth) dialogue—essential for real-world conversational agents.
3. Novel evaluation metrics for generative models of two-way visual dialogue to quantify answer-generation quality, question relevance, and the models's generative capacity.

4.2 Preliminaries

Here we present a brief treatment of the preliminaries for deep generative models—a conglomerate of deep neural networks and generative models. In particular, we discuss the VAE [171] which, given a dataset $\mathcal{X}$ with elements $\boldsymbol{x} \in \mathcal{X}$, simultaneously learns i) a variational approximation $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$[2] to the unknown posterior distribution $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ for latent variable $\boldsymbol{z}$, and ii) a generative model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$ over data and latent variables. These are both highly attractive prospects as the ability to approximate the posterior distribution helps *amortise* inference for any given data point $\boldsymbol{x}$ over the entire dataset $\mathcal{X}$, and learning a generative model helps effectively capture the underlying abstractions in the data. Learning in this model is achieved through a unified objective, involving the marginal likelihood (or *evidence*)

[2]Following the literature, the terms recognition model or inference network may also be used to refer to the posterior variational approximation.

of the data, namely:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}) &= \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}, \boldsymbol{z}) - \log q_\phi(\boldsymbol{z} \mid \boldsymbol{x})] \\ &\geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| p_\theta(\boldsymbol{z})) \end{aligned} \tag{4.2.1}$$

The unknown true posterior $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ in the first KL divergence is intractable to compute making the objective difficult to optimise directly. Rather a lower-bound of the marginal log-likelihood $\log p_\theta(\boldsymbol{x})$, referred to as the evidence lower bound (ELBO), is maximised instead.

By introducing a condition variable $\boldsymbol{y}$, we capture a *conditional* posterior approximation $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$ and a *conditional* generative model $p_\theta(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{y})$, thus deriving the CVAE [310]. Similar to Eq. 4.2.1, the conditional ELBO is:

$$\log p_\theta(\boldsymbol{x} \mid \boldsymbol{y}) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y})] - \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{y})) \tag{4.2.2}$$

where the first term is referred to as the reconstruction or negative cross entropy (CE) term, and the second, the regularisation or KL divergence term. Here too, similar to the VAE, $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$ and $p_\theta(\boldsymbol{z} \mid \boldsymbol{y})$ are typically taken to be isotropic multivariate Gaussian distributions, whose parameters $(\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2)$ and $(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)$ are provided by deep neural networks (DNNs) with parameters ϕ and θ, respectively. The generative model likelihood $p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y})$, whose form varies depending on the data type – Gaussian or Laplace for images and Categorical for language models—is also parametrised similarly. In this work, we employ the CVAE model for the task of eliciting dialogue *given* contextual information from vision (images) and language (captions and dialogue history).

4.3 Generative models for visual dialogue

In applying deep generative models to visual dialogue, we begin by characterising a preliminary step toward it, visual question-answering (VQA). In VQA, the goal is to answer a single question in the context of a visual cue, typically an image. The primary goal for such a model is to ensure that the elicited answer conforms to a stronger notion of relevance than simply answering the given question—it must also relate to the visual cue provided. This notion can

be extended to one-way visual dialogue (1VD) which we define as the task of answering a *sequence* of questions contextualised by an image (and a short caption describing its contents), similar to [77]. Being able to exclusively answer questions, however, is not fully encompassing of true conversational agents. We therefore extend 1VD to the more general and realistic task of two-way visual dialogue (2VD). Here the model must elicit not just answers given questions, but questions given answers as well—generating *both* components of a dialogue, contextualised by the given image and caption. Generative 1VD and 2VD models introduce stochasticity in the latent representations.

As such, we begin by characterising our generative approach to 2VD using a CVAE. For a given image $\boldsymbol{i}$ and associated caption $\boldsymbol{c}$, we define a dialogue as a sequence of question-answer pairs $\boldsymbol{d}_{1:T} = \langle(\boldsymbol{q}_t, \boldsymbol{a}_t)\rangle_{t=1}^{T}$, simply denoted $\boldsymbol{d}$ when sequence indexing is unnecessary. Additionally, we denote a dialogue context $\boldsymbol{h}$. When indexed by step as $\boldsymbol{h}_t$, it captures the dialogue subsequence $\boldsymbol{d}_{1:t}$.

With this formalisation, we characterise a generative model for 2VD under latent variable $\boldsymbol{z}$ as $p_\theta(\boldsymbol{d}, \boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) = p_\theta(\boldsymbol{d} \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})\ p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})$, with the corresponding recognition model defined as $q_\phi(\boldsymbol{z} \mid \boldsymbol{d}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})$. We refer to this model as FLIPDIAL. Note that with relation to Eq. 4.2.2, data $\boldsymbol{x}$ is dialogue $\boldsymbol{d}$ and the condition variable is $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}\}$, giving:

$$\begin{aligned}\log p_\theta(\boldsymbol{d} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) \geq\ & \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{d},\boldsymbol{i},\boldsymbol{c},\boldsymbol{h})}[\log p_\theta(\boldsymbol{d} \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})] \\ & - \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{d}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h})),\end{aligned} \tag{4.3.1}$$

with the graphical model structures shown in Figure 4.2.

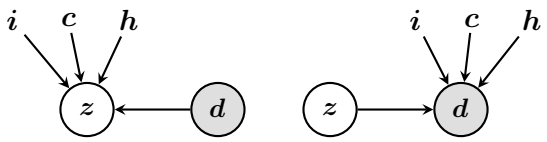


Figure 4.2 left: conditional recognition model and **right:** conditional generative model for 2VD.

FLIPDIAL's formulation in Eq. 4.3.1 is general enough to be applied to single question-answering (VQA) all the way to full two-way dialogue generation (2VD). Taking a step back from generative 2VD, we can re-frame the formulation for generative 1VD (i.e. sequential answer generation) by considering the generated component to be the answer to a particular question at step t,

given context from the image, caption and the sequence of previous question-answers. Simply put, this corresponds to the data $\boldsymbol{x}$ being the answer $\boldsymbol{a}_t$, conditioned on the image, its caption, the dialogue history until $t-1$, and the current question, or $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_{t-1}, \boldsymbol{q}_t\}$. For simplicity, we denote a compound context as $\boldsymbol{h}_t^+ = \langle \boldsymbol{h}_{t-1}, \boldsymbol{q}_t \rangle$ and reformulate Eq. 4.3.1 for 1VD as:

$$\log p_\theta(\boldsymbol{d} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) = \sum_{t=1}^{T} \log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+),$$

$$\begin{aligned} \log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \geq\ & \mathbb{E}_{q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)}[\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)] \\ & - \mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) \,\|\, p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)), \end{aligned} \tag{4.3.2}$$

with the graphical model structures shown in Figure 4.3.

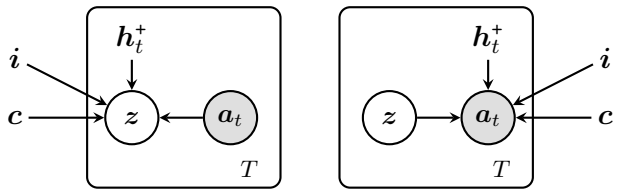


Figure 4.3 reft: conditional recognition model and **right:** conditional generative model for 1VD.

Our baseline [77] for the 1VD model can also be represented in our formulation by taking the variational posterior and generative prior to be conditional Dirac-Delta distributions. That is, $q_\phi(\boldsymbol{z} \mid \boldsymbol{a}_t, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) = p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) = \delta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)$. This transforms the objective from Eq. 4.3.2 by i) replacing the expectation of the log-likelihood over the recognition model by an evaluation of the log-likelihood for a *single* encoding (one that satisfies the Dirac-Delta), and ii) ignoring the $\mathbb{D}_{\mathrm{KL}}$ regulariser, which is trivially 0. This computes the marginal likelihood directly as just the model likelihood $\log p_\theta(\boldsymbol{a}_t \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)$, where $\boldsymbol{z} \sim \delta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)$.

Note that while such models can "generate" answers to questions by sampling from the likelihood function, we typically don't call them generative since they effectively make the encoding of the data and conditions fully deterministic. We explore and demonstrate the benefit of a fully generative treatment of 1VD in Section 4.4.3. It also follows trivially that the basic VQA model (for single question-answering) itself can be obtained from this 1VD model by simply assuming there is no dialogue history (i.e. step length $T = 1$).

4.3.1 "Colouring" visual dialogue with convolutions

FlipDial's convolutional formulation allows us to *implicitly* capture the sequential nature of sentences, and sequences of sentences. Here we introduce how we encode questions, answers, and whole dialogues with CNNs.

We begin by noting the prevalence of recurrent approaches (e.g. LSTM [136], GRU [66]) in modelling both visual dialogue and general dialogue to date [77, 78, 152, 381]. Typically recurrence is employed at two levels—at the lower level to sequentially generate the words of a sentence (a question or answer in the case of dialogue), and at a higher level to sequence these sentences together into a dialogue.

Recently however, there has been considerable interest in convolutional models of language [36, 141, 162, 267], which have shown to perform at least as well as recurrent models, if not better, on a number of tasks. They are also computationally more efficient, and typically suffer less from issues of exploding or vanishing gradients for which recurrent networks are known [257].

In modelling sentences with convolutions, the tokens (words) of the sentence are transformed into a stack of fixed-dimensional embeddings (e.g. using *word2vec* [229] or *GloVe* [264], or those learned for a specific task). For a given sentence, say question $\boldsymbol{q}_t$, this results in an embedding $\mathring{\boldsymbol{q}}_t \in \mathbb{R}^{E \times L}$ for embedding size E and sentence length L, where L can be bounded by the maximum sentence length in the corpus, with padding tokens employed where required. This two-dimensional stack is essentially a single-channel "image" on which convolutions can be applied in the standard manner in order to encode the entire sentence. Note this similarly applies to the answer $\boldsymbol{a}_t$ and caption $\boldsymbol{c}$, producing embedded $\mathring{\boldsymbol{a}}_t$ and $\mathring{\boldsymbol{c}}$, respectively.

We then extend this idea of viewing sentences as "images" to whole dialogues, producing a *multi-channel* language embedding. Here, a sequence of sentences can be seen as a stack of (a stack of) word embeddings $\mathring{\boldsymbol{d}} \in \mathbb{R}^{E \times L \times 2T}$, where now the number of channels accounts for the number of questions and answers in the dialogue. We refer to this process as "colouring" dialogue, by analogy to the most common meaning given to image channels—colour.

Our primary motivation for adopting a convolutional approach here is to explore its efficacy in extending from simpler language tasks [141, 162] to full visual dialogue. We hence instantiate

the following FlipDial models for 1VD and 2VD:

Answer [1VD]: We employ the CVAE formulation from Eq. 4.3.2 and Figure 4.3 to iteratively generate answers, conditioned on the image, caption and current dialogue history.

Block [1VD, 2VD]: Using the CVAE formulation from Eq. 4.3.1 and Figure 4.2 we generate entire *blocks* of dialogue directly (i.e. $\boldsymbol{h} = \emptyset$ since dialogue context is implicit rather than explicit). We allow the convolutional model to *implicitly* supply the context instead. We consider this 2VD, although this block architecture can also generate iteratively, and can be evaluated on 1VD (see Section 4.4.2).

Block Auto-Regressive [1VD, 2VD]: We introduce an auto-regressive component to our model following auto-regressive generative models for images [120, 335]. We augment the **Block** model by feeding its output through an auto-regressive (AR) module which explicitly enforces sequentiality in the generation of the dialogue blocks. This effectively factorises the likelihood in Eq. 4.3.1 as $p_\theta(\boldsymbol{d} \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}) = p_\theta\big(\boldsymbol{d}^1 \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}\big) \prod_{n=2}^{N} p_\theta\big(\boldsymbol{d}^n \mid \boldsymbol{d}^{1:n-1}\big)$ where N is the number of AR layers, and $\boldsymbol{d}^1$ is the (intermediate) output from the standard **Block** model. Note, again $\boldsymbol{h} = \emptyset$, and $\boldsymbol{d}^n$ refers to an entire dialogue at the n^{th} AR layer (rather than the t^{th} dialogue exchange as is denoted by $\boldsymbol{d}_t$).

4.4 Experiments

We present an extensive quantitative and qualitative analysis of FlipDial's performance in both 1VD, which requires answering a sequence of image-contextualised questions, and full 2VD, where both questions *and* answers must be generated given a specific visual context. Our proposed generative models are denoted as follows:

A – **a**nswer architecture for 1VD

B – **b**lock dialogue architecture for 1VD & 2VD

$\mathbf{B_{AR}}$ – auto-regressive extension of **B** for 1VD & 2VD

A is a generative convolutional extension of our baseline [77] and is used to validate our methods against a standard benchmark in the 1VD task. **B** and $\mathbf{B_{AR}}$, like **A**, are generative, but are extensions capable of doing full dialogue generation, a much more difficult task. Importantly, **B** and $\mathbf{B_{AR}}$ are flexible in that despite being trained to generate a block of

questions *and* answers ($\boldsymbol{h} = \emptyset$), they can be *evaluated* iteratively for both 1VD and 2VD (see Section 4.4.2). We summarise the data and condition variables for all models in Table 4.1. To evaluate performance on both tasks, we propose novel evaluation metrics which augment those of our baseline [77]. To the best of our knowledge, we are the first to report models that can generate both questions and answers given an image and caption, a necessary step toward a truly conversational agent.

Table 4.1 Data ($\boldsymbol{x}$) and condition ($\boldsymbol{y}$) variables for models **A** and $\mathbf{B}/\mathbf{B_{AR}}$ for 1VD and 2VD. Models $\mathbf{B}/\mathbf{B_{AR}}$ can be evaluated as a block or iteratively (see Section 4.4.2), accepting ground-truth ($\boldsymbol{h}_{t-1}$) or previously generated ($\hat{\boldsymbol{h}}_{t-1}$) dialogue history. For 1VD, $\hat{\boldsymbol{h}}_{t-1}$ contains ground-truth questions and previously generated answers, while for 2VD $\hat{\boldsymbol{h}}_{t-1}$ contains previously generated questions *and* answers (see Table 4.2). For clarity, $\boldsymbol{h}_t^+$ has been expanded to $\langle \boldsymbol{h}_{t-1}, \boldsymbol{q}_t \rangle$.

Task	Model	Train		Evaluate			
		data $\boldsymbol{x}$	condition $\boldsymbol{y}$	data $\boldsymbol{x}$	condition $\boldsymbol{y}$	generation $\hat{\boldsymbol{x}}$	eval method
1VD	**A**	$\boldsymbol{a}_t$	$\boldsymbol{i}, \boldsymbol{c}, \boldsymbol{q}_t, \boldsymbol{h}_{t-1}$	$\emptyset$	$\boldsymbol{i}, \boldsymbol{c}, \boldsymbol{q}_t, \boldsymbol{h}_{t-1}$	$\hat{\boldsymbol{a}}_t$ each turn	-
	B, $\mathbf{B_{AR}}$	$\boldsymbol{d}$	$\boldsymbol{i}, \boldsymbol{c}$	$\boldsymbol{q}_t, \{\boldsymbol{h}_{t-1}$ or $\hat{\boldsymbol{h}}_{t-1}\}$	$\boldsymbol{i}, \boldsymbol{c}$	$\hat{\boldsymbol{a}}_t$ each turn	iterative
2VD	**B**, $\mathbf{B_{AR}}$	$\boldsymbol{d}$	$\boldsymbol{i}, \boldsymbol{c}$	$\emptyset$	$\boldsymbol{i}, \boldsymbol{c}$	$\hat{\boldsymbol{d}}$ as a whole	block
				alternating $\hat{\boldsymbol{h}}_{t-1}$ and $\hat{\boldsymbol{q}}_t, \hat{\boldsymbol{h}}_{t-1}$		alternating $\hat{\boldsymbol{q}}_t$ and $\hat{\boldsymbol{a}}_t$ each turn	iterative

Our key results are:

- We set SOTA results in the 1VD task on the *VisDial* dataset, improving the mean rank of the generated answers by 5.66 (Table 4.3, $\mathcal{S}_{w2v}$) compared to Das et al. [77].
- Our block models are able to generate both questions and answers, a more difficult but more realistic task (2VD).
- Since our models are generative, we are able to show highly diverse and plausible question and answer generations based on the provided visual context.

Datasets We use the *VisDial* [77] dataset (*v0.9*) which contains Microsoft COCO [193] images each paired with a caption and a dialogue of 10 question-answer pairs. The train/test split is 82,783/40,504 images, respectively. The dialogues are collected by pairs of annotators on Amazon Mechanical Turk (AMT). One "blind" annotator (provided only the image caption, not the image itself) acts as the questioner and the other "oracle" annotator (provided both image and caption) acts as the answerer.

Baseline Das et al. [77]'s best model, `MN-QIH-G`, is a recurrent encoder-decoder architecture which encodes the image $\boldsymbol{i}$, the current question $\boldsymbol{q}_t$ and the attention-weighted *ground truth* dialogue history $\boldsymbol{d}_{1:t-1}$. The output conditional likelihood distribution is then used to (token-

wise) predict an answer. Our **A** model is a generative and convolutional extension, evaluated using existing ranking-based metrics [77] on the generated and candidate answers. We also (iteratively) evaluate our **B**/$\mathbf{B_{AR}}$ for 1VD as detailed in Section 4.4.2 (see Table 4.3).

4.4.1 Network architectures and training

Following the CVAE formulation (Section 4.3) and its convolutional interpretation (Section 4.3.1), all our models (**A**, **B** and $\mathbf{B_{AR}}$) have three core components: an encoder network, a prior network and a decoder network. Figure 4.4 (top) shows the encoder and prior networks, and Figure 4.4 (middle, bottom) show the standard, and auto-regressive decoder networks.

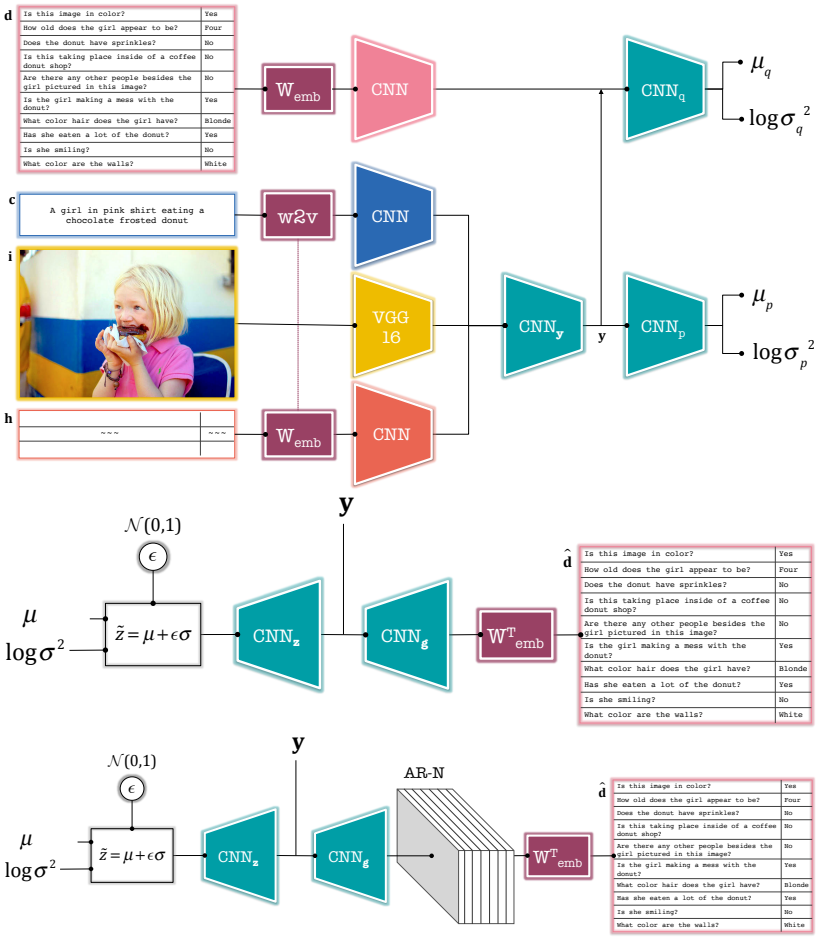


Figure 4.4 Convolutional **(top)** conditional encoder and prior architecture, **(middle)** conditional decoder, and **(bottom)** auto-regressive conditional decoder architectures, for one- and two-way visual dialogue (1VD and 2VD).

Prior network The prior neural network, parametrised by θ, takes as input the image $\boldsymbol{i}$, the caption $\boldsymbol{c}$ and the dialogue context. Referring to Table 4.1, for model $\mathbf{A}$, recall $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+\}$ where the context $\boldsymbol{h}_t^+$ is the dialogue history up to t–1 and the current question $\boldsymbol{q}_t$. For models $\mathbf{B}/\mathbf{B_{AR}}$, $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}\}$ (note $\boldsymbol{h} = \emptyset$). To obtain the image representation, we pass $\boldsymbol{i}$ through *VGG-16* [305] and extract the penultimate (4096-dimensional) feature vector. We pass caption $\boldsymbol{c}$ through a pre-trained *word2vec* [229] module (we do not learn these word embeddings). If $\boldsymbol{h} \neq \emptyset$, we pass the one-hot encoding of each word through a *learnable* word embedding module and stack these embeddings as described in Section 4.3.1. We encode these condition variables convolutionally to obtain $\boldsymbol{y}$, and pass this through a convolutional block to obtain $\boldsymbol{\mu}_p$ and $\log \boldsymbol{\sigma}_p^2$, the parameters of the conditional prior $p_\theta(\boldsymbol{z} \mid \boldsymbol{y})$.

Encoder network The encoder network, parametrised by ϕ, takes $\boldsymbol{x}$ and the encoded condition $\boldsymbol{y}$ (obtained from the prior network) as input. For model $\mathbf{A}$, $\boldsymbol{x} = \boldsymbol{a}_t$ while for $\mathbf{B}/\mathbf{B_{AR}}$, $\boldsymbol{x} = \boldsymbol{d} = \langle (\boldsymbol{q}_t, \boldsymbol{a}_t) \rangle_{t=1}^{T}$. In all models, $\boldsymbol{x}$ is transformed through a word embedding module into a single-channel answer "image" for $\mathbf{A}$, or a multi-channel "image" of alternating questions and answers for $\mathbf{B}/\mathbf{B_{AR}}$. The embedded output is then combined with $\boldsymbol{y}$ to obtain $\boldsymbol{\mu}_q$ and $\log \boldsymbol{\sigma}_q^2$, the parameters of the conditional latent posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$.

Decoder network The decoder network takes as input a latent $\boldsymbol{z}$ and the encoded condition $\boldsymbol{y}$. The sample is transpose-convolved, combined with $\boldsymbol{y}$ and further transformed to obtain an intermediate output volume of dimension $E \times L \times M$, where E is the word embedding dimension, L is the maximum sentence length and M is the number of dialogue entries in $\boldsymbol{x}$ ($M = 1$ for $\mathbf{A}$, $M = 2T$ for $\mathbf{B}$ variants). Following this, our models diverge in their architectures: $\mathbf{A}$ and $\mathbf{B}$ employ a standard linear layer, projecting the E dimension to the vocabulary size V (Figure 4.4 (middle)), whereas $\mathbf{B_{AR}}$ employs an autoregressive module followed by this standard linear layer (Figure 4.4 (bottom)). At train time, the *softmax* of the V-dimensional output is used to compute the CE term of the ELBO. At test time, the *argmax* of the output provides the predicted word index. The weights of the encoder and prior's learnable word embedding module and the decoder's final linear layer are shared.

Autoregressive module Inspired by *PixelCNN* [334] which sequentially predicts image pixels, and similar to [120], we apply $N = \{8, 10\}$ size-preserving autoregressive layers to the intermediate output of model $\mathbf{B}$ (size $E \times L \times 2T$), and then project E to vocabulary size V. Each layer employs masked convolutions to ensure only past embeddings (in the dialogue history) are considered, and sequentially predicts $2T * L$ embeddings of size E. This approach thus enforce sequentiality at both the sentence- and dialogue-level.

KL annealing Motivated by [49] in learning continuous latent embedding spaces for language, we employ KL annealing in the loss objectives of Eq. 4.3.1 and Eq. 4.3.2. We weight the KL term by α which is linearly interpolated from $[0, 1]$ over 100 epochs. We then train for a further 50 epochs with $\alpha = 1$.

Network and training hyper-parameters When embedding sentences, we pad to a maximum sequence length of $L = 64$ and use a word embedding dimension of $E = 256$ (for *word2vec*, $E = 300$). After pre-processing and filtering, the vocabulary size is $V = 9710$ (see supplementary for further details). We use the Adam optimiser [170] with default parameters, a latent dimensionality of 512 and employ batch normalisation with momentum= 0.001 and learnable parameters. For model $\mathbf{A}$ we use a batch size of 200, and 40 for $\mathbf{B}/\mathbf{B_{AR}}$. We implement our pipeline using PyTorch [275].

4.4.2 Evaluation methods for block models

Although $\mathbf{B}/\mathbf{B_{AR}}$ generate whole blocks of dialogue directly ($\boldsymbol{h} = \emptyset$), they can be evaluated iteratively, lending them to both 1VD and 2VD (see supplementary for descriptions of generation/reconstruction pipelines).

- **Block evaluation [2VD]**. The generation pipeline generates whole blocks of dialogue directly, conditioned on the image and caption (i.e. $\boldsymbol{x} = \emptyset$ and $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}\}$ for $\mathbf{B}/\mathbf{B_{AR}}$ evaluation in Table 4.1). This is 2VD since the model must generate a coherent block of both questions *and* answers.
- **Iterative evaluation**. The reconstruction pipeline can generate dialogue items iteratively. At time t, the input dialogue block is filled with zeros (`PAD` token) and the ground-

truth/predicted dialogue history to $< t$ is slotted in (see below and Table 4.2). This future-padded block is then encoded with the conditional inputs, and then reconstructed. The t^{th} dialogue item is extracted (whether an answer if 1VD or a question/answer if 2VD), and this is repeated T (for 1VD) or $2T$ (for 2VD) times. Variations are:

- **1VD with ground-truth history**. At time t, the input dialogue block is filled with *ground-truth* dialogue history up to t–1, along with the current ground-truth question. All future entries are padded—equivalent to [77] using the ground-truth dialogue history.
- **1VD with generated history**. Similar to above, except that the input block is filled with the history of ground-truth questions and *previously predicted* answers along with the current ground-truth question. This is a more realistic 1VD.
- **2VD with generated history**. The most challenging and realistic condition in which the input block is filled with the history of previously predicted questions *and* answers, alternating with a generated question or answer each turn.

Table 4.2 Iterative evaluation of $\mathbf{B}/\mathbf{B_{AR}}$ for 1VD and 2VD. Under each condition, the input dialogue block is filled with ground-truth or predicted history ($\boldsymbol{q}/\boldsymbol{a}$ or $\hat{\boldsymbol{q}}/\hat{\boldsymbol{a}}$, respectively), while future entries are filled with the PAD token.

Task		$\mathbf{B}/\mathbf{B_{AR}}$ iterative evaluation			
		data $\boldsymbol{x}$	$> t$	$= t$	$> t$
1VD	with ground-truth history:	$\boldsymbol{q}_t, \boldsymbol{h}_{t-1}$	$(\boldsymbol{q}, \boldsymbol{a})$	$(\boldsymbol{q}$, PAD$)$	(PAD, PAD)
	with generated history:	$\boldsymbol{q}_t, \hat{\boldsymbol{h}}_{t-1}$	$(\boldsymbol{q}, \hat{\boldsymbol{a}})$	$(\boldsymbol{q}$, PAD$)$	(PAD, PAD)
2VD	with generated history:	alternating $\hat{\boldsymbol{h}}_{t-1}$ and $\hat{\boldsymbol{q}}_t, \hat{\boldsymbol{h}}_{t-1}$	$(\hat{\boldsymbol{q}}, \hat{\boldsymbol{a}})$	alternating (PAD, PAD) and $(\hat{\boldsymbol{q}}$, PAD$)$	(PAD, PAD)

4.4.3 Evaluation and analysis

We evaluate our **A**, **B**, and $\mathbf{B_{AR}}$ models on the 1VD and 2VD tasks. Under 1VD, we predict an answer with each time step, given an image, caption and the current dialogue history (Section 4.4.3 and Table 4.3), while under 2VD, we predict both questions *and* answers (Section 4.4.3 and Table 4.4). All three models are able to perform the first task , while only **B** and $\mathbf{B_{AR}}$ are capable of the second task.

One-way visual dialogue (1VD) task

We evaluate the performance of **A** and $\mathbf{B}/\mathbf{B_{AR}}$ on 1VD using the candidate ranking metric of [77] as well as an extension of this which assesses the *generated* answer quality (Table 4.3). Figure 4.1 and Figure 4.5 show our qualitative results for 1VD.

Table 4.3 1VD evaluation for **A** and **B**/$\mathbf{B_{AR}}$ on *VisDial (v0.9)* test set. Results show ranking of answer candidates based on the score functions $\mathcal{S}_M$ and $\mathcal{S}_{w2v}$. **B**/$\mathbf{B_{AR}}$ are evaluated iteratively for the 1VD task, using either ground-truth or previously generated dialogue history (see Table 4.2).

Score function		Model	MR	MRR	R@1	R@5	R@10
$\mathcal{S}_M$		RL-QAbot [78]	21.13	0.4370	-	53.67	60.48
		`MN-QIH-G` [77]	17.06	0.5259	42.29	62.85	68.88
		A (LW)	23.87	0.4220	30.48	53.78	57.52
		A (ELBO)	20.38	0.4549	34.08	56.18	61.11
$\mathcal{S}_{w2v}$		`MN-QIH-G` [77]	31.31	0.2215	16.01	22.42	34.76
		A (RECON)	15.36	0.4952	41.77	54.67	66.90
		A (GEN)	**25.65**	0.3227	25.88	33.43	47.75
	with GT history	iterative **B**	28.45	0.2927	23.50	29.11	42.29
		iterative $\mathbf{B_{AR}}8$	25.87	**0.3553**	**29.40**	**36.79**	**51.19**
		iterative $\mathbf{B_{AR}}10$	26.30	0.3422	28.00	35.34	50.54
	with gen. history	iterative **B**	30.57	0.2188	16.06	20.88	35.37
		iterative $\mathbf{B_{AR}}8$	29.10	0.2864	22.52	29.01	48.43
		iterative $\mathbf{B_{AR}}10$	29.15	0.2869	22.68	28.97	46.98

Candidate ranking by model log-likelihood [$\mathcal{S}_\mathbf{M}$] The *VisDial* dataset [77] provides a set of 100 candidate answers $\{\boldsymbol{a}_t^c\}_{c=1}^{100}$ for each question-answer pair at time t per image. The set includes the ground-truth answer $\boldsymbol{a}_t$ as well as similar, popular, and random answers. Das et al. [77] rank these candidates using the log-likelihood value of each under their model (conditioned on the image, caption and dialogue history, including the current question), and then observe the position of the ground-truth answer in the ranked list (where closer to 1 is better). This position is averaged over the dataset to obtain the mean rank (MR). In addition, the mean reciprocal rank (MRR; 1/MR) and recall rates at $k = \{1, 5, 10\}$ are computed.

To compare against their baseline, we rank the 100 candidates answers by estimates of their *marginal* likelihood from **A**. This can be done with i) the conditional ELBO (Eq. 4.3.2), and by ii) likelihood weighting (LW) in the conditional generative model $p_\theta(\boldsymbol{a}_t \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) = \int p_\theta(\boldsymbol{a}_t, \boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) dz = \int p_\theta(\boldsymbol{z} \mid \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+) p_\theta(\boldsymbol{a} \mid \boldsymbol{z}, \boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+)\, dz$. Ranking by both these approaches is shown in the $\mathcal{S}_\mathbf{M}$ section of Table 4.3, indicating that we are comparable to SOTA models in sequential VQA [77, 78].

Candidate ranking by *word2vec* cosine distance [$\mathcal{S}_{w2v}$] The evaluation protocol of [77] scores and ranks a given set of candidate answers, without being a function of the actual answer *predicted* by the model, $\hat{\boldsymbol{a}}_t$. This results in the rank of the ground-truth answer candidate reflecting its score under the model *relative* to the rest of the candidates' scores, rather than capturing the quality of the answer output by the model, which is left unobserved. To remedy this, we instead score each candidate by the cosine distance between the *word2vec*

Question	Ground-truth answer	z_1	z_2	z_3
How old is the girl?	Maybe three	Looks about six	I can't tell	Yes
What race is the girl?	White	Yes	White	Caucasian
Is she outside?	Yes	No	Yes	Yes
Is her hair long or short?	Short	Short	Short	Short
What color is her hair?	Blonde	Blonde	Brown	Brown
Is her hair curly or straight?	It's straight	Straight	Straight	Straight
What is she wearing?	Pink shirt, white pants	T-shirt and jacket and pants	Shirt and pants	Jeans like a pajamas
Is the teddy bear in her lap?	Yes	No	Yes	Yes
What color is the teddy bear?	White	Brown	Tan	Tan and white
Is it nice outside?	Yes sunny	Yes	It looks lovely	Yes

Question	Ground-truth answer	z_1	z_2	z_3
How old does she look?	Around seven or eight	I cannot tell about her	Looks about six	She is about teenager around
Any other people?	No	No	Yes	Yes
Any buildings?	No	Yes	No	No
Is it day or night?	Day	It looks like it in image is in so	Daytime	Day
Is it raining?	No	Yes	Yes	No
What color umbrella?	Pink and clear	Dark colored color	White	White
Is it open or closed?	Open	Yes	Open	Open
Is it sunny?	I can't tell	Yes	No	Yes
What color is her hair?	Dark brown	Brown	Brown	Brown
Is it long or short?	Long	Short	Short	I'd say long

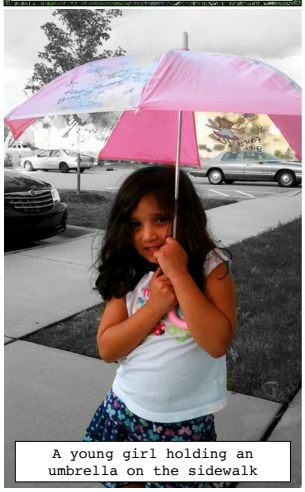


Figure 4.5 Qualitative examples of generated answers from **A**'s conditional prior – conditioned on an image, caption, question and dialogue history. See supplementary material for further examples.

embedding of the predicted answer $\hat{\boldsymbol{a}}_t$ and that candidate's *word2vec* embedding. We take the embedding of a sentence to be the average embedding over word tokens following [30]. In addition to accounting for the predicted answer, this method also allows semantic similarities to be captured such that if the predicted answer is similar (in meaning and/or words generated) to the ground-truth candidate answer, then the cosine distance will be small, and hence the ground-truth candidate's rank closer to 1.

We report these numbers for **A**, iteratively-evaluated $\mathbf{B}/\mathbf{B_{AR}}$, and also our baseline model `MN-QIH-G` [77], which we re-evaluate using the *word2vec* cosine distance ranking (see $\mathcal{S}_{w2v}$ in Table 4.3). In the case of **A** (GEN), we evaluate answer *generations* from **A** whereby we condition on $\boldsymbol{i}, \boldsymbol{c}$ and $\boldsymbol{h}_t^+$ via the prior network, sample $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)$ and generate an answer via the decoder network. Here we show an improvement of 5.66 points in MR over

the baseline. On the other hand, **A** (RECON) evaluates answer *reconstructions* in which $\boldsymbol{z}$ is sampled from $\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2)$ (where ground-truth answer $\boldsymbol{a}_t$ is provided). We include **A** (RECON) merely as an "oracle" autoencoder, observing its good ranking performance, but do not explicitly compare against it.

We also note that the ranking scores of the block models are worse (by 3-4 MR points) than those of **A**. This is expected since **A** is explicitly trained for 1VD which is not the case for $\mathbf{B}/\mathbf{B_{AR}}$. Despite this, the performance gap between **A** (GEN) and $\mathbf{B}/\mathbf{B_{AR}}$ (with ground-truth history) is not large, bolstering our iterative evaluation method for the block architectures. Note finally that the $\mathbf{B}/\mathbf{B_{AR}}$ models perform better when given the ground-truth rather than generated dialogue history (by 2-3 MR points). This is also expected as answering is easier with access to the ground-truth dialogue history rather than only the previously *predicted* answers (and ground-truth questions).

Two-way visual dialogue (2VD) task

Our flexible CVAE formulation for visual dialogue allows us to move from 1VD to the generation of both questions *and* answers (2VD). Despite this being inherently more challenging, $\mathbf{B}/\mathbf{B_{AR}}$ are able to generate diverse sets of questions and answers contextualised by the given image and caption. Figure 4.6 shows snippets of our two-way dialogue generations.

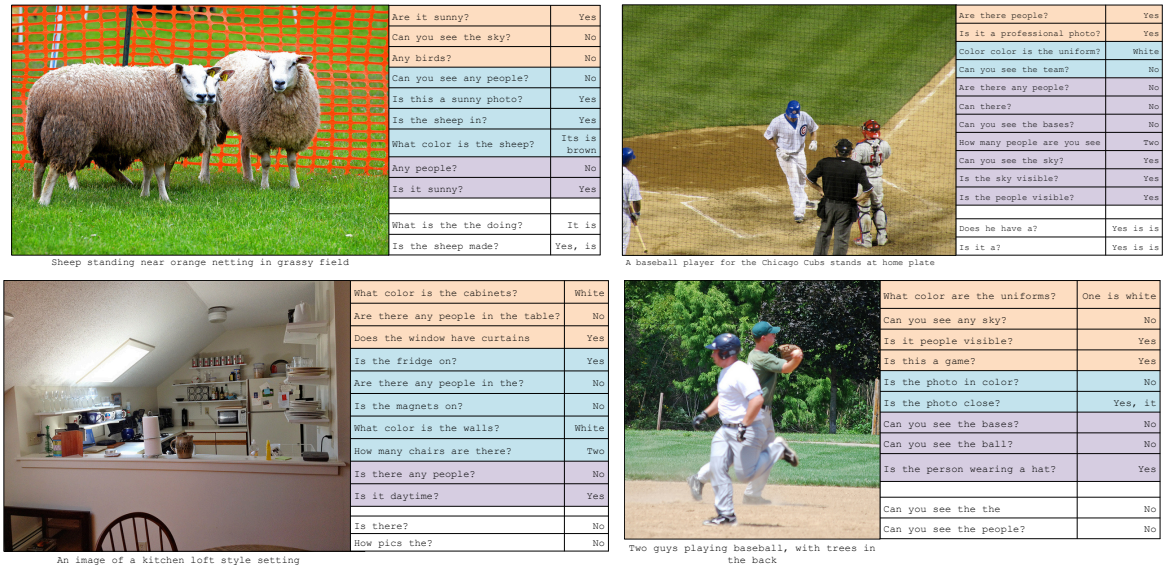


Figure 4.6 Qualitative examples of two-way dialogue generation from the $\mathbf{B}/\mathbf{B_{AR}}$ models. Different colours indicate different generations—coherent sets with a single colour, and failures in white. See supplementary for further examples.

In evaluating our models for 2VD, the candidate ranking protocol of [77] which relies on a *given* question to rank the answer candidates, is no longer usable when the questions themselves are being generated. This is the case for $\mathbf{B}/\mathbf{B_{AR}}$ block evaluation, which has no access to the ground-truth dialogue history, and the iterative evaluation when only the generated history of questions and answers is used (Table 4.2). We therefore look directly to the CE and KL terms of the ELBO as well as propose two new metrics, $sim_{\boldsymbol{c},\boldsymbol{q}}$ and $sim_{\circlearrowleft}$, to compare our methods in the 2VD task:

- **Question relevance ($sim_{\boldsymbol{c},\boldsymbol{q}}$)**. We expect a generated question to query an aspect of the image, and we use the presence of semantically similar words in both the question and image caption as a proxy of this. We compute the cosine similarity between the (average) *word2vec* embedding of each predicted question $\boldsymbol{q}_t$ and that of the caption $\boldsymbol{c}$, and average over all T questions in the dialogue. A value closer to 1 would indicate higher semantic similarity between the generated questions and the visual component.
- **Latent dialogue dispersion ($sim_{\circlearrowleft}$)**. For a generated dialogue block $\boldsymbol{d}^g$, $sim_{\circlearrowleft}$ computes the KL divergence $\mathbb{D}_{\text{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{d}^g, \boldsymbol{i}, \boldsymbol{c}) \parallel q_\phi(\boldsymbol{z}|\boldsymbol{d}, \boldsymbol{i}, \boldsymbol{c}))$. This measures how close the generated dialogue is to the true dialogue $\boldsymbol{d}$ in the latent space, given the same image $\boldsymbol{i}$ and caption $\boldsymbol{c}$.

From Table 4.4, we observe a decrease in the loss terms as the auto-regressive capacity of the model increases (none $\rightarrow$ 8 $\rightarrow$ 10), suggesting that explicitly enforcing sequentiality in the dialogue generations is useful. For $\text{sim}_{\circlearrowleft}$ within a particular model, the dispersion values are typically larger for the harder task (without dialogue context). We also observe that dispersion increases with number of AR layers, suggesting AR improves the diversity of the model outputs, and avoids simply recovering data observed at train time.

While the proposed metrics provide a novel means to evaluate dialogue in a generative framework, like all language-based metrics, they are not complete. The question-relevance metric, $\text{sim}_{\boldsymbol{c},\boldsymbol{q}}$, can stagnate, and neither metric precludes redundant or nonsensical questions. We intend for these metrics to *augment* the bank of metrics available to evaluate dialogue and language models. Further evaluation, including i) using auxiliary tasks, as in the image-retrieval task of [78], to drive and evaluate the dialogues, and ii) turning to human evaluators to rate the generated dialogues, can be instructive in painting a better picture of these models.

Table 4.4 2VD evaluation for $\mathbf{B}/\mathbf{B_{AR}}$ on *VisDial* (v0.9) test set, with either whole dialogues generated at once (block evaluation), or questions/answers generated iteratively and in alternation (using previously generated history) (see Section 4.4.2). Arrows indicate which direction is better.

Model	**Eval Method**	**Cross Entropy** ⇓	**KL divergence** ⇓	$\text{sim}_{\boldsymbol{c,q}}$ ⇑	$\text{sim}_{\circlearrowleft}$ ⇓
B	block	31.18	4.34	0.4931	14.20
	iterative	25.40	4.01	0.4091	1.86
$\mathbf{B_{AR}}8$	block	28.81	2.54	0.4878	31.50
	iterative	26.60	2.29	0.3884	2.39
$\mathbf{B_{AR}}10$	block	28.49	1.89	0.4927	44.34
	iterative	24.93	1.80	0.4101	2.35

4.5 Conclusion

In this work we propose FLIPDIAL, a generative convolutional model for visual dialogue which is able to generate answers (1VD) as well as generate both questions *and* answers (2VD) based on a visual context. In the 1VD task, we set new SOTA results with the answers generated by our model, and in the 2VD task, we are the first to establish a baseline, proposing two novel metrics to assess the quality of the generated dialogues. In addition, we propose and evaluate our models under a much more realistic setting for both visual dialogue tasks in which the *predicted* rather than ground-truth dialogue history is provided at test time. This challenging setting is more akin to real-world situations in which dialogue agents must be able to evolve with their predicted exchanges. We emphasize that research focus must be directed here in the future. Finally, under all cases, the sets of questions and answers generated by our models are qualitatively good: diverse and plausible given the visual context. Looking forward, we are interested in exploring additional methods for enforcing diversity in the generated questions and answers, as well as extending this work to explore recursive models of reasoning for visual dialogue.

Acknowledgements DM, NS, PKD and PHST were supported by ERC grant ERC-2012-AdG 321162-HELIOS, EPSRC grant Seebibyte EP/M013774/1, EPSRC/MURI grant EP/N019474/1. DM was additionally supported by the Skye Foundation.

4.6 Supplementary

4.6.1 Glossary

block dialogue/architecture Models $\mathbf{B}/\mathbf{B_{AR}}$ are built and trained for the task of two-way visual dialogue (2VD) with data $\boldsymbol{x} = \boldsymbol{d}$ and condition variable $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}\}$. Since $\boldsymbol{d}$ is a whole dialogue sequence/block $\langle(\boldsymbol{q}_t, \boldsymbol{a}_t)\rangle_{t=1}^{T}$ we refer to $\mathbf{B}/\mathbf{B_{AR}}$ as *block* architectures.

generation This represents the scenario when only the condition variable $\boldsymbol{y}$ is available at test time. In this case, the decoder network receives a sample $\boldsymbol{z} \sim p_\theta(\boldsymbol{z} \mid \boldsymbol{y})$, a multivariate Gaussian parametrised by $(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)$ learned using the prior network. We call the decoded output $\hat{\boldsymbol{x}}$ a *generation*.

reconstruction Differing from a generation, both $\boldsymbol{y}$ *and* $\boldsymbol{x}$ are available. The decoder network receives a sample $\boldsymbol{z} \sim q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$, a multivariate Gaussian parametrised by $(\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2)$ learned using the encoder network. We call the decoded output $\hat{\boldsymbol{x}}$ a *reconstruction*. The reconstruction pipeline is used during training when the input $\boldsymbol{x}$ *and* the condition variable $\boldsymbol{y}$ are available. Note, this pipeline is also used when $\mathbf{B}/\mathbf{B_{AR}}$ are evaluated iteratively (see Section 4.4.2).

4.6.2 Extended quantitative results on 1VD task

Table 4.3 in the main paper evaluates $\mathbf{A}$ and $\mathbf{B}/\mathbf{B_{AR}}$ in the task of one-way visual dialogue (1VD). Here we shed light on these numbers and the metrics used to obtain them. We also present a more extensive quantitative analysis of $\mathbf{B}/\mathbf{B_{AR}}$ in the 1VD task (see Table 4.5).

Evaluating $\mathbf{B}/\mathbf{B_{AR}}$ on 1VD We extend Table 4.3 with Table 4.5, which further compares the iteratively-evaluated $\mathbf{B}/\mathbf{B_{AR}}$ models when using the ground-truth or previously generated dialogue history. We look at the CE and KL terms of the ELBO, and $\text{sim}_{\circlearrowleft}$. We observe that $\mathbf{B}/\mathbf{B_{AR}}$ with ground-truth dialogue history beats $\mathbf{B}/\mathbf{B_{AR}}$ with generated history by 7-10 points in MR, and also improves in MRR and recall rates. The $\text{sim}_{\circlearrowleft}$ metric, on the other hand, show very little performance difference across the two evaluation settings. We also note that ranking performance is worse when both image $\boldsymbol{i}$ and caption $\boldsymbol{c}$ are excluded from condition variable. This does not, however, correlate with the CE and KL terms of the loss which are

lower for a condition-less setting. We attribute this to the model being transformed from a CVAE to a VAE, hence lifting the burden of capturing the conditional posterior distribution (i.e. the KL is now between an unconditional $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ and $\mathcal{N}(0, 1)$).

Table 4.5 1VD iterative evaluation for $\mathbf{B}/\mathbf{B_{AR}}$ on *VisDial (v0.9)* test set. Results show ranking of answer candidates based on the $\mathcal{S}_{w2v}$ scoring function. Note that *GT history* indicates the iterative evaluation method when the ground-truth dialogue history is provided, while *gen. history* indicates the iterative evaluation method when the ground-truth question and previously generated answer history is provided (see Section 4.4.2). The +/- indicate models *trained* with/without respective conditions, image $\boldsymbol{i}$ and caption $\boldsymbol{c}$. Arrows indicate which direction is better.

Model	$\boldsymbol{i}$	$\boldsymbol{c}$		CE ⇓	KLD ⇓	MR ⇓	MRR ⇑	R@1 ⇑	R@5 ⇑	R@10 ⇑	$\text{sim}_{\circlearrowleft}$ ⇓
B	+	+	GT history	18.87	4.36	28.45	0.2927	23.50	29.11	42.29	2.68
	+	+	gen. history	25.10	4.02	30.57	0.2188	16.06	20.88	35.37	2.42
	-	+		16.80	3.13	27.76	0.3243	26.59	33.21	47.65	4.48
	+	-		21.02	4.71	29.82	0.2144	15.25	21.07	34.96	5.44
	-	-		19.35	13.34	29.00	0.3026	24.36	30.70	47.62	6.17
B$_{\mathbf{AR}}$8	+	+	GT history	15.11	2.53	25.87	0.3553	29.40	36.79	51.19	4.30
	+	+	gen. history	25.70	2.21	29.10	0.2864	22.52	29.01	48.43	3.47
	-	+		16.19	2.80	26.04	0.3566	29.62	36.75	50.62	4.17
	+	-		20.39	2.89	28.99	0.3024	24.33	30.74	47.17	8.16
	-	-		20.92	2.84	28.79	0.3045	24.46	30.99	48.10	0.18
B$_{\mathbf{AR}}$10	+	+	GT history	16.04	1.89	26.30	0.3422	28.00	35.34	50.54	4.84
	+	+	gen. history	24.77	1.81	29.15	0.2869	22.68	28.97	46.98	2.85
	-	+		19.97	2.58	26.84	0.3212	25.90	32.92	47.68	5.95
	+	-		20.39	2.79	27.27	0.3157	25.45	32.26	47.87	13.22
	-	-		19.17	0.00	29.00	0.3026	24.36	30.70	47.62	0.00

4.6.3 Extended quantitative results on 2VD task

Extending Table 4.4 in the main paper, Table 4.6 here shows results for $\mathbf{B}/\mathbf{B_{AR}}$ trained with permutations of the image $\boldsymbol{i}$ and caption $\boldsymbol{c}$) (denoted by + if included in the condition, and - otherwise). We note the decrease in CE and KL as conditions $(\boldsymbol{i}, \boldsymbol{c})$ are excluded from the model. This is expected since the task of dialogue generation is made simpler without the constraints of an explicit visual/textual condition.

4.6.4 Detailed network architectures and training

The following section provides detailed descriptions of the architectures of our models **A**, **B** and $\mathbf{B_{AR}}$. The descriptions are dense but thorough. We also include details of our training procedure. Where not explicitly noted, each convolutional layer is proceeded by batch normalisation (with momentum = 0.001 and learnable parameters) and a *ReLU* activation.

Table 4.6 2VD evaluation for **B**/**B**$_{\mathbf{AR}}$ models on *VisDial (v0.9)* test set. Note that the block evaluation method indicates when a whole dialogue is generated given only an image and its caption, while the iterative evaluation indicates when questions and answers are generated iteratively and in alternation, conditioned on the image, caption and previously generated dialogue history (see Section 4.4.2). The + and - indicate models *trained* with and without respective conditions, image $\boldsymbol{i}$ and caption $\boldsymbol{c}$. Arrows indicate which direction is better.

Model	$\boldsymbol{i}$	$\boldsymbol{c}$	Eval Method	CE ⇓	KLD ⇓	$\text{sim}_{\boldsymbol{c},\boldsymbol{q}}$ ⇑	$\text{sim}_{\circlearrowleft}$ ⇓
B	+	+	block	31.18	4.34	0.4931	14.20
			iterative	25.40	4.01	0.4091	1.86
	-	+	block	29.09	3.26	0.4889	11.23
			iterative	24.59	3.05	0.3877	3.45
	+	-	block	28.60	4.26	0.4634	15.56
			iterative	29.85	4.66	0.4221	3.54
	-	-	block	19.92	6.42	0.4590	6.34
			iterative	19.34	0.00	0.4638	0.00
B$_{\mathbf{AR}}$8	+	+	block	28.81	2.54	0.4878	31.50
			iterative	26.60	2.29	0.3884	2.39
	-	+	block	30.59	2.72	0.4889	43.17
			iterative	26.15	2.77	0.3758	3.57
	+	-	block	31.51	2.91	0.4602	24.75
			iterative	21.41	2.68	0.4453	5.49
	-	-	block	20.32	2.77	0.4464	0.26
			iterative	21.53	2.99	0.4419	0.10
B$_{\mathbf{AR}}$10	+	+	block	28.49	1.89	0.4927	44.34
			iterative	24.93	1.80	0.4101	2.35
	-	+	block	30.83	2.53	0.4951	38.60
			iterative	28.59	2.52	0.3903	1.91
	+	-	block	30.18	2.89	0.4592	100.81
			iterative	28.32	2.44	0.4334	6.73
	-	-	block	19.60	0.00	0.4585	0.00
			iterative	19.17	0.00	0.4614	0.00

Prior network The prior neural network, parametrised by θ, takes as input the image $\boldsymbol{i}$, the caption $\boldsymbol{c}$ and the dialogue context. For the model **A**, this context is $\boldsymbol{h}_t^+$, containing the dialogue history up to t–1 and the current question $\boldsymbol{q}_t$. For models **B**/**B**$_{\mathbf{AR}}$, the dialogue context is the null set ($\boldsymbol{h} = \emptyset$). To obtain the image representation, we scale and centre-cropped each image to $3 \times 224 \times 224$ and feed it through *VGG-16* [305]. The output of the penultimate layer is extracted and ℓ_2-normalised (as in [77]) to obtain a 4096-dimensional image feature vector. For the caption, we pass $\boldsymbol{c}$ through a pre-trained *word2vec* [229] model (we do not learn these word embeddings) to obtain $\mathring{\boldsymbol{c}} \in \mathbb{R}^{300 \times L}$ where L is the maximum sentence length ($L = 64$). For the dialogue context (relevant only in the case of **A**) we pass the one-hot encoding of each word through a learnable word embedding module. We stack these

embeddings as described in Section 4.3.1 of the main paper to obtain $\mathring{\boldsymbol{h}}_t^+ \in \mathbb{R}^{E \times L \times K}$, where E is the word embedding dimension ($E = 256$), L is the maximum sentence length ($L = 64$), and K is the number of dialogue entries at time t. We encode these inputs convolutionally to obtain $\boldsymbol{y}$ (the encoded condition) as follows: $\mathring{\boldsymbol{c}}$ is passed through a convolutional block (output size $64 \times 8 \times 8$) and concatenated with the image feature vector (reshaped to $64 \times 8 \times 8$). The concatenated output is passed through a convolutional block to obtain the jointly encoded image-caption (output size $64 \times 8 \times 8$). If $\boldsymbol{h} \neq \emptyset$, then the context is passed through a convolutional block (output size $64 \times 8 \times 8$), is concatenated with the encoded image-caption, and passed through another convolutional block to get the encoded image-caption-context (output size $64 \times 8 \times 8$). We call this the encoded condition $\boldsymbol{y}$, and it is then passed through a further convolutional block (output size $256 \times 4 \times 4$) followed by two final convolutional layers (in parallel) to obtain $\boldsymbol{\mu}_p$ and $\log \boldsymbol{\sigma}_p^2$, respectively, the parameters of the conditional prior $p_\theta(\boldsymbol{z} \mid \boldsymbol{y})$. At this stage, $\boldsymbol{\mu}_p$ and $\log \boldsymbol{\sigma}_p^2$ are both of size $512 \times 1 \times 1$ (the latent dimensionality). At test time, a sample is obtained via $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)$ and is passed to the decoder to generate a sample $\hat{\boldsymbol{a}}_t$ (for $\mathbf{A}$) or $\hat{\boldsymbol{d}}$ (for $\mathbf{B}/\mathbf{B_{AR}}$).

Encoder network The encoder network, parametrised by ϕ, takes $\boldsymbol{x}$ as input along with the encoded condition, $\boldsymbol{y}$, obtained from the prior network. For model $\mathbf{A}$, $\boldsymbol{x} = \boldsymbol{a}_t$ and $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+\}$. For models $\mathbf{B}/\mathbf{B_{AR}}$, $\boldsymbol{x} = \boldsymbol{d} = \langle (\boldsymbol{q}_t, \boldsymbol{a}_t) \rangle_{t=1}^T$ and $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}\}$. In all models, $\boldsymbol{x}$ is passed through a learnable word embedding module, and the word embeddings stacked (see Section 4.3.1 in the main paper) to obtain $\mathring{\boldsymbol{x}} \in \mathbb{R}^{E \times L \times M}$, where $E = 256$, $L = 64$, and M is the number of entries in $\boldsymbol{x}$ (for $\mathbf{A}$, $M = 1$ and for $\mathbf{B}/\mathbf{B_{AR}}$, $M = 2T$). In this way, we transform $\boldsymbol{x}$ into a single-channel answer "image" in the case of $\mathbf{A}$, and a multi-channel "image" of alternating questions and answers in the case of $\mathbf{B}/\mathbf{B_{AR}}$. $\mathring{\boldsymbol{x}}$ is then passed through a convolutional block (output size $64 \times 8 \times 8$), the output of which is concatenated with $\boldsymbol{y}$ and forwarded through another convolutional block (output size $256 \times 4 \times 4$). This output is passed through two final convolutional layers (in parallel) to obtain $\boldsymbol{\mu}_q$ and $\log \boldsymbol{\sigma}_q^2$, the parameters of the conditional latent posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$. Here $\boldsymbol{\mu}_q$ and $\log \boldsymbol{\sigma}_q^2$ are both of size $512 \times 1 \times 1$. At train time, the KL divergence term of the ELBO is computed using $(\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q)$ (from the encoder network) and $(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p)$ (from the prior network).

Decoder network The decoder network (for simplicity, the parameters of the prior and decoder network are subsumed into θ) takes as input a latent $\boldsymbol{z}$ and the encoded condition $\boldsymbol{y}$. During training, $\boldsymbol{z}$ is sampled from a Gaussian parametrised by the $(\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2)$ output of the encoder network. This distribution is $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$. At test time, $\boldsymbol{z}$ is sampled from a Gaussian parametrised by the $(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)$ output of the prior network. This distribution is $p_\theta(\boldsymbol{z} \mid \boldsymbol{y})$. At both train and test time, we employ the commonly-used re-parametrisation trick [171] to compute the latent sample as $\boldsymbol{z} = \mu + \epsilon\sigma$ where $\epsilon \sim \mathcal{N}(0, 1)$ and μ and σ correspond to those derived from the encoder or prior network as described above.

The sample $\boldsymbol{z}$ is then transformed through a transpose-convolutional block (output size $64 \times 8 \times 8$), concatenated with $\boldsymbol{y}$ and passed through a convolutional block (output size $64 \times 8 \times 8$). This output is passed through a second transpose-convolutional block, producing an intermediate output volume of dimension $M \times E \times L$ which we permute to match the size of $\mathring{\boldsymbol{x}}$. As before, $E = 256$, $L = 64$, and $M = 1$ (for $\mathbf{A}$) or $M = 2T$ (for $\mathbf{B}/\mathbf{B_{AR}}$).

Following this, our models diverge in architecture: $\mathbf{A}$ and $\mathbf{B}$ employ a standard linear layer which projects the E dimension of the intermediate output to the vocabulary size V. The $\mathbf{B_{AR}}$ model instead employs an autoregressive module (detailed below) followed by this standard linear layer. At train time, the V-dimensional network output is *softmax*'ed and used in the computation of the CE term of the ELBO. At test time, the *argmax* of the (*softmax*-ed) output is taken to be the index of the word token predicted. We share the weight matrices of the decoder's final linear layer and the encoder and prior's learnable word embedding module (which are the same size by virtue of our network architecture) with the motivation that language encoders and decoders should share common word representations.

Autoregressive block The autoregressive (AR) block (AR-N in Figure 4.4 (bottom)) in $\mathbf{B_{AR}}$'s decoder is inspired by *PixelCNN* [334] which sequentially predicts image pixels along the two spatial dimensions. In the same fashion, we use an autoregressive approach to sequentially predict the next sentence (question or answer) in a dialogue. Since our framework is convolutional, our approach can similarly be adapted from that of [120, 334]. We first reshape the intermediate output of the decoder to $E \times L{*}M$ (essentially "unravelling" the dialogue sequentially into a stack of its word embeddings). We then apply a size-preserving

masked convolution to the reshaped output (followed by a learnable batch normalisation and a $ReLU$ activation). We call this triplet an AR layer. The masked convolution of the AR layer ensures that future rows (i.e. future E-dimensional word embedding) are hidden in the prediction of the current row/word embedding. We apply N AR layers in this way with each layer taking in the output of the previous AR layer. Following the AR-N block, a linear layer projects the final output's E dimension to the vocabulary size V. We report numbers for $N = \{8, 10\}$. We base our AR block on a publicly-available implementation of *PixelCNN*[3].

4.6.5 Dialogue preprocessing

The word vocabulary is constructed from the *VisDial v0.9* [77] training dialogues (not including the candidate answers). The dialogues are preprocessed as follows: apostrophes are removed, numbers are converted to their worded equivalents, and all exchanges are made lower-case and either padded or truncated to a maximum sequence length ($L = 64$). The vocabulary is also filtered such that words with a frequency of <5 are removed and replaced with the `UNK` token. After pre-processing and filtering, the vocabulary size is $V = 9710$.

4.6.6 Extended qualitative results

We present additional qualitative results for the **A** model in Figures 4.7 and 4.8 (1VD task) and for the $\mathbf{B_{AR}}10$ model (under the block evaluation setting) in Figures 4.9 and 4.10 (2VD task). Note that for both, different colours indicate generations ($\hat{\boldsymbol{a}}_t$ for **A** and $\hat{\boldsymbol{d}}$ for $\mathbf{B}/\mathbf{B_{AR}}$) from different samples of $\boldsymbol{z}$. In Figures 4.9 and 4.10, whole generated dialogue blocks are shown with coloured sections indicating subsets exhibiting coherent question-answering and white sections indicating subsets that are not entirely coherent.

[3] https://github.com/jzbontar/pixelcnn-pytorch

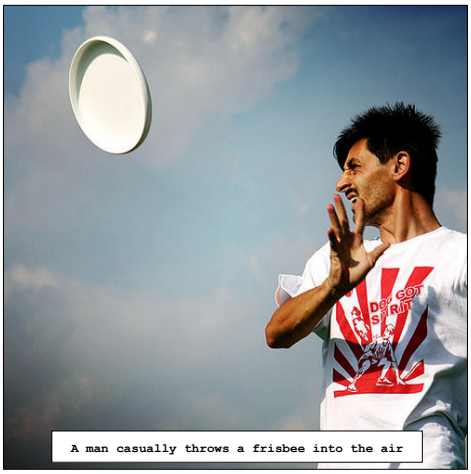
Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5
How old is the girl?	Maybe three	I can't tell	Looks about six	She is maybe	I can't tell	Yes
What race is the girl?	White	All	Yes	White	White	Caucasian
Is she outside?	Yes	Yes	No	Yes	Yes	Yes
Is her hair long or short?	Short	Short	Short	Yes	Short	Short
What color is her hair?	Blonde	Blonde	Blonde	Blonde	Brown	Brown
Is her hair curly or straight?	It's straight	Straight	Straight	Short	Straight	Straight
What is she wearing?	Pink shirt, white pants	Clothes, shirt and top	T-shirt and jacket and pants	T-shirt and pants	Shirt and pants	Jeans like a pajamas
Is the teddy bear in her lap?	Yes	There	No	Yes	Yes	Yes
What color is the teddy bear?	White	Brown	Brown	Tan	Tan	Tan and white
Is it nice outside?	Yes sunny	Yes	Yes	It looks sunny	It looks lovely	Yes

Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5
How old does she look?	Around seven or eight	I cannot tell about her	Looks about six	She is about teenager around	I can't tell	I can't tell
Any other people?	No	No	Yes	Yes	No	I don't see anyone
Any buildings?	No	Yes	No	No	Secret can't tell	Yes
Is it day or night?	Day	It looks like it in image is in so	Daytime	Day	Day	Yes
Is it raining?	No	Yes	Yes	No	No	No
What color umbrella?	Pink and clear	Dark colored color	White	White	Grey	White
Is it open or closed?	Open	Yes	Open	Open	Open	It is open
Is it sunny?	I can't tell	Yes	No	Yes	Yes	No
What color is her hair?	Dark brown	Brown	Brown	Brown	Brown	Brown
Is it long or short?	Long	Short	Short	I'd say long	Short	Yes

Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5
Are there any animals present?	No animals	No	No	Yes	No	No
Are there any kids?	Two kids	Yes	No	Yes	Yes	I
Boys or girls?	Boys	Boy	Boys	I	Boys	Yes
Are there any adults?	One adult	Yes	Yes	Yes	Yes	Yes
What are they doing?	Adult is flying kite and boys are watching	Playing	I can't see their faces	She is watching parasols	Spectating	They are playing watching
Is it sunny?	Yes	Yes	No	I can't tell	Yes	Yes
Are there trees?	Yes	Yes	Yes	Yes	Yes	There are no background
Is kite colorful?	Yes	No	Yes	Yes	Yes	Yes
Is kite in air?	Yes	Yes	Yes	No	Yes	No
Are boys smiling?	Yes	Yes	Yes	No	Yes	Yes

Figure 4.7 Qualitative examples of diverse answer generations from the **A** model for the 1VD task.

Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5
Is the man in a park?	I can't tell	It appears so	Yes	Yes	Yes	He is in
How old is the man?	Thirtys maybe	Locate	Maybe late thirtys	Thirtys	Thirty	Twentys
Has the frisbee left his hand?	Yes	Yes	Yes	Yes	Yes	Yes
Can you tell what he's throwing it towards?	No	No	I can't tell, is ball	I can't tell	No	Yes
What does 'casually thrown' mean?	It looks like he just tossed it	Casually	At least relief casually	I	He has in a skateboard shirt	Casually
What color is the frisbee?	White	White	White	Blue	It is white	White
What is the man wearing?	A white screen print tee	T-shirt	T-shirt and shorts	He looks casual	Shorts	T-shirt and shorts shorts
Is he wearing jeans?	I can't see his pants	No	Shorts	Yes	No	Yes
Are there other people around?	No	No	No	No	No	No
Do you see any animals?	No	No	No	No	No	No



Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5
Is it a tall wall?	Yes, it goes underwater	Yes	Yes	Yes can't tell	No	No
Is there only one polar bear?	Yes, just one	Yes	One	Yes	One	Yes
Any people there?	No	Yes	I can't see any	No	No	Yes
Do you see quite a lot of water?	Yes the polar bear is underwater	Yes	I can't see the ground	No	Yes can't see any	Yes can't tell
Can you see the sky?	No	No	Yes	Yes	Yes	Yes
Is this picture in color?	Yes it is	Yes	Yes	Yes	Yes	Yes
Is this a professional picture?	I don't think it is	Looks like relaxed	I	Yes	No	Yes
Do you see any buildings?	No	No	Yes	There are no buildings	I	No
Any trees?	No	No	There see zero	Yes	Yes	No
Is this a close-up type of photo?	No	Yes	Yes	Yes	Yes	It looks like it is in frame

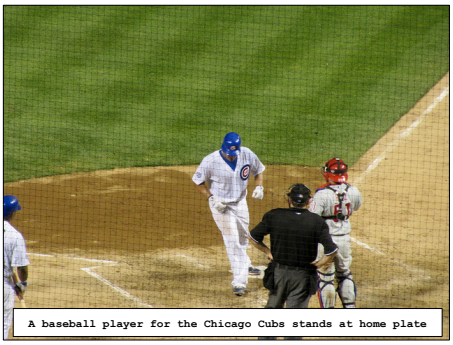
Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5
How many children are there?	Many	There are six	One	Three	There	More than ten ten
All boys or girls?	Both	I can't tell	Boy	Girl and each	It appears boy girl	They are so
How old do they look?	Six	Five	Ten-twelve	Twenty	Yes	Five looks six
Are they outside?	Yes	I	Yes	Yes	Yes	Yes
Is it sunny out?	Maybe	Yes	Yes	Yes	Yes	Yes
Are they in a field?	Yes	Yes	No	No	Yes	No
Are there many adults around?	Zero	No	Yes	No	Yes	No
Is there only one ball?	Yes	Yes	Yes	Yes	Yes	Yes
Are there soccer nets?	Not visible	No	No	Yes	I can't say	Yes
What's in the background?	Fence	Trees	I see the grass	There are numerous snow trees	Lots	I see some grass

Figure 4.8 Qualitative examples of diverse answer generations from the **A** model for the 1VD task—continued.

How many sheep are there?	Four
Are you any any water?	No
What color are they?	Brown
Any people?	No
What are the the?	Grazing
Are there people?	No
Is it sunny?	Yes
Can any animals?	No
Are any any?	No
Is any sunny?	Yes

Is this a color?	Yes
Can you see any?	No
Can you see trees?	No
Any people?	No
Is it sunny?	No
Are it sunny?	Yes
Any people?	No
Any people?	No
What time the day?	Can't tell
Are they eating?	Yes

Can you see any people?	No
Is the photo in color?	Yes
What is the fence made?	I is
Can you see any?	No
How many sheep are there?	I are
What color is the shacks?	It
Does the grass have?	Yes
Can you see any sky?	No
Is the grass green?	Yes
Do you see any sky?	No



Is the people male?	Yes
Is the person wearing?	Yes
Can you see any sky?	No
What color is the uniforms?	White
Is there grass in?	Yes
How many people are there?	Two
Is it sunny sunny?	Yes
Can you see any?	Yes
What color is the sky?	The
Is it sunny?	Yes

Can you see the ball?	Yes, I
Is the player a?	Yes, is is
Is it a a a?	Yes, is is
Can you see the baseball?	Yes, is is
What is the the wearing?	It is wearing
How is the the bat?	I is is
Is there any other people?	Yes
Can you see the pitcher?	No, the see
Can you see the scoreboard?	No
Can you see the sky?	No

Is this a professional game?	Yes
Any many players are?	One
Is the a?	Yes
Can you see the pitcher?	No
Is the a in?	Yes
Is it a in the?	Yes
What color is the batters?	White
Is there people in the?	No
Is the pitcher in?	Yes
Is it sunny?	Yes

What color is?	White
Is there a fridge?	Yes
Is there any people?	No
Is it a?	Yes
Is there a window?	No
Is there a?	No
Is there a?	Yes
Is the fridge on?	Yes
Can you see the window?	No
Is it a?	Yes

What color is the fridge?	White
What color is?	White
Can you see the sink?	Yes
Is there a window?	No
Any people?	No
Is there?	No
How pics the?	No
What color is the walls?	White
Any windows?	No
How about on?	Two

Can you see the?	No
What appliances?	Microwave
Is the fridge on?	Yes
Is it a?	Yes
Is the fridge have magnets?	Yes
Can you see what time?	Colgate
Is there a on the?	No
Can you tell what time?	No
Is there a?	Yes
Are there any people?	No

Is the photo in color?	Yes
Is it a professional photo?	Yes
Are the men wearing or?	Yes
Is any of wearing?	Yes
Are there any people in the?	Yes
How many men?	Two
Are the men wearing wearing uniforms?	Yes
Are there see other people?	No
What color are the uniforms?	White
Do the have the the?	Yes

What color are the uniforms?	One is white
Are they wearing?	Yes
Are there a?	Yes
Is there a?	Yes
Is there trees?	Yes
Is it sunny?	Yes
Are the wearing wearing?	Yes they does
Are they wearing cleats?	Yes
Are there a fence?	Yes
Can you see the?	No

What is the the of?	One
Are there see other	Yes
Are they people uniforms?	Yes
Is they wearing uniforms?	Yes
What color are the uniforms?	White
Does the have wearing hair	Yes
What color the wearing?	A
Can you see any ball?	No
Can you see any coaches?	No
Do you see the color?	Yes

Figure 4.9 Extended qualitative examples of diverse two-way dialogue generations from the $\mathbf{B_{AR}}10$ model (block evaluation) for the 2VD task.

Is the picture in color?	Yes
Is the photo in?	Yes
What color is the frisbee?	White
Can there?	No
Is the grass visible?	Yes
Is the sunny?	Yes
Is it sunny?	Yes
Is there other people?	No
What color is the frisbee?	White
Is the man green?	Yes

How many people are there?	Two
What are they?	I
Can you see the frisbee?	Yes
What color of the frisbee?	Blue
Is there people?	Yes
Is it sunny?	Yes
What color is the?	I
Is there sunny?	Yes
How many people are there?	I
Is they having fun?	Yes

What color is the man's?	White
What is the man man?	I
How old is the man?	I
Can you see any ball?	Yes
Is there any other people?	No
Does the man have a?	No
Is the man wearing a hat?	No
Does he man have a?	No
Are there see any other people?	No
Can you see any sky?	No

Is it big?	Yes
How is of the is?	I
Can you see the breed?	No
Are there any people?	No
Is the dog on?	No
Can you tell the breed of it?	No
What is the color?	It is
Any there see the driver?	No
Is it a Harley?	Yes
Any other?	No

Are there only one?	Yes
What kind of dog is it?	I is like
Does it look like be?	It, it
Is there any people in the boat?	No
What time of day is it?	It is light
What color is the boat?	White
How it is like is?	I is
Is it a or or?	It Is
What is of day is?	It is
What color is the boat?	It

What kind of dog is it?	It looks like a mutt
Does he have a collar?	Yes
Can you see what breed?	No
Is it people boat the the boat?	No
Is there oars?	No
How the boat the the?	I
Is it sunny?	Yes
How many barges?	No
Any people?	No
Any?	No

What color is the man?	White
What color is his hair?	Black
How old is he?	Thirtys
Is he have other?	Yes
Is it sunny?	Yes
Can you see any?	Yes
Is it court?	Yes
Is it sunny?	Yes
Is it sunny?	Yes
Can you see sky?	No

Is this a color photo?	Yes
What color is the tennis?	It
Are there any other people in the?	No
Can you see any tennis?	No
Can you see any people in the?	No
What time of day is it?	It
Is the appear like be be?	It
Are there any other in the the?	No
Can you see the time?	No
Can you see the season?	No

What is the color?	Is
Is the player?	Yes
Is it a?	Yes
Is it outdoors?	Yes
Is there people?	No
Is he playing?	Yes
Is it a?	Yes
Is it a?	Yes
Can you see the net?	No
Is it man?	Yes

Figure 4.10 Extended qualitative examples of diverse two-way dialogue generations from the $\mathbf{B_{AR}}10$ model (block evaluation) for the 2VD task–continued.

Chapter 5

On the Evaluation of Visual Dialogue

Daniela Massiceti Viveka Kulharia Puneet K. Dokania N. Siddharth Philip H.S. Torr

Engineering Science, University of Oxford

Abstract

Visual Dialogue (VD)—the task of answering a sequence of questions relating to a visual input—involves subjective human judgements and presents a challenge to evaluate. *VisDial* [77], a widely used dataset, evaluates performance based on an answer-ranking task with a trained model. In exploring some of the implicit correlations in the dataset, we observe that an extremely simple baseline model employing canonical correlation analysis (CCA) between given questions and answers, manages to perform comparably to state-of-the-art models on mean rank (MR). An analysis of this result leads to insights about the evaluation protocol for *VisDial*, revealing several issues, most crucially the limitations imposed by rank-based metrics. To address these issues, we present a suite of alternate evaluations taking inspiration from the NLP literature, which we posit are better suited to the nature of the VD task, by capturing consensus between a generated answer and a set of relevant answers. To extract relevant answer sets, we propose a semi-supervised method based on correlation, which allows us to automatically extend and scale sparse human annotations to the entire dataset, for use in evaluation and further model development. We intend this alternate paradigm as a next-best solution in the face of constraints on the dataset and its experimental design choices, and hope that the community adopts these evaluations going forward.

Under review[1]

[1] A precursor to this work was published as a *NeurIPS* workshop paper [217].

5.1 Introduction

Recent years have seen a great deal of interest in autonomous conversational agents with the goal of facilitating natural language interaction between humans and machines. Early pioneering efforts of this include ELIZA [351] and SHRDLU [360]. This resurgence of interest builds on the ubiquitous successes of neural network-based approaches in the last decade, particularly in the perceptual domains of vision and language.

A particularly thriving sub-area of interest in this domain is that of visually-grounded dialogue, termed visual dialogue (VD), in which an agent converses with a human about visual content [77, 78, 219]. Specifically, VD involves answering questions about an image, given some dialogue history—a fragment of previous questions and answers. As is standard practice in machine learning, learning VD involves defining an objective to achieve, procuring data with which to learn, and establishing a measure of success at the stated objective.

The objective for VD is reasonably clear at first glance—answer in sequence, a set of questions about an image. The primary choice of dataset, *VisDial* [77], addresses precisely this criterion, involving a large set of images, each with a dialogue—a set of question-answer pairs—collected by pairs of human annotators playing a game to understand the images. And finally, the evaluation measures used judge how well the specified objective is being achieved, generally assess the ability to match a human-derived "ground-truth" answer.

However, quirks in the choices of the above factors can lead to algorithms with unintended behaviour [3, 18, 217] (Figure 5.1) due to implicit biases in data and evaluations. These quirks may conspire against one another—a poorly defined evaluation metric can mask underlying biases in the data, and an ill-formed objective may in turn make faithful evaluation difficult.

To determine the extent to which these quirks exist in the *VisDial* dataset, we establish a simple baseline which applies canonical correlation analysis (CCA) between the questions and answers alone, to perform the traditional answer-ranking evaluation called for by the dataset. Intriguingly, we find that this simple CCA-based approach is comparable in mean rank (MR)—one of the dataset's primary metrics—to state-of-the-art (SOTA) approaches all of which employ complex neural network architectures, complicated training schemes over

millions of parameters, and many hours of time and expensive GPU resources. In comparison, CCA uses standard off-the-shelf feature extractors, avoids computing gradients, involves few parameters and requires just a few seconds on CPU—all without the image or dialogue! This surprising finding suggests that, within the current formulation, there is a mismatch between the VD task and the evaluation employed by the *VisDial* dataset in measuring the ability to faithfully answer a given question.

Caption: A man and a woman sit on the street in front of a large mural painting.

Question	**Answer**
How old is the baby?	About 2 years old
What color is the remote?	White
Where is the train?	On the road
How many cows are there?	Three

Figure 5.1 Failures in *visual* dialogue.
Visually-unrelated questions, and their visually-unrelated plausible answers[2]

Motivated by this, we explore the factors giving rise to this behaviour, and in particular address limitations imposed by single-candidate answer ranking schemes that are currently the primary evaluation paradigms of the dataset. Here, for a given question, the model scores a corresponding set of 100 candidate answers, and the position, or rank, of the labelled "ground-truth" answer in the sorted set is observed. Specifically, we show that within each set of candidates there exists an essential answer equivalence class—any one of which would feasibly answer the given question. Rank-based metrics therefore lose meaning when computed within this set, having knock-on effects for model comparison and performance quantification.

A recent update to *VisDial* introduces relevance scores for answers within the candidate sets, obtained from human annotators. While a step in the right direction, only a fraction of the

[2]From online demos of SOTA models–VisDial [77] and FlipDial [219]

dataset is annotated with such scores, a function of the cost of such an exercise at scale, and the difficulties with extracting consistent human judgements on natural language expressions.

To address these issues, we propose changes to how the *VisDial* dataset is evaluated. We adopt community-accepted automatic metrics with roots in the natural language processing (NLP) literature to quantify the *consensus* between an answer *generated* by a model and a reference set of relevant answers given the image and question. This more closely aligns with the VQA [24] evaluation set-up in which the predicted answer is compared to a *set* of answers provided by multiple human annotators. Since reference answer sets are only available for a fraction of the *VisDial* dataset, we propose a semi-supervised approach to automatically extract groups of relevant answers from candidate sets, using CCA applied to the questions and answers. We then apply a simple heuristic on the correlations between the ground-truth answer and other answers in the candidate set, to construct a cluster of relevant answers, thus expanding reference sets to the rest of the *VisDial* dataset.

We consider this alternate evaluation suite to be a solution for improved evaluation on the *VisDial* dataset, as far as the experimental design under which the data was collected allows. Going forwards, evaluation metrics should look beyond syntactic-based matching schemes, and toward metrics where performance on a downstream task, necessitating use of language, is measured [31, 78, 80, 301]. More generally, however, we acknowledge that the evaluation of language is an open research challenge, and we recognise that all efforts, including this work, are with the intention to move the field forwards.

To summarise, our contributions are:

1. An alternate evaluation suite for *VisDial* drawing on existing metrics from NLP to measure the consensus between a *generated* answer and a reference set of answers.
2. A semi-supervised method for automatically extracting reference sets of answers from given candidate sets for questions and images which are as high quality as those provided by human annotators.
3. A subsequent expansion of the *VisDial* dataset, including reference set annotations, which we will release as a baseline for future evaluation and model development.

5.2 Preliminaries

5.2.1 Canonical Correlation Analysis (CCA)

We begin with the preliminaries for CCA [138] and its multi-view extension [169]. Given paired observations $\{\mathbf{x}_1 \in \mathbb{R}^{n_1}, \mathbf{x}_2 \in \mathbb{R}^{n_2}\}$, (2-view) CCA jointly learns projections $W_1 \in \mathbb{R}^{n_1 \times k}$ and $W_2 \in \mathbb{R}^{n_2 \times k}$, $k \leq \min(n_1, n_2)$, which are maximally correlated.

Multi-view CCA, a generalisation of two-view CCA, extends the above formulation to associated data across m domains, learning projections $W_i \in \mathbb{R}^{n_i \times k}, i = 1, \ldots, m$. Kettenring [169] shows one formulation (among several) which minimises the Frobenius norm between each pair of views, with constraints over the projection matrices themselves [124]. Optimising multi-view CCA then reduces to solving a generalised eigenvalue decomposition, $A\boldsymbol{v} = \lambda B\boldsymbol{v}$ [32], where A and B are derived from the inter- and intra-view correlation matrices (see Section 5.7.1).

The top k (eigenvalue-sorted) eigenvectors are chosen, and for the i^{th} view, their i^{th} sub-components are column-wise stacked to construct the projection matrix $W_i \in \mathbb{R}^{n_i \times k}$. This matrix is then used to embed a sample $\mathbf{x}_i$ from view i as $\phi(\mathbf{x}_i; W_i) = D_{\boldsymbol{\lambda}}^p W_i^\top \mathbf{x}_i$, where D_λ is a diagonal matrix of the top k (sorted) eigenvalues $\boldsymbol{\lambda}$, and p is a scalar controlling the extent of weighting—reducing to the typical case at $p = 0$[3].

With this formulation, a variety of tasks can be tackled at test time—ranking and retrieval across all possible combinations of multiple views—simply by computing correlation between projections of any pair of inputs $\{\mathbf{x}_i, \mathbf{x}_j\}$:

$$\text{corr}(\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)) = \frac{\psi(\mathbf{x}_i)^\top \psi(\mathbf{x}_j)}{\|\psi(\mathbf{x}_i)\|_2 \, \|\psi(\mathbf{x}_j)\|_2} \tag{5.2.1}$$

where $\psi(\cdot)$ is a mean-centred (over *train* set projections) version of projection $\phi(\cdot)$.

5.2.2 Visual Dialogue

The visual dialogue (VD) task involves answering a sequence of visually-grounded questions. Given image I and dialogue history $[(Q_1, A_1), (Q_2, A_2), \ldots, Q_i]$, the goal is to produce A_i.

[3] $p > 0$ has been shown to give better performance sometimes [109].

The most prominent push towards this goal has been facilitated by the *VisDial* dataset [77], a large corpus of images paired with question-answer dialogues, sequentially collected by pairs of annotators in an interactive game on Amazon Mechanical Turk (AMT). *VisDial v1.0* comprises $123,287/2064/8000$ train/val/test images, each paired with dialogues of up to 10 exchanges[4]. Each question is coupled with a candidate set of 100 answers $\boldsymbol{A}$ including a ground-truth answer $A_{\text{gt}} \in \boldsymbol{A}$, and human-annotated relevance scores $\rho(A) \in [0, 1]$ where $A \in \boldsymbol{A}$, for a subset of the data where one question per image is annotated (2000/2064/8000 images respectively).

The VD task is formulated by coupling each (I, Q, A) triplet with the candidate answers $\boldsymbol{A}$. Two learning paradigms arise from this formulation [77]. In the first, termed the *discriminative* setting, the model has access to $(I, Q, \boldsymbol{A})$ at train time, hence framing the predictive task as a classification problem of selecting A_{gt} out of $\boldsymbol{A}$. In the second setting, candidate answers are *not* provided. Instead, the model is given only (I, Q, A_{gt}) during training, and must learn to *generate* an answer conditioned on an image-question pair. This is termed the *generative* setting, and is the setting of primary interest here.

At test time, for an image-question pair, the model scores its associated candidate answers. The rank of A_{gt} and a normalised discounted cumulative gain (NDCG) on the candidate answers' human-annotated relevance scores is then used to judge the model's effectiveness at the VD task. A model targeting the discriminative paradigm scores candidates directly by the classifier's *softmax* probabilities, while a model targeting the generative paradigm scores each candidate under its learned likelihood.

5.3 CCA for Visual Dialogue

CCA is chosen as our diagnostic model for VD for two principal reasons: i) it explicitly targets correlations, which is ideally suited to probing the relationship between questions and answers (and images) for the VD task, and ii) as stated in Section 5.2.1, computing CCA is extremely simple, involving a single eigendecomposition, and avoiding complex networks, gradient computation, and large GPU resources and time. A stark representation of the latter is shown in Table 5.1 for our application of CCA (on CPU) between just questions and answers.

[4] 10 exchanges for train/val, and $\leq$ 10 exchanges for test.

Table 5.1 CCA vs. SOTA computation comparison.

Model	Number of Parameters	Train time (s)
HCIAE-G-DIS [206]	2.12×10^7	–
HREA-QIH-G [77]	2.42×10^7	1.2×10^5
FlipDial [219]	1.70×10^7	2.0×10^5
CCA A-Q	1.80×10^5	2.0
Factor (≈)	**90**	$\mathbf{10^5}$

We first transform the images I, questions Q, and answers A into lower-dimensional fixed-length feature vectors using pre-trained feature extractors, as is commonly done [77, 206, 364], and use CCA to learn joint embeddings for different subsets of questions, answers, and images. We learn joint embeddings for two settings: **A-Q** between answers and questions, using two-view CCA, and **A-QI** between answers, questions, and images, using three-view CCA. The first explores the (lack of) utility of the image in performing the VD task, and the second serves as a useful indicator of how unique any question-image pairing is. Improved performance of **A-QI** relative to **A-Q** would indicate that the image *does* contextualise correlations between questions and answers, with the converse indicating otherwise. Note that in both cases, dialogue history is ignored and the answer is ever-present in service of the VD task. We include a further ablation **NCCA A-Q** (no-CCA), by computing correlation, as centred cosine distance, directly between the features for questions and answers. As these models do not include a traditional likelihood to score candidates with, as in current SOTA models [77, 219], we score using correlation between the question and answer embeddings using Eq. 5.2.1, a schematic of which is shown in Figure 5.2.

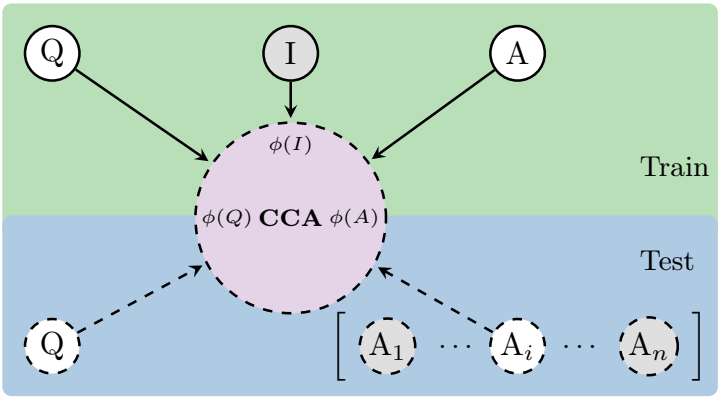


Figure 5.2 Experimental set-up for VD on *VisDial*: CCA is used to learn a joint embedding between train questions and answers (and images), with $k = 300, p = 1$ chosen through a simple grid search.

5.3.1 Experimental details

Feature extractors For the images, we employ pre-trained *ResNet34* [132], extracting a 512-dimensional feature—the output of the *avg pool* layer after *conv5*. For the questions and answers, we employ *FastText* [45] to extract 300-dimensional embeddings for the words. To obtain (300-dimensional) sentence representations, we average word embeddings, following generally received intuition [30, 358], with suitable padding or truncation (up to 16 words). Our choice of feature extractors is largely arbitrary, however, to rule out gains from one feature extractor over another, and to compare against SOTA models [77, 206, 364] which use others, we also employ *VGG-16/19* [305] (4096-dimensional, extracted after the second *fc-4096* layer) and *GloVe* [264] (300-dimensional, from Common Crawl-42B tokens) features for the images and questions/answers, respectively.

SOTA models We compare results against SOTA models [77, 206, 364]. For each, we select their top-performing model variant under the generative setting (see Section 5.2.2), train the models on *VisDial v1.0* train set, cross-validate on mean reciprocal rank (MRR), and select the best for evaluation.

5.3.2 Experimental analysis

Table 5.2 shows the results of running CCA with different pre-trained feature extractors on *VisDial (v1.0)*, comparing against SOTA models. Note that for A-QI (Q) at test time, correlation is computed between questions and candidate answers $\boldsymbol{A}$ using projection matrices learned using images (I) as well. An indicative selection of results are shown here, with extended results on the older *VisDial (v0.9)* dataset and ablative baselines shown in Section 5.7.2.

Surprisingly, we observe that CCA A-Q, which is a simple, lightweight model that completely *ignores the image and dialogue sequence*, performs favourably in mean rank (MR) with current SOTA models (16.60 MR), all of which require at least an order of magnitude more learnable parameters and many hours on GPU (versus CCA's seconds on CPU) to train (Table 5.1). This supports the impression, from Figure 5.1, that there exist implicit correlations between just the questions and answers in the data, rendering it possible to perform visual dialogue without heed to either the visual or the sequential dialogue aspects.

Table 5.2 Results of CCA vs. SOTA on the *VisDial v1.0* dataset. CCA achieves comparable performance in mean rank (MR) while ignoring both image and dialogue sequence.

	Model	I/QA features	MR	R@1	R@5	R@10	MRR	NDCG
SOTA	HCIAE-G-DIS (*v0.9*) [206]	VGG-19/learned	14.23	44.35	65.28	71.55	0.5467	N/A
	CoAtt-GAN (*v0.9*) [364]	VGG-16/learned	14.43	46.10	65.69	71.74	0.5578	N/A
	HREA-QIH-G [77]	”	19.15	34.73	56.55	63.18	0.4555	0.5189
CCA	A-Q	GloVe	16.60	16.10	39.38	54.68	0.2824	0.3504
		FastText	17.07	16.18	40.18	55.35	0.2845	0.3493
	NCCA A-Q	FastText	55.12	5.53	11.89	16.35	0.1015	0.1760
	A-QI (Q)	ResNet34/FastText	19.25	12.63	32.88	48.68	0.2379	0.3077
		VGG-16/GloVe	19.11	13.53	32.43	47.13	0.2415	0.3071
		VGG-19/GloVe	19.29	13.38	32.73	47.23	0.2415	0.3000

However, while CCA's performance is comparable on the MR of the ground-truth answer A_{gt}, it does not appear to do as well on other related metrics such as mean reciprocal rank (MRR) and recall at the top 1, 5, and 10 answers, which measures how often the ground-truth answer has rank within 1, 5, and 10 respectively. To understand this disparity across metrics, we performed a simple qualitative analysis, actively *anti-cherry-picking* results where the CCA model gave A_{gt} a poor rank. For such examples, we wished to see how qualitatively good the top-ranked answers actually were. A selection of these results is shown in Figure 5.3, which surprisingly show that even though the rank for A_{gt} is poor, the top ranked answers are perfectly feasible answers to the given question with regard to the intended image. We observed that this is in fact a consistent pattern across our results, where the top-ranked answers through CCA were indeed feasible answers to the given question, if not actually matching the intended ground-truth answer A_{gt}.

Q: What colour is the bear?
Ranked Ans
(61) Floral white (A_{gt})
(1) White and brown
(2) Brown and white
(3) Brown, black & white

Q: Does she have long hair?
Ranked Ans
(41) No (A_{gt})
(1) No, it is short hair
(2) Short
(3) No it's short

Q: Can you see any passengers?
Ranked Ans
(48) Not really (A_{gt})
(1) No
(2) Zero
(3) No I cannot

Q: Are there people not on bus?
Ranked Ans
(22) Few (A_{gt})
(1) No people
(2) No, there are no people around
(3) I don't see any people

Figure 5.3 Qualitative results for the A-Q model showing the top 3 ranked answers for questions where the ground-truth answer is given a low rank—showing them to be perfectly feasible.

Q: Are they adult giraffe?
GT Yes
Ranked Ans
1. Yes the giraffe seems to be an adult
2. It seems to be an adult, yes
3. The giraffe is probably an adult, it looks big
4. Young adult

Q: Are there other animals?
GT No
Ranked Ans
1. No, there are no other animals
2. No other animals
3. There are no other animals around
4. Don't see any animals

Q: Any candles on cake?
GT Just a large "number one"
Ranked Ans
1. There are no candles on the cake
2. I actually do not see any candles on the cake
3. No, no candles
4. No candles

Q: Is the cake cut?
GT No, but the boy has sure had his hands in it!
Ranked Ans
1. No it's not cut
2. No the cake has not been cut
3. Nothing is cut
4. No, the cake is whole

Figure 5.4 Example answers generated by CCA-AQ-G using the nearest-neighbours approach.

While these results indicate that CCA is able to select relevant answers from the candidate set $\boldsymbol{A}$, there still remains the question of CCA's general utility to serve the generative setting (see Section 5.2.2) of the VD task. To explore this ability, we adopt a nearest-neighbour approach based on CCA-derived correlations to "generate" answers. For a given test question, we select the 100 closest questions across the entire train set using correlations computed with the A-Q model. Using their 100 corresponding ground-truth answers, we construct a pseudo-candidate set that we then order based on correlation to the test question. We then "generate" answers by sampling from this set in proportion to the correlation, denoting the model CCA-AQ-G. Results in Figure 5.4, indicating the top 4 answers with highest correlation, suggest that CCA can leverage the correlations quite well.

Taken together, the quality of results indicated through Figures 5.3 and 5.4, and the disparity to SOTA models on non-MR metrics seem to indicate a mismatch between the task of VD and the evaluations actually being employed. We follow this line of reasoning further, prompting an exploration into the factors unsuspectingly affecting *VisDial*'s evaluation, from which we make the case that to better capture a model's ability to answer a visually-grounded question—the very goal of VD—necessitates changes to the evaluation.

5.4 Analysing *VisDial* evaluations

A particular observation from our analysis above is that there are many ways to satisfactorily answer a given question, even within the selection of 100 candidate answers. Indeed, by construction, the candidate sets contain multiple plausible answers including up to 50 nearest neighbours to A_{gt} in *GloVe* [264] space. With this in mind, the standard evaluations which focus on the rank of a single, privileged, ground-truth answer would appear to be unnecessarily constrictive as far as targeting the actual *task* of VD. This contrasts established practice in the VQA literature [24] which evaluates by comparing the predicted answer to answers collected from 10 human annotators for a single question.

While the effect of this constraint is obvious in the case of MR and recall (R@1,5,10), it is a little more subtle in the case of MRR. MRR, being the inverse harmonic mean of ground-truth answer ranks, by extension of the properties of the harmonic mean weighs instances with better rank (closer to 1) more than those with worse rank (further from 1)—a bias that is not necessarily well motivated given the underlying task, and the observations above. It is indeed instructive that these metrics have ubiquitously been used to evaluate performance on the *VisDial* dataset since its inception.

Single-candidate ranking The key limitation of single-candidate based ranking metrics is that they are insufficient to deal with a task as complex as VD, and are unable to *rule out* poor models. In other words, models with poor MRR or recall aren't necessarily poor at VD. Even MR is only weakly indicative of performance as the following heuristic demonstrates: using the A-Q model, we first compute correlation within a candidate answer set between A_{gt} and $A \in \boldsymbol{A} \setminus A_{\text{gt}}$, giving $\mathcal{C} = (\phi(A_{\text{gt}}, A_1), \ldots, \phi(A_{\text{gt}}, A_{100}))$. We then select the subset of answers with correlations in $[\mathcal{C}_{\max} - \sigma, \mathcal{C}_{\max}]$, where $\mathcal{C}_{\max} = \max(\mathcal{C}), \sigma = \text{stdev}(\mathcal{C})$, roughly estimating answers which are plausibly similar to A_{gt}. Statistics are then computed to measure how small and tightly packed these clusters are—the mean and standard deviation of the correlations within the cluster as well as cluster size. We average these across all candidate sets, giving an average mean correlation of 0.58, an average standard deviation of 0.22, and an average cluster size of 12.30. This indicates an *equivalence class* of plausible answers, within which ordering in terms of rank can be meaningless with regard to the VD task.

Multi-candidate ranking & NDCG To ameliorate issues with the single-candidate ranking, the updated *VisDial v1.0* dataset tasked 4-5 human annotators with labelling whether each answer in a candidate set answers a given image-question pair (as a hard 0/1 choice). For each candidate answer A, average judgement across annotators becomes a *relevance* score $\rho(A) \in [0, 1]$. With these scores, a modified evaluation metric is introduced, the normalised discounted cumulative gain (NDCG), which weighs predicted rankings by the scores, excluding irrelevant answers ($\rho(A) = 0$). Further details on NDCG are given in Section 5.7.3. While undoubtedly a step in the right direction, reliance on NDCG-weighted ranks can still be fraught with difficulty, although to a lesser degree than the MR, MRR, and recall.

Two aspects in particular can be troublesome: i) the sheer scale of the *VisDial* dataset makes eliciting reliable human judgements a daunting task—reflected in the fact that only a small fraction of the data, one question per image, actually has such judgements (see Section 5.3.1), and ii) the conversion of binary *validity* judgements into relevance scores over a small handful (4-5) of annotators can lead to quirks—especially for perceptual judgements on natural language expressions. For the first, having annotations for just a single question per image itself ignores the notion of dialogue sequence. And the second leads to quirks such as 18.15%/47.14% of the validation/train annotated subsets do not have a *single* candidate answer with relevance score 1.0, not even the ground-truth, indicating relatively poor consensus across annotators. Moreover, 20.69%/9.01% of samples, respectively, do not assign a relevance score to the ground-truth answer at all ($\rho(A_{\text{gt}}) = 0$).

Generating answers A more nuanced issue with current evaluation methods, however, lies in ranking the answers of a candidate set. The ultimate task of VD is to provide an answer to a given question, not to pick an answer from a set. This is why we are primarily interested in the *generative* setting of the VD task, as stated in Section 5.2.2. While it is expected that the likelihood of a valid answer is a reasonable measure of a model's ability to *generate* a good answer, this may not necessarily be the case when there are multiple potential answers, some not even in the candidate set. It is possible that the likelihood serves well as a *relative* measure between candidates, but the highest-probability answers are entirely different or unrelated—indicating a poorly learnt model. This supports the idea that a better metric should instead directly evaluate an answer *generated* from the model.

5.5 An alternate evaluation for the *VisDial* dataset

The analysis in Section 5.4 indicates that an evaluation that is well matched to the underlying goals of VD should:

i) use more than just a single valid answer as reference,
ii) directly use generated answers from the model, and
iii) do the above at scale over the entire dataset.

In the following, we propose an alternate suite of evaluations which attempts to cover the above desiderata for the *VisDial* dataset and for the task of VD.

Multiple reference answers Based on the discussion in Section 5.4 on multiple-candidate ranking and NDCG, using human judgements of answer validity is a step in the right direction. Here, we propose employing validity directly to construct reference sets. That is, for a given question, any candidate answer chosen by a annotator to be valid automatically qualifies as a reference answer, thus defining the human reference set as $H = \{A : \rho(A) > 0, \forall A \in \boldsymbol{A}\} \cup \{A_{\text{gt}}\}$. We union with A_{gt} to protect against cases where $\rho(A_{\text{gt}}) = 0$.

Evaluating generations Armed with this reference set, we turn to established metrics from the NLP literature to measure *consensus* between a *generated* answer and a reference set of valid answers. Crucially, such consensus-driven metrics require a reference set with more than one element; the more, the better. Here, we explore two classes of metric for capturing consensus, based on overlap and embedding distances.

Overlap-based metrics compute overlap or co-occurrence of n-grams (word couplets of size n) between pairs of sentences—here, the generated answer and $h \in H$. We explore the use of three such metrics: CIDER [338], BLEU [255], and METEOR [186]—all, particularly CIDER, known to be well correlated with human judgements. CIDER computes the cosine similarity between a pair of vectors, each of which is composed of the term-frequency inverse-document-frequencies (tf-idf) of the sentence's n-grams. For $0 < n \leq i$, similarities are averaged over all n-grams up to length i. BLEU and METEOR are similar,

but are computed on the raw term-frequencies, favouring precision and recall respectively. For METEOR, a uni-gram matching function precedes the overlap computation.

Embedding distance-based metrics arise from a rich literature in capturing semantic similarity between natural language expressions [45, 84, 264, 265, 302]. They guard against limitations of overlap-based metrics regarding short (one- or two-word) sentences, which are frequency in VD and VQA data. The recent successes of BERT [84] and *FastText* [45] on a host of downstream NLP tasks including question-answering and semantic similarity, make them apposite choices. The L_2 or cosine similarity (CS) between sentences in embedding space is typically computed.

Table 5.3 shows overlap and embedding-distance scores for models trained on the full training set of *VisDial v1.0*, and evaluated on the validation subset for which human-annotated relevance scores are available (denoted $\mathcal{H}_v$) averaged over answers in reference sets H. For each model, scores are additionally averaged over their top 10 generations—with the HREA-QIH-G* model employing a beam search to generate answers [77]. We evaluate on the validation set since human-annotated relevance scores are not publicly available for the test set.

We further define a reference baseline for the overlap metrics, estimating upper bounds for the respective scores as Γ_H, which iteratively scores answers in H against H itself, taking the maximum over the resulting scores. We do not compute the reference baseline for the embedding distance metrics since lower bounds on average L_2 distance, and upper bounds on average cosine similarity are hard to define on arbitrarily scaled continuous embedding spaces such as those of BERT and *FastText*.

Table 5.3 Overlap and embedding distance metrics on the human-annotated validation set $\mathcal{H}_v$, averaged over the top 10 generated answers for each model, against H. For HREA-QIH-G, on average ∼6 answers are the empty string, which are excluded from the computation. Metrics marked ↑ indicate higher values are better, and those marked ↓ indicate lower values are better.

Model	CIDER↑				BLEU↑				METEOR↑	BERT		FASTTEXT	
	n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4		L_2↓	CS↑	L_2↓	CS↑
Γ_H	0.1915	0.1437	0.1165	0.0962	1.0000	0.7633	0.5974	0.3371	1.0000				
CCA-QA-G	0.0708	0.0420	0.0291	0.0221	0.4156	0.1539	0.0516	0.0149	0.2721	7.0301	0.8702	2.8850	0.4889
HREA-QIH-G	0.0558	0.0303	0.0206	0.0156	0.3919	0.0648	0.0169	0.0039	0.2958	6.6661	0.8796	2.9599	0.4907
HREA-QIH-G*	0.0937	0.0600	0.0438	0.0337	0.5825	0.2739	0.1423	0.0461	0.4310	6.0202	0.9031	2.8534	0.5117
HCIAE-G-DIS	0.1053	0.0570	0.0390	0.0294	0.6352	0.1031	0.0443	0.0108	0.5449	5.4297	0.9120	2.3339	0.6067

The results highlight an interesting aspect of the proposed alternative evaluations: different models appear to be doing well on the two classes of metrics. While the CCA model serves as a useful blind baseline on both sets of metrics, HCIAE-G-DIS typically performs better on the embedding distance metrics, whereas HREA-QIH-G performs better on the overlap metrics. This difference, we believe, suggests that the different models are better suited to different downstream tasks—with models that performs better on the overlap metrics being potentially better suited for human-interactive use, and models that do better on the embedding distance metrics potentially better suited to further downstream NLP tasks such as sentiment analysis or summarisation. Indeed, this is the sort of flexibility of purpose we would expect when dealing with evaluations for complex multi-modal tasks such as visual dialogue.

Scaling-up through weak supervision While the results of Table 5.3 support the utility of our alternative evaluations, they were derived from a small-scale experiment. Extending the paradigm to the whole dataset is problematic since human annotations of relevance are only available for a fraction of the data—less than 1% of the questions in the dataset. In general, given the scale of the *VisDial* dataset, eliciting human judgements for all the data is a challenge—there are order 10^6 questions, each with 100 candidate answers! In order to acquire valid reference sets at scale, circumventing the cost and idiosyncrasies associated with humans, we propose a semi-supervised approach that harnesses the annotations we *do* have, to derive an automated method to construct the reference sets from given candidate sets.

We learn a CCA model between questions and answers in the relevance-annotated subset $\mathcal{H}_t$ of the full train set, pairing each question with *all* answers where $\rho(A) > 0$. With this model, denoted CCA A-Q*, we compute the correlation within a candidate set between A_{gt} and $A \in \boldsymbol{A} \setminus A_{\text{gt}}$, giving $\mathcal{C} = (\phi(A_{\text{gt}}, A_1), \ldots, \phi(A_{\text{gt}}, A_{100}))$ similar to Section 5.4. We then evaluate 3 approaches to constructing a reference set C of answers similar to A_{gt}, using $\mathcal{C}$:

Simple: $\Sigma = \{A : \phi(A_{\text{gt}}, A) \in [\mathcal{C}_{\max} - \sigma, \mathcal{C}_{\max}]\} \cup \{A_{\text{gt}}\}$, where $\mathcal{C}_{\max} = \max(\mathcal{C}), \sigma = \text{stdev}(\mathcal{C})$.

Meanshift: choosing the best-ranked cluster M' after running meanshift [70] on $\mathcal{C}$ to derive $M = M' \cup \{A_{\text{gt}}\}$.

Agglomerative: choosing the best-ranked cluster G' after running agglomerative clustering on $\mathcal{C}$, with number of clusters set to 5, to derive $G = G' \cup \{A_{\text{gt}}\}$.

Table 5.4 shows estimated intersection-over-union (IOU), precision, recall, and set size for the different methods C on the corresponding validation set $\mathcal{H}_v$, against the human-annotated reference set H. We select one with *both* the best precision, i.e., one that selects answers that are maximally in the reference set H, *and* the smallest cluster size $|C|$, i.e., one that likely excludes irrelevant answers, giving us $C = \Sigma$. A detailed and thorough evaluation of these approaches, including tests over hyper-parameters, is provided in Section 5.7.5.

Table 5.4 Evaluation of alternate methods for automated reference set construction on the human-annotated validation set $\mathcal{H}_v$, against H. Values in parentheses denote standard deviation across the set.

C	$\frac{\lvert H \cap C \rvert}{\lvert H \cup C \rvert}$	$\frac{\lvert H \cap C \rvert}{\lvert C \rvert}$	$\frac{\lvert H \cap C \rvert}{\lvert H \rvert}$	$\lvert C \rvert$
Σ	24.13 (16.73)	62.48 (31.24)	32.91 (23.52)	7.17 (6.94)
M	25.01 (18.40)	59.19 (31.55)	39.35 (28.86)	12.00 (16.07)
G (n=5)	25.59 (17.69)	59.20 (30.71)	35.74 (24.47)	7.91 (6.18)

The overlap of the constructed reference sets $C = \Sigma$ and human-annotated reference sets H supports the observation that our semi-supervised method is capable of extracting clusters of similar answers. We additionally validate the constructed sets by testing their utility, relative to H, in improving performance on the VD task—if the clusters are meaningful and contain answers similar to the correct answer, then a model trained on these clusters should show improved answering performance. Short of turning to human evaluators to manually validate the constructed sets, which would be both cost- and time-intensive, we employ this as an automatic test of the feasibility of the sets relative to the original ground-truth answers.

As a baseline, we first pair questions in the $\mathcal{H}_t$ subset with each of the answers from their human-annotated reference set H, and train a CCA A-Q model. We then repeat the experiment, but this time pairing the questions with answers from the automatically computed reference set Σ instead of H. As shown in the top two rows of Table 5.5, the model trained using Σ performs better than that employing the human-annotated reference sets H across the battery of metrics, including NDCG. As a further sanity check against potential quirks with using the subset $\mathcal{H}_t$, we train a CCA model on that subset, but only between questions and their single ground-truth answers A_{gt} (as opposed to *all* answers in H or Σ). As we address in Section 5.4, the single-candidate ranking metrics actually show that this model outperforms the baseline using H as the reference, but, as expected, NDCG paints a better picture, showing reduced performance.

Finally, having compared the utility of automated reference set construction against human annotations, we conduct an experiment across the whole dataset, learning a CCA model between questions and answers in reference set Σ, over the entire training data of *VisDial (v1.0)*. The last two rows of Table 5.5 compare this model against the standard CCA Q-A model (Table 5.2) trained on questions and ground-truth answers. We observe a substantial improvement in NDCG, with what is effectively a simple data augmentation procedure, using Σ. These experiments suggest that the automated reference set construction is indeed fit for purpose, being well-matched to human judgements of answer validity in the candidate sets.

Table 5.5 Evaluation of the utility of automated reference set construction Σ on the standard VD evaluation. Models were trained using CCA on the indicated subsets ($\mathcal{H}_t$ or all) of *VisDial (v1.0)*, using answers drawn from different sets ('Ref'), and tested on the evaluation test server to compute standard metrics. Arrows indicate which direction is better.

Train Set	Train #QA pairs	Ref	MR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MRR ↑	NDCG ↑
$\mathcal{H}_t$	15,317	H	26.49	6.05	21.50	35.53	0.1550	0.3647
	17,055	Σ	**20.36**	8.35	32.88	**48.78**	0.2066	**0.3715**
	1996	$\{A_{\text{gt}}\}$	23.71	**13.13**	**34.05**	46.90	**0.2428**	0.2734
all	10,419,489	Σ	17.20	10.73	34.20	51.80	0.2312	**0.4023**
	1,232,870	$\{A_{\text{gt}}\}$	**17.07**	**16.18**	**40.18**	**55.35**	**0.2845**	0.3493

Putting it all together Having validated the method to construct reference sets, we now take the final step of evaluating both CCA and SOTA models on the complete *VisDial (v1.0)* dataset. Table 5.6 shows the overlap and embedding distance scores for models trained on the whole training set, and evaluated on the whole validation set, averaged over answers in the automatically constructed reference sets Σ. As with Table 5.3, for each of the models, scores are additionally averaged over the top 10 generations, with HREA-QIH-G* employing a beam-search to generate answers. Again, we evaluate only on the validation set since ground-truth answers are not publicly available for the test set—something we require in order to construct the reference sets Σ. Note that the reference baseline for this experiment Γ_Σ is different to that in Table 5.3 since the reference set is different: Σ instead of H.

As with Table 5.3, albeit at a much larger scale, the results here follow a similar pattern appearing to indicate that model preference ought to be contextualised by what that model is to be employed for. A particularly interesting observation is the performance of HCIAE-G-DIS on the overlap metrics for n=1—interesting due the fact that a sizeable proportion of answers

in the data do consist of one word answers (primarily "Yes"/"No" answers), which the model appears to be well-tuned to. Separation of evaluation across different n-gram lengths allows us to tease apart such characteristics, potentially providing further insight into models.

Table 5.6 Overlap and embedding distance metrics on the *entire* validation set, averaged over the top 10 generated answers for each model, against $C = \Sigma$. For HREA-QIH-G, on average ∼6 answers are the empty string, which are excluded from the computation. Metrics marked ↑ indicate higher values are better, and those marked ↓ indicate lower values are better.

Model	CIDER↑				BLEU↑				METEOR↑	BERT		FastText	
	n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4		L_2↓	CS↑	L_2↓	CS↑
Γ_Σ	0.3454	0.2734	0.2320	0.1957	1.0000	0.7586	0.6371	0.3816	1.0000	—			
CCA-QA-G	0.0825	0.0489	0.0342	0.0259	0.3280	0.1041	0.0370	0.0074	0.2255	7.0290	0.8704	2.8408	0.5150
HREA-QIH-G	0.0254	0.0136	0.0092	0.0069	0.3134	0.0432	0.0099	0.0017	0.2387	6.7067	0.8778	2.8971	0.5046
HREA-QIH-G*	0.0988	0.0624	0.0456	0.0351	0.4225	0.1580	0.0801	0.0264	0.3031	6.1834	0.8978	2.8438	0.5210
HCIAE-G-DIS	0.1306	0.0709	0.0487	0.0368	0.5084	0.0714	0.0283	0.0069	0.4221	6.0375	0.9001	2.9231	0.5360

5.6 Discussion

In this paper, we propose an alternate evaluation suite for *VisDial* drawing on existing metrics from the NLP community that measures consensus between answers generated by a model and a given reference set of answers. We arrive at the need for alternate evaluations through an analysis of existing evaluation metrics on the *VisDial* dataset, which we show can suffer from a number of issues to do with a mismatch between the task of VD and an evaluation for it that depends on ranking metrics. This disparity is highlighted through the use of an exceedingly simple baseline model based on CCA, which needs orders of magnitude fewer parameters and compute time, but still matches SOTA on mean rank, indicating its ability to leverage correlations in the data between questions and answers, and ignore the visual stimulus and sequential aspect of the dialogue.

While a recent update to the evaluation paradigm of *VisDial* incorporates both human judgements of answer validity and multiple plausible answers into a final score, issues relating to ranking persist, albeit to a lesser extent. Here, we advocate use of answers directly *generated* by a model, in concert with the consensus-based metrics we propose, evaluated against sets of answers which have been marked as valid by human annotators.

Given the scale of the dataset, however, eliciting human judgements becomes practically infeasible, restricting the extent to which such metrics could be applied. To address this issue,

we develop a semi-supervised automated mechanism, using existing sparse human annotations and correlations through CCA to obtain sets of valid answers for the entire dataset. We perform multiple experiments to evaluate the quality of such a mechanism, particularly focussing on its relevance to the task of VD, and the similarity of these automated reference sets with respect to those marked by humans. Based on such experiments, we expand the *VisDial* dataset with these reference set annotations, which we shall release as a baseline for future evaluation and model development.

We intend this alternate evaluation suite, along with the expanded data, and the semi-supervised process by which such expansion can happen, as one possible solution in the face of constraints on the dataset and its experimental design choices. We hope that the community adopts these evaluations going forward.

Acknowledgements DM, PKD, NS and PHST were supported by the ERC grant ERC-2012-AdG 321162-HELIOS, EPSRC grant Seebibyte EP/M013774/1, EPSRC/MURI grant EP/N019474/1, FAIR *ParlAI* grant. DM was additionally supported by the Skye Foundation.

5.7 Supplementary

5.7.1 Multi-view Canonical Correlation Analysis

Among several possible ways to formulate the canonical correlation analysis (CCA) objective for multiple variables, we choose the Frobenius norm-based objective of Hardoon et al. [124]. Let us assume that there are m views and $\mathbf{x}_i \in \mathbb{R}^{n_i}$ represents an observation from the i^{th} view. $X_i \in \mathbb{R}^{n_i \times N}$ represents the column-wise stack of N observations from the i^{th} view. The objective is to jointly learn projection matrices $W_i \in \mathbb{R}^{n_i \times k}$ for all the m views such that the embeddings in the $k(\leq n_i \forall i)$-dimensional space are maximally correlated. This is achieved by optimising the following problem:

$$\begin{aligned} &\min_{W_1,\ldots,W_m} \sum_{i,j=1,i\neq j}^{m} \left\| X_i^\top W_i - X_j^\top W_j \right\|_F^2 \\ &s.t.\ W_i^\top C_{ii} W_i = \mathbb{I}, \qquad \mathbf{w}_i^{l\top} C_{ij} \mathbf{w}_j^n = 0, \\ &i,j = 1,\ldots,m, i\neq j, l,n = 1,\ldots,k,\ l \neq n \end{aligned} \tag{5.7.1}$$

where $\mathbf{w}_i^l$ is the l^{th} column of W_i, and C_{ij} is the correlation matrix between the i^{th} and j^{th} views, or $C_{ij} = \frac{1}{N-1} X_i X_j^T$. Bach and Jordan [32] show that optimising Eq. 5.7.1 reduces to solving a generalised eigenvalue decomposition problem of the following form:

$$A\boldsymbol{v} = \lambda B \boldsymbol{v}$$

$$\begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1m} \\ C_{21} & C_{22} & \cdots & C_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ C_{m1} & C_{m2} & \cdots & C_{mm} \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_m \end{pmatrix} = \lambda \begin{pmatrix} C_{11} & 0 & \cdots & 0 \\ 0 & C_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & C_{mm} \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_m \end{pmatrix}$$

The top k (eigenvalue-sorted) eigenvectors $\boldsymbol{v}_i \in \mathbb{R}^{n_i}$ are column-wise stacked to construct projection matrix W_i for view $i \in \{1,\ldots,m\}$.

5.7.2 Detailed CCA results

We conduct a more detailed rank-based performance analysis of CCA versus SOTA approaches across both *v0.9* and *v1.0* of the *VisDial* dataset, extending Table 5.2 in the main paper with Table 5.7 here. Note, NDCG scores are not computed for *v0.9* since human annotations on answer relevance are not available. We compare against ablative versions (i.e. when the image and dialogue history are removed) of the SOTA models, as well as two nearest-neighbour baselines, as established in [77]:

NN-A-Q: given a test question, we find the k nearest-neighbour questions (by average *GloVe* embedding) from the training set. We take the mean of their k corresponding answers (again in GloVe embedding) to represent a "canonical" answer to that question, ranking the test question's candidate answers by their L_2 distances to it.

NN-A-QI: given a test question and image, we first draw the k nearest-neighbour questions to the test question from the training set. From this set, we draw the k' questions whose corresponding image features are most similar to the test image feature. Taking the mean of their k' corresponding answers, we then rank the test question's candidate answers as above ($k = 100$, $k' = 20$ as per [77]).

As in Table 5.2, we observe that the MR achieved by CCA models is similar to that of SOTA approaches, despite the approach's light-weight nature. Comparing to the nearest-neighbour baselines, CCA is superior in MR, and additionally in computation and storage requirements since a nearest-neighbour approach requires the train data (including images) at test time.

5.7.3 NDCG details

The NDCG is the ratio of the discounted cumulative gain (DCG) of a model's predicted ranking to the DCG of the "ideal" ranking, obtained by sorting the relevance scores in descending order:

$$\text{NDCG@}m = \frac{\text{DCG@}m}{\text{ideal DCG@}m}$$

where m is the number of answers with human-derived relevance scores in the set of 100, and DCG@m is defined as:

$$\text{DCG@}m = \sum_{i}^{m} \frac{\rho_i}{\log_2(i+1)}$$

where i is the rank of the answer candidate, and ρ_i is the human-assigned relevance score of the i^{th} ranked answer.

Table 5.7 Extended results for SOTA vs. CCA on the *VisDial v0.9* and *v1.0* dataset. CCA achieves comparable performance in mean rank (MR) while ignoring both image and dialogue sequence.

		Model	I/QA features	MR	R@1	R@5	R@10	MRR	NDCG
SOTA	*v0.9*	HCIAE-G-DIS [206]	VGG-19/learned	**14.23**	44.35	65.28	71.55	0.5467	-
		CoAtt-GAN [364]	VGG-16/learned	**14.43**	46.10	65.69	71.74	0.5578	-
		HREA-QIH-G [77]	"	**16.60**	42.13	62.44	68.42	0.5238	-
		LF-QIH-G [77]	"	**16.76**	40.86	62.05	68.28	0.5146	-
		HRE-QIH-G [77]	"	**16.97**	42.23	62.28	68.11	0.5237	-
		LF-QI-G [77]	"	17.06	42.06	61.65	67.60	0.5206	-
		LF-Q-G [77]	"	17.80	39.74	60.67	66.49	0.5048	-
	v1.0	HRE-QIH-G [77]	VGG-16/learned	18.78	34.78	56.18	63.72	0.4561	0.5245
		LF-QIH-G [77]	"	18.81	35.08	55.92	64.02	0.4568	0.5121
Baselines	*v0.9*	NN A-Q	GloVe	19.67	29.88	47.07	55.44	0.3898	-
			FastText	25.92	19.86	34.74	43.55	0.2830	-
		NN A-QI	VGG-16/GloVe	20.14	29.93	46.42	54.76	0.3873	-
			ResNet34/FastText	25.88	21.19	35.78	44.31	0.2941	-
		NCCA A-Q	FastText	57.18	4.13	9.67	13.89	0.0837	-
CCA	*v0.9*	A-Q	GloVe	**15.86**	16.93	44.83	58.44	0.3044	-
			FastText	**16.21**	16.85	44.96	58.10	0.3041	-
		A-QI (Q)	ResNet34/FastText	18.27	12.24	35.55	50.88	0.2439	-
			VGG-16/GloVe	26.03	12.24	30.96	42.63	0.2237	-
			VGG-19/GloVe	18.88	12.42	34.52	48.47	0.2409	-
	v1.0	A-Q	GloVe	**16.60**	16.10	39.38	54.68	0.2824	0.3504
			FastText	**17.07**	16.18	40.18	55.35	0.2845	0.3493
		A-QI (Q)	ResNet34/FastText	19.25	12.63	32.88	48.68	0.2379	0.3077
			VGG-16/GloVe	19.11	13.53	32.43	47.13	0.2415	0.3071
			VGG-19/GloVe	19.29	13.38	32.73	47.23	0.2415	0.3000

5.7.4 Consensus performance with H and C

We conduct a thorough analysis of the consensus measures between an answer generation and a set of relevant answers. Table 5.8, extending Table 5.3 in the main paper, measures the consensus of an answer generation with reference to H, the set of relevant answers taken from the human-annotated validation set $\mathcal{H}_v$. Table 5.9, extending Table 5.6, measures the consensus with reference to C, the relevant answer set automatically extracted using Σ. We measure consensus scores using the overlap and embedding distance-based metrics described

Table 5.8 Overlap and embedding distance metrics on the human-annotated validation set $\mathcal{H}_v$, for answer generations sampled from each model, against H. For HREA-QIH-G, on average ∼6 answers are the empty string, which are excluded from the computation. Metrics marked ↑ indicate higher values are better, and those marked ↓ indicate lower values are better. When $k=10$, 10 answer generations are sampled from the model—$\mu, \sigma,$ andγ are the mean, standard deviation and maximum of the k scores, respectively, averaged over the dataset. Otherwise, 1 answer generation is sampled and the mean μ is shown.

Model		CIDER↑				BLEU↑				METEOR↑	BERT		FastText	
		n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4		L_2↓	CS↑	L_2↓	CS↑
A_{gt}	μ	0.1889	0.1253	0.0986	0.0819	0.9961	0.4840	0.3934	0.2916	0.9971				
Γ_H	μ	0.1915	0.1437	0.1165	0.0962	1.0000	0.7633	0.5974	0.3371	1.0000				
	σ	(0.1490)	(0.1051)	(0.0869)	(0.0772)	(0.0000)	(0.1440)	(0.1538)	(0.1675)	(0.0000)				
	γ	0.2765	0.2151	0.1810	0.1513	1.0000	0.9898	0.9826	0.9300	1.0000				
CCA-QA-G	μ	0.0687	0.0414	0.0285	0.0215	0.4323	0.1461	0.0345	0.0138	0.2713	7.1231	0.8690	3.1251	0.4555
CCA-QA-G (k=10)	μ	0.0708	0.0420	0.0291	0.0221	0.4156	0.1539	0.0516	0.0149	0.2721	7.0301	0.8702	2.8850	0.4889
	σ	(0.1071)	(0.0603)	(0.0418)	(0.0317)	(0.2931)	(0.2183)	(0.1248)	(0.0500)	(0.2410)	(1.1316)	(0.0393)	(0.6072)	(0.1077)
	γ	0.1320	0.0840	0.0609	0.0470	0.7251	0.4445	0.2276	0.0960	0.5047	5.7117	0.9132	2.3446	0.5984
CCA-IQA-G	μ	0.0649	0.0400	0.0277	0.0211	0.4107	0.1407	0.0348	0.0153	0.2581	7.2405	0.8641	3.1807	0.4411
HREA-QIH-G	μ	0.0880	0.0483	0.0333	0.0252	0.5557	0.0948	0.0411	0.0136	0.4813	6.2875	0.8927	2.9724	0.5079
HREA-QIH-G (k=10)	μ	0.0558	0.0303	0.0206	0.0156	0.3919	0.0648	0.0169	0.0039	0.2958	6.6661	0.8796	2.9599	0.4907
	σ	(0.0716)	(0.0405)	(0.0290)	(0.0232)	(0.2691)	(0.1190)	(0.0663)	(0.0346)	(0.2452)	(1.1225)	(0.0415)	(0.5827)	(0.0906)
	γ	0.1194	0.0656	0.0450	0.0341	0.7154	0.1975	0.0607	0.0154	0.5915	5.6003	0.9149	2.5639	0.5600
HREA-QIH-G*	μ	0.1359	0.0721	0.0494	0.0372	0.7646	0.0614	0.0374	0.0064	0.7149	5.5727	0.9149	3.2664	0.4971
HREA-QIH-G* (k=10)	μ	0.0937	0.0600	0.0438	0.0337	0.5825	0.2739	0.1423	0.0461	0.4310	6.0202	0.9031	2.8534	0.5117
	σ	(0.0656)	(0.0409)	(0.0300)	(0.0231)	(0.2271)	(0.1959)	(0.1412)	(0.0808)	(0.1868)	(0.7250)	(0.0241)	(0.4241)	(0.0761)
	γ	0.2456	0.1701	0.1326	0.1042	0.9780	0.8198	0.6472	0.2946	0.9562	4.9441	0.9295	2.0886	0.6584
HCIAE-G-DIS	μ	0.1337	0.0719	0.0495	0.0374	0.7512	0.0965	0.0553	0.0117	0.6724	5.6709	0.9122	3.1451	0.5047
HCIAE-G-DIS (k=10)	μ	0.1053	0.0570	0.0390	0.0294	0.6352	0.1031	0.0443	0.0108	0.5449	5.4297	0.9120	2.3339	0.6067
	σ	(0.1004)	(0.0528)	(0.0361)	(0.0273)	(0.2980)	(0.1473)	(0.0950)	(0.0423)	(0.3156)	(1.0121)	(0.0254)	(0.4738)	(0.0850)
	γ	0.1921	0.1130	0.0806	0.0615	0.9278	0.4261	0.2462	0.0742	0.8654	4.7000	0.9334	1.8757	0.6992

Table 5.9 Overlap and embedding distance metrics on the *entire* validation set, for answer generations sampled from each model, against $C = \Sigma$. For HREA-QIH-G, on average ∼6 answers are the empty string, which are excluded from the computation. Metrics marked ↑ indicate higher values are better, and those marked ↓ indicate lower values are better. When $k=10$, 10 answer generations are sampled from the model—$\mu, \sigma,$ andγ are the mean, standard deviation and maximum of the k scores, respectively, averaged over the dataset. Otherwise, 1 answer generation is sampled and the mean μ is shown.

Model		CIDER↑				BLEU↑				METEOR↑	BERT		FastText	
		n=1	n=2	n=3	n=4	n=1	n=2	n=3	n=4		L_2↓	CS↑	L_2↓	CS↑
A_{gt}	μ	0.3502	0.2479	0.2004	0.1692	0.9948	0.4827	0.3935	0.2940	0.9955				
Γ_H	μ	0.3454	0.2734	0.2320	0.1957	1.0000	0.7586	0.6371	0.3816	1.0000				
	σ	(0.2241)	(0.1906)	(0.1682)	(0.1485)	(0.0000)	(0.1949)	(0.2320)	(0.2822)	(0.0000)				
	γ	0.4212	0.3429	0.2991	0.2583	1.0000	0.9915	0.9767	0.7904	1.0000				
CCA-QA-G	μ	0.0789	0.0461	0.0313	0.0235	0.3123	0.0752	0.0129	0.0024	0.1864	7.1873	0.8673	3.0908	0.4782
CCA-QA-G (k=10)	μ	0.0825	0.0489	0.0342	0.0259	0.3280	0.1041	0.0370	0.0074	0.2255	7.0290	0.8704	2.8408	0.5150
	σ	(0.1397)	(0.0850)	(0.0608)	(0.0462)	(0.3142)	(0.1908)	(0.1062)	(0.0342)	(0.2411)	(1.1715)	(0.0393)	(0.6733)	(0.1279)
	γ	0.1617	0.1025	0.0743	0.0571	0.5902	0.3192	0.1663	0.0477	0.4282	5.7403	0.9118	2.2635	0.6349
HREA-QIH-G	μ	0.1109	0.0597	0.0409	0.0308	0.4521	0.0656	0.0243	0.0058	0.3710	6.2743	0.8924	2.8815	0.5334
HREA-QIH-G (k=10)	μ	0.0254	0.0136	0.0092	0.0069	0.3134	0.0432	0.0099	0.0017	0.2387	6.7067	0.8778	2.8971	0.5046
	σ	(0.0324)	(0.0178)	(0.0122)	(0.0092)	(0.2770)	(0.0961)	(0.0493)	(0.0207)	(0.2444)	(1.2885)	(0.0448)	(0.6059)	(0.1029)
	γ	0.1569	0.0850	0.0579	0.0436	0.5901	0.1349	0.0365	0.0068	0.4742	5.6594	0.9119	2.4015	0.6059
HREA-QIH-G*	μ	0.1580	0.0835	0.0568	0.0428	0.5923	0.0418	0.0221	0.0047	0.5269	5.7023	0.9097	3.1888	0.5196
HREA-QIH-G* (k=10)	μ	0.0988	0.0624	0.0456	0.0351	0.4225	0.1580	0.0801	0.0264	0.3031	6.1834	0.8978	2.8438	0.5210
	σ	(0.0847)	(0.0579)	(0.0436)	(0.0338)	(0.2609)	(0.1621)	(0.1110)	(0.0634)	(0.1986)	(0.9805)	(0.0326)	(0.4866)	(0.0895)
	γ	0.3105	0.2122	0.1653	0.1302	0.8918	0.6207	0.4368	0.1742	0.7991	4.8802	0.9287	2.0014	0.6934
HCIAE-G-DIS	μ	0.1611	0.0874	0.0602	0.0455	0.5989	0.0708	0.0337	0.0080	0.5148	5.7358	0.9088	3.0385	0.5345
HCIAE-G-DIS (k=10)	μ	0.1306	0.0709	0.0487	0.0368	0.5084	0.0714	0.0283	0.0069	0.4221	6.0375	0.9001	2.9231	0.5360
	σ	(0.1319)	(0.0744)	(0.0526)	(0.0404)	(0.3388)	(0.1279)	(0.0775)	(0.0359)	(0.3376)	(1.2567)	(0.0375)	(0.5814)	(0.1104)
	γ	0.2503	0.1482	0.1061	0.0812	0.8066	0.3072	0.1603	0.0461	0.6956	5.1979	0.9230	2.2556	0.6507

in the main paper. We sample 1 and 10 answer generations from the respective models, and compute the mean μ, standard deviation σ, and maximum γ of the k scores. We also include two baselines intended as upper bounds: i) A_{gt}, takes the ground-truth answer to be the generated answer, and ii) Γ_H, cycles through H, treating each answer as the generated answer. Since each answer in the set could be a plausible one (as marked by humans), we take the best-case score (minimum score for all except embedding-based L_2 distance for which we take the maximum), and then average over the dataset.

5.7.5 Automated evaluation of reference sets

We analyse the performance of different methods for extracting answer clusters C, by comparing to the answer reference set H obtained from the human-annotated validation set H_v. Extending the analysis of Table 5.4 in the main paper, in Table 5.10, between C and H, we measure IOU, precision, recall, and set size for the different methods, as well as, within cluster C, its average correlation, the standard deviation of the correlations, and the likelihood of C containing A_{gt}.

Table 5.10 Extended evaluation of alternate methods for automated reference set construction on the human-annotated validation set $\mathcal{H}_v$, against H. Values in parentheses denote standard deviation across the set. For G, n is the number of clusters specified for the agglomerative clustering.

C			$\frac{\lvert H \cap C \rvert}{\lvert H \cup C \rvert}$	$\frac{\lvert H \cap C \rvert}{\lvert C \rvert}$	$\frac{\lvert H \cap C \rvert}{\lvert H \rvert}$	$\lvert C \rvert$	$A_{\text{gt}} \in C$	corr (C)	std (corr(C))
H		———	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)	12.77 (7.24)	100.00 (0.00)	0.2393 (0.1942)	0.1670 (0.0917)
M	CCA A-Q	$(Q, \boldsymbol{A})_{\text{gt}}$	17.59 (13.04)	31.01 (27.31)	57.67 (32.17)	39.85 (31.51)	100.00 (0.00)	0.2998 (0.2610)	0.0856 (0.0374)
		$(Q, \boldsymbol{A})_{\text{max}}$	15.06 (14.31)	37.02 (34.72)	38.43 (35.62)	21.53 (26.14)	54.89 (49.77)	0.4794 (0.2379)	0.0852 (0.0424)
		$(A_{\text{gt}}, \boldsymbol{A})$	13.14 (12.74)	89.10 (26.06)	18.57 (23.21)	4.37 (10.36)	100.00 (0.00)	0.9466 (0.1016)	0.1586 (0.0844)
		$(A_{\text{gt}}, \tilde{\boldsymbol{A}})$	19.05 (12.65)	42.68 (30.16)	50.47 (35.70)	27.96 (32.01)	100.00 (0.00)	0.5069 (0.2734)	0.2068 (0.0939)
	CCA A-Q*	$(Q, \boldsymbol{A})_{\text{gt}}$	18.66 (14.83)	25.33 (23.22)	62.46 (27.37)	45.89 (28.70)	100.00 (0.00)	0.1843 (0.2146)	0.0735 (0.0360)
		$(Q, \boldsymbol{A})_{\text{max}}$	19.89 (16.99)	34.58 (30.69)	49.39 (34.08)	26.74 (25.82)	53.49 (49.89)	0.3270 (0.2186)	0.0710 (0.0367)
		$(A_{\text{gt}}, \boldsymbol{A})$	13.15 (12.84)	93.96 (20.12)	16.64 (20.51)	3.12 (8.11)	100.00(0.00)	0.9660 (0.0959)	0.1207 (0.0686)
		$(A_{\text{gt}}, \tilde{\boldsymbol{A}})$	25.01 (18.40)	59.19 (31.55)	39.35 (28.86)	12.00 (16.07)	100.00 (0.00)	0.5841 (0.2167)	0.2231 (0.0990)
Σ	CCA A-Q	$(Q, \boldsymbol{A})_{\text{gt}}$	19.92 (12.94)	32.64 (26.49)	64.25 (31.71)	39.45 (29.86)	100.00 (0.00)	0.3334 (0.2310)	0.1162 (0.0637)
		$(Q, \boldsymbol{A})_{\text{max}}$	14.42 (13.89)	40.72 (35.20)	27.33 (28.47)	11.32 (13.38)	47.14 (49.93)	0.5193 (0.2034)	0.0661 (0.0295)
		$(A_{\text{gt}}, \boldsymbol{A})$	12.67 (11.97)	89.83 (25.66)	18.17 (22.76)	4.01 (9.08)	100.00 (0.00)	0.9658 (0.0753)	0.0960 (0.0355)
		$(A_{\text{gt}}, \tilde{\boldsymbol{A}})$	21.16 (13.60)	49.21 (28.34)	38.60 (28.03)	12.30 (12.17)	100.00 (0.00)	0.5859 (0.1898)	0.2092 (0.0949)
	CCA A-Q*	$(Q, \boldsymbol{A})_{\text{gt}}$	21.94 (15.26)	27.31 (22.21)	77.53 (24.32)	50.37 (29.80)	100.00 (0.00)	0.2194 (0.1879)	0.1002 (0.0647)
		$(Q, \boldsymbol{A})_{\text{max}}$	19.64 (17.18)	40.95 (33.44)	32.89 (26.85)	11.93 (9.73)	37.02 (48.30)	0.3739 (0.2000)	0.0505 (0.0264)
		$(A_{\text{gt}}, \boldsymbol{A})$	13.15 (12.89)	93.32 (19.91)	15.49 (17.80)	2.10 (3.72)	100.00 (0.00)	0.9764 (0.0596)	0.0897 (0.0369)
		$(A_{\text{gt}}, \tilde{\boldsymbol{A}})$	24.13 (16.73)	62.48 (31.24)	32.91 (23.52)	7.17 (6.94)	100.00 (0.00)	0.6206 (0.1773)	0.2269 (0.0975)
G	CCA A-Q	$(Q, \boldsymbol{A})_{\text{gt}}$ n=3	19.56 (13.04)	27.14 (21.46)	58.03 (26.41)	33.78 (18.68)	100.00 (0.00)	0.2888 (0.2430)	0.0851 (0.0398)
		$(Q, \boldsymbol{A})_{\text{gt}}$ n=4	19.15 (12.81)	31.61 (24.59)	46.77 (25.22)	23.96 (14.68)	100.00 (0.00)	0.3193 (0.2488)	0.0679 (0.0355)
		$(Q, \boldsymbol{A})_{\text{gt}}$ n=5	10.49 (6.31)	17.08 (13.61)	28.42 (17.82)	21.25 (8.64)	100.00 (0.00)	0.5656 (0.1375)	0.0287 (0.0100)
		$(Q, \boldsymbol{A})_{\text{max}}$ n=4	17.65 (14.16)	36.87 (29.85)	35.27 (27.86)	14.75 (11.63)	57.17 (49.50)	0.4716 (0.2189)	0.0758 (0.0380)
		$(Q, \boldsymbol{A})_{\text{max}}$ n=5	9.47 (8.99)	20.73 (24.04)	19.14 (17.22)	14.94 (8.64)	20.78 (40.59)	0.7673 (0.0605)	0.0336 (0.0121)
		$(A_{\text{gt}}, \boldsymbol{A})$ n=5	15.53 (13.11)	81.47 (29.65)	21.82 (22.55)	4.70 (7.58)	100.00 (0.00)	0.8879 (0.1561)	0.1513 (0.0692)
		$(A_{\text{gt}}, \tilde{\boldsymbol{A}})$ n=5	22.13 (13.18)	47.40 (27.16)	38.93 (25.76)	11.58 (9.13)	100.00 (0.00)	0.5633 (0.1943)	0.1987 (0.0834)
	CCA A-Q*	$(Q, \boldsymbol{A})_{\text{gt}}$ n=5	10.49 (6.31)	17.08 (13.61)	28.42 (17.82)	21.25 (8.64)	100.00 (0.00)	0.5656 (0.1375)	0.0287 (0.0100)
		$(Q, \boldsymbol{A})_{\text{max}}$ n=5	9.47 (8.99)	20.73 (24.04)	19.14 (17.22)	14.94 (8.64)	20.78 (40.59)	0.7673 (0.0605)	0.0336 (0.0121)
		$(A_{\text{gt}}, \boldsymbol{A})$ n=5	15.34 (14.45)	89.81 (23.18)	18.42 (19.76)	2.74 (4.20)	100.00 (0.00)	0.9342 (0.1189)	0.1369 (0.0672)
		$(A_{\text{gt}}, \tilde{\boldsymbol{A}})$ n=4	27.74 (19.02)	54.15 (30.33)	42.51 (26.81)	11.08 (8.49)	100.00 (0.00)	0.5290 (0.1968)	0.2123 (0.0721)
		$(A_{\text{gt}}, \tilde{\boldsymbol{A}})$ n=5	25.59 (17.69)	59.20 (30.71)	35.74 (24.47)	7.91 (6.18)	100.00 (0.00)	0.5874 (0.1892)	0.2188 (0.0877)

In the main paper, the methods we explore to construct C were simple (Σ), meanshift (M), and agglomerative clustering (G). We employ CCA A-Q*, learned on solely those (Q, A) pairs in $\mathcal{H}_t$ for which humans scored $\rho(A) > 0$, to compute the correlations $\mathcal{C}$ between A_{gt} and each $A \in \tilde{\boldsymbol{A}}$, where $\tilde{\boldsymbol{A}} = \boldsymbol{A} \setminus A_{\text{gt}}$. Cluster C is then constructed by applying Σ, M, or G to correlations $\mathcal{C}$, and unioning A_{gt} with the resulting set. We denote this $(A_{\text{gt}}, \tilde{\boldsymbol{A}})$.

In Table 5.10 we additionally show the results of the above, but using i) CCA A-Q, learned on *all* train (Q, A) pairs, and ii) correlations $\mathcal{C}$ computed between A_{gt} or the question Q, and the *full* candidate set—$(A_{\text{gt}}, \boldsymbol{A})$ and $(Q, \boldsymbol{A})$, respectively. In the case of $(Q, \boldsymbol{A})$, we can construct C in two ways: either by selecting all those answers in $\boldsymbol{A}$ with the same cluster label as A_{gt}, or those with the same label as the answer with the maximum correlation to Q. We denote this $(Q, \boldsymbol{A})_{\text{gt}}$ and $(Q, \boldsymbol{A})_{\text{max}}$, respectively. This does not apply to $(A_{\text{gt}}, \boldsymbol{A})$ since A_{gt} and the answer with the maximum correlation will always be the same, nor for $(A_{\text{gt}}, \tilde{\boldsymbol{A}})$, since A_{gt}, by definition, is excluded from $\tilde{\boldsymbol{A}}$, and simply unioned with the resulting cluster afterwards. As described in the main paper, we cross-validate and select the method (in our case, Σ) giving us good precision, $|H \cap C|/|C|$, and a small cluster size, $|C|$. Using this method, we show in Figure 5.5, some qualitative examples of the answer clusters which our method extracts from the candidate sets associated with given questions and images. The majority of the answers are relevant both to the image and the question.

Q: Are there any other people?
- Way in the background
- There are few people way off in background
- I see a few in the background
- There are a few in the background

Q: Is the driver of the truck nearby?
- I can't see anyone in the picture
- No people
- Can't see anyone else

Q: Is the broccoli raw or cooked?
- It's raw.
- Raw

Q: Is the mountain large or small?
- It's large
- Fairly large
- It's medium size
- Large
- Pretty large
- Medium size I would say not small not large
- I would say large

Figure 5.5 Qualitative examples of the relevant answers our semi-supervised approach (Σ) extracts from given candidate answer sets. Note, we show *all* answers which our method extracts from the sets.

5.7.6 Utility evaluation of reference sets for VD

To demonstrate the utility of the automatically extracted sets of relevant answers, we show the improved performance of the CCA A-Q model when provided with an augmented train set including these automatic relevant answer sets extracted via weak supervision. We measure utility by performance on the downstream task of VD, where we look primarily to the NDCG score. For completeness, we also show performance in the other rank-based metrics. We extend Table 5.5 in the main paper by showing results for all automated methods (M, Σ, G) on the validation and test set in Table 5.11 and Table 5.12, respectively.

Table 5.11 Extended evaluation of the utility of automated reference set construction methods (M, Σ, G) on the standard VD evaluation. Models were trained using CCA on the indicated subsets ($\mathcal{H}_t$, all train, or all trainval) of *VisDial v1.0*, using answers drawn from different sets ('Ref'), and tested on the evaluation test server[5].

Train Set	Train #QA pairs	Ref	MR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MRR ↑	NDCG ↑
$\mathcal{H}_t$	15,317	H	26.49	6.05	21.50	35.53	0.1550	0.3647
	17,055	Σ	**20.36**	8.35	32.88	**48.78**	0.2066	**0.3715**
	26,318	M	21.53	6.83	29.80	45.63	0.1862	0.3503
	16,923	G (n=5)	20.66	8.08	30.35	46.33	0.1981	0.3657
	1996	$\{A_{\text{gt}}\}$	23.71	**13.13**	**34.05**	46.90	**0.2428**	0.2734
all train	10,419,489	Σ	17.20	10.73	34.20	51.80	0.2312	0.4023
	17,600,151	M	20.67	9.38	24.45	39.93	0.1905	0.3339
	10,614,163	G (n=5)	17.79	9.78	31.40	48.93	0.2171	0.3918
	1,232,870	$\{A_{\text{gt}}\}$	**17.07**	**16.18**	**40.18**	**55.35**	**0.2845**	0.3493
all trainval	10,599,533	Σ	17.31	10.20	33.30	51.45	0.2242	**0.4050**
	17,931,897	M	20.47	9.38	24.83	40.55	0.1917	0.3380
	10,798,877	G (n=5)	17.60	9.85	31.67	49.20	0.2184	0.3927
	1,253,510	$\{A_{\text{gt}}\}$	17.10	16.10	40.05	55.07	0.2833	0.3486

Table 5.12 Extended evaluation of the utility of automated reference set construction methods (M, Σ, G) on the standard VD evaluation. Models were trained using CCA on the indicated subsets ($\mathcal{H}_t$, all train, or all trainval) of *VisDial v1.0*, using answers drawn from different sets ('Ref'), and tested on the validation set.

Train Set	Train #QA pairs	Ref	MR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MRR ↑	NDCG ↑
$\mathcal{H}_t$	15,317	H	25.34	6.20	22.58	37.84	0.1598	0.3755
	17,055	Σ	**20.94**	8.54	32.16	**47.69**	0.2049	**0.3884**
	26,318	M	21.84	7.16	28.78	44.95	0.1858	0.3669
	16,923	G (n=5)	21.47	7.93	30.29	45.89	0.1942	0.3779
	1996	$\{A_{\text{gt}}\}$	23.80	**13.50**	**34.06**	46.64	**0.2442**	0.2816
all train	10,419,489	Σ	17.39	10.27	34.01	51.54	0.2264	**0.4099**
	17,600,151	M	20.96	9.11	22.92	39.30	0.1850	0.3354
	10,614,163	G (n=5)	18.03	9.68	31.16	48.85	0.2136	0.4005
	1,232,870	$\{A_{\text{gt}}\}$	**17.04**	**16.00**	**41.21**	**55.16**	**0.2860**	0.3547

[5] https://visualdialog.org/challenge/2018

"*Once I knew only darkness and stillness, but a little word fell into my hand that clutched at emptiness, and my heart leaped to the rapture of living.*"

Helen Keller

Discussion

The principle contributions of this thesis are in developing methods for i) learning visual representations with low-cost, easy-to-acquire labels, ii) relaying information to VI users, and iii) robustly evaluating models which deliver information via natural language. Here we discuss the implications of each in developing real-world assistive technologies for VI users.

6.1 Learning visual scene representations

Real-world assistive devices may need to learn scene and object representations with only minimal or weakly-labelled examples. Motivated by this, in Chapter 2 we developed a weakly-supervised method for the semantic segmentation of images—one type of scene representation—which uses only classification labels. The algorithmic contributions of our work lie in coupling an EM-based method with a simple-to-complex curriculum learning paradigm. With this approach, our method achieved SOTA performance on the *PASCAL VOC 2012* semantic segmentation task compared to existing methods which use the same degree of weak supervision.

Assistive devices will likely require a human-in-the-loop since they will need to dynamically tune themselves to users' inputs, interactions, and environments. Our method facilitates this as classification labels can easily be collected verbally from a VI user, converted to text with an off-the-shelf speech-to-text engine, and then used to download further class-specific examples from the internet, providing a virtually limitless source of training data for on-the-fly learning. This set-up, therefore, makes it possible not only to learn completely novel object categories for which no training samples are available, but also to improve the learned representations of existing categories by integrating further training samples. To achieve this, models will need to draw on continual [256, 324], and active learning [58, 233, 323] paradigms, both active

areas of research. Continual learning endows models with the ability to acquire, fine-tune, and transfer knowledge from a continuous stream of information over time. Models will need to be robust to non-stationary data distributions as the number of concepts and scenarios encountered by the user expand. This will involve leveraging existing knowledge to learn new concepts, while simultaneously remembering the old [172, 222]. Active learning paradigms, on the other hand, allow models to interactively request a human to label (a subset of) samples. The majority of incoming data from a mobile or wearable camera will be unlabelled. It will be useful, therefore, for the model to identify the important or relevant samples for which labels are needed so that new and existing representations can be learned and improved, respectively.

In the future it will also be important to consider which *types* of representations will be most useful, given the user's current task and the computational constraints of the device. Assigning class labels down to the pixel level, for example, may be unnecessarily fine-grained for a navigation task that simply requires objects' locations, regardless of their class. Furthermore, obtaining real-time segmentations, or training a segmentation model on-the-fly, may be infeasible on resource-constrained mobile devices[1]. On the other hand, a representation which delivers class-agnostic bounding boxes or coarse blobs may not be fine-grained enough if the task is to locate a specific object. In fact, any task relying on information which goes beyond an object's class label, of which there are many in day-to-day life, would face similar limitations. Learning representations which factor in additional detail will thus also be necessary in some scenarios. Choosing the representation and when to deliver it so that it is most informative to the user are open research challenges, and will likely need to be human-driven, for example, by obtaining the user's goal explicitly, or inferring it from the environment.

Finally, while visual cues dominate much of our perception of the world, it is possible that integrating other information, like sound cues or locations, may help to deliver more robust scene representations. Furthermore, while object-centric representations from object detection, semantic segmentation, or instance segmentation pipelines may intrinsically encode the spatial and functional relationships between objects, they do not explicitly encode this knowledge. A structured representation, in the vein of scene graphs [368], may be richer, and additionally help downstream interpretability and reasoning, which are important in medical devices.

[1] Although mobile-friendly segmentation models are available, with real-time inference possible in some scenarios [297, 379].

6.2 Relaying scene information

The second contribution of this thesis lies in two methods for relaying information about environments to VI users. In Chapter 3 we developed mappings for transforming 3D VR environments into soundscapes which encode the 3D positions of objects and are updated in real-time as the user navigates. With a novel locomotion-controlled testing method, we experimentally validated these mappings by getting human participants to perform simple navigation and obstacle avoidance tasks using only the audio soundscapes. Our findings showed that participants' navigational efficiency and strategy improved after only a short training period, which is promising for using these mappings in navigational tasks.

In Chapter 4 we proposed delivering information by answering questions asked by the VI user about their environment. Here, users are able to more directly probe their environment and gain specific, finer-grained information which may not be captured by the previous method. We explored this through the task of VD, answering a sequence of questions about an image, where our contribution was a fully generative model which learns conditional distributions over responses under two settings: 1) a one-way VD task, where the model generates a *set* of feasible answers to a question, based on the intuition that a sentence can be phrased in many ways, and 2) a two-way VD task, where the model generates (sets of) questions *and* answers, based on the intuition that an assistant should be able to both answer *and* ask questions. Our generative approach performs on-par with SOTA models on these VD tasks, while avoiding the need for a computationally intensive beam search as a post-processing step.

While these works take exploratory steps forward, the best way to relay information to VI users is still an open research question. This is because vision is very efficient at filtering relevant information and relaying it at a high bandwidth. Developing efficient alternate methods requires choices in 1) mapping and 2) delivery, which are both non-trivial tasks.

Mapping involves identifying and extracting relevant elements of a scene, and transforming them into non-visual representations. This is challenging because relevancy varies as a function of the task at hand, and because the encoding to non-visual representations may involves losses due to time and other delivery constraints. For example, both echolocation and humming soundscapes encode the 3D positions of objects, but none of their other features. While

sufficient for navigation, they may not be suited to finer-grained tasks. There is also the question of how best to aggregate non-visual representations when multiple scene elements are relevant, constrained by the cognitive load the aggregated representation would place on the user. These challenges were evident in the soundscape mapping which attached humming sounds to objects: users reported hums were difficult to disentangle when they were nearby multiple objects. While this information overload could be alleviated by the user querying their environment with specific questions, hence filtering for relevance manually, questions and answers may not be suited to all scenarios. When navigating, for example, it is infeasible for the user to repeatedly ask *'Is there a hazard in front of me?'*. Instead, a simple spatial audio alert when nearing a hazard would be more appropriate.

Once a non-visual mapping is obtained, delivery involves presenting the non-visual representation to the VI user. The primary challenge here is ensuring that the delivered form does not interfere with the user's otherwise adapted abilities to perceive the world. Both of the methods we explored use audio cues, in the forms of spatial sound and verbal answers, which would likely be presented via head/earphones. The delivery of these cues may, however, impede the user's hearing, for which there is strong evidence to show is especially informative for VI people across a range of tasks [113, 187, 202, 288, 318, 343, 361]. To circumvent these issues, alternate equipment for delivery could be investigated, for example, bone-conduction headphones[2]. Other delivery forms could also be used, for example, electro- or vibro-tactile cues, though these may not allow as much information to be encoded, or to the same specificity, as audio cues. Making the correct choice for both mapping and delivery is important to ensure visual information is efficiently delivered to VI users in real-world scenarios.

In the future, it will be essential to validate these hypotheses through controlled user studies. Our spatial audio encoding methods were tested on blind-folded sighted participants, rather than VI participants, as a preliminary proof-of-concept test of the vision-to-audio mappings and the testing paradigm. Our natural language delivery method was validated not by its usefulness to VI participants, but by the model's answering performance using established metrics of the dataset on which it was trained. To truly assess the effectiveness of information relay methods, they will need to be deployed and tested in real-world scenarios by VI users.

[2]Bone-phones transmit sound through vibrating pads placed on the jaw bones, resulting in sound sources perceived externally rather than in-ear.

6.3 Evaluating effectiveness of vision-language models

Relaying information about an environment to a VI user via language requires on a model that is able to understand, reason and make inferences about the scene, before parsing the user's question and then delivering the correct answer. In Chapter 5 we investigated the extent to which the *VisDial* dataset [77], used in Chapter 4, and its evaluation metrics, facilitate progress toward this goal. We established a simple CCA baseline without any notion of the image or dialogue history, and found that it performs as well as significantly more complex SOTA models on the dataset's established metrics. This finding, we demonstrated, is a combination of i) dominant linguistic statistics in the dataset which a model based on correlation can exploit, and ii) a rank-based evaluation paradigm which does not account for the answer generated by the model, nor that the correct answer could be one of many. Together, these factors mask true progress in the VD task. To partly address this, we proposed a revised evaluation which instead measures the consensus between an answer *generated* by a model and a *set* of equally-plausible answers to a given question, rather than just a single one. We developed a weakly-supervised and easily scalable method to construct sets of plausible answers for the entire dataset. Our principle recommendation is that the revised evaluation paradigm be used for model development and evaluation on the *VisDial* dataset in the future.

Robustly evaluating natural language at scale, however, remains an open research question. Turning to human evaluators is infeasible, and existing automatic evaluation metrics, including those used in our revised paradigm, have their shortfalls. Overlap-based metrics [186, 191, 255], for example, rely on pattern matching word n-grams between sentences, and hence may penalise short answers which are prevalent in VQA and VD datasets. Distance-based metrics in sentence embedding space [84, 302], on the other hand, may not distinguish between sentences which are syntactically and semantically very similar (*"it is on the left"* versus *"it is on the right"*). Additionally, key informational content may be lost in the mapping from word to sentence embedding space, although strong baselines have been proposed [29, 87]. In the absence of a single robust metric suited to all scenarios, the crux of our revised evaluation paradigm is a *suite* of these metrics, making model performance on the *VisDial* dataset more robust to the shortcomings of any one single metric. We advocate that aggregate approaches such as these are better suited to the evaluation of natural language.

The linguistic biases in the *VisDial* dataset, however, remain present, with the revised evaluation paradigm only helping up to a limit. These strong language priors can lead to models that do not truly understand and reason about visual and linguistic content. Accepting the design choices made in the dataset's collection, rebalancing the dataset could be one method to overcome these issues, as has been done in existing vision-language datasets [4, 5, 115, 160, 378]. Rebalancing could adopt one of many forms. Following [115, 378], each question in each dialogue could be coupled with a pair of images for which the correct answer would be different. Another approach [5] could be to differ the answer distributions between train and test settings, thus exposing models which simply exploit linguistic correlations. Complementary to these efforts, it will also be important to explore more effective methods of learning joint embeddings between vision and language modalities, like those based on bilinear pooling [100, 375], and shallow-layer fusion in network architectures [81].

Finally, relinquishing focus on the *VisDial* dataset, motivated by the *need* for language in many tasks, we posit that the curation of a task-specific VQA or VD dataset, where the task itself relies on the visual and linguistic input being understood, would naturally mitigate dominant linguistic statistics, and aid the design of better evaluation metrics. This departs from the many non-task specific VQA and VD datasets [24, 77, 385] which simply pair annotators in free-form question-answering. Not only does this often give rise to arbitrary questions, but it also hamstrings the evaluation to settings where the only way to evaluate predicted answers is to match them to human-provided ones, for which known challenges exist. Through carefully designing the collection method, and carefully selecting the task (for example, navigation or object localisation) the resulting dataset may improve the interrogation of joint vision and language reasoning, and the development of VI assistive technologies. Specifying a task would also allow the generated language to instead be evaluated by its *utility* in accomplishing the specified task, with model performance quantified by the speed and efficiency of task completion in the spirit of [79, 336]. In this direction, a host of works have already investigated the use of auxiliary tasks to evaluate, or supplement the evaluation of, language [21, 78, 79, 80], supporting downstream task performance as a proxy for model performance going forwards.

6.4 Deploying in the real-world

A crucial component in developing data-driven assistive technologies for VI people is to ensure that they work on real-world portable devices. In academic settings, including this thesis, these factors are often not considered. To remedy this we now turn to specific areas that should be focussed on to realise real-world assistive devices.

6.4.1 Curating datasets specific for VI assistance

A host of datasets in the research community have been used to benchmark performance and establish machine capabilities in a range of perceptual visual and language tasks [2, 24, 77, 79, 83, 88, 91, 179, 193, 311, 326]. Most of these datasets, however, were collected and annotated by sighted individuals without an assistance task in mind and, therefore, fundamentally differ from the scenarios seen by assistive devices in the real-world. Computer vision datasets [83, 88, 179, 193, 326], for example, are mainly composed of images taken from the third-person perspective, with the target objects/s often centred and focussed, and only small variations in image quality. This is significantly different from images that would be streamed from a hand-held or head-mounted camera, which may be blurry or exclude the target object altogether. Video datasets [2, 91, 166, 291, 311] partially address these issues, with some specifically collected from a first-person perspective [75, 188, 286], however, many remain in the third-person, and do not cover scenarios regularly encountered by VI people. From a linguistic perspective, language and vision-language datasets also present several differences. Given the wide range of possible tasks, and the differences between sighted and VI people's interactions, existing language corpuses may not be specific enough to deliver useful assistance. For example, visual question-answering datasets [24, 77, 80, 385], often contain questions which are irrelevant to a VI person because of the non-task specific way in which they were created (e.g. *'Is the image black and white?'* or *'What colour is the plate?'*).

Developing assistive devices which will be robust and effective in the real-world will require that these differences in data be addressed. This can be jointly tackled through i) curating datasets tailored to VI assistance, and ii) using continual and active learning paradigms to integrate knowledge learned from offline and online datasets, each of which we discuss below.

Curating datasets which are tuned to real-world scenes and challenges faced by VI people, requires that VI people are brought into the data collection loop [7, 39, 121]. *VizWiz* [121] is the first public dataset to do this. It pairs questions asked by VI people about images taken on their mobile phones, with answers from (offline) sighted annotators. As a result, the dataset contains highly relevant questions (e.g. *'Is the light on?'*, and *'Is the washing machine set to the rinse cycle?*) with corresponding images which are often blurry and off-target. Subsequently, mobile calling apps, *Be My Eyes* [39] and *Aira* [7], have established similar pipelines but instead collect answers from *online* sighted helpers based on videos captured by VI video callers[3].

Relying on a human-in-the-loop for large-scale dataset collection, however, is time- and cost-intensive, and still may not account for the great diversity in scenarios faced by VI people. To address this, the highly photo-realistic and semi-interactive VR environments which have become increasingly available [57, 177, 365] may serve as good real-world approximations. In these environments, a simulated VI mechanical turker (e.g. by obscuring their FOV) and a sighted turker could be paired to scalably collect (visuo-linguistic or other) data under similar settings as with physical humans. Furthermore, such a data collection method could flexibly be used to simulate a range of conditions, like the FOVs of different eye conditions, or the view-points from different devices (e.g. wearable or mobile phone). Additionally, a great diversity of environments could be created to cover the tasks faced by VI people, like cooking and locating objects in a home, navigating public spaces and transport, or shopping for groceries. As discussed in Section 6.3, a task-specific dataset may further facilitate the design of better evaluation metrics where the performance on the *task* in the VR environment serves as a proxy for a trained model's performance. Of course, it will be important to validate the VR scenarios and tasks against small subsets of real-world samples collected under similar conditions with real VI users.

The world is constantly changing however, which makes it difficult for even a VR simulation to account for all possible scenarios. It will therefore be essential to equip assistive devices with the ability to learn novel concepts on-the-fly. As discussed in Section 6.1, this will involve using knowledge learned from both VI-specific and general datasets in novel situations, and will

[3] At the time of writing this thesis, however, neither company has, or intends to, publicly share their data.

draw on advances in continual and active learning [58, 203, 233, 256]. Specific challenges in this direction lie in handling temporal shifts in the distribution of data so that new knowledge is not favoured in place of existing knowledge, referred to as forgetting, and existing knowledge facilitates the acquisition of new knowledge, referred to as transfer learning [172, 222].

6.4.2 Developing mobile-compatible models

Most of today's research in machine learning is conducted under resource-unconstrained settings, with few limitations placed on the available compute or storage/memory requirements of the model at train and test time. SOTA models often constitute millions of parameters (∼300-500MB), and rely on a GPU (sometimes multiple) to be trained and to deliver predictions within reasonable time frames [132, 142, 305, 320]. These considerations, however, have crucial implications for mobile assistive devices which will not only be defined by their performance, but by their size, weight, battery life, form factor, and user interface. Another consideration is the speed at which the model can parse and process scenes, ideally at real-time or close to it. The time required to (re)train or fine-tune a model, either locally or remotely, may also be important, particularly in active/continual learning settings.

Meeting these requirements will involve both hardware and software considerations. In terms of hardware, the miniaturisation of electronics is already enabling better cameras and more powerful processors on portable devices. Many high-end smart-phones, for example, ship with on-board graphics cards and neural network packages. The resource-constrained nature of mobile devices and wearables, however, will require also optimisation of the models themselves. This will draw on active research in neural network-based model compression which aims to reduce the space and time complexity of models, without significant loss in their performance [62]. It is also possible that rather than locally processing incoming data streams, devices could instead delegate the processing to remote servers. The increasing affordability and availability of cloud computing services and high internet speeds may indeed make this a feasible solution in the near future. This, however, may not be a silver bullet. Many third-world and developing countries, for example, have slow or only limited access to internet.

6.4.3 Designing devices

The trade-off between form and functionality also manifests in the *design* of assistive devices, in particular their form factors and user interfaces, both of which have been shown to heavily influence the uptake of an assistive tool [105]. A device, whether wearable or hand-held, should be inconspicuous, lightweight, well-fitting when worn or carried, and overall aesthetically pleasing. It should additionally not hamper safety, for example, by requiring the user to enter commands which are difficult or dangerous in real-world scenarios, or by blocking natural sounds. Importantly, the device should be intuitive to operate, drawing on functions and features which VI users may already be accustomed to using, like screen readers, accessible keyboards, tactile buttons, and voice commands.

Regional prevalence, heterogeneity in eye condition, and age must also be considered when designing assistive devices. Solutions will need to be tailored to the particular condition of the user: an assistive device which enhances the edges or contrast of objects/people in the central FOV might be unhelpful for an AMD sufferer, who typically has limited central vision, and indeed would not be helpful at all to a fully blind person. Additionally, the majority of those with vision impairment are over 50 years of age [362]. Given the low uptake of and familiarity with technological devices by elderly people [248], an unintuitive design may further reduce its usefulness to large portions the global VI population.

6.4.4 Final word

These many considerations, from understanding diverse scenarios, to effectively mapping and delivering information about them in a form that is easily understood for the task at hand, will be essential in developing data-driven assistive technologies for VI users which are robust, efficient, portable, and importantly safe. These technologies have the potential to dramatically improve the quality of life of hundreds of millions of people around the world.

Assistive Technology Review

Many assistive tools and devices have been developed to ease the challenges faced by VI people, ranging from simple clear-path indicators and obstacle detectors, like the white cane or guide dog, to wearable cameras which can identify objects and read text on a page aloud. Here we discuss a sampling of relevant assistive devices for VI people, addressing their capabilities and shortcomings. Following the categorisation presented in Chapter 1, we group the aids by those that provide long-, medium-, and short-range assistance.

A.1 Long-range aids

Many assistive technologies aim to provide navigational assistance to ease the inherent challenge of independent navigation for VI people. Successful navigation constitutes two parts: way-finding, and mobility. Way-finding involves moving to a long-range intended destination, requiring self-orientation, typically via landmarks and the ultimate destination, route decision, route monitoring, and destination recognition [190, 207]. In this section, we discuss long-range tools to assist with way-finding.

The ubiquity of global positioning system (GPS) and global information system (GIS), particularly on mobile devices, has changed day-to-day way-finding for both sighted and non-sighted populations. To VI people, however, the visual displays and turn-by-turn instructions of GPS-based localisation services like *Google Maps* may not always be accessible. GPS-based assistive tools have therefore been specifically developed for the VI community. The first was the Personal Guidance System (PGS) [37, 201] which employed GPS tracking and a GIS

database to relay positional information about the environment via haptic cues, synthesized speech, and spatialised sound. Similar systems have followed, delivering navigational instructions and information about the environment via refreshable Braille interfaces and speech [117, 316]. These systems were initially developed for portable digital assistants (PDAs), however, have since been re-engineered into stand-alone devices [118] and smart-phone apps [119, 182, 227, 231].

In addition to providing step-by-step navigational assistance, some aids also focus on exploring and discovering aspects along the route [72, 90, 148, 231, 304]. *Talking Signs* [72, 148] uses infrared transmitters and directional voice descriptions to guide VI users toward public beacons and landmarks, for example, restrooms, and entrances/exits. A mobile app, *BlindSquare* [231], connects an online global map with *FourSquare* [96], another app of crowd-sourced check-in points, and reads information about points of interest (POIs) aloud as the user moves along their selected route. Other services [90, 304] provide curated voice descriptions about architectural and cultural sites around major cities.

Voice descriptions, however, are difficult to i) spatially localise, and ii) disentangle if multiple POIs are present and described simultaneously. Some aids have therefore shifted to the use of spatial sound cues as a more intuitive way to guide users to their end destinations [182, 227, 359]. *Autour* [182], a mobile app, augments the turn-by-turn directions from a mapping service like *Google Maps* with ambient spatial audio cues which reveal information about POIs along the route—for example, signs, shop fronts, and restaurants. Additional verbal descriptions are available upon request. Microsoft's *Soundscape* [227] does away with step-by-step instructions altogether, and instead attaches binaural sound cues to beacons which are, in sequence, used to incrementally guide the user to their destination.

These GPS-based services are useful because they provide global or long-range context about the environment. They also, however, have their limitations: GPS offers only coarse-grained localisation accuracy, and has limited functionality indoors. GPS-based devices also require consistent internet connectivity, which may not always be available. Map and landmark databases must additionally be kept up-to-date, and furthermore, only register static information about the environment—for example buildings, roads, and shops. This alone is insufficient for navigation for which an understanding of the dynamic parts of an environment is also

required—for example, a moving bus, a red traffic light, or a temporarily open man-hole. For these reasons, VI users often pair GPS-based tools with mobility aids, like a white cane, guide dog, or others (see Appendix A.2) for hazard and obstacle detection at closer range.

A.2 Medium-range aids

Alongside way-finding, a second component of navigation is mobility. Mobility involves identifying obstacles and adeptly manoeuvring around them along the route to a destination. These tasks occur at mid-range distance to the user, and we therefore consider aids for mobility as medium-range aids. We broadly group these aids into two classes: the first constitutes simple clear-path indicators and obstacle detectors, and the second constitutes tools which relay richer representations of the environment, referred to as environmental imagers [105].

The effectiveness of the white cane as an obstacle detector and identifier has inspired many technological mobility aids, or electronic travel aids (ETAs). These ETAs aim to address the cane's primary limitations, namely its proximal range and hence limited information bandwidth, by employing sonar or infrared technology to detect objects further afield. The first ETA developed was the *Kay Sonic Torch* [164], which has been followed by several others: the *UltraCane* [312], the *MiniGuide* [102], the *Bat K Sonar Cane* [204], the *Sonic Guide* [165], the *TriSensor* [85], the *EyeCane* [211], and others [8, 48, 127, 146, 228, 309, 348]. Alternatives based on laser technology have also been developed [46, 76]. These aids, a selection of cane-like and hand-held devices, estimate obstacles' distance [8, 8, 102, 127, 211] and sometimes also their azimuthal position [48, 85, 146, 164, 165, 204, 228, 309, 312, 348], and feed this information back to the user via (electro- or vibro-) tactile or audio cues. In the simplest cases, the frequency (pitch, for audio) of the cue is mapped to the user's distance to the obstacle, with higher frequencies corresponding to closer obstacles. Its azimuthal position is conveyed through either where on the device the vibration is delivered [48, 127, 312], or through spatial stereo or binaural sound [85, 146, 165, 204, 228, 309, 348]. These devices thus not only extend the reach of the traditional white cane, but also provide information about off-course and head-height obstructions, which the cane and guide dog do not. Furthermore, their similarity to the cane may make them preferable options as mobility aids.

Despite their extended range, however, these devices simply detect the positions of obstacles, and thus capture only a portion of the full environment. In addition, since they are based on line-of-sight technologies, the pulse reflections are prone to being blocked or distorted, for example, by a passing person or opening door. As a result, these aids are insufficient for safely navigating in all situations.

Much exploration, therefore, has gone into devices which convey the contents of full images of a scene to VI users [1, 34, 35, 54, 224, 244]. Pioneered by Bach-y-Rita in the 1960s [34] with a vibrating pin-array placed on users' backs to convey simple shapes, these devices have the goal of substituting vision (or other damaged senses) with an intact sense, and have thus come to be called Sensory Substitution Devices (SSDs). Research into SSDs has been motivated by findings in the neuroscience literature which support cross-modality plasticity in the visual cortex of blind subjects [12, 13, 14, 113, 225, 343]. Extending this initial work, a number of works have investigated "seeing" with vibro-tactile and electro-tactile stimulation applied to different body surfaces [33, 35, 47, 137, 161], including most interestingly the tongue [356]! Other SSDs have gone the spatial auditory route [1, 54, 224]. The *vOICe*, perhaps the most widely investigated vision-to-audio SSD, converts a 2D grayscale image of the scene into an audio frequency spectrum or "soundscape", played left to right [224][1], with other SSDs employing similar mapping functions from image to spatial audio. Despite the promise of "seeing" with another sense for assistance with mobility and other tasks, uptake of SSDs in the VI community has been low [105, 210, 289]. This is because extracting task-relevant information and intuitively mapping it to another sense is difficult, and the alternate representations take a long time for VI users to learn [105].

Navigation, however, is a 3D task and is often carried out in dynamic environments. ETAs capture only slices of 3D environments [102, 164, 165, 204, 312], and most of the SSDs transform static 2D images of the environment [1, 54, 224], without accounting for depth or temporal information. To address this, range sensors have thus been paired with optical cameras to extract 3D information about environments for the purposes of VI mobility assistance [215, 271, 272, 359, 376]. Yuan and Manduchi [376] employ structured light to detect the presence of obstacles, steps, and drop-offs which are then relayed via audio cues, similar to

[1] A real-time web-cam demo is available at `www.seeingwithsound.com`

the ultrasound-based ETAs of [85, 146, 165, 204]. Others employ visual SLAM [215, 271, 272] to fuse camera pose estimates with dense 3D data from stereo triangulation to build a map of the user's local surroundings. Localising a user relative to the map, and updating this map as the user moves, therefore, allows for nearby obstacles to be detected [215], and steering cues can thus be delivered to guide users around them [272].

The proliferation of range sensors, particularly on mobile platforms, as well as high-fidelity real-time SLAM algorithms [189, 240, 274, 352] and semantic understanding abilities [108], is opening the door to the improved mapping and understanding of 3D environments. This is promising for VI assistive devices which will need to deliver navigation-relevant information about 3D environments at a high frame rate in order to provide mobility assistance.

A.3 Short-range aids

Interactions with objects and people characterise most of our day-to-day activities. For the VI community, these interactions present many challenges, from reading small text on a medication box, to setting thermostat controls on a home heating system, to recognising people in public spaces. A range of assistive tools, therefore, have been developed to help VI people with these short-range, interactional tasks. Below we discuss these aids, grouped by those for users with low vision, and those for users with no vision.

Low vision aids Low vision aids aim to augment what remains of a user's sight to facilitate easier interaction with objects and people. Most do this via magnification/zoom features, which enlarge text [315], screens [97, 99] or scenes [74, 86, 149, 246]. Others aim to filter the scene in meaningful ways, for example, by reducing background clutter, or highlighting objects in the central/peripheral FOV [86, 246, 246, 252]. *NuEyes* [246], *IrisVision* [149], and *eSight* [86], for example, are wearable headsets with zoom functionalities, with the additional ability to manipulate the colour, focus, and contrast of the incoming video. OXSIGHT's *Crystal* [252] provides similar functionalities through a pre-specified set of filters which highlights salient parts of the scene, for example by increasing contrast, cartoonising the image, or stripping out colour. The enhanced image is then projected onto the headset's lenses in real-time as part of an augmented reality display. These aids, therefore, assist with many short-range tasks, including reading, identifying people and their emotions/gestures, and locating objects.

No vision aids Low vision aids rely on a degree of existing vision which is not be the case for all in the VI community. Here, therefore, we discuss aids for interactional, short-range tasks which assume the user has no vision. These aids can easily be used by both partially-sighted and fully blind individuals alike.

Accessing written information is central to many of our daily short-range interactions, and is a core challenge faced by VI people [307]. Historically, Braille has been used by VI people to read, however, in recent years its usage has declined [355]. This is attributed to Braille's steep learning curve, the fact that Braille translations may not always be available, and the ubiquity of digital content, for which text-to-speech engines and screen reader software are nowadays highly effective. Reading printed information, however, remains a challenge for VI people, with difficulties often reported with reading instructions on medication boxes and food packaging, information on voting ballots, and information from banks and healthcare providers [307]. Tools which assist with these tasks typically rely on OCR technology: they work by capturing printed text or handwriting as an image, and then relaying it to the user via synthesised speech, or a refreshable Braille display [246, 250, 279, 287]. Advances in OCR have made it possible to handle menus, bank statements, credit cards, and other formats, however some cases still present difficulties, for example complex formatting, or when multimedia content is present. OCR-based tools are also generally not suited to longer-length pieces of printed text, for example, books, magazines, and newspapers. Instead, expansive libraries of large-print, audio, and Braille resources exist, often provided at no cost [292]. In the case of audio resources, technical standards have been developed to formalise the conversion to audio [71], with both audio playback software and purpose-built hardware in popular use [273, 303, 317].

Our daily interaction with technological devices like mobile phones, laptops and computers, has made it important to consider their accessibility to VI people. Screen readers—platform-specific software packages [23, 26, 98] which relay the content and layout of a digital screen as a user interacts with it—have filled this need, and are one of the most commonly used tools for accessing digital information by VI people [307]. Like the OCR-based tools, feedback is typically delivered via synthesised speech, or a refreshable Braille display. In this way, tasks like emailing, web browsing, accessing documents, and navigating the devices, can easily and independently be done by VI people. Screen readers, however, are limited by their prescriptive

set of actions. Furthermore, the increase in digital information with integrated multimedia, like images and videos, and highly-formatted visual displays, particularly online, poses challenges for screen readers. While some websites have accessible versions with which screen readers can interface (for example some social media platforms allow textual descriptions to be uploaded with an image or video), these are expensive to develop, and as a result, are not always available.

Digital devices are not the only types of objects with which we regularly interact. Consider, for instance, the home environment: we heat our food in a microwave, we do our laundry using a washing machine and dryer, and we watch the television or listen to the radio. To cater for VI access, therefore, many household devices and appliances ship with accessible features. In the kitchen for instance, a host of "talking tools" have been developed: a verbal tin tagger to locate a specific tin at a later stage, an audible/vibrating tool to measure the liquid level in a mug or pot, a metal disc to alert the VI user when the water has boiled, and talking kitchen scales, microwaves, hobs, and ovens, all relaying interface information via synthesised voice descriptions [293]. Beyond the kitchen, other products include talking clocks and calendars, hand-held colour and light intensity detectors, and devices which allow voice recordings to be attached to sticky labels and placed on objects, food, and clothing, for later playback. Many televisions and radios also have features which read out on-device information such as channels and menus using voice-over technology. Each of these aids, therefore, conveys fine-grained information about objects and allows VI users to operate and interact with them[2].

For entertainment, many television programmes and films are also accompanied by audio descriptions which verbally narrate the visual content[3]. This goes beyond closed captions which simply convert the spoken portions of the programme to text. Additionally, podcasts and music streaming services have serendipitously provided alternate routes of information delivery and entertainment for the VI community.

Fusing many of the functionalities offered by the above aids, and leveraging recent advances in machine learning, computer vision, and natural language processing (NLP), a host of aids with basic capabilities in speech, object recognition, and scene understanding are now available.

[2]For a full scope of available products see www.shop.rnib.org.uk.

[3]UK regulations require at least 10% of programming content to be available in audio-described format.

One of the most widely know is Microsoft's *SeeingAI* [226], a camera-based mobile app that is able to i) recognise and describe people nearby including their emotions, ii) describe the user's surroundings, iii) describe perceived colours, and iv) identify currency bills, along with reading out printed and handwritten text using standard OCR. Other mobile apps like *TapTapSee* [67] and *iDentifi* [328] perform object recognition on an image or video segment, and deliver spoken descriptions of the visual contents, including details like brand and colour. Wearables like *OrCam MyEye2* [249] and *CyberEyez* [74] offer similar capabilities but in the form of a wearable headset or pair of spectacles, allowing for complete hands-free operation. These wearables, however, are generally not as affordable as a mobile app. In addition, advances in NLP and speech processing have spurred the development of home voice assistants like Google's *Home* [111] and Amazon's *Echo* [11]. These assistants control appliances around the house, purchase supplies, play the news, podcasts, and audio-books, manage calendars on behalf of the user, and thereby, as a fortuitous by-product, can assist with some of the in-house challenges faced by VI people [44]. They, however, do no visual processing of the scene, and are also not portable. This limits their use cases to only specific scenarios for VI people.

The dynamic environment and range of possible tasks makes it difficult for these assistive devices to work in all cases. Mobile services like *Be My Eyes* [39] and *Aira* [7] address this by pairing sighted human helpers with VI video callers via a live video connection. In this way, real-time *human* assistance is provided across a great diversity of tasks. These services, however, rely on around-the-clock human helpers who may need to be formally trained, which can be expensive and time-consuming. With recent advances in machine learning, computer vision, and NLP, it is therefore a promising direction that one day these human assistants may be replaced by data-driven models with similar capabilities, thus allowing some of these issues to be circumvented. These data-driven assistive tools hold great promise for the VI community, not only in helping with short-range tasks, but with many others too.

Bibliography

[1] Sami Abboud, Shlomi Hanassy, Shelly Levy-Tzedek, Shachar Maidenbaum, and Amir Amedi. EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution. *Restorative Neurology and Neuroscience*, 32(2):247–257, 2014. 8, 9, 10, 11, 16, 46, 73, 149

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016. URL http://arxiv.org/abs/1609.08675. 13, 14, 142

[3] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 19, 111

[4] Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. C-VQA: A compositional split of the visual question answering (VQA) v1.0 dataset. *CoRR*, abs/1704.08243, 2017. URL http://arxiv.org/abs/1704.08243. 18, 19, 141

[5] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 18, 19, 141

[6] Faruk Ahmed, Dany Tarlow, and Dhruv Batra. Optimizing expected intersection-over-union with candidate-constrained CRFs. In *International Conference on Computer Vision (ICCV)*, 2015. 15, 26, 33

[7] Aira Tech Corp. Aira, 2018. URL www.aira.io. Accessed 12 April 2019. 8, 11, 17, 143, 153

[8] Junichi Akita, Takanori Komatsu, Kiyohide Ito, Tetsuo Ono, and Makoto Okamoto. Cyarm: Haptic sensing device for spatial localization on basis of exploration by arms. *Advances in Human-Computer Interaction*, 2009:6, 2009. 8, 9, 148

[9] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(11):2189–2202, 2012. 36, 41

[10] Amazon, Inc. Alexa, 2014. URL www.developer.amazon.com/alexa. Accessed 9 February 2018. 11, 82

[11] Amazon, Inc. Amazon Echo, 2019. URL www.amazon.co.uk/Echo. Accessed on 18 April 2019. 8, 11, 153

[12] Amir Amedi, Rafael Malach, Talma Hendler, Sharon Peled, and Ehud Zohary. Visuo-haptic object-related activation in the ventral visual pathway. *Nature Neuroscience*, 4(3):324, 2001. 16, 149

[13] Amir Amedi, Noa Raz, Pazit Pianka, Rafael Malach, and Ehud Zohary. Early 'visual' cortex activation correlates with superior verbal memory performance in the blind. *Nature Neuroscience*, 6(7):758, 2003. 16, 149

[14] Amir Amedi, Lotfi B. Merabet, Felix Bermpohl, and Alvaro Pascual-Leone. The occipital cortex in the blind: Lessons about plasticity and vision. *Current Directions in Psychological Science*, 14(6):306–311, 2005. 16, 149

[15] Amir Amedi, Shir Hofstetter, Shachar Maidenbaum, and Benedetta Heimler. Task selectivity as a comprehensive principle for brain organization. *Trends in Cognitive Sciences*, 21(5):307–310, 2017. 46

[16] American Printing House for the Blind. Nearby Explorer, 2019. URL www.aph.org/nearby-explorer. Accessed on 10 May 2019. 8

[17] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin. In *International Conference on Machine Learning (ICML)*, 2016. 2

[18] Ankesh Anand, Eugene Belilovsky, Kyle Kastner, Hugo Larochelle, and Aaron C. Courville. Blindfold baselines for embodied QA. *CoRR*, abs/1811.05013, 2018. URL http://arxiv.org/abs/1811.05013. 19, 111

[19] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *European Conference on Computer Vision (ECCV)*, 2016. 19

[20] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 18

[21] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 141

[22] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 18

[23] Android. TalkBack, 2019. URL www.support.google.com/accessibility/android/answer/6007100. Accessed on 11 April 2019. 5, 8, 9, 10, 151

[24] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 2, 18, 19, 82, 113, 120, 141, 142

[25] Apple, Inc. iOS Siri, 2011. URL www.developer.apple.com/sirikit. Accessed 9 February 2018. 11, 82

[26] Apple, Inc. VoiceOver, 2019. URL www.apple.com/uk/accessibility/mac/vision. Accessed on 11 April 2019. 5, 8, 9, 10, 151

[27] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T. Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 26, 36, 37, 41

[28] Anurag Arnab and Philip H.S. Torr. Bottom-up instance segmentation using deep higher-order CRFs. In *British Machine Vision Conference (BMVC)*, 2016. 14, 78

[29] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations (ICLR)*, 2016. 140

[30] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations (ICLR)*, 2017. 96, 117

[31] Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: A corpus for adding memory to goal-oriented dialogue systems. *CoRR*, abs/1704.00057, 2017. URL http://arxiv.org/abs/1704.00057. 113

[32] Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research (JMLR)*, 3(Jul):1–48, 2002. 114, 129

[33] Paul Bach-y-Rita. Tactile sensory substitution studies. *Annals of the New York Academy of Sciences*, 1013(1):83–91, 2004. 16, 46, 73, 149

[34] Paul Bach-y-Rita, Carter C. Collins, Frank A. Saunders, Benjamin White, and Lawrence Scadden. Vision substitution by tactile image projection. *Nature*, 221(5184):963–964, 1969. 5, 8, 9, 16, 46, 73, 149

[35] Paul Bach-y-Rita, Kurt A. Kaczmarek, Mitchell E. Tyler, and Jorge Garcia-Lara. Form perception with a 49-point electrotactile stimulus array on the tongue: A technical note. *Journal of Rehabilitation Research and Development*, 35(4):427–430, 1998. 8, 9, 16, 46, 73, 149

[36] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, abs/1803.01271, 2018. URL http://arxiv.org/abs/1803.01271. 88

[37] University College Santa Barbara. Personal Guidance System, 2008. URL www.geog.ucsb.edu/pgs. Accessed on 11 April 2019. 146

[38] Cassie Barton and Lucy Pullicino. Choosing central heating controls and saving energy, 2015. URL www.ridc.org.uk/sites/default/files/documents/pdfs/home-tech/choosing-central-heating-controls-saving-energy.pdf. Research Institute for Disabled Consumers. 5

[39] Be My Eyes. Be My Eyes, 2017. URL www.bemyeyes.com. Accessed 12 April 2019. 8, 11, 17, 143, 153

[40] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision (ECCV)*, 2016. 24, 26, 27, 30, 36

[41] Durand R. Begault. *3D sound for virtual reality and multimedia*. Academic Press Professional, Inc., San Diego, CA, 1994. 73

[42] BeltMap. BeltMap, 2019. URL www.beltmap.com. Accessed on 7 May 2019. 8

[43] Alex Black, Jan E. Lovie-Kitchin, Russell L. Woods, Nicole Arnold, John Byrnes, and Jane Murrish. Mobility performance with retinitis pigmentosa. *Clinical and Experimental Optometry*, 80(1):1–12, 1997. 45

[44] Ian Bogost. Alexa is a revelation for the blind. The Atlantic, 2018. URL www.theatlantic.com/magazine/archive/2018/05/what-alexa-taught-my-father/556874. Accessed 12 April 2019. 11, 153

[45] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017. 117, 123

[46] D. Ridgely Bolgiano and Elinor Meeks. A laser cane for the blind. *IEEE Journal of Quantum Electronics*, 3(6):268–268, 1967. 148

[47] Johann Borenstein. The NavBelt—a computerized multi-sensor travel aid for active guidance of the blind. In *CSUN Conference on Technology and Persons with Disabilities*, 1990. 8, 16, 149

[48] Johann Borenstein and Iwan Ulrich. The guidecane—a computerized travel aid for the active guidance of blind pedestrians. In *International Conference on Robotics and Automation (ICRA)*, 1997. 8, 9, 16, 148

[49] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *SIGNLL Conference on Computational Language Learning (CoNLL)*, 2016. 93

[50] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 14

[51] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. Visual challenges in the everyday lives of blind people. In *SIGCHI Conference on Human Factors in Computing Systems*, 2013. 13, 14

[52] Brian Brown, Lesley Brabyn, Leslie Welch, Gunilla Haegerstrom-Portnoy, and August Colenbrander. Contribution of vision variables to mobility in age-related maculopathy patients. *American Journal of Optometry and Physiological Optics*, 63(9):733–739, 1986. 45

[53] Sheryl E. Burgstahler and Richard E. Ladner. Increasing the participation of people with disabilities in computing fields. *Computer*, 40(5):94–97, 2007. 6

[54] Christian Capelle, Charles Trullemans, Patricia Arno, and Claude Veraart. A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution. *IEEE Transactions on Biomedical Engineering*, 45(10):1279–1293, 1998. 8, 9, 10, 16, 46, 149

[55] Sarah S. Chance, Florence Gaunet, Andrew C. Beall, and Jack M. Loomis. Locomotion mode affects the updating of objects encountered during travel: The contribution of vestibular and proprioceptive inputs to path integration. *Presence*, 7(2):168–178, 1998. 49, 73

[56] Siddhartha Chandra and Iasonas Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep Gaussian CRFs. In *European Conference on Computer Vision (ECCV)*, 2016. 24

[57] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision (3DV)*, 2017. 143

[58] Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H.S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European Conference on Computer Vision (ECCV)*, 2018. 136, 144

[59] Daniel-Robert Chebat, Shachar Maidenbaum, and Amir Amedi. Navigation using sensory substitution in real and virtual mazes. *PLOS ONE*, 10(6):e0126307, 2015. 73

[60] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *International Conference on Learning Representations (ICLR)*, 2015. 10, 13, 14, 24, 35

[61] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 10, 13, 14

[62] Jian Cheng, Peisong Wang, Gang Li, Qinghao Hu, and Hanqing Lu. Recent advances in efficient computation of deep convolutional neural networks. *Frontiers of Information Technology and Electronic Engineering*, 19(1):64–77, 2018. 144

[63] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip H.S. Torr. BING: Binarized normed gradients for objectness estimation at 300FPS. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 36, 41

[64] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H.S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 37 (3):569–582, 2015. 29, 41

[65] Hyo Geun Choi, Min Joung Lee, and Sang-Mok Lee. Visual impairment and risk of depression: A longitudinal follow-up study using a national sample cohort. *Scientific reports*, 8(1):2083, 2018. 7

[66] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 88

[67] Cloudsight, Inc. TapTapSee, 2013. URL www.taptapseeapp.com. Accessed on 15 April 2019. 2, 8, 11, 12, 16, 153

[68] Michael Cogswell, Xiao Lin, Senthil Purushwalkam, and Dhruv Batra. Combining the best of graphical models and ConvNets for semantic segmentation. *CoRR*, abs/1412.4313, 2014. URL http://arxiv.org/abs/1412.4313. 15, 26, 33

[69] Chetz Colwell, Helen Petrie, Diana Kornbrot, Andrew Hardwick, and Stephen Furner. Haptic virtual reality for blind computer users. In *International ACM Conference on Assistive Technologies (ASSETS)*, 1998. 79

[70] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, (5):603–619, 2002. 124

[71] Daisy Consortium. Daisy, 2019. URL www.daisy.org. Accessed on 22 April 2019. 151

[72] William Crandall, John Brabyn, Billie Louise Bentzen, and Linda Myers. Remote infrared signage evaluation for transit stations and intersections. *Development*, 36(4), 1999. 8, 147

[73] Laura A. Cushman, Karen Stein, and Charles J. Duffy. Detecting navigational deficits in cognitive ageing and Alzheimer disease using virtual reality. *Neurology*, 71(12):888–895, 2008. 17, 74

[74] CyberTimez. Cyber Eyez, 2019. URL www.cybertimez.com. Accessed on 18 April 2019. 2, 8, 11, 150, 153

[75] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-Kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 142

[76] Quoc Dang, Youngjoon Chee, Duy Pham, and Young Suh. A virtual blind cane using a line laser-based vision system and an inertial measurement unit. *Sensors*, 16(1):95, 2016. 148

[77] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 18, 19, 20, 82, 83, 86, 87, 88, 89, 90, 91, 94, 95, 96, 98, 102, 105, 110, 111, 112, 115, 116, 117, 118, 123, 130, 131, 140, 141, 142

[78] Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *International Conference on Computer Vision (ICCV)*, 2017. 18, 83, 88, 95, 98, 111, 113, 141

[79] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 18, 19, 141, 142

[80] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. GuessWhat?! Visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 18, 113, 141, 142

[81] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C. Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 141

[82] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 24, 27

[83] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 25, 34, 41, 142

[84] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805. 2, 10, 123, 140

[85] Randolph D. Easton and Richard M. Jackson. Pilot test of the Trisensor, a new generation sonar sensory aid. *Journal of Visual Impairment and Blindness*, 1983. 8, 9, 148, 150

[86] eSight Corp. eSight, 2018. URL www.esighteyewear.com/homex. Accessed on 20 April 2019. 8, 9, 10, 11, 14, 150

[87] Kawin Ethayarajh. Unsupervised random walk sentence embeddings: A strong but simple baseline. In *Annual Meeting of the Association for Computational Linguistics (ACL) Workshop on Representation Learning for NLP*, 2018. 140

[88] Mark Everingham, Luc van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. In *International Journal of Computer Vision (IJCV)*, 2010. 142

[89] Mark Everingham, S.M. Ali Eslami, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, 2015. 15, 23, 26, 29, 35, 41

[90] Vocal Eyes. London Beyond Sight, 2013. URL www.vocaleyes.co.uk/audio/london-beyond-sight. Accessed on 11 April 2019. 147

[91] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 142

[92] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision (IJCV)*, 59(2):167–181, 2004. 36, 41

[93] Ione Fine, Connie L. Cepko, and Michael S. Landy. Vision research special issue: Sight restoration: Prosthetics, optogenetics and gene therapy. *Vision Research*, 111(Pt B):115, 2015. 2, 45

[94] Seth R. Flaxman, Rupert R.A. Bourne, Serge Resnikoff, Peter Ackland, Tasanee Braithwaite, Maria V. Cicinelli, Aditi Das, Jost B. Jonas, Jill Keeffe, John H. Kempen, et al. Global causes of blindness and distance vision impairment 1990–2020: A systematic review and meta-analysis. *Lancet Global Health*, 5 (12):e1221–e1234, 2017. 2, 45

[95] Kate Flynn, Sue Keil, and Rory Cobb. Provision of accessible GCSE exam papers, 2016. URL www.rnib.org.uk/knowledge-and-research-hub/research-reports/education-research/accessible-gcse-exam-papers. Royal National Institute of Blind People. 6

[96] FourSquare Labs, Inc. FourSquare City Guide, 2017. URL www.foursquare.com. Accessed on 9 April 2019. 147

[97] Freedom Scientific, Inc. MAGic, 2019. URL www.freedomscientific.com/Products/software/MAGic. Accessed on 20 April 2019. 5, 8, 9, 150

[98] Freedom Scientific, Inc. JAWS, 2019. URL www.freedomscientific.com/Products/software/JAWS. Accessed on 11 April 2019. 5, 8, 9, 10, 151

[99] Freedom Scientific Inc. ZoomText, 2019. URL www.zoomtext.com. Accessed on 20 April 2019. 5, 8, 9, 150

[100] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 18, 19, 141

[101] Rebeca I. García-Betances, María Teresa Arredondo Waldmeyer, Giuseppe Fico, and María Fernanda Cabrera-Umpiérrez. A succinct overview of virtual reality technology use in Alzheimer's disease. *Frontiers in Aging Neuroscience*, 7:80, 2015. 17, 74

[102] GDP Research. MiniGuide, 2005. URL www.gdp-research.com.au/minig_1.htm. Accessed on 9 April 2019. 8, 9, 16, 46, 148, 149

[103] Duane R. Geruschat, Kathleen A. Turano, and Julie W. Stahl. Traditional measures of mobility performance and retinitis pigmentosa. *Optometry and Vision Science*, 75(7):525–537, 1998. 45

[104] Nicholas A. Giudice. Navigating without vision: Principles of blind spatial cognition. In D.R. Montello, editor, *Handbook of Behavioural and Cognitive Geography*, chapter 15, pages 260–288. Edward Elgar Publishing, Cheltenham, UK, 1 edition, 2018. 4, 45

[105] Nicholas A. Giudice and Gordon E. Legge. Blind navigation and the role of technology. *The Engineering Handbook of Smart Technology for Aging, Disability, and Independence*, pages 479–500, 2008. 7, 9, 10, 46, 145, 148, 149

[106] Nicholas A. Giudice, Jonathan Z. Bakdash, Gordon E. Legge, and Rudrava Roy. Spatial learning and navigation using a virtual verbal display. *ACM Transactions on Applied Perception*, 7(1):3:1–3:22, 2010. 49

[107] Reginald G. Golledge. Geography and the disabled: A survey with special reference to vision impaired and blind populations. *Transactions of the Institute of British Geographers*, pages 63–85, 1993. 4, 45

[108] Stuart Golodetz, Michael Sapienza, Julien P.C. Valentin, Vibhav Vineet, Ming-Ming Cheng, Anurag Arnab, Victor Adrian Prisacariu, Olaf Kähler, Carl Yuheng Ren, David W. Murray, Shahram Izadi, and Philip H.S. Torr. SemanticPaint: A framework for the interactive segmentation of 3D scenes. *CoRR*, abs/1510.03727, 2015. URL http://arxiv.org/abs/1510.03727. 2, 13, 150

[109] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision (IJCV)*, 106(2):210–233, 2014. 114

[110] Gregory L. Goodrich and Richard Ludt. Assessing visual detection ability for mobility in individuals with low vision. *Visual Impairment Research*, 5(2):57–71, 2003. 62

[111] Google. Google Home, 2019. URL store.google.com/gb/product/google_home. Accessed on 18 April 2019. 8, 11, 153

[112] Frédéric Gougoux, Franco Lepore, Maryse Lassonde, Patrice Voss, Robert J. Zatorre, and Pascal Belin. Neuropsychology: Pitch discrimination in the early blind. *Nature*, 430(6997):309, 2004. 77

[113] Frédéric Gougoux, Robert J. Zatorre, Maryse Lassonde, Patrice Voss, and Franco Lepore. A functional neuroimaging study of sound localization: Visual cortex activity predicts performance in early-blind individuals. *PLOS Biology*, 3(2):e27, 2005. 16, 77, 139, 149

[114] Neda F. Gould, M. Kathleen Holmes, Bryan D. Fantie, David A. Luckenbaugh, Daniel S. Pine, Todd D. Gould, Neil Burgess, Husseini K. Manji, and Carlos A. Zarate Jr. Performance on a virtual reality spatial memory navigation task in depressed patients. *American Journal of Psychiatry*, 164(3):516–519, 2007. 17, 74

[115] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 18, 19, 141

[116] H. Paul Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics*, volume 3, pages 41–58. Academic Press, Cambridge, MA, 1975. 17

[117] Sendero Group. BrailleNote GPS, 2006. URL www.senderogroup.com/products/shopgps.html. Accessed on 11 April 2019. 8, 9, 147

[118] Sendero Group. Mobile Geo, 2008. URL www.senderogroup.com/products/shopmgeo.html. Accessed on 11 April 2019. 147

[119] Sendero Group. Seeing Eye GPS, 2010. URL www.senderogroup.com/products/SeeingEyeGPS. Accessed on 11 April 2019. 8, 9, 147

[120] Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. PixelVAE: A latent variable model for natural images. In *International Conference on Learning Representations (ICLR)*, 2017. 89, 93, 104

[121] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz Grand Challenge: Answering visual questions from blind people. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 18, 143

[122] David Guth and Robert LaDuke. Veering by blind pedestrians: Individual differences and their implications for instruction. *Journal of Visual Impairment and Blindness*, 89(1):28–37, 1995. 4, 45

[123] Alex Hadwen-Bennett, Sue Sentance, and Cecily Morrison. Making programming accessible to learners with visual impairments: A literature review. *International Journal of Computer Science Education in Schools*, 2(2):n2, 2018. 6

[124] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical Correlation Analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004. 114, 129

[125] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *International Conference on Computer Vision (ICCV)*, 2011. 35

[126] Eric Harris. Choosing cookers, ovens, hobs, and microwaves, 2015. URL `www.ridc.org.uk/sites/default/files/documents/pdfs/home-tech/cookers.pdf`. Research Institute for Disabled Consumers. 5

[127] Jess Hartcher-O'Brien, Malika Auvray, and Vincent Hayward. Perception of distance-to-obstacle through time-delayed tactile feedback. In *Conference on World Haptics Conference (WHC)*, 2015. 9, 46, 73, 148

[128] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, Cambridge, 2003. 13

[129] Sharon A. Haymes, Daryl Guest, Anthony D. Heyes, and Alan W. Johnston. Mobility of people with retinitis pigmentosa as a function of vision and psychological variables. *Optometry and Vision Science*, 73(10):621–637, 1996. 45

[130] Sharon A. Haymes, Alan W. Johnston, and Anthony D. Heyes. Relationship between vision impairment and ability to perform activities of daily living. *Ophthalmic and Physiological Optics*, 22(2):79–91, 2002. 1, 5, 45

[131] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 37 (9):1904–1916, 2015. 2, 10, 13, 14

[132] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 117, 144

[133] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *International Conference on Computer Vision (ICCV)*, 2017. 14

[134] Anthony D. Heyes. The sonic pathfinder—a new travel aid for the blind. In *High Technology Aids for the Disabled*, pages 165–171. 1983. 8

[135] Stephen L. Hicks, Iain R. Wilson, Louwai Muhammed, John Worsfold, Susan M. Downes, and Christopher Kennard. A depth-based head-mounted visual display to aid navigation in partially sighted individuals. *PLOS ONE*, 8(7):e67695, 2013. 50

[136] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. 2, 88

[137] Samantha Horvath, John Galeotti, Bing Wu, Roberta L. Klatzky, Mel Siegel, and George Stetten. FingerSight: Fingertip haptic sensing of the visual environment. *IEEE Journal of Translational Engineering in Health and Medicine*, 2:1–9, 2014. 8, 16, 149

[138] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 1936. 114

[139] Qibin Hou, Ming-Ming Cheng, Xiao-Wei Hu, Ali Borji, Zhuowen Tu, and Philip H.S. Torr. Deeply supervised salient object detection with short connections. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 25, 29, 30, 34, 36, 37

[140] Qibin Hou, Daniela Massiceti, Puneet K. Dokania, Yunchao Wei, Ming-Ming Cheng, and Philip H.S. Torr. Bottom-up top-down cues for weakly-supervised semantic segmentation. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, 2017. 22

[141] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 18, 83, 88

[142] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 10, 13, 14, 144

[143] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 18, 19

[144] Barry Hughes. Active artificial echolocation and the nonvisual perception of aperture passability. *Human Movement Science*, 20(4):371–400, 2001. 48, 73

[145] Tina Iachini, Gennaro Ruggiero, and Francesco Ruotolo. Does blindness affect egocentric and allocentric frames of reference in small and large scale spaces? *Behavioural Brain Research*, 273:73–81, 2014. 77

[146] Tohru Ifukube, Tadayuki Sasaki, and Chen Peng. A blind mobility aid modeled after echolocation of bats. *IEEE Transactions on Biomedical Engineering*, 38(5):461–465, 1991. 48, 73, 148, 150

[147] Rubén Iglesias, Sara Casado, Teresa Gutiérrez, JI Barbero, Carlo A. Avizzano, Simone Marcheschi, and Massimo Bergamasco. Computer graphics access for blind people through a haptic and audio virtual environment. In *International Workshop on Haptic, Audio and Visual Environments and Their Applications (HAVE)*, 2004. 79

[148] Smith-Kettlewell Eye Research Institute. Talking Signs, 2009. URL www.talkingsignsservices.com. Accessed on 11 April 2019. 8, 9, 147

[149] IrisVision. IrisVision, 2019. URL www.irisvision.com. Accessed on 20 April 2019. 8, 9, 150

[150] Rebecca Q. Ivers, Robert G. Cumming, Paul Mitchell, and Karin Attebo. Visual impairment and falls in older adults: The Blue Mountain's eye study. *Journal of the American Geriatrics Society*, 46(1): 58–64, 1998. 1, 4, 45

[151] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In *European Conference on Computer Vision (ECCV)*, 2016. 19

[152] Unnat Jain, Ziyu Zhang, and Alexander Schwing. Creativity: Generating diverse questions using variational autoencoders. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 18, 84, 88

[153] Saumya Jetley, Michael Sapienza, Stuart Golodetz, and Philip H.S. Torr. Straight to shapes: Real-time detection of encoded shapes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 13, 14, 78, 79

[154] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, pages 675–678. ACM, 2014. 35

[155] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 36, 37, 41

[156] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: The winning entry to the VQA challenge 2018. *CoRR*, abs/1807.09956, 2018. URL http://arxiv.org/abs/1807.09956. 18

[157] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. DenseCap: Fully convolutional localization networks for dense captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 82

[158] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *International Conference on Computer Vision (ICCV)*, 2017. 18, 19

[159] Karen Sparck Jones and Julia R. Galliers. *Evaluating natural language processing systems: An analysis and review*, volume 1083. Springer Science & Business Media, 1 edition, 1995. 19

[160] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *International Conference on Computer Vision (ICCV)*, 2017. 18, 141

[161] Hiroyuki Kajimoto, Yonezo Kanno, and Susumu Tachi. Forehead retina system. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH) Emerging Technologies*, 2006. 16, 149

[162] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014. 18, 83, 88

[163] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 82

[164] Leslie Kay. An ultrasonic sensing probe as a mobility aid for the blind. *Ultrasonics*, 2(2):53–59, 1964. 8, 9, 16, 46, 148, 149

[165] Leslie Kay. A sonar aid to enhance spatial perception of the blind: Engineering design and evaluation. *Radio and Electronic Engineer*, 44(11):605–627, 1974. 8, 9, 16, 46, 148, 149, 150

[166] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. *CoRR*, abs/1705.06950, 2017. URL http://arxiv.org/abs/1705.06950. 142

[167] Sue Keil. Local authority VI education service provision for blind and partially sighted children and young people in 2015, 2016. URL www.rnib.org.uk/knowledge-and-research-hub-research-reports-education-research/vi-service-provision-2015. Royal National Institute of Blind People. 6

[168] Gertrudis I.J.M. Kempen, Judith Ballemans, Adelita V. Ranchor, Ger H.M.B. van Rens, and G.A. Rixt Zijlstra. The impact of low vision on activities of daily living, symptoms of depression, feelings of anxiety and social support in community-living older adults seeking vision rehabilitation services. *Quality of Life Research*, 21(8):1405–1411, 2012. 1, 45

[169] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 1971. 114

[170] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014. 93

[171] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 18, 83, 84, 104

[172] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 137, 144

[173] Charles Daniel Kish. Evaluation of an echo-mobility program for young blind people. Master's thesis, California State University, San Bernardino, 1995. 17, 54

[174] Charles Daniel Kish. Human echolocation: How to "see" like a bat. *New Scientist*, 202(2703):31–33, 2009. URL www.sciencedirect.com/science/article/pii/S0262407909609970. 47

[175] Andrew J. Kolarik, Silvia Cirstea, Shahina Pardhan, and Brian C.J. Moore. A summary of research investigating echolocation abilities of blind and sighted humans. *Hearing Research*, 310:60–68, 2014. 17, 47, 54

[176] Alexander Kolesnikov and Christoph H. Lampert. Seed, Expand and Constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision (ECCV)*, 2016. 24, 26, 35, 36, 39

[177] Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3D environment for visual AI. *CoRR*, abs/1712.05474, 2017. URL http://arxiv.org/abs/1712.05474. 143

[178] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011. 36, 37, 41

[179] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 142

[180] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. 25

[181] Thomas Kuyk, Jeffry L. Elliott, and Patti S. Fuhr. Visual correlates of mobility in real world settings in older adults with low vision. *Optometry and Vision Science*, 75(7):538–547, 1998. 45

[182] McGill Shared Reality Lab. Autour, 2014. URL autour.mcgill.ca/en. Accessed on 9 April 2019. 8, 11, 16, 147

[183] Ecosse L. Lamoureux, Jennifer B. Hassell, and Jill E. Keeffe. The determinants of participation in activities of daily living in people with impaired vision. *American Journal of Ophthalmology*, 137(2): 265–270, 2004. 1, 45

[184] Ecosse L. Lamoureux, Elaine W. Chong, Julian Thumboo, Hwee Lin Wee, Jie Jin Wang, Seang-Mei Saw, Tin Aung, and Tien Y. Wong. Vision impairment, ocular conditions, and vision-specific function: The Singapore Malay eye study. *Ophthalmology*, 115(11):1973–1981, 2008. 1, 4, 45

[185] Jean François Lapointe, Pascal Savard, and Norman G. Vinson. A comparative study of four input devices for desktop virtual walkthroughs. *Computers in Human Behaviour*, 27(6):2186–2191, 2011. 73

[186] Alon Lavie and Michael J. Denkowski. The METEOR metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115, 2009. 19, 20, 122, 140

[187] Nadia Lessard, Michael Par, Franco Lepore, and Maryse Lassonde. Early-blind human subjects localize sound sources better than sighted subjects. *Nature*, 395(6699):278–280, 1998. 77, 139

[188] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *European Conference on Computer Vision (ECCV)*, 2018. 142

[189] Jingtang Liao, Bert Buchholz, Jean-Marc Thiery, Pablo Bauszat, and Elmar Eisemann. Indoor scene reconstruction using near-light photometric stereo. *IEEE Transactions on Image Processing*, 26(3): 1089–1101, 2017. 11, 13, 16, 48, 150

[190] William Lidwell, Kritina Holden, and Jill Butler. *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design*. Rockport Publishers, Beverly, MA, 2010. 146

[191] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics (ACL) Workshop on Text Summarization Branches Out*, 2004. 19, 140

[192] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 24, 26

[193] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, 2014. 24, 41, 90, 142

[194] Nian Liu and Junwei Han. DHSNet: Deep hierarchical saliency network for salient object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 40, 41

[195] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 41

[196] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 24, 35

[197] Richard G. Long. Orientation and mobility research: What is known and what needs to be known. *Peabody Journal of Education*, 67(2):89–109, 1990. 4, 45

[198] Richard G. Long and Nicholas A. Giudice. Establishing and maintaining orientation for mobility. In B.B. Blasch, W.R. Wiener, and R.W. Welsh, editors, *Foundations of Orientation and Mobility*, volume 1, pages 45–62. American Foundation for the Blind, New York, NY, 3 edition, 2010. 4, 45

[199] Jack M. Loomis, Reginald G. Golledge, and Roberta L. Klatzky. Navigation system for the blind: Auditory display modes and guidance. *Presence*, 7(2):193–203, 1998. 49

[200] Jack M. Loomis, Roberta L. Klatzky, and Reginald G. Golledge. Navigating without vision: Basic and applied research. *Optometry and Vision Science*, 78(5):282–289, 2001. 4, 45

[201] Jack M. Loomis, James R. Marston, Reginald G. Golledge, and Roberta L. Klatzky. Personal guidance system for people with visual impairment: A comparison of spatial displays for route guidance. *Journal of Visual Impairment and Blindness*, 99(4):219, 2005. 146

[202] Jack M. Loomis, Roberta L. Klatzky, and Nicholas A. Giudice. Sensory substitution of vision: Importance of perceptual and cognitive processing. In R. Manduchi and S. Kurniawan, editors, *Assistive Technology for Blindness and Low Vision*, pages 162–191. CRC Press, Boca Raton, FL, 2012. 17, 46, 139

[203] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 144

[204] Bay Advanced Technologies Ltd. K-Sonar Cane, 2005. URL ableddata.acl.gov/product/k-sonar-model-1-07000-00. Accessed on 9 April 2019. 8, 9, 16, 46, 148, 149, 150

[205] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 18, 19

[206] Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 18, 19, 116, 117, 118, 131

[207] Kevin Lynch. *The image of the city*, volume 11. MIT Press, Cambridge, MA, 1960. 146

[208] Justin A. MacDonald, Paula P. Henry, and Tomasz R. Letowski. Spatial audio through a bone conduction interface. *International Journal of Audiology*, 45(10):595–599, 2006. 79

[209] Shachar Maidenbaum, Shelly Levy-Tzedek, Daniel-Robert Chebat, and Amir Amedi. Increasing accessibility to the blind of virtual environments, using a virtual mobility aid based on the "EyeCane": Feasibility study. *PLOS ONE*, 8(8):e72555, 2013. 73, 79

[210] Shachar Maidenbaum, Sami Abboud, and Amir Amedi. Sensory substitution: Closing the gap between basic research and widespread practical visual rehabilitation. *Neuroscience and Biobehavioral Reviews*, 41:3–15, 2014. 9, 46, 149

[211] Shachar Maidenbaum, Daniel Robert Chebat, Shelly Levy-Tzedek, and Amir Amedi. Depth-to-audio sensory substitution for increasing the accessibility of virtual environments. In *International Conference on Universal Access in Human-Computer Interaction (UAHCI)*, 2014. 8, 9, 16, 46, 73, 148

[212] Roberto Manduchi and Sri Kurniawan. Mobility-related accidents experienced by people with visual impairment. *AER Journal: Research and Practice in Visual Impairment and Blindness*, 4(2):44–54, 2011. 1, 4, 45

[213] James A. Marron and Ian L. Bailey. Visual factors and orientation-mobility performance. *American Journal of Optometry and Physiological Optics*, 59(5):413–426, 1982. 1, 45

[214] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *International Conference on Computer Vision (ICCV)*, 2001. 41

[215] Juan Manuel Saez Martinez and Francisco Escolano Ruiz. Stereo-based aerial obstacle detection for the visually impaired. In *European Conference on Computer Vision (ECCV) Workshop on Computer Vision Applications for the Visually Impaired*, 2008. 149, 150

[216] Daniela Massiceti, Alexander Krull, Eric Brachmann, Carsten Rother, and Philip H.S. Torr. Random forests versus neural networks—what's best for camera localization? In *International Conference on Robotics and Automation (ICRA)*, 2017. 22

[217] Daniela Massiceti, Puneet K. Dokania, N. Siddharth, and Philip H.S. Torr. Visual dialogue without vision or dialogue. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop on Correcting and Critiquing Trends in Machine Learning*, 2018. 19, 22, 110, 111

[218] Daniela Massiceti, Stephen L. Hicks, and Joram J. van Rheede. Stereosonic vision: Exploring visual-to-auditory sensory substitution mappings in an immersive virtual reality navigation paradigm. *PLOS ONE*, 13(7):e0199389, 2018. 22

[219] Daniela Massiceti, N. Siddharth, Puneet K. Dokania, and Philip H.S. Torr. FlipDial: A generative model for two-way visual dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 22, 111, 112, 116

[220] Daniela Massiceti, Viveka Kulharia, Puneet K. Dokania, N. Siddharth, and Philip H.S. Torr. On the evaluation of visual dialogue. [under review], 2019. 22

[221] Robert J. Matheis, Maria T. Schultheis, Lana A. Tiersky, John DeLuca, Scott R. Millis, and Albert Rizzo. Is learning and memory different in a virtual environment? *The Clinical Neuropsychologist*, 21 (1):146–161, 2007. 17, 74

[222] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989. 137, 144

[223] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007. 24, 27

[224] Peter B.L. Meijer. An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering*, 39(2):112–121, 1992. 8, 9, 10, 11, 16, 46, 73, 149

[225] Lotfi B. Merabet, Jascha D. Swisher, Stephanie A. McMains, Mark A. Halko, Amir Amedi, Alvaro Pascual-Leone, and David C. Somers. Combined activation and deactivation of visual cortex during tactile sensory processing. *Journal of Neurophysiology*, 97(2):1633–1641, 2007. 16, 149

[226] Microsoft Corporation. Seeing AI, 2016. URL www.microsoft.com/seeing-ai. Accessed on 12 April 2019. 2, 8, 11, 16, 153

[227] Microsoft Corporation. Microsoft Soundscape, 2018. URL www.microsoft.com/en-us/research/product/soundscape. Accessed on 10 April 2019. 8, 11, 16, 147

[228] Péter Mihajlik, M. Guttermuth, K. Seres, and Peter Tatai. DSP-based ultrasonic navigation aid for the blind. In *IEEE Instrumentation and Measurement Technology Conference (IMTC)*, 2001. 48, 73, 148

[229] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. 88, 92, 102

[230] Darwin Minassian and Angela Reidy. Future sight loss in the decade 2010 to 2020: An epidemiological and economic model, 2009. URL www.rnib.org.uk/sites/default/files/FSUK_Report_2_0.doc. Royal National Institute of Blind People. 1

[231] MIPsoft. BlindSquare, 2017. URL www.blindsquare.com. Accessed on 18 April 2019. 8, 147

[232] Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 15

[233] Ishan Misra, Ross Girshick, Rob Fergus, Martial Hebert, Abhinav Gupta, and Laurens van der Maaten. Learning by asking questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 18, 136, 144

[234] Betty J. Mohler, William B. Thompson, Sarah H. Creem-Regehr, Herbert L. Pick, and William H. Warren. Visual flow influences gait transition speed and preferred walking speed. *Experimental Brain Research*, 181(2):221–228, 2007. 65, 74

[235] Marian Morris and Paula Smith. Educational provision for blind and partially sighted children and young people in Britain: 2007, 2008. National Foundation for Educational Research. 6

[236] Cecily Morrison, Nicolas Villar, Anja Thieme, Zahra Ashktorab, Eloise Taysom, Oscar Salandin, Daniel Cletheroe, Greg Saul, Alan F. Blackwell, Darren Edge, Martin Grayson, and Haiyan Zhang. Torino: A tangible programming language inclusive of children with visual disabilities. *Human–Computer Interaction*, 0(0):1–49, 2018. 6

[237] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016. 18

[238] R. Mottaghi, X. Chen, X. Liu, N. G. Cho, S. W. Lee, S. Fidler, R. Urtasun, and A.L. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 10, 13, 14

[239] Andrew Nash. National population projections: 2014-based statistical bulletin, 2015. URL www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/bulletins/nationalpopulationprojections/2015-10-29. Office of National Statistics. 1

[240] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011. 2, 11, 13, 16, 48, 78, 150

[241] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 11, 13

[242] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 18

[243] Claire L. Nollett, Nathan Bray, Catey Bunce, Robin J. Casten, Rhiannon T. Edwards, Mark T. Hegel, Sarah Janikoun, Sandra E. Jumbe, Barbara Ryan, Julia Shearn, et al. Depression in Visual Impairment Trial (DEPVIT): A randomized clinical trial of depression treatments in people with low vision. *Investigative Ophthalmology and Visual Science*, 57(10):4247–4254, 2016. 7

[244] Scott D. Novich and David M. Eagleman. A vibrotactile sensory substitution device for the deaf and profoundly hearing impaired. In *IEEE Haptics Symposium (HAPTICS)*, 2014. 8, 9, 16, 46, 73, 149

[245] Sebastian Nowozin. Optimal decisions from probabilistic models: The intersection-over-union case. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 15, 26, 33

[246] NuEyes. NuEyes Pro, 2019. URL www.nueyes.com/products/nueyes-pro. Accessed on 20 April 2019. 5, 8, 9, 150, 151

[247] Donata Oertel and Allison J. Doupe. The auditory central nervous system. In E.R. Kandel, J.H. Schwartz, T.M. Jessell, S.A. Siegelbaum, and A.J. Hudspeth, editors, *Principles of Neural Science*, pages 682–711. McGraw-Hill, New York, 5 edition, 2013. 47

[248] Patrick Emeka Okonji. Use of computer assistive technologies by older people with sight impairment: Perceived state of access and considerations for adoption. *British Journal of Visual Impairment*, 36(2): 128–142, 2018. 145

[249] OrCam. OrCam MyEye 2, 2019. URL www.orcam.com/gb/myeye2. Accessed on 18 April 2019. 2, 8, 11, 153

[250] OrCam. OrCam MyReader 2, 2019. URL www.orcam.com/en/myreader2. Accessed on 18 April 2019. 5, 8, 9, 10, 151

[251] World Health Organisation. 11$^{\text{th}}$ Revision of the International Classification of Diseases (ICD-11), 2018. URL www.who.int/classifications/icd/en/. 3

[252] OXSIGHT Ltd. OXSIGHT Crystal, 2019. URL www.oxsight.co.uk. Accessed on 15 April 2019. 8, 9, 10, 11, 14, 150

[253] Mustafa Ozuysal, Vincent Lepetit, and Pascal Fua. Pose estimation for category specific multiview object localization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 14

[254] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly- and semi-supervised learning of a DCNN for semantic image segmentation. In *International Conference on Computer Vision (ICCV)*, 2015. 24, 25, 26, 28, 32, 35, 36

[255] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002. 19, 20, 122, 140

[256] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. 136, 144

[257] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning (ICML)*, 2013. 88

[258] Achille Pasqualotto, Mary Jane Spiller, Ashok S. Jansari, and Michael J. Proulx. Visual experience facilitates allocentric spatial representation. *Behavioural Brain Research*, 236:175–179, 2013. 77

[259] Romedi Passini and Guyltne Proulx. Wayfinding without vision: An experiment with congenitally totally blind people. *Environment and Behaviour*, 20(2):227–252, 1988. 4, 45

[260] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. In *International Conference on Learning Representations (ICLR)*, 2014. 26, 39

[261] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *International Conference on Computer Vision (ICCV)*, 2015. 24, 26, 36, 39

[262] Aftab E. Patla. Understanding the roles of vision in the control of human locomotion. *Gait and Posture*, 5(1):54–69, 1997. 4, 45

[263] Sue Pavey, Andrew Dodgson, Graeme Douglas, and Ben Clements. Travel, transport, and mobility of people who are blind and partially sighted in the UK, 2009. URL www.rnib.org.uk/sites/default/files/2009_4_Travel_Transport_Mobility.doc. Royal National Institute of Blind People. 5

[264] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 88, 117, 120, 123

[265] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018. 123

[266] Lynne Pezzullo, Jared Streatfeild, Philippa Simkiss, and Darren Shickle. The economic impact of sight loss and blindness in the UK adult population. *BMC Health Services Research*, 18(1):63, 2018. 1

[267] Ngoc-Quan Pham, German Kruszewski, and Gemma Boleda. Convolutional neural network language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 18, 83, 88

[268] Lorenzo Picinali, Amandine Afonso, Michel Denis, and Brian F.G. Katz. Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge. *International Journal of Human-Computer Studies*, 72(4):393–407, 2014. 73

[269] Pedro O. Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 24, 26, 36

[270] Martin Pinquart and Jens P. Pfeiffer. Psychological well-being in visually impaired and unimpaired individuals: A meta-analysis. *British Journal of Visual Impairment*, 29(1):27–45, 2011. 7

[271] Vivek Pradeep, Gerard Medioni, and James Weiland. Piecewise planar modeling for step detection using stereo vision. In *European Conference on Computer Vision (ECCV) Workshop on Computer Vision Applications for the Visually Impaired*, 2008. 149, 150

[272] Vivek Pradeep, Gerard Medioni, and James Weiland. Robot vision for the visually impaired. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2010. 149, 150

[273] Pratsam. Pratsam Reader Web, 2019. URL www.pratsam.com/pratsam-reader-web-product. Accessed on 22 April 2019. 5, 151

[274] Victor Adrian Prisacariu, Olaf Kähler, Stuart Golodetz, Michael Sapienza, Tommaso Cavallari, Philip H.S. Torr, and David W. Murray. InfiniTAM v3: A framework for large-scale 3D reconstruction with loop closure. *CoRR*, abs/1708.00783, 2017. URL http://arxiv.org/abs/1708.00783. 11, 13, 78, 150

[275] PyTorch. PyTorch, 2017. Accessed on 4 November 2017. 93

[276] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum PointNets for 3D object detection from RGB-D data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 13

[277] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. Augmented feedback in semantic segmentation under image level supervision. In *European Conference on Computer Vision (ECCV)*, 2016. 24, 26, 36, 39

[278] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 18

[279] Giraffe Reader. Giraffe Reader, 2014. URL www.giraffe-reader.com. Accessed on 22 April 2019. 5, 8, 9, 151

[280] D. Raj Reddy. Speech understanding systems: A summary of results of the five-year research effort, 1977. 18

[281] Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 10, 13, 14, 78, 79

[282] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You Only Look Once: Unified, real-time object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 78

[283] Rebecca J. Reed-Jones, Guillermina R. Solis, Katherine A. Lawson, Amanda M. Loya, Donna Cude-Islas, and Candyce S. Berger. Vision and falls: A multidisciplinary review of the contributions of visual impairment to falls among older adults. *Maturitas*, 75(1):22–28, 2013. 1, 4, 45

[284] Lior Reich, Shachar Maidenbaum, and Amir Amedi. The brain as a flexible task machine: Implications for visual rehabilitation using noninvasive vs. invasive approaches. *Current Opinion in Neurology*, 25(1): 86–95, 2012. 16, 46

[285] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 13, 14

[286] Jose Rivera-Rubio, Saad Idrees, Ioannis Alexiou, Lucas Hadjilucas, and Anil A. Bharath. A dataset for hand-held object recognition. In *International Conference on Image Processing (ICIP)*, 2014. 142

[287] RNFB Reader. KNRB Reader, 2018. URL www.knfbreader.com. Accessed on 18 April 2019. 6, 8, 9, 10, 151

[288] Brigitte Röder, Wolfgang Teder-Sälejärvi, Anette Sterr, Frank Rösler, Steven A. Hillyard, and Helen J. Neville. Improved auditory spatial tuning in blind humans. *Nature*, 400(6740):162–166, 1999. 77, 139

[289] Uta R. Roentgen, Gert Jan Gelderblom, Mathijs Soede, and Luc P. de Witte. Inventory of electronic mobility aids for persons with visual impairments: A literature review. *Journal of Visual Impairment and Blindness*, 102(11):702–723, 2008. 9, 46, 149

[290] Bernardino Romera-Paredes and Philip H.S. Torr. Recurrent instance segmentation. In *European Conference on Computer Vision (ECCV)*, 2016. 14

[291] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew C. Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru. AVA-ActiveSpeaker: An audio-visual dataset for active speaker detection. *CoRR*, abs/1901.01342, 2019. URL http://arxiv.org/abs/1901.01342. 142

[292] Royal National Institute of Blind People. RNIB Library, 2019. URL www.rniblibrary.com. Accessed on 22 April 2019. 151

[293] Royal National Institute of Blind People. Everyday living solutions 2019, 2019. URL shop.rnib.org.uk/media/instructions/2019_Everyday_living_solutions_product_catalogue_RE190115%20(1).pdf. Accessed on 22 April 2019. 9, 152

[294] Royal National Institute of Blind People. Technology resource hub: The latest facts, tips and guides, 2019. URL www.rnib.org.uk/practical-help/technology/resource-hub. Accessed on 12 April 2019. 7

[295] Roy A. Ruddle, Ekaterina Volkova, and Heinrich H. Bülthoff. Learning to walk in virtual reality. *ACM Transactions on Applied Perception*, 10(2):11:1–11:17, 2013. 73

[296] Jaime Sánchez, Mauricio Sáenz, Alvaro Pascual-Leone, and Lotfi Merabet. Navigation for the blind through audio-based virtual environments. In *Extended Abstracts on Human Factors in Computing Systems (CHI EA)*, 2010. 49

[297] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 79, 137

[298] Hélène Sauzéon, Prashant Arvind Pala, Florian Larrue, Gregory Wallet, Marie Déjos, Xia Zheng, Pascal Guitton, and Bernard N'kaoua. The use of virtual reality for episodic memory assessment: Effects of active navigation. *Experimental Psychology*, 59(2):99–108, 2012. 17, 74

[299] Bo N. Schenkman and Gunnar Jansson. The detection and localization of objects by the blind with the aid of long-cane tapping sounds. *Human Factors*, 28(5):607–618, 1986. 7

[300] Victor R. Schinazi, Tyler Thrash, and Daniel-Robert Chebat. Spatial navigation by congenitally blind individuals. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(1):37–58, 2016. 45, 77

[301] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. Visual reference resolution using attention memory for visual dialog. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 113

[302] Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799, 2017. URL http://arxiv.org/abs/1706.09799. 19, 20, 123, 140

[303] Shinano Kenshi Co., Ltd. PlexTalk Linio Pocket, 2011. URL www.plextalk.eu/en/top/products/liniopocket. Accessed on 22 April 2019. 6, 151

[304] Art Beyond Sight. New York Beyond Sight, 2007. URL www.nybeyondsight.org. Accessed on 11 April 2019. 147

[305] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 2, 10, 13, 14, 92, 102, 117, 144

[306] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR)*, 2014. 36, 41

[307] John Slade and Rose Edwards. My Voice 2015: The views and experiences of blind and partially sighted people in the UK, 2015. URL www.rnib.org.uk/knowledge-and-research-hub-research-reports-general-research/my-voice. Royal National Institute of Blind People. 1, 2, 4, 5, 6, 7, 9, 151

[308] Audrey J. Smith, William de L'Aune, and Duane R. Geruschat. Low vision mobility problems: Perceptions of O&M specialists and persons with low vision. *Journal of Visual Impairment and Blindness*, 86(1):58–62, 1992. 4, 45

[309] Jascha Sohl-Dickstein, Santani Teng, Benjamin M. Gaub, Chris C. Rodgers, Crystal Li, Michael R. DeWeese, and Nicol S. Harper. A device for human ultrasonic echolocation. *IEEE Transactions on Biomedical Engineering*, 62(6):1526–1534, 2015. 48, 73, 148

[310] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 18, 83, 85

[311] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. URL http://arxiv.org/abs/1212.0402. 142

[312] Sound Foresight Technology Ltd. UltraCane, 2005. URL www.ultracane.com. Accessed on 9 April 2019. 8, 9, 16, 46, 148, 149

[313] Joan Stelmack. Quality of life of low-vision patients and outcomes of low-vision rehabilitation. *Optometry and Vision Science*, 78(5):335–342, 2001. 1, 45

[314] Gretchen A. Stevens, Richard A. White, Seth R. Flaxman, Holly Price, Jost B. Jonas, Jill Keeffe, Janet Leasher, Kovin Naidoo, Konrad Pesudovs, Serge Resnikoff, et al. Global prevalence of vision impairment and blindness: Magnitude and temporal trends, 1990–2010. *Ophthalmology*, 120(12):2377–2384, 2013. 1, 45

[315] Humanware Store. Explore 8 handheld electronic magnifier, 2018. URL store.humanware.com/heu/explore-8-handheld-electronic-magnifier.html. Accessed on 22 April 2019. 6, 150

[316] Humanware Store. Victor Reader Trek, 2018. URL store.humanware.com/heu/victor-reader-trek-talking-book-player-gps.html. Accessed on 11 April 2019. 8, 9, 147

[317] Humanware Store. Victor Reader Stream, 2018. URL store.humanware.com/hus/victor-reader-stream-new-generation.html. Accessed on 22 April 2019. 6, 151

[318] Michael Supa, Milton Cotzin, and Karl M. Dallenbach. "Facial vision": The perception of obstacles by the blind. *The American Journal of Psychology*, 57(2):133–183, 1944. 5, 45, 77, 139

[319] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 2, 10, 18

[320] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence*, 2017. 2, 10, 13, 14, 144

[321] Santani Teng, Amrita Puri, and David Whitney. Ultrafine spatial acuity of blind expert human echolocators. *Experimental Brain Research*, 216(4):483–488, 2012. 47, 54

[322] Lore Thaler and Melvyn A. Goodale. Echolocation in humans: An overview. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(6):382–393, 2016. 17, 47, 54

[323] Sebastian Thrun. Exploration in active learning. In Michael A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 381–384. MIT Press, Cambridge, MA, 1998. 136

[324] Sebastian Thrun and Tom M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems*, 15 (1-2):25–46, 1995. 136

[325] Mhairi Thurston, Allen Thurston, and John McLeod. Socio-emotional effects of the transition from sight to blindness. *British Journal of Visual Impairment*, 28(2):90–112, 2010. 7

[326] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(11):1958–1970, 2008. 142

[327] Alexander Toshev and Christian Szegedy. DeepPose: Human pose estimation via deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 10, 13

[328] Anmol Tukrel. iDentifi, 2016. URL www.getidentifi.com. Accessed on 15 April 2019. 2, 8, 11, 16, 153

[329] Kathleen A. Turano, Gary S. Rubin, and Harry A. Quigley. Mobility performance in glaucoma. *Investigative Ophthalmology and Visual Science*, 40(12):2803–2809, 1999. 45

[330] Kathleen A. Turano, Aimee T. Broman, Karen Bandeen-Roche, Beatriz Munoz, Gary S. Rubin, and Sheila K. West. Association of visual field loss and mobility performance in older adults: Salisbury eye evaluation study. *Optometry and Vision Science*, 81(5):298–307, 2004. 1, 4, 45

[331] Jasper R.R. Uijlings, Koen E.A. van de Sande, Theo Gevers, and Arnold W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013. 36, 41

[332] World Blind Union. Status of the white cane, 2014. URL www.worldblindunion.org/English/resources/Documents/Status%20of%20the%20White%20Cane.doc. World Blind Union. 7

[333] Martin Usoh, Kevin Arthur, Mary C. Whitton, Rui Bastos, Anthony Steed, Mel Slater, and Frederick P. Brooks Jr. Walking > walking-in-place > flying, in virtual environments. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1999. 49, 73

[334] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 93, 104

[335] Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with PixelCNN decoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 89

[336] Joram J. van Rheede, Iain R. Wilson, Rose I. Qian, Susan M. Downes, Christopher Kennard, and Stephen L. Hicks. Improving mobility performance in low vision with a distance-based representation of the visual scene: A distance-based representation for low vision. *Investigative Ophthalmology and Visual Science*, 56(8):4802–4809, 2015. 50, 52, 62, 141

[337] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 10

[338] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 122

[339] Claude Veraart and Marie-Chantal Wanet-Defalque. Representation of locomotor space by the blind. *Perception and Psychophysics*, 42(2):132–139, 1987. 77

[340] Tiziana Vercillo, Alessia Tonelli, and Monica Gori. Early visual deprivation prompts the use of body-centered frames of reference for auditory localization. *Cognition*, 170:263–269, 2017. 78

[341] Vibhav Vineet, Ondrej Miksik, Morten Lidegaard, Matthias Niener, Stuart Golodetz, Victor A. Prisacariu, Olaf Khler, David W. Murray, Shahram Izadi, and Patrick Prez. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In *International Conference on Robotics and Automation (ICRA)*, 2015. 13, 16, 48, 78

[342] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 10

[343] Patrice Voss, Maryse Lassonde, Frédéric Gougoux, Madeleine Fortin, Jean-Paul Guillemot, and Franco Lepore. Early-and late-onset blind individuals show supra-normal auditory abilities in far-space. *Current Biology*, 14(19):1734–1738, 2004. 16, 77, 139, 149

[344] Bruce N. Walker and Jeffrey Lindsay. Navigation performance in a virtual environment with bonephones. In *International Conference on Auditory Display (ICAD)*, 2005. 49, 73, 79

[345] Bruce N. Walker and Jeffrey Lindsay. Using virtual environments to prototype auditory navigation displays. *Assistive Technology*, 17(1):72–81, 2005. 49, 73

[346] Bruce N. Walker and Jeffrey Lindsay. Navigation performance with a virtual auditory display: Effects of beacon sound, capture radius, and practice. *Human Factors*, 48(2):265–278, 2006. 49, 73

[347] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H.S. Torr. Fast online object tracking and segmentation: A unifying approach. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 13, 14, 78

[348] Dean A. Waters and Husam H. Abulula. Using bat-modelled sonar as a navigational tool in virtual environments. *International Journal of Human-Computer Studies*, 65(10):873–886, 2007. 48, 73, 148

[349] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 26, 34, 36, 37

[350] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. STC: A simple to complex framework for weakly-supervised semantic segmentation. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(11):2314–2320, 2017. 24, 25, 26, 35, 36

[351] Joseph Weizenbaum. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966. 111

[352] Thomas Whelan, Renato F. Salas-Moreno, Ben Glocker, Andrew J. Davison, and Stefan Leutenegger. ElasticFusion: Real-time dense SLAM and light source estimation. *The International Journal of Robotics Research*, 35(14):1697–1716, 2016. 11, 13, 78, 150

[353] Adam Whitaker. Special educational needs in England: January 2015, 2015. URL www.gov.uk/government/statistics/special-educational-needs-in-england-january-2015. Department for Education. 6

[354] Gareth R. White, Geraldine Fitzpatrick, and Graham McAllister. Toward accessible 3D virtual environments for the blind and visually impaired. In *International Conference on Digital Interactive Media in Entertainment and Arts (DIMEA)*, 2008. 79

[355] Jaroslaw Wiazowski. Can Braille be revived? A possible impact of high-end Braille and mainstream technology on the revival of tactile literacy medium. *Assistive Technology*, 26(4):227–230, 2014. 9, 151

[356] Wicab, Inc. BrainPort Vision Pro, 2007. URL www.wicab.com/brainport-vision-pro. Accessed on 11 April 2019. 149

[357] William R. Wiener, Richard L. Welsh, and Bruce B. Blasch. *Foundations of orientation and mobility*, volume 1. American Foundation for the Blind, New York, NY, 3 edition, 2010. 4, 7, 45

[358] John Wieting and Douwe Kiela. No training required: Exploring random encoders for sentence classification. In *International Conference on Learning Representations (ICLR)*, 2019. 117

[359] Jeff Wilson, Bruce N. Walker, Jeffrey Lindsay, Craig Cambias, and Frank Dellaert. SWAN: System for Wearable Audio Mavigation. In *International Symposium on Wearable Computers (ISWC)*, 2007. 8, 49, 73, 147, 149

[360] Terry Winograd. Procedures as a representation for data in a computer program for understanding natural language. Technical report, Massachusetts Institute of Technology, Cambridge, MA, 1971. 111

[361] Philip Worchel, Jack Mauney, and John G. Andrew. The perception of obstacles by the blind. *Journal of Experimental Psychology*, 40(6):746–751, 1950. 5, 45, 77, 139

[362] World Health Organisation. Fact Sheet 282: Visual impairment and blindness, 2017. URL www.who.int/mediacentre/factsheets/fs282/en. 1, 45, 145

[363] C.F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1): 95–103, 1983. 28

[364] Qi Wu, Peng Wang, Chunhua Shen, Ian D. Reid, and Anton van den Hengel. Are you talking to me? Reasoned visual dialog generation through adversarial learning. *CoRR*, abs/1711.07613, 2017. URL http://arxiv.org/abs/1711.07613. 18, 19, 116, 117, 118, 131

[365] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3D environment. *CoRR*, abs/1801.02209, 2018. URL http://arxiv.org/abs/1801.02209. 143

[366] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond Pascal: A benchmark for 3D object detection in the wild. In *Winter Conference on Applications of Computer Vision (WACV)*, 2014. 14

[367] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In *Royal Statistical Society (RSS)*, 2018. 14

[368] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 137

[369] Jia Xu, Alexander G. Schwing, and Raquel Urtasun. Learning to segment under various forms of weak supervision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 24, 26

[370] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Richard Zemel Ruslan Salakhutdinov, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015. 2, 10

[371] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 41

[372] Ming-Der Yang, Chih-Fan Chao, Kai-Siang Huang, Liang-You Lu, and Yi-Ping Chen. Image-based 3D scene reconstruction and exploration in augmented reality. *Automation in Construction*, 33:48–60, 2013. 13, 16, 48

[373] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 18, 19

[374] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron C. Courville. Describing videos by exploiting temporal structure. In *International Conference on Computer Vision (ICCV)*, 2015. 2, 10, 13

[375] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, (99):1–13, 2018. 18, 141

[376] Dan Yuan and Roberto Manduchi. Dynamic environment exploration using a virtual white cane. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 149

[377] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision (IJCV)*, 126(10):1084–1102, 2018. 25, 26, 29, 30, 34, 36, 40, 41

[378] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 19, 141

[379] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 79, 137

[380] Zhengyou Zhang. Microsoft Kinect sensor and its effect. *IEEE Multimedia*, 19(2):4–10, 2012. 78

[381] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017. 84, 88

[382] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H.S. Torr. Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision (ICCV)*, 2015. 24

[383] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *CoRR*, abs/1512.02167, 2015. URL http://arxiv.org/abs/1512.02167. 19

[384] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 36, 41

[385] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded question answering in images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 18, 141, 142