

End-to-End Learning, and Audio-Visual Human-Centric Video Understanding

Andrew Brown



Linacre College

University of Oxford

Supervisor: Professor Andrew Zisserman

A thesis presented for the degree of

Doctor of Philosophy

Michaelmas 2022

Acknowledgements

I am very grateful for the guidance and support of many important people throughout this journey.

First, to my supervisor Andrew Zisserman. Your creative, exciting approach to problem-solving has been inspiring, and has shaped me into the researcher that I am today. It has been a privilege to work with you for the last five years. Thank you for your guidance and for seeing the potential in a Matlab-wielding fledgling all of those years ago.

Thank you to the collaborators that I have been lucky enough to work with. Weidi Xie, who taught me the value and importance of solving hard problems. Thank you for guiding me through my first research paper, for providing clarity with your vast knowledge of the field, and for answering my hundreds of questions without hesitation, even as a new and tired parent. To Andrea Vedaldi, for being so generous with his insights and time. Thank you for showing me the importance of detail, and for the iconic COVID-19 memories of having equations explained to me on virtual whiteboards written on from Venetian public transport. To Arsha Nagrani, who has been an invaluable mentor to me throughout my time at VGG and guided me up the many steep learning curves. To Ernesto Coto, thank you for your unwavering patience in helping me with the many engineering mysteries encountered on this journey, such as compiling Caffe. To all my collaborators: Vicky Kalogeiton, Joon Son Chung, Max Bain, Jaesung Huh, Samuel Albanie, Yang Liu, Bruno Korbar, Cheng-Yang Fu, Omkar Parkhi, and Tamara Berg. Thank you sharing your ideas and time with me, and for making research meetings so exciting and inspiring.

To all the members of the VGG family: Tengda Han, Chuhan Zhang, Yuki Asano, Olivia Wiles, Sagar Vaze, Shangzhe Wu, Mandela Patrick, Gul Varol, Lili Momeni, Rhydian Windsor, Daffy Afouras, Tomas Jakab, Xu Ji, Christian Rupprecht, Erika Lu, and João Henriques. Thank you for creating such a welcoming and supportive environment, for the lunches at St. Anne's, and the good evenings at the Royal Oak. I feel very privileged to have learnt from all of you, and I am excited to see you all grow to be giants in the research field in the coming years.

To my family (Jo, Phil, Will, Katie, and Candice), thank you for supporting me when I raised the prospect of spending yet another few years in education. To my friends in London who were still there

for me after I disappeared off the face of the earth for some time. You all mean the world to me and I am so lucky to have you all in my life.

Finally, to Maya. Thank you for your endless encouragement, support, and understanding. Thank you for believing in my dream before I did, and for subsequently telling a young and lost undergraduate what was coming if he instead decided to go down a career-path in the City. Thank you for pushing me to follow this dream, and for encouraging me every time I fell along the way. I could not have done this without her, and I am so excited for our next journey.

This is dedicated to her.

Abstract

The field of machine learning has seen tremendous progress in the last decade, largely due to the advent of deep neural networks. When trained on large-scale labelled datasets, these machine learning algorithms can learn powerful semantic representations directly from the input data, *end-to-end*. End-to-end learning requires the availability of three core components: useful input data, target outputs, and an objective function for measuring how well the model's predictions match the target outputs. In this thesis, we explore and overcome a series of challenges as related to assembling these three components in the sufficient format and scale for end-to-end learning.

The first key idea presented in this thesis is to learn representations by enabling end-to-end learning for tasks where such challenges exist. We first explore whether better representations can be learnt for the image retrieval task by directly optimising the evaluation metric, Average Precision. This is notoriously challenging task, because such rank-based metrics are non-differentiable. We introduce a simple objective function that optimises a smoothed approximation of Average Precision, termed *Smooth-AP*, and demonstrate the benefits of training end-to-end over prior approaches. Secondly, we explore whether a representation can be learnt end-to-end for the task of image editing, where target data does not exist in sufficient scale. We propose a self-supervised approach that simulates target data by augmenting off-the-shelf image data, giving remarkable benefits over prior work.

The second idea presented in this thesis is focused on how to use the rich multi-modal signals that are essential for human perceptual systems as input data for deep neural networks. More specifically, we explore the use of audio-visual input data for the human-centric video understanding task. Here, we first explore if highly optimised speaker verification representations can transfer to the domain of movies where humans intentionally disguise their voice. We do this by collecting an audio-visual dataset of humans speaking in movies. Second, given strong identity discriminating representations, we present two methods that harness the complementarity and redundancy between multi-modal signals in order to build robust perceptual systems for determining who is present in a scene. These methods include an automated pipeline for labelling people in unlabelled video archives, and an approach for clustering people by identity in videos.

Table of Contents

1	Introduction and Background	4
1.1	Motivation	6
1.2	Key Ideas and Contributions	11
1.2.1	Enabling End-to-End Learning	11
1.2.2	Audio-Visual Human-Centric Video Understanding.	13
1.3	Thesis Outline	15
1.3.1	Publications not included:	16
I	Enabling End-to-End Learning	17
2	Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval	18
2.1	Introduction	19
2.2	Related Work	20
2.3	Background	22
2.4	Approximating Average Precision (AP)	24
2.5	Experimental Setup	28
2.6	Results	31
2.7	Conclusions	36
3	End-to-End Visual Editing with a Generatively Pre-Trained Artist	38
3.1	Introduction	39
3.2	Related Work	41
3.3	Method	42
3.4	Experiments	49
3.5	Conclusions, Limitations and Future work	55

II	Audio-Visual Human-Centric Video Understanding	57
4	Playing a Part: Speaker Verification at the Movies	58
4.1	Introduction	59
4.2	Related Work	61
4.3	Cross-Domain Data	62
4.4	Dataset Collection Pipeline	63
4.5	Experiments	65
4.6	Results	69
4.7	Conclusion	70
5	Automated Video Labelling: Identifying Faces by Corroborative Evidence	71
5.1	Introduction	72
5.2	Related Work	75
5.3	Stage 1 – Using image-search engines as sole/corroborating evidence	76
5.4	Stage 2 – Boosting the Number of Tags	79
5.5	Datasets, Evaluations and Implementation	80
5.6	Results	83
5.7	Conclusions	87
6	Face, Body, Voice: Video Person-Clustering with Multiple Modalities	88
6.1	Introduction	89
6.2	Related Work	91
6.3	Method	94
6.4	Video Person-Clustering Dataset	99
6.5	Experiments	102
6.6	Conclusions	107
7	Discussion	109
7.1	Achievements and Impact	109
7.2	What Comes Next?	113

7.3 Conclusion	115
Appendix A Statements of Authorship	148

1 | Introduction and Background

The focus of this thesis is the field of artificial machine learning; where algorithms are designed that learn patterns from data, and can in turn make inferences about new data. In 2012, a scientific breakthrough was made when researchers in Toronto released AlexNet [Krizhevsky et al., 2012]; an algorithm that was able to classify images at a significantly higher accuracy than previous approaches. The key to this advance was the successful training on a large image dataset of a deep neural network (DNN); a machine learning algorithm which is able to learn complex, semantic representations directly from input data, *end-to-end*. Although deep neural networks had been successfully trained before for object [LeCun et al., 1989], and speech recognition [Hochreiter & Schmidhuber, 1997], AlexNet harnessed the increased speed of Graphics Processing Units (GPUs), making it computationally feasible to train on large datasets. Tremendous progress followed, with deep neural networks demonstrating expert capabilities on complex tasks ranging from natural language understanding [Devlin et al., 2019b; Radford et al., 2019], to image generation [Ramesh et al., 2021a].

The recipe for training deep neural networks is seemingly a simple one. Paired with modern deep learning frameworks for implementing the iterative gradient-based learning process, training a Deep Neural Network end-to-end requires just (I) useful **input data** from the world that is necessary for the task at hand, (II) **target outputs**, and (III) an **objective function** to measure how well the model's predictions match the target outputs. For learning powerful representations, it is essential that these three components are assembled in the sufficient scale, and correct format. However, doing so can raise significant challenges. For example:

Input Data: Most deep neural networks learn from input data from just a single modality, such as images, or audio. In contrast, as humans, our understanding of the world is inherently multi-modal. For example, when deciding whether to put on a coat as we leave the house, not only do we see the cloudless blue sky above (which when taken alone would be ambiguous and could indicate warm weather), but crucially we hear the sound of a howling wind, and feel the chill against our skin. For solving real-world, complex tasks, we must therefore also be training neural networks with the rich *multi-modal* sensory information in the world.

Target Outputs: Large and varied datasets exist where target outputs, or *labels*, have been collected via manual annotation. This includes image classification [J. Deng et al., 2009], object detection [Everingham, Gool, et al., 2009; T.-Y. Lin et al., 2014], and image-text tasks [Schuhmann et al., 2021]. However, this manual annotation process is expensive and slow; a problem that compounds with the complexity of the annotation task. For many tasks, collecting manual annotation at the scale required for learning useful representations with data-hungry neural networks is simply unfeasible. We must hence find another option for assembling target outputs, that does not rely on manual annotation.

Objective Functions: The objective function provides the training signal back to the neural network; pointing out to the model what it can learn to improve its predictions. The optimal choice of function is therefore one that is coupled to the task you are training the neural network for, and is an essential component for encouraging the correct learnt representation. It is crucial that we explore whether our standard objective functions are suited for the tasks we are training them for.

In this thesis, we explore and overcome a series of challenges arising from these three components:

Without well-suited objective functions or sufficient target outputs, it is not possible to train deep neural networks end-to-end for given tasks, or to learn effective representations as a result. Our first research contribution is **Enabling End-to-End Learning** in settings where these challenges exist. Specifically, we explore the end-to-end optimisation of non-differentiable evaluation metrics, and the self-supervised synthesis of target data for tasks where manual annotation is unfeasible.

Without rich multi-modal input data, it is not possible to learn human-level representations of the real world. Our second research contribution is to explore representations from **Audio-Visual** data. We focus our research here on the task of **Human-Centric Video Understanding**; where the goal is to understand human-centred stories and interactions in videos.

Next we explain the motivations for exploring these research directions in Section 1.1. We present the key ideas and contributions proposed in this thesis in Section 1.2, and we overview the thesis outline in Section 1.3.

1.1 Motivation

Objective Functions for Retrieval. Here, we motivate our research on optimising well-suited objective functions for the task of image retrieval in Chapter 2. We first present the motivation for studying retrieval in computer vision, followed by a discussion on objective functions.

When interacting with and experiencing the outside world, what happens when we come across something that we have never seen before, such as a rare and exotic animal, a flamingo. We may know and be familiar with similar animals, but not this one. If we could only conceive of the animals that we already know, this situation would leave us dumbfounded. Luckily though, as humans, we are gifted with the notion of similarity, and this is key to how we learn. For example, when exposed to new objects that they have not seen before, children demonstrate their learnt notion of similarity to determine for example that the flamingo is similar to a crow (something that they have seen before, and that is from the same category), whereas it is less similar to a bat (something they also know, from a different category) [Gelman & Markman, 1986]. Interestingly, children do this by learning rich abstracted notions of object likeness, where, rather than relying on simple perceptual similarities, they may use abstracted representations of object shape [Landau et al., 1988].

In Computer vision, the task of learning similarities between an open-world of unseen objects and classes is explored through the field of *retrieval*. Due to these real-world applications, this field has attracted much interest, with researchers exploring how models can learn similarities between instances using just a single modality *e.g.* vision [Sivic & Zisserman, 2003; Jégou et al., 2008], or even exploring similarity learning *cross-modality*, *e.g.* from audio to images [Arandjelović & Zisserman, 2018], or text to video [Otani et al., 2016].

A key aspect when training and comparing neural networks is quantifying how well they can perform a task. In retrieval, the performance of a model is measured by scoring how well the model can rank a retrieval set (a large database of instances) in terms of similarity to a query instance with the hope that similar instances from the same category are ranked highly. This scoring is done using evaluation metrics such as Average Precision. When training neural networks end-to-end, the default choice is to use such well-defined metrics as the objective function. However, seeing as ranking metrics are

non-differentiable, and so cannot be directly optimised with gradient-based methods, much attention has instead been placed on designing differentiable objective functions for retrieval that will give good rankings once optimised.

For example, the contrastive [Chopra et al., 2005], and triplet [K. Q. Weinberger & Saul, 2009] objectives consider pairs or triplets of elements, respectively, and encourage a good ranking by forcing all positive instances to be *close*, as measured in the deep neural network’s learnt representation, while separating negatives by a fixed margin. Similarly, the InfoNCE objective [Oord et al., 2018] maximises the mutual information between similar instances relative to dissimilar ones.

In Chapter 2 we explore whether these objective functions are well-suited replacements for metrics such as Average Precision. Here, we instead focus on the end-to-end optimisation of the ranking metric. Relevant to our work in this thesis is therefore prior work focused on directly optimising differentiable versions of an evaluation metric. Within the retrieval field, researchers have attempted to directly optimise the Average Precision metric by designing objectives that are smooth upper bounds [McFee & Lanckriet, 2010; Mohapatra et al., 2018] to the metric, or by implementing differentiable versions of histogram binning [Cakir et al., 2019a; K. He, Cakir, et al., 2018]. Similarly, outside of retrieval, examples include directly optimising an approximation to Average Precision for object detection [H. Chen et al., 2019], or optimising a smooth version of the top-k metric for classification [Berrada et al., 2018].

Learning using self-supervision. Here, we motivate our research in Chapter 3 on using self-supervision instead of relying on manual annotation.

Deep neural networks have been successfully trained on a wide variety of complex tasks by harnessing labels from manual annotation. This is however an expensive and slow process, and the data-hungry nature of deep neural networks means that there is much to be gained from learning without the requirement for manual labelling. While the standard paradigm in *supervised learning* is to learn a representation by learning to predict the labels directly from the input data, the key idea in *self-supervised learning* is to instead learn from *proxy tasks*. Here, the proxy task holds out some part of the input data. The representation is then learnt by predicting the held out part, given the remaining data, hence removing the requirement for manual annotation.

Early examples of proxy tasks include training a model to predict the relative positioning of two patches [Doersch et al., 2015], or to solve a jigsaw puzzle by predicting the correct layout of a grid of image patches [Noroozi & Favaro, 2016; D. Kim et al., 2018], or predicting the correct rotation of an input image [Gidaris et al., 2018]. A key point here is that in these examples of self-supervision, we are not actually interested in how well the model can solve the proxy task, but instead in the semantic representation that the model has been forced to learn in the process of solving the proxy task. For example, in the process of solving a jigsaw puzzle, a model must have learnt meaningful representations of objects and scenes. These useful representations are then evaluated using methods such as training a small set of extra parameters on top of the representation for downstream tasks such as image classification.

More relevant to our proposed work in Chapter 3 are the set of self-supervised approaches where the *proxy task itself* corresponds to a useful downstream application of the model. Here, rather than teasing out the useful semantic representations through fine-tuning, the trained model can be used off-the-shelf to solve the proxy task in the real world.

For image-based tasks, examples of this include training a neural network for predicting colorized images given their grey-scale versions [R. Zhang et al., 2016; Larsson et al., 2016], or learning to in-paint by having the model predict hidden parts of an input image from their context [Pathak et al., 2016]. Whereas for videos, tracking capabilities have emerged from learning cross-frame pixel correspondences by either using the temporal coherency of colors in video [Vondrick et al., 2018; Lai & Xie, 2019] or the cycle-consistency of time [X. Wang, Jabri, & Efros, 2019] as proxy tasks. A further example includes training a neural network to detect audio-visual objects by exploiting the likely audio-visual synchronisation of online videos [Afouras, Owens, et al., 2020].

Learning from Multiple Modalities. Here, we motivate our research in Chapters 4 to 6, where we focus on the use of multi-modal data for computer vision tasks.

Learning from uni-modal data has drawbacks. An example is model robustness. When training on images, perturbing a single pixel in an image can change a model’s class prediction [Su et al., 2019], and when training on text data, inserting just small spelling errors in sentences can fool text classification

systems [Ebrahimi et al., 2018]. In contrast, when learning from multi-modal data, a model is less vulnerable to corruption or distortion in one of the modalities, hence improving robustness.

Learning from multiple modalities is motivated from the human learning process [Barlow, 1989]. To explain, we introduce two concepts that will appear in this thesis, *redundancy* and *complementarity*. These concepts are motivated by the way humans learn from multi-modal data, and provide the motivation for exploring multi-modality in machine learning research. Redundancy is the concept that multi-modal signals coming from the same stimulus contain overlapping information. For example, when exposed to just one of these signals, humans are instantly able to recall the memory of another [Edelman, 1987]. On the other hand, multi-modal signals also contain non-overlapping, complementary information, which gives human perceptual systems robustness. When exposed to such multi-modal signals, we can still identify their source, even when one of the signals is corrupted.

Progress has been made in the Computer Vision community in harnessing these advantages of multi-modal learning. For example, the redundancy between modalities has been utilised to train networks that can predict one modality from another, including generating images from text captions [Mansimov et al., 2016; Reed et al., 2016] (and vice versa [Karpathy & Fei-Fei, 2015; Vinyals et al., 2015]), or predicting the sound of someone’s voice from the look of their face [Nagrani et al., 2018b].

By instead harnessing the complementary information in multi-modal data, robust text-to-video retrieval systems have been trained [Y. Liu et al., 2019]. Here the authors exploit the fact that the audio and visual aspects of a video may contain non-overlapping information contained in the text query. For multi-modal signals arising from the same object in a video, the only overlapping information may be an abstract notion of object category. By learning a correspondence between two such modalities, researchers have developed methods for learning high-level semantic representations without the need for any manual annotation [Arandjelović & Zisserman, 2018; Han et al., 2020; Miech et al., 2019].

In Chapters 5 and 6, we harness the redundancy and complementarity between visual and audio signals coming from the same person in videos for building more robust and effective identification systems. Similar approaches use face recognition to build datasets of voice utterances for training speaker recognition systems [Nagrani et al., 2017], or learning a shared embedding space for face and voice for retrieval [Nagrani et al., 2018a].

Human-Centric Video Understanding. In Chapters 4 to 6 we apply our research to human-centric video understanding, a research field that we provide motivation for here.

Humans do not perceive the world in distinct, independent images, but rather in video; consecutive frames linked together in our local understanding of the arrow of time. Moving beyond image to video understanding is hence a key research area within Computer Vision, and is crucial for building systems for analysing the growing amount of video content available online.

Interestingly, most video understanding tasks are *human-centric*. This is for two reasons: Firstly, many online videos are centred on humans. These videos in turn make for plentiful resources for building datasets that are focused on for example, human action classification [Kuehne et al., 2011; Kay et al., 2017; Heilbron et al., 2015], localisation [C. Gu et al., 2018; Ke et al., 2005], or ego-centric human-vision [Damen et al., 2020]. Secondly, because of our natural interest in human-centred stories and interactions. This interest has led researchers to pursue for example the understanding of human emotions in videos [Albanie et al., 2018], the classification of human interactions and relationships [Kukleva et al., 2020; Lv et al., 2018], or the understanding of human-centred story-lines. The latter is a fascinating and challenging task, and as a result, it has been formulated in several different ways: MovieGraphs [Vicol et al., 2018] focuses on understating motivations and interactions between characters in movies. Other approaches train models to answer questions about characters and events [Lei et al., 2018; Tapaswi et al., 2016], or to retrieve video clips given high-level captions describing story-lines [Bain et al., 2020].

Particularly related to the research presented in Chapters 4 to 6, are those works that focus on understanding *who* is in each frame/scene. This is an important stepping stone for understanding story-lines, as demonstrated by the importance of introducing character annotations in [Bain et al., 2020], and as discussion in [Vicol et al., 2018]. However, the task poses many challenges due to the many factors that can vary a character’s appearance in videos *e.g.* pose, lighting, and emotion. The seminal research in this area focused on labelling face-tracks in videos using the names found in transcripts [Everingham et al., 2006].

An effective system for understanding who is in each scene can alleviate a tremendous cost of identity-annotation that would otherwise be required when building a story understanding or video indexing

system. As a result, much attention has been placed on designing identity-annotation systems that assume smaller amounts of prior annotation accompanying the videos. Weakly supervised identity-annotation methods have been proposed that rely on subtitles or transcripts [Parkhi et al., 2020; Bauml et al., 2013], or images of each character [Nagrani & Zisserman, 2017a]. Other approaches rely on pre-trained face-discriminating representations [Q. Cao et al., 2018a] for clustering the faces by identity in videos [Tapaswi et al., 2019; Kalogeiton & Zisserman, 2020]

1.2 Key Ideas and Contributions

In this Section, we give details on the specific key ideas and contributions presented in this thesis, namely (I) enabling end-to-end learning, and (II) audio-visual human-centric video understanding. Where not previously explained, we outline the computer vision tasks that our ideas are applied to and place them within the relevant prior work.

1.2.1 Enabling End-to-End Learning

Here, we present two key ideas for enabling end-to-end learning. First, we explore the end-to-end optimisation of non-differentiable objective functions in image retrieval. Second, we present a method for synthesising target-data using self-supervision for targeted visual editing.

Smoothing Objective Functions. Where the ideal objective function is not conducive to gradient-based learning methods because it is non-differentiable, end-to-end training can be enabled by instead optimising a smoothed version of the objective function.

In Chapter 2, we focus on the task of ‘query by example’ image retrieval; the task of, given a query image, ranking all of the instances in a retrieval set in accordance with relevance to the query. The benchmark metric for evaluating retrieval systems is Average Precision, which cannot be used for an objective function and directly optimised with gradient-based learning methods because it is non-differentiable. As a result, previous approaches do not train end-to-end for Average Precision, and instead optimise other differentiable metrics [Burges et al., 2007; Z. Cao et al., 2007a] that are designed

to encourage a good ranking.

Our key idea here is that we can enable end-to-end training for the evaluation metric by instead optimising a smoothed version of the rank-based metric Average Precision. Our key insight in Chapter 2 is that differentiable alternatives to an Average Precision loss that are used in prior work are not well-suited replacements because they are not optimising a rank-based metric, but instead a distance-based metric. As a result, we hypothesise that prior work are instead learning representations that are not optimised for Average Precision. Instead we design a smoothed approximation of Average Precision that is in fact optimising a rank-based metric, and as a result, we say that we can train models end-to-end for the Average Precision metric.

Our contributions in Chapter 2 include an analysis as to why *Smooth-AP* is a closer approximation to Average Precision than previous approximations and deep metric learning objectives. Furthermore, our proposed smoothing method can be readily applied to any other non-differentiable ranking metric. Finally, we demonstrate the quantitative benefits of training end-to-end for Average Precision by evaluating on several large-scale real-world retrieval datasets, where *Smooth-AP* outperforms prior work.

Synthesising Target-Data with Self Supervision. Where target-data is not available in the sufficient quantity required for training deep neural networks, end-to-end learning can be enabled by instead using self-supervision to synthesise that target-data.

In Chapter 3, we focus on the task of conditional image editing; the task of editing a region in an image using some signal that specifies the desired change. Obtaining target-data for this task is prohibitively expensive, as it would require the detailed intervention of skilled human-artists. As a result, previous approaches do not train models end-to-end. Instead, prior works learn a task-agnostic image-prior [Bau et al., 2021; Chai et al., 2021a], which is not necessarily suited for the conditional image editing task, resulting in undesired artifacts in edited images, and limited control of the editing process. We hypothesise that end-to-end training for the task of image editing would result in higher quality edits that are simultaneously more natural looking and more dependant on the conditioning signal.

Our key idea is here is that we can enable end-to-end training by using self-supervision to synthesise target data. Here, following the self-supervised paradigm, we train the image editing model to predict

some piece of the input data that has been hidden from the model. We use this concept to train a model end-to-end for the conditional image editing task for the first time.

A different perspective exists for this key idea: We show that a self-supervised learning proxy task can be formulated that directly corresponds to the downstream task of conditional image editing. We demonstrate that a model trained end-to-end for this proxy task, can be used off-the-shelf for conditional image editing. This perspective places our key insight amongst those prior work in self-supervised learning where the proxy task itself corresponds to a useful downstream application of the model [Afouras, Owens, et al., 2020; R. Zhang et al., 2016] (see Section 1.1).

Our contributions in Chapter 3 include the presentation of a novel self-supervised approach for synthesising target data for the task of targeted image editing, which allows for the end-to-end learning of representations for this task. We analyse how different editing effects can be learnt by an intuitive control of the augmentation in the self-supervised process. Finally, we demonstrate the advantage of training models end-to-end, by showing significant improvements over all prior work in extensive qualitative and quantitative experiments, including human studies.

1.2.2 Audio-Visual Human-Centric Video Understanding.

Here, we outline the key ideas presented in this thesis for learning from audio-visual data for human-centric video understanding. In detail, our ideas focus on the challenge of determining *who* is in a scene in a video.

It is important when determining *who* is present in a scene, that identity discriminating deep representations are effective in the target domain of human-centric videos. State-of-the-art voice identity discriminating features are trained on large datasets of audio clips taken from interviews; a domain which can differ considerably from that of human-centric videos such as movies and TV-shows where there exist greater variation in background noise, and actors often intentionally disguise their voice. In Chapter 4 we ask the question: *how well do state-of-the-art identity discriminating voice representations adapt to human-centric videos such as movies and TV-shows*. Identity information is present in both the face and the voice, so here we use face recognition in unlabelled videos in order to collect a

dataset of voice segments from movies, and in doing so we create a dataset for exploring this question.

Our key contributions in Chapter 4 include the collection of a challenging dataset for speaker verification, *VoxMovies*, containing speech segments from movies, where actors display strong emotion, varying accents, and are often intentionally trying to disguise their voice to play a character. We introduce several challenging domain adaptation evaluation sets, and demonstrate the difficulty that state-of-the-art speaker verification systems find in adapting to this new domain. By experimenting with several simple domain adaptation paradigms, we show that performance on *VoxMovies* can be improved, but that there is still a lot of work to be done by the community on this task.

Given strong identity discriminating representations, in Chapters 5 and 6 we focus on using these representations to identify *who* is present in each scene in human-centric videos. A challenge when determining identities in videos is to build systems that require the least supervision possible. In Chapter 5 we focus on the realistic setting where an effective system must not rely on any human annotation beyond the names of the videos. Our key idea here is that although we have very little human annotation to rely on, there is still rich multi-modal data found throughout the videos that a human would use to solve this task. We hence take a human-inspired approach to this task. We build a system which looks for clues for people’s names, either found in the videos, when their name is mentioned in speech, or in lists of cast members found in online archives. We then verify whether these names correspond to co-occurring faces on screen by looking for corroborating evidence on image search engines, and then annotate these people throughout the videos using visual and audio identity discriminating signals.

Our contributions in Chapter 5 include the presentation of a fully automated system for labelling faces in video archives such as TV broadcasts, without the need for any human annotation beyond simply the name of the program. As a result, we present a scalable automatic annotation solution for the contemporary problem of growing, unwieldy and unlabelled online video archives. As a key component of our system, we present a novel method for robustly and automatically searching for corroborating visual evidence for a person’s identity on image search engines. We demonstrate that the proposed system sets new state-of-the-art on all public benchmarks for this task.

A different approach to identifying *who* is present in a video, is to group all of the people found in a video by identity; a task referred to as *person clustering*. This is a challenging task due to the large

variation in someone’s appearance throughout a video. Our key idea here is that we can rely on the complementarity between multi-modal signals coming from the same person in a video to build a robust and effective person-clustering system. Differently to prior work which just focuses on clustering face-tracks, in Chapter 6 we cluster *all* people in videos by using all available cues from face, body-appearance, and voice. A further key insight that we present is that the resulting *person-level* reasoning that person-clustering offers, is more suited than prior approaches for downstream applications such as story understanding.

Our contributions in Chapter 6 include a novel method for clustering person-tracks rather face-tracks, by relying on all available cues, including those from the face, body-appearance and the voice of each of the characters in the video. We explore how the resulting *person-level* reasoning can be useful for downstream applications such as story understanding via character co-occurrence. Finally, we collect a large and varied dataset for evaluating *person-clustering* and demonstrate the significant quantitative benefits of our approach over prior work.

1.3 Thesis Outline

Chapters 2 to 6 each contain a paper which has been peer-reviewed and accepted for publication in a conference. The papers have been left unmodified from their published forms, with the exception of formatting changes. For each publication, we also provide a statement of authorship in Appendix 2. The publications included in this thesis are:

Chapter 2: “Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval”.

Andrew Brown, Weidi Xie, Vicky Kalogeiton, Andrew Zisserman. In Proc. *European Conference on Computer Vision (ECCV)* 2020.

Chapter 3: “End-to-End Visual Editing with a Generatively Pre-Trained Artist”.

Andrew Brown, Cheng-Yang Fu, Omkar Parkhi, Tamara L. Berg, Andrea Vedaldi. In Proc. *European Conference on Computer Vision (ECCV)* 2022.

Chapter 4: “Playing a Part: Speaker Verification at the Movies”.

Andrew Brown, Jaesung Huh, Arsha Nagrani, Joon Son Chung, Andrew Zisserman. In Proc. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2021.

Chapter 5: “Automated Video Labelling: Identifying Faces by Corroborative Evidence”.

Andrew Brown, Ernesto Coto, Andrew Zisserman. In Proc. *Multimedia Information Processing and Retrieval (MIPR)* 2021 (Oral Presentation, Best Student Paper Award).

Chapter 6: “Face, Body, Voice: Video Person-Clustering with Multiple Modalities”.

Andrew Brown, Vicky Kalogeiton, Andrew Zisserman. In Proc. *International Conference on Computer Vision Workshops (ICCVW)* 2021 (Spotlight Presentation).

Chapter 7 includes an outline on the achievements and impact of each of the contributions made in this thesis, followed by a discussion on directions for future research, and finally a conclusion.

1.3.1 Publications not included:

“Condensed Movies: Story Based Retrieval with Contextual Embeddings”

Max Bain, Arsha Nagrani, **Andrew Brown**, Andrew Zisserman. In Proc. *Asian Conference on Computer Vision (ACCV)* 2020 (Oral Presentation)

“SeeHear: Signer Diarisation and a new dataset”

Samuel Albanie, Gul Varol, Liliane Momeni, Triantafyllos Afouras, **Andrew Brown**, Chuhan Zhang, Ernesto Coto, Necati Cihan Camgoz, Ben Saunders, Abhishek Dutta, Neil Fox, Richard Bowden, Bencie Woll, Andrew Zisserman. In Proc. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2021.

Part I

Enabling End-to-End Learning

2 | Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval

Andrew Brown Weidi Xie Vicky Kalogeiton Andrew Zisserman

VGG, Oxford

Abstract

Optimising a ranking-based metric, such as Average Precision (AP), is notoriously challenging due to the fact that it is non-differentiable, and hence cannot be optimised directly using gradient-descent methods. To this end, we introduce an objective that optimises instead a *smoothed approximation* of AP, coined *Smooth-AP*. Smooth-AP is a plug-and-play objective function that allows for end-to-end training of deep networks with a simple and elegant implementation. We also present an analysis for why directly optimising the ranking based metric of AP offers benefits over other deep metric learning losses.

We apply Smooth-AP to standard retrieval benchmarks : Stanford Online products and VehicleID, and also evaluate on larger-scale datasets: INaturalist for fine-grained category retrieval, and VGGFace2 and IJB-C for face retrieval. In all cases, we improve the performance over the state-of-the-art, especially for larger-scale datasets, thus demonstrating the effectiveness and scalability of Smooth-AP to real-world scenarios.

Published in the Proceedings of the European Conference on Computer Vision, 2020

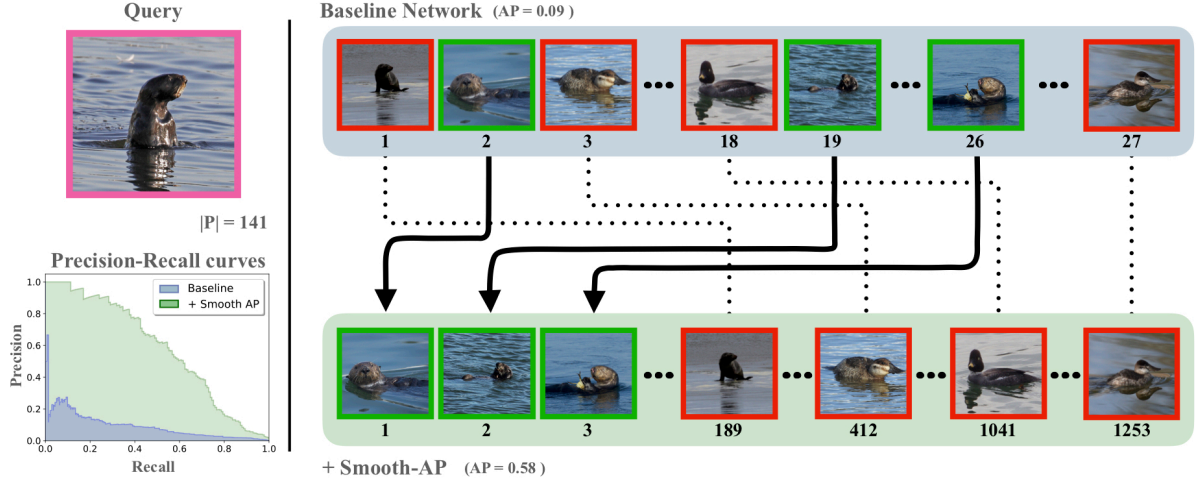


Figure 2.1: **Ranked retrieval sets** before (top) and after (bottom) applying Smooth-AP on a baseline network (*i.e.* ImageNet pre-trained weights) for a given query (pink image). The precision-recall curve is shown on the left. Smooth-AP results in large boost in AP, as it moves positive instances (green) high up the ranks and negative ones (red) low down. $|\mathcal{P}|$ is the number of positive instances in the retrieval set for this query. Images are from the INaturalist dataset.

2.1 Introduction

Our objective in this paper is to improve the performance of ‘query by example’, where the task is: given a query image, rank all the instances in a retrieval set according to their relevance to the query. For instance, imagine that you have a photo of a friend or family member, and want to search for all of the images of that person within your large smart-phone image collection; or on a photo licensing site, you want to find all photos of a particular building or object, starting from a single photo. These use cases, where high recall is premium, differ from the ‘Google Lens’ application of identifying an object from an image, where only one ‘hit’ (match) is sufficient.

The benchmark metric for retrieval quality is Average Precision (AP) (or its generalized variant, Normalized Discounted Cumulative Gain, which includes non-binary relevance judgements). With the resurgence of deep neural networks, end-to-end training has become the *de facto* choice for solving specific vision tasks with well-defined metrics. However, the core problem with AP and similar metrics is that they include a discrete ranking function that is neither differentiable nor decomposable. Consequently, their direct optimization, *e.g.* with gradient-descent methods, is notoriously difficult.

In this paper, we introduce a novel differentiable AP approximation, *Smooth-AP*, that allows end-to-end training of deep networks for ranking-based tasks. Smooth-AP is a simple, elegant, and scalable method that takes the form of a plug-and-play objective function by relaxing the Indicator function in the non-differentiable AP with a sigmoid function. To demonstrate its effectiveness, we perform experiments on two commonly used image retrieval benchmarks, Stanford Online Products and VehicleID, where Smooth-AP outperforms all recent AP approximation approaches [Cakir et al., 2019a; RolÅnek et al., 2020] as well as recent deep metric learning methods. We also experiment on three further large-scale retrieval datasets (VGGFace2, IJB-C, INaturalist), which are orders of magnitude larger than the existing retrieval benchmarks. To our knowledge, this is the first work that demonstrates the possibility of training networks for AP on datasets with millions of images for the task of image retrieval. We show large performance gains over all recently proposed AP approximating approaches and, somewhat surprisingly, also outperform strong verification systems [J. Deng et al., 2019; W. Liu et al., 2017] by a significant margin, reflecting the fact that metric learning approaches are indeed inefficient for training large-scale retrieval systems that are measured by global ranking metrics.

2.2 Related Work

As an essential component of information retrieval [Manning et al., 2008a], algorithms that optimize rank-based metrics have been the focus of extensive research over the years. In general, the previous approaches can be split into two lines of research, namely metric learning, and direct approximation of Average Precision.

Image Retrieval. This is one of the most researched topics in the vision community. Several themes have been explored in the literature, for example, one theme is on the speed of retrieval and explores methods of approximate nearest neighbors [Chum et al., 2011; Jégou et al., 2008; Jégou, Douze, & Schmid, 2011; Jégou, Perronin, et al., 2011; Philbin et al., 2007; Sivic & Zisserman, 2003]. Another theme is on how to obtain a compact image descriptor for retrieval in order to reduce the memory footprint. Descriptors were typically constructed through an aggregation of local features, such as

Fisher vectors [Perronnin et al., 2010] and VLAD [Arandjelovic & Zisserman, 2013; Jégou et al., 2010]. More recently, neural networks have made impressive progress on learning representations for image retrieval [Arandjelović et al., 2016; Babenko et al., 2014; Gordo et al., 2016; Radenovic et al., 2016; J. Wang et al., 2014], but common to all is the choice of the loss function used for training; in particular, it should ideally be a loss that will encourage a “good” ranking.

Metric Learning. To avoid the difficulties from directly optimising rank-based metrics, such as Average Precision, there is a great body of work that focuses on metric learning [Arandjelović et al., 2016; Arandjelovic & Zisserman, 2013; Burges et al., 2007; Z. Cao et al., 2007b; H. Chen et al., 2019; Chopra et al., 2005; Jégou, Perronnin, et al., 2011; Law et al., 2017; Movshovitz-Attias et al., 2017; Oh Song et al., 2016; Y. Rao et al., 2018; Ustinova & Lempitsky, 2016; X. Wang, Hua, et al., 2019; K. Weinberger et al., 2006]. For instance, the contrastive [Chopra et al., 2005] and triplet [K. Weinberger et al., 2006] losses, which consider pairs or triplets of elements, and force all positive instances to be close in the high-dimensional embedding space, while separating negatives by a fixed distance (margin). However, due to the limited rank-positional awareness that a pair/triplet provides, a model is likely to waste capacity on improving the order of positive instances at low (poor) ranks at the expense of those at high ranks, as was pointed out by Burges *et al.* [Burges et al., 2007]. Of more relevance, the list-wise approaches [Burges et al., 2007; Z. Cao et al., 2007b; Movshovitz-Attias et al., 2017; Oh Song et al., 2016; X. Wang, Hua, et al., 2019] look at many examples from the retrieval set, and have been proven to improve training efficiency and performance. Despite being successful, one drawback of metric learning approaches is that they are mostly driven by minimizing distances, and therefore remain ignorant of the importance of shifting ranking orders – the latter is essential when evaluating with a rank-based metric.

Optimizing Average Precision (AP). The trend of directly optimising the non-differentiable AP has been recently revived in the retrieval community. Sophisticated methods [Cakir et al., 2019a; H. Chen et al., 2019; Engilberge et al., 2019; K. He, Cakir, et al., 2018; K. He, Lu, & Sclaroff, 2018; Henderson & Ferrari, 2016; McFee & Lanckriet, 2010; Revaud et al., 2019; RolÁnek et al., 2020; Song et al., 2016; Taylor et al., 2008; Ustinova & Lempitsky, 2016; H.-F. Yang et al., 2016; Yue et al., 2007] have been developed to overcome the challenge of non-decomposability and non-differentiability in

optimizing AP. Methods include: creating a distribution over rankings by treating each relevance score as a Gaussian random variable [Taylor et al., 2008], loss-augmented inference [McFee & Lanckriet, 2010], direct loss minimization [Henderson & Ferrari, 2016; Song et al., 2016], optimizing a smooth and differentiable upper bound of AP [McFee & Lanckriet, 2010; Mohapatra et al., 2018; Yue et al., 2007], training a LSTM to approximate the discrete ranking step [Engilberge et al., 2019], differentiable histogram binning [Cakir et al., 2019a; K. He, Cakir, et al., 2018; K. He, Lu, & Sclaroff, 2018; Revaud et al., 2019; Ustinova & Lempitsky, 2016], error driven update schemes [H. Chen et al., 2019], and the very recent blackbox optimization [RolÁnek et al., 2020]. Significant progress on optimizing AP was made by the information retrieval community [Burges et al., 2007; Chapelle et al., 2007; Guiver & Snelson, 2008; K. Li et al., 2014; Qin et al., 2010; Taylor et al., 2008], but the methods have largely been ignored by the vision community, possibly because they have never been demonstrated on large-scale image retrieval or due to the complexity of the proposed smooth objectives. One of the motivations of this work is to show that with the progress of deep learning research, *e.g.* auto-differentiation, better optimization techniques, large-scale datasets, and fast computation devices, it is possible and in fact very easy to directly optimize a close approximation to AP.

2.3 Background

In this section, we define the notations used throughout the paper.

2.3.0.1 Task definition.

Given an input query, the goal of a retrieval system is to rank all instances in a retrieval set $\Omega = \{I_i, i = 0, \dots, m\}$ based on their relevance to the query. For *each* query instance I_q , the retrieval set is split into the positive $\mathcal{P}_q$ and negative $\mathcal{N}_q$ sets, which are formed by all instances of the same class and of different classes, respectively. Note that there is a different positive and negative set for each query.

2.3.0.2 Average Precision (AP).

AP is one of the standard metrics for information retrieval tasks [Manning et al., 2008a]. It is a single value defined as the area under a Precision-Recall curve. For a query I_q , the predicted relevance scores of all instances in the retrieval set are measured via a chosen metric. In our case, we use the cosine similarity (though the Smooth-AP method is independent of this choice):

$$\mathcal{S}_\Omega = \left\{ s_i = \left\langle \frac{v_q}{\|v_q\|} \cdot \frac{v_i}{\|v_i\|} \right\rangle, i = 0, \dots, n \right\}, \quad (2.1)$$

where $\mathcal{S}_\Omega = \mathcal{S}_P \cup \mathcal{S}_N$, and $\mathcal{S}_P = \{s_\zeta, \forall \zeta \in \mathcal{P}_q\}$, $\mathcal{S}_N = \{s_\xi, \forall \xi \in \mathcal{N}_q\}$ are the positive and negative relevance score sets, respectively, v_q refers to the query vector, and v_i to the vectorized retrieval set. The AP of a query I_q can be computed as:

$$AP_q = \frac{1}{|\mathcal{S}_P|} \sum_{i \in \mathcal{S}_P} \frac{\mathcal{R}(i, \mathcal{S}_P)}{\mathcal{R}(i, \mathcal{S}_\Omega)}, \quad (2.2)$$

where $\mathcal{R}(i, \mathcal{S}_P)$ and $\mathcal{R}(i, \mathcal{S}_\Omega)$ refer to the rankings of the instance i in $\mathcal{P}$ and Ω , respectively. Note that, the rankings referred to in this paper are assumed to be *proper rankings*, meaning no two samples are ranked equally.

2.3.0.3 Ranking Function ($\mathcal{R}$).

Given that AP is a ranking-based method, the key element for direct optimisation is to define the ranking $\mathcal{R}$ of one instance i . Here, we define it in the following way [Qin et al., 2010]:

$$\mathcal{R}(i, \mathcal{S}) = 1 + \sum_{j \in \mathcal{S}, j \neq i} \mathbb{I}\{(s_i - s_j) > 0\}, \quad (2.3)$$

where $\mathbb{I}\{\cdot\}$ acts as an Indicator function, and $\mathcal{S}$ any set, *e.g.* Ω . Conveniently, this can be implemented by computing a difference matrix $D \in \mathbb{R}^{m \times m}$:

$$D = \begin{bmatrix} s_1 & \dots & s_m \\ \vdots & \ddots & \vdots \\ s_1 & \dots & s_m \end{bmatrix} - \begin{bmatrix} s_1 & \dots & s_1 \\ \vdots & \ddots & \vdots \\ s_m & \dots & s_m \end{bmatrix} \quad (2.4)$$

The exact AP for a query instance I_q from Eq. 2.2 becomes:

$$AP_q = \frac{1}{|\mathcal{S}_P|} \sum_{i \in \mathcal{S}_P} \frac{1 + \sum_{j \in \mathcal{S}_P, j \neq i} \mathbb{I}\{D_{ij} > 0\}}{1 + \sum_{j \in \mathcal{S}_P, j \neq i} \mathbb{I}\{D_{ij} > 0\} + \sum_{j \in \mathcal{S}_N} \mathbb{I}\{D_{ij} > 0\}} \quad (2.5)$$

2.3.0.4 Derivatives of Indicator.

The particular Indicator function used in computing AP is a Heaviside step function $\mathcal{H}(\cdot)$ [Qin et al., 2010], with its distributional derivative defined as Dirac delta function:

$$\frac{d\mathcal{H}(x)}{dx} = \delta(x),$$

This is either flat everywhere, with zero gradient, or discontinuous, and hence cannot be optimized with gradient based methods (Figure 2.2).

2.4 Approximating Average Precision (AP)

As explained, AP and similar metrics include a discrete ranking function that is neither differentiable nor decomposable. In this section, we first describe *Smooth-AP*, which essentially replaces the discrete indicator function with a sigmoid function, and then we provide an analysis on its relation to other ranking losses, such as triplet loss [Hermans et al., 2017; K. Weinberger et al., 2006], FastAP [Cakir et al., 2019a] and Blackbox AP [RolÅnek et al., 2020].

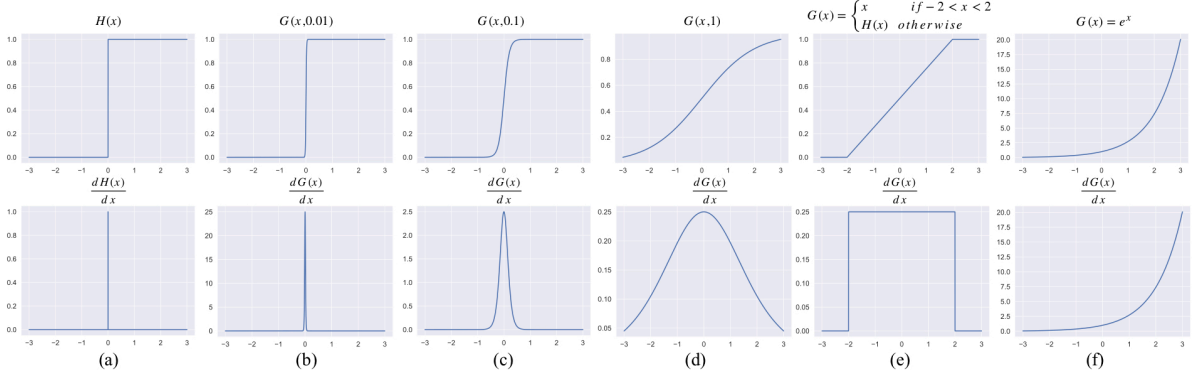


Figure 2.2: The possible **different approximations** to the discrete Indicator function. First row: Indicator function (a), three sigmoids with increasing temperatures (b, c, d), linear (e), exponential (f). Second row: their derivatives.

2.4.1 Smoothing AP

To smooth the ranking procedure, which will enable direct optimization of AP, Smooth-AP takes a simple solution which is to replace the Indicator function $\mathbb{I}\{\cdot\}$ by a sigmoid function $\mathcal{G}(\cdot; \tau)$, where the τ refers to the temperature adjusting the sharpness:

$$\mathcal{G}(x; \tau) = \frac{1}{1 + e^{\frac{-x}{\tau}}}. \quad (2.6)$$

Substituting $\mathcal{G}(\cdot; \tau)$ into Eq. 2.5, the true AP can be approximated as:

$$AP_q \approx \frac{1}{|\mathcal{S}_P|} \sum_{i \in \mathcal{S}_P} \frac{1 + \sum_{j \in \mathcal{S}_P} \mathcal{G}(D_{ij}; \tau)}{1 + \sum_{j \in \mathcal{S}_P} \mathcal{G}(D_{ij}; \tau) + \sum_{j \in \mathcal{S}_N} \mathcal{G}(D_{ij}; \tau)}$$

with tighter approximation and convergence to the indicator function as $\tau \rightarrow 0$. The objective function during optimization is denoted as:

$$\mathcal{L}_{AP} = \frac{1}{m} \sum_{k=1}^m (1 - AP_k) \quad (2.7)$$

Smoothing parameter τ governs the temperature of the sigmoid that replaces the Indicator function $\mathbb{I}\{\cdot\}$. It defines an operating region, where terms of the difference matrix are given a gradient by the Smooth-AP loss. If the terms are mis-ranked, Smooth-AP will attempt to shift them to the correct or-

der. Specifically, a small value of τ results in a small operating region (Figure 2.2 (b) – note the small region with gradient seen in the sigmoid derivative), and a tighter approximation of true AP. The strong acceleration in gradient around the zero point (Figure 2.2 (b)-(c) second row) is essential to replicating the desired qualities of AP, as it encourages the shifting of instances in the embedding space that result in a change of rank (and hence change in AP), rather than shifting instances by some large distance but not changing the rank. A large value of τ offers a large operating region, however, at the cost of a looser approximation to AP due to its divergence from the indicator function.

Relation to Triplet Loss. Here, we demonstrate that the triplet loss (a popular surrogate loss for ranking) is in fact optimising a distance metric rather than a ranking metric, which is sub-optimal when evaluating using a ranking metric. As shown in Eq. 2.5, the goal of optimizing AP is equivalent to minimizing all the $\sum_{i \in \mathcal{S}_P, j \in \mathcal{S}_N} \mathbb{I}\{D_{ij} < 0\}$, *i.e.* the violating terms. We term these as such because these terms refer to cases where a negative instance is ranked above a positive instance in terms of relevance to the query, and optimal AP is only acquired when all positive instances are ranked above all negative instances.

For example, consider one query instance with predicted relevance score and ground-truth relevance labels as:

Instances ordered by score : $(s_0 \ s_4 \ s_1 \ s_2 \ s_5 \ s_6 \ s_7 \ s_3)$

Ground truth labels : $(1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1)$

the violating terms are: $\{(s_4 - s_1), (s_4 - s_2), (s_4 - s_3), (s_5 - s_3), (s_6 - s_3), (s_7 - s_3)\}$. An ideal AP loss would actually treat each of the terms unequally, *i.e.* the model would be forced to spend more capacity on shifting orders between s_4 and s_1 , rather than s_3 and s_7 , as that makes a larger impact on improving the AP.

Another interpretation of these violating cases can also be drawn from the triplet loss perspective. Specifically, if we treat the query instance as an “anchor”, with s_j denoting the similarity between

the “anchor” and negative instance, and s_i denoting the similarity between the “anchor” and positive instance. In this example, the triplet loss tries to optimize a margin hinge loss:

$$\begin{aligned}\mathcal{L}_{\text{triplet}} = & \max(s_4 - s_1 + \alpha, 0) + \max(s_4 - s_2 + \alpha, 0) \\ & + \max(s_4 - s_3 + \alpha, 0) + \max(s_5 - s_3 + \alpha, 0) \\ & + \max(s_6 - s_3 + \alpha, 0) + \max(s_7 - s_3 + \alpha, 0)\end{aligned}$$

This can be viewed as a differentiable approximation to the goal of optimizing AP where the Indicator function has been replaced with the margin hinge loss, thus solving the gradient problem. Nevertheless, using a triplet loss to approximate AP may suffer from two problems: *First*, all terms are linearly combined and treated equally in $\mathcal{L}_{\text{triplet}}$. Such a surrogate loss may force the model to optimize the terms that have only a small effect on AP, *e.g.* optimizing $s_4 - s_1$ is the same as $s_7 - s_4$ in the triplet loss, however, from an AP perspective, it is important to correct the mis-ordered instances at high rank. *Second*, the linear derivative means that the optimization process is purely based on distance (not ranking orders), which makes it sub-optimal when evaluating AP. For instance, in the triplet loss case, reducing the distance $s_4 - s_1$ from 0.8 to 0.5 is the same as from 0.2 to -0.1 . In practise, however, the latter case (shifting orders) will clearly have a much larger impact on the AP computation than the former.

Comparison to other AP-optimising methods. The two key differences between Smooth-AP and the recently introduced FastAP and Blackbox AP, are that Smooth-AP (i) provides a closer approximation to Average Precision, and (ii) is far simpler to implement. Firstly, due to the sigmoid function, Smooth-AP optimises a ranking metric, and so has the same objective as Average Precision. In contrast, FastAP and Blackbox AP linearly interpolate the non-differentiable (piecewise constant) function, which can potentially lead to the same issues as triplet loss, *i.e.* optimizing a distance metric, rather than rankings. Secondly, Smooth-AP simply needs to replace the indicator function in the AP objective with a sigmoid function. While FastAP uses abstractions such as Histogram Binning, and Blackbox AP uses a variant of numerical derivative. These differences are positively affirmed through the improved performance of Smooth-AP over several datasets (Section 2.6.5).

Table 2.1: **Datasets** used for training and evaluation.

dataset		# Images	# Classes	# Ims/Class
object/animal retrieval datasets	SOP train	59,551	11,318	5.3
	SOP test	60,502	11,316	5.3
	VehicleID train	110,178	13,134	8.4
	VehicleID test	40,365	4,800	8.4
	INaturalist train	325,846	5,690	57.3
	INaturalist test	136,093	2,452	55.5
face retrieval datasets	VGGFace2 train	3.31 M	8,631	363.7
	VGGFace2 test	169,396	500	338.8
	IJB-C	148,824	3,531	42.1

2.5 Experimental Setup

In this section, we describe the datasets used for evaluation, the test protocols, and the implementation details. The procedure followed here is to take a pre-trained network and fine-tune with Smooth-AP loss. Specifically, ImageNet pretrained networks are used for the object/animal retrieval datasets, and high-performing face-verification models for the face retrieval datasets.

2.5.1 Datasets

We evaluate the Smooth-AP loss on five datasets containing a wide range of domains and sizes. These include the commonly used retrieval benchmark datasets, as well as several additional *large-scale* (>100K images) datasets. Table 2.1 describes their details.

Stanford Online Product (SOP) [Song et al., 2016] was initially collected for investigating the problem of metric learning. It includes 120K images of products that were sold online. We use the same evaluation protocol and train/test split as [X. Wang, Hua, et al., 2019].

VehicleID [Wah et al., 2011] contains 221, 736 images of 26, 267 vehicle categories, 13, 134 of which are used for training (containing 110, 178 images). By following the same test protocol as [Wah et al.,

2011], three test sets of increasing size are used for evaluation (termed small, medium, large), which contain 800 classes (7,332 images), 1600 classes (12,995 images) and 2400 classes (20,038 images) respectively.

INaturalist [Van Horn et al., 2018] is a large-scale animal and plant species classification dataset, designed to replicate real-world scenarios through 461,939 images from 8,142 classes. It features many visually similar species, captured in a wide variety of environments. We construct a new image retrieval task from this dataset, by keeping 5,690 classes for training, and 2,452 unseen classes for evaluating image retrieval at test time, according to the same test protocols as existing benchmarks [X. Wang, Hua, et al., 2019]. We will make the train/test splits publicly available.

VGGFace2 [Q. Cao et al., 2018b] is a large-scale face dataset with over 3.31 million images of 9,131 subjects. The images have large variations in pose, age, illumination, ethnicity and profession, *e.g.* actors, athletes, politicians. For training, we use the pre-defined *training set* with 8,631 identities, and for testing we use the *test set* with 500 identities, totalling 169K testing images.

IJB-C [Maze et al., 2018] is a challenging public benchmark for face recognition, containing images of subjects from both still frames and videos. Each video is treated as a single instance by averaging the CNN-produced vectors for each frame to a single vector. Identities with less than 5 instances (images or videos) are removed.

2.5.2 Test Protocol

Here, we describe the protocols for evaluating retrieval performance, mean Average Precision (mAP) and Recall@K (R@K). For all datasets, every instance of each class is used in turn as the query I_q , and the retrieval set Ω is formed out of all the remaining instances. We ensure that each class in all datasets contains several images (Table 2.1), such that if an instance from a class is used as the query, there are plenty of remaining positive instances in the retrieval set. For object/animal retrieval evaluation, we

use the $Recall@K$ metric in order to compare to existing works. For face retrieval, AP is computed from the resulting output ranking for each query, and the mAP score is computed by averaging the APs across every instance in the dataset, resulting in a single value.

2.5.3 Implementation Details

Object/animal retrieval (SOP, VehicleID, INaturalist). In line with previous works [Cakir et al., 2019a; RolÁnek et al., 2020; Roth et al., 2019; Sanakoyeu et al., 2019; X. Wang et al., 2020; C.-Y. Wu et al., 2017], we use ResNet50 [K. He et al., 2016a] as the backbone architecture, which was pretrained on ImageNet [Russakovsky et al., 2015]. We replace the final softmax layer with one linear layer (following [Cakir et al., 2019a; RolÁnek et al., 2020], with dimension being set to 512). All images are resized to 256×256 . At training time, we use random crops and flips as augmentations, and at test time, a single centre crop of size 224×224 is used. For all experiments we set τ to 0.01 (Section 2.6.4).

Face retrieval datasets (VGGFace2, IJB-C). We use two high performing face verification networks: the method from [Q. Cao et al., 2018b] using the SENet-50 architecture [Hu et al., 2018a] and the state-of-the-art ArcFace [J. Deng et al., 2019] (using ResNet-50), both trained on the VGGFace2 training set. For SENet-50, we follow [Q. Cao et al., 2018b] and use the same face crops (extended by the recommended amount), resized to 224×224 and we L2-normalize the final 256D embedding. For ArcFace, we generate normalised face crops (112×112) by using the provided face detector [J. Deng et al., 2019], and align them with the predicted 5 facial key points, then L2-normalize the final 512D embedding. For both models, we set the batch size to 224 and τ to 0.01 (Section 2.6.4).

Mini-batch training. During training, we form each mini-batch by randomly sampling classes such that each represented class has $|\mathcal{P}|$ samples per class. For all experiments, we L2-normalize the embeddings, use cosine similarity to compute the relevance scores between the query and the retrieval set, set $|\mathcal{P}|$ to 4, and use an Adam [Kingma & Ba, 2014] optimiser with a base learning rate of 10^{-5} with weight decay $4e^{-5}$. We employ the same hard negative mining technique as [Cakir et al., 2019a; RolÁnek et al., 2020] only for the Online Products dataset. Otherwise we use no special sampling

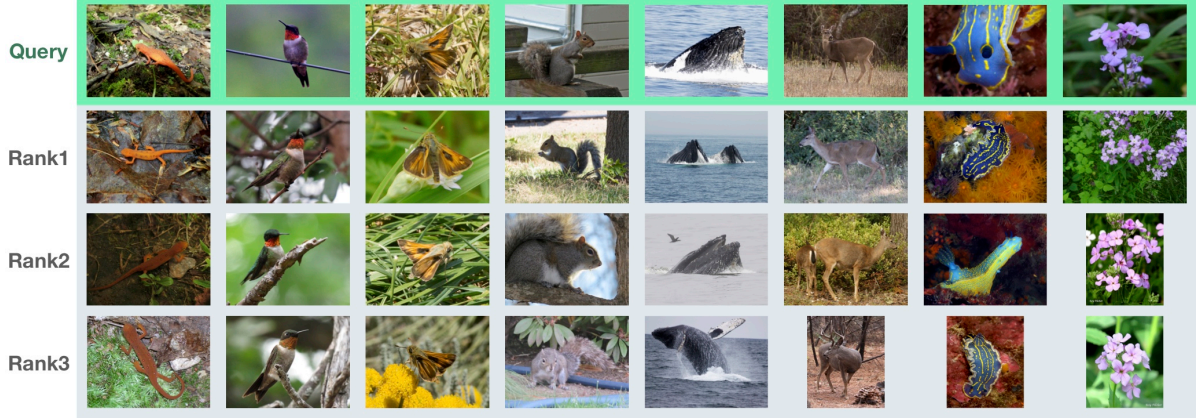


Figure 2.3: **Qualitative results for the INaturalist dataset using Smooth-AP loss.** For each query image (top row), the top 3 instances from the retrieval set are shown ranked from top to bottom. Every retrieved instance shown is a true positive.

strategies.

2.6 Results

In this section, we first explore the effectiveness of the proposed Smooth-AP by examining the performance of various models on the five retrieval datasets. Specifically, we compare with the recent AP optimization and broader metric learning methods on the standard benchmarks SOP and VehicleID (Section 2.6.1), and then shift to further large-scale experiments, *e.g.* INaturalist for animal/plant retrieval, and IJB-C and VGGFace2 for face retrieval (Sections 2.6.2-2.6.3). Then, we present an ablation study of various hyper-parameters that affect the performance of Smooth-AP: the sigmoid temperature, the size of the positive set, and the batch size (Section 2.6.4). Finally, we discuss various findings and analyze the performance gaps between various models (Section 2.6.5).

Note that, although there has been a rich literature on metric learning methods [Duan et al., 2018; Ge et al., 2018; Harwood et al., 2017; W. Kim et al., 2018; Law et al., 2017; Lu et al., 2019; Movshovitz-Attias et al., 2017; Oh Song et al., 2017; Opitz et al., 2017; Qian et al., 2019; RolÅnek et al., 2020; Song et al., 2016; Suh et al., 2019; Ustinova & Lempitsky, 2016; J. Wang et al., 2017; X. Wang, Hua,

et al., 2019; X. Wang, Han, et al., 2019; X. Wang et al., 2020; C.-Y. Wu et al., 2017; Xuan et al., 2018; Yuan et al., 2017] using these image retrieval benchmarks, we only list the very recent state-of-the-art approaches, and try to compare with them as fairly as we can, *e.g.* no model ensemble, and using the same backbone network and image resolution. However, there still remain differences on some small experimental details, such as embedding dimensions, optimizer, and learning rates. Qualitative results for the INaturalist dataset are illustrated in Figure 2.3.

2.6.1 Evaluation on Stanford Online Products (SOP)

We compare with a wide variety of state-of-the-art image retrieval methods, *e.g.* deep metric learning methods [Roth et al., 2019; Sanakoyeu et al., 2019; X. Wang et al., 2020; C.-Y. Wu et al., 2017], and AP approximation methods [Cakir et al., 2019a; RolÅnek et al., 2020]. As shown in Table 2.2, we observe that Smooth-AP achieves state-of-the-art results on the SOP benchmark. In particular, our best model outperforms the very recent AP approximating methods (Blackbox AP and FastAP) by a 1.5% margin for Recall@1. Furthermore, Smooth-AP performs on par with the concurrent work (Cross-Batch Memory [X. Wang et al., 2020]). This is particularly impressive as [X. Wang et al., 2020] harnesses memory techniques to sample from many mini-batches simultaneously for each weight update, whereas Smooth-AP only makes use of a single mini-batch on each training iteration.

Figure 2.4 provides a quantitative analysis into the effect of sigmoid temperature τ on the tightness of the AP approximation, which can be plotted via the AP approximation error:

$$AP_e = |AP_{pred} - AP| \quad (2.8)$$

where AP_{pred} is the predicted approximate AP when the sigmoid is used in place of the indicator function in Equation 5, and AP is the true AP. As expected, a lower value of τ leads to a tighter approximation to Average Precision, shown by the low approximation error.

Recall@K	SOP			
	1	10	100	1000
Margin [C.-Y. Wu et al., 2017]	72.7	86.2	93.8	98.0
Divide [Sanakoyeu et al., 2019]	75.9	88.4	94.9	98.1
FastAP [Cakir et al., 2019a]	76.4	89.0	95.1	98.2
MIC [Roth et al., 2019]	77.2	89.4	95.6	-
Blackbox AP [RolĀnek et al., 2020]	78.6	90.5	96.0	98.7
Cont. w/M [X. Wang et al., 2020]	80.6	91.6	96.2	98.7
Smooth-AP BS=224	79.2	91.0	96.5	98.9
Smooth-AP BS=384	80.1	91.5	96.6	99.0

Table 2.2: **Results on Stanford Online Products.** Deep metric learning and recent AP approximating methods are compared to using the ResNet50 architecture. BS: mini-batch size.

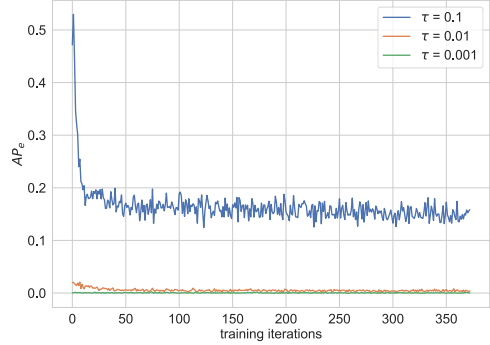


Figure 2.4: **The AP approximation error, AP_e over one training epoch for Online Products for different values of sigmoid annealing temperature, τ .**

2.6.2 Evaluation on VehicleID and INaturalist

In Table 2.3, we show results on the VehicleID and INaturalist dataset. We observe that Smooth-AP achieves state-of-the-art results on the challenging and large-scale VehicleID dataset. In particular, our model outperforms FastAP by a significant 3% for the Small protocol Recall@1. Furthermore, Smooth-AP exceeds the performance of [X. Wang et al., 2020] on 4 of the 6 recall metrics.

As we are the first to report results on INaturalist for *image retrieval*, in addition to Smooth-AP, we re-train state-of-the-art metric learning and AP approximating methods, with the respective official code, *e.g.* Triplet and ProxyNCA [Roth & Brattoli, 2019], FastAP [Cakir et al., 2019b], Blackbox AP [Vlastelica et al., 2020]. As shown in Table 2.3. Smooth-AP outperforms all methods by 2–5% on Recall@1 for the experiments when the same batch size is used (224). Increasing the batch size to 384 for Smooth-AP leads to a further boost of 1.4% to 66.6 for Recall@1. These results demonstrate that Smooth-AP is particularly suitable for *large-scale* retrieval datasets, thus revealing its scalability to *real-world* retrieval problems. We note here that these large-scale datasets (>100k images) are less influenced by hyper-parameter tuning and so provide ideal test environments to demonstrate improved image retrieval techniques.

Table 2.3: **Results on the VehicleID (left) and INaturalist (right).** All experiments are conducted using ResNet50 as backbone. All results for INaturalist are from publicly available official implementations in the PyTorch framework with a batch size of 224. † refers to the recent re-implementation [Roth & Brattoli, 2019] - we make the design choice for Proxy NCA loss to keep the number of proxies equal to the number of training classes. The VehicleID results are obtained with a batch-size of 384.

	VehicleID						INaturalist			
	Small		Medium		Large		1	4	16	32
Recall@K	1	5	1	5	1	5				
Divide [Sanakoyeu et al., 2019]	87.7	92.9	85.7	90.4	82.9	90.2				
MIC [Roth et al., 2019]	86.9	93.4	-	-	82.0	91.0				
FastAP [Cakir et al., 2019a]	91.9	96.8	90.6	95.9	87.5	95.1				
Cont. w/M [X. Wang et al., 2020]	94.7	96.8	93.7	95.8	93.0	95.8				
Smooth-AP	94.9	97.6	93.3	96.4	91.9	96.2				
<i>Recall@K</i>										
Triplet Semi-Hard [C.-Y. Wu et al., 2017]							58.1	75.5	86.8	90.7
Proxy NCA† [Movshovitz-Attias et al., 2017]							61.6	77.4	87.0	90.6
FastAP [Cakir et al., 2019a]							60.6	77.0	87.2	90.6
Blackbox AP [RolĀnek et al., 2020]							62.9	79.0	88.9	92.1
Smooth-AP BS=224							65.9	80.9	89.8	92.7
Smooth-AP BS=384							67.2	81.8	90.3	93.1

Table 2.4: **mAP results on face retrieval datasets.** Smooth-AP *consistently* boosts the AP performance for both VGGFace2 and ArcFace, while outperforming other standard metric learning losses (Pairwise contrastive and Triplet).

VGGFace2	VF2 Test	IJB-C	ArcFace	VF2 Test	IJB-C
Softmax	0.828	0.726	ArcFace	0.858	0.772
+Pairwise	0.828	0.728	+Pairwise	0.861	0.775
+Triplet	0.845	0.740	+Triplet	0.880	0.787
+Smooth-AP	0.850	0.754	+Smooth-AP	0.902	0.803

2.6.3 Evaluation on Face Retrieval

Due to impressive results [Q. Cao et al., 2018b; J. Deng et al., 2019], face retrieval is considered saturated. Nevertheless, we demonstrate here that Smooth-AP can further boost the face retrieval performance. Specifically, we append Smooth-AP on top of modern methods (VGGFace2 and ArcFace) and evaluate mAP on IJB-C and VGGFace2, *i.e.* one of the largest face recognition datasets.

As shown in Table 2.4, when appending the Smooth-AP loss, retrieval metrics such as mAP can be significantly improved upon the baseline model for both datasets. This is particularly impressive as both baselines have already shown very strong performance on facial verification and identification tasks, yet Smooth-AP is able to increase mAP by up to 4.4% on VGGFace2 and 3.1% on ArcFace. Moreover, Smooth-AP strongly outperforms both the pairwise [Chopra et al., 2005] and triplet [Schroff et al., 2015] losses, *i.e.* the two most popular surrogates to a ranking loss. As discussed in Section 2.4.1, these surrogates optimise a distance metric rather than a ranking metric, and the results show that the latter is optimal for AP.

τ	mAP	$ \mathcal{P} $	mAP	B	mAP
0.1	0.824 / 0.726	4	0.844 / 0.736	64	0.824 / 0.726
0.01	0.844/0.736	8	0.833 / 0.734	128	0.844 / 0.736
0.001	0.839 / 0.733	16	0.824 / 0.726	256	0.853/0.754
$ \mathcal{P} = 4, B = 128$		$\tau = 0.01, B = 128$		$\tau = 0.01, \mathcal{P} = 4$	

Table 2.5: Ablation study over different parameters: temperature τ , size of positive set during mini-batch sampling $|\mathcal{P}|$, and batch size B . Performance is benchmarked on VGGFace2-Test and IJB-C.

2.6.4 Ablation study

To investigate the effect of different hyper-parameter settings, *e.g.* the sigmoid temperature τ , the size of the positive set $|\mathcal{P}|$, and batch size B (Table 2.5), we use VGGFace2 and IJB-C with SE-Net50 [Q. Cao et al., 2018b], as the large-scale datasets are unlikely to lead to overfitting, and therefore provide a fair understanding about these hyper-parameters. Note that we only vary one parameter at a time.

Effect of sigmoid temperature τ . As explained in Section 2.4.1, τ governs the smoothing of the sigmoid that is used to approximate the indicator function in the Smooth-AP loss. The ablation shows that a value of 0.01 leads to the best mAP scores, which is the optimal trade-off between AP approximation and a large enough operating region in which to provide gradients. Surprisingly, this value (0.01) corresponds to a small operating region. We conjecture that a tight approximation to true AP is the key, and when partnered with a large enough batch size, enough elements of the difference matrix will lie within the operating region in order to induce sufficient re-ranking gradients. The sigmoid temperature can further be viewed from the margin perspective (inter-class margins are commonly used in metric learning to help generalisation [Chopra et al., 2005; Oh Song et al., 2016; K. Weinberger et al., 2006]). Smooth-AP only stops providing gradients to push a positive instance above a negative instance once they are a distance equal to the width of the operating region apart, hence enforcing a margin that equates to roughly 0.1 for this choice of τ .

Effect of positive set $|\mathcal{P}|$. In this setting, the positive set represents the *instances* that come from the same class in the mini-batch during training. We observe that a small value (4) results in the highest mAP scores, this is because mini-batches are formed by sampling at the class level, where a low value

for $|\mathcal{P}|$ means a larger number of sampled classes and a higher probability of sampling hard-negative instances that violate the correct ranking order. Increasing the number of classes in the batch results in a better batch approximation of the true class distribution, allowing each training iteration to enforce a more optimally structured embedding space.

Effect of batch size B . Table 2.5 shows that large batch sizes result in better mAP, especially for VGGFace2. This is expected, as it again increases the chance of getting hard-negative samples in the batch.

2.6.5 Further discussion

There are several important observations in the above results. Smooth-AP outperforms all previous AP approximation approaches, as well as the metric learning techniques (pair, triplet, and list-wise) on *three* image retrieval benchmarks, SOP, VehicleID, Inaturalist, with the performance gap being particularly apparent on the large-scale *INaturalist* dataset. Similarly, when scaled to face datasets containing millions of images, Smooth-AP is able to improve the retrieval metrics for state-of-the-art face verification networks. We hypothesise that these performance gains upon the previous AP approximating methods come from a tighter approximation to AP than other existing approaches, hence demonstrating the effectiveness and scalability of Smooth-AP. Furthermore, many of the properties that deep metric learning losses handcraft into their respective methods (distance-based weighting [Song et al., 2016; C.-Y. Wu et al., 2017], inter-class margins [Chopra et al., 2005; Schroff et al., 2015; Song et al., 2016; X. Wang, Hua, et al., 2019], intra-class margins [X. Wang, Hua, et al., 2019]), are naturally built into our AP formulation, and result in improved generalisation capabilities.

2.7 Conclusions

We introduce *Smooth-AP*, a novel loss that directly optimizes a smoothed approximation of AP. This is in contrast to modern contrastive, triplet, and list-wise deep metric learning losses which act as surrogates to encourage ranking. We show that Smooth-AP outperforms recent AP-optimising methods, as

well as the deep metric learning methods, and with a simple and elegant, plug-and-play style method. We provide an analysis for the reasons why Smooth-AP outperforms these other losses, *i.e.* Smooth-AP preserves the goal of AP which is to optimise ranking rather than distances in the embedding space. Moreover, we also show that fine-tuning face-verification networks by appending the Smooth-AP loss can strongly improve the performance. Finally, in an effort to bridge the gap between experimental settings and real-world retrieval scenarios, we provide experiments on several large-scale datasets and show Smooth-AP loss to be considerably more scalable than previous approximations.

2.7.0.1 Acknowledgements.

We are grateful to Tengda Han, Olivia Wiles, Christian Rupprecht, Sagar Vaze, Quentin Pleple and Maya Gulieva for proof-reading, and to Ernesto Coto for the initial motivation for this work. Funding for this research is provided by the EPSRC Programme Grant Seebibyte EP/M013774/1. AB is funded by an EPSRC DTA Studentship.

Appendices

Appendices can be accessed online.¹

Statement of Authorship

A statement of authorship for this work can be found in Appendix [A](#).

¹<https://arxiv.org/pdf/2007.12163.pdf>

3 | End-to-End Visual Editing with a Generatively Pre-Trained Artist

Andrew Brown¹ Cheng-Yang Fu² Omkar Parkhi²
Tamara L. Berg² Andrea Vedaldi^{1,2}

¹VGG, Oxford ²Meta AI Research

Abstract

We consider the targeted image editing problem, namely blending a region in a source image with a driver image that specifies the desired change. Differently from prior works, we solve this problem by learning a conditional probability distribution of the edits, *end-to-end* in code space. Training such a model requires addressing the lack of example edits for training. To this end, we propose a self-supervised approach that simulates edits by augmenting off-the-shelf images in a target domain. The benefits are remarkable: implemented as a state-of-the-art auto-regressive transformer, our approach is simple, sidesteps difficulties with previous methods based on GAN-like priors, obtains significantly better edits, and is efficient. Furthermore, we show that different blending effects can be learned by an intuitive control of the augmentation process, with no other changes required to the model architecture. We demonstrate the superiority of this approach across several datasets in extensive quantitative and qualitative experiments, including human studies, significantly outperforming prior work.

Published in the Proceedings of the European Conference on Computer Vision, 2022.



Figure 3.1: E2EVE combines a driver and source image (resp. to the left and right of the $\oplus$ symbol), generating a new version of the source that resembles the driver in the edit region (marked in blue). The generated output looks realistic while faithfully resembling the driver. Our method can be trained to work well on different types of images, including bedrooms, dresses, and faces, and can use regions of arbitrary shape, from rough rectangles to pixel-accurate segmentations (bottom-right).

3.1 Introduction

A key part of the creative process is the ability to combine known factors in novel ways. For instance, we can imagine how a dress would look like with a different v-neck, or our bedroom would look like with the large windows we have seen in a magazine. In this paper, we thus consider the problem of generating new variants of a source image, guided by another image containing a feature, such as a component of a dress or window style, that we wish to change in the source. For additional control, we wish the edit operation to focus on a particular target region of the source, leaving the context as unchanged as possible while maintaining the realism of the edit (see fig. 3.1).

Prior works consider image editing tasks, but often guided by a *textual* description of the desired change [Ramesh et al., 2022; Nichol et al., 2021; Gafni et al., 2022]. We argue that specifying edits visually rather than textually offers a far more fine grained and explicit level of control, ultimately resulting in a more useful editor¹. Formally, we can describe the editing process as drawing a sample from a conditional image distribution $P(\hat{x}|x, y, R)$, where x is the source image, y is the driver image,

¹After all, a picture is worth a thousand words!

R the edit region and $\hat{x}$ is an edited version of the source x . The goal of the edit $\hat{x}$ is to look natural while being close to the source x everywhere except for the region R , where it should resemble the driver image y .

A main challenge in learning the model $P(\hat{x}|x, y, R)$ is the lack of suitable training data, namely quadruplets $(\hat{x}, x, y, R)$, that exemplify the desired mapping. Most authors have thus proposed to focus on learning an unconditional prior distribution $P(x)$ of images, for which abundant training data is usually available, and then seeking an edit $\hat{x}$ that is both likely according to the prior and close in some sense to the source x and the driver y . This can be achieved in a pre-processing [J. Zhu et al., 2020; Chai et al., 2021b] or post-processing stage [Bau et al., 2021], and often uses a Generative Adversarial Network (GAN) to model the prior $P(x)$. Although demonstrating some impressive results, such approaches offer limited control on the edit $\hat{x}$, which either shows only a weak dependency on the driver image y , or does not stay on the image manifold $P(x)$, resulting in undesirable artifacts.

In this work, we overcome these challenges by considering a different approach where we learn the conditional distribution $P(\hat{x}|x, y, R)$ directly, *end-to-end* in code space. For this, we propose new ways of *synthesising* suitable training quadruplets $(\hat{x}, x, y, R)$ on a large scale and without requiring manual intervention. We do this in a self-supervised manner: given an image $\hat{x}$, we select an edit region R at random and use it to decompose the image into source x and driver y images, so that the edit can be written as a (random) function $(x, y, R) = f(\hat{x})$ of $\hat{x}$. A shortcoming is that such x and y are statistically correlated, whereas in a “creative” edit process a user must be able to choose y independently of x . A key contribution is to show that, if the process f is carefully designed, then the resulting images x and y are independent *enough*, meaning that they can be used to learn a high-quality conditional generator $P(\hat{x}|x, y, R)$ which works even when x and y are sampled independently.

We pair this intuition with the adoption of state-of-the-art auto-regressive image modelling using transformers for learning and sampling the distribution $P(\hat{x}|x, y, R)$. Overall, our End-to-End Visual Editor (E2EVE) approach has significant advantages over prior image editing works: (1) E2EVE learns to *directly* map the input to the output representations; (2) hence, based on extensive qualitative, quantitative and human-analysis experiments on several datasets, it results in higher quality edits that are simultaneously more dependent on the driver image and more natural looking than prior works based on GANs

and attention which train task-agnostic priors not necessarily optimal for editing; (3) it is generally easier to implement and tune than GAN-based alternatives; and (4) it is efficient because it allows the direct sampling of edits without involving the expensive pre- or post-processing steps required by some prior methods.

3.2 Related Work

Targeted Generative Image Editing. Approaches for editing images in targeted locations include spatial manipulation of objects [J.-Y. Zhu et al., 2016], adding or removing a closed-set of objects [Bau et al., 2020; Bau, Strobel, et al., 2019], or text-driven manipulation using CLIP [Bau et al., 2021; Shi et al., 2021]. These GAN-based approaches have some downsides: First, they require inverting GANs to represent the input image — a difficult problem [Abdal et al., 2020, 2019; J.-Y. Zhu et al., 2016; Lipton & Tripathi, 2017; Bau, Zhu, et al., 2019; Guan et al., 2020] which can limit the *editability* of the images [Tov et al., 2021; Abdal et al., 2020]. Second, they are not trained end-to-end. Third, text-driven approaches offer limited fine-grained control of shape and texture [Bau et al., 2021]. We address such shortcomings by training a sequence-to-sequence model end-to-end for the task of targeted image-based visual editing.

Note that the targeted image manipulation task which we address, differs from spatially-conditioned generation [Isola et al., 2017; Park et al., 2019; T.-C. Wang et al., 2018; H. Kim et al., 2021; W. Xia et al., 2021; Zhao et al., 2021], or global image manipulation, where an entire image is generated/manipulated [Radford et al., 2016; Karras et al., 2019]. In global manipulation works, interpolations in the latent space are located which correspond to edits over the entire image, either via visual attribute classifiers [Shen et al., 2020; Jahanian et al., 2020; C. Yang et al., 2020; Choi et al., 2018], unsupervised disentanglement [Schwettmann et al., 2021; Härkönen et al., 2020; Voynov & Babenko, 2020; Peebles et al., 2020; Z. Wu et al., 2020], or via image-text similarity [Patashnik et al., 2021; Crowson, n.d.; Galatolo et al., 2021; Schaldenbrand et al., 2021; X. Liu, Park, et al., 2021; Abdal et al., 2021; W. Xia et al., 2021]. Different to in-painting [Iizuka et al., 2017; G. Liu et al., 2018], in our task, generation depends on the driver image as well as image context.

Image Composition. Image composition combines possibly inconsistent images into a single cohesive output. Previous approaches include collaging nearest neighbors [Isola & Liu, 2013], composing foregrounds and backgrounds [Tsai et al., 2017], composing a closed set of visual attributes [Mokady et al., 2018; Press et al., 2019], or using semantic pyramids [Shocher et al., 2020]. Others fuse images by projecting their composition on the manifold generated by a GAN via inversion [Xu et al., 2021; Chai et al., 2021b; Ghosh et al., 2021; Abdal et al., 2020; J. Zhu et al., 2020; Issenhuth et al., 2021]. These methods work well if the driver image is sufficiently aligned to the source (which usually requires manual intervention), but worse than our end-to-end model when this is not the case.

Two Stage Image Synthesis. We adopt state-of-the-art two stage auto-regressive models [Dai & Wipf, 2019; Esser et al., 2020; Z. Xiao et al., 2019; M. Chen et al., 2020; Ramesh et al., 2021b; Esser, Rombach, & Ommer, 2021; Ding et al., 2021] for the image generator. These models scale better than sequence-to-sequence [Radford et al., 2019; Devlin et al., 2019a] models applied directly to pixels by reducing first the dimensionality of images via a discrete autoencoder [van den Oord et al., 2017; Razavi et al., 2019; Fauw et al., 2019; Esser, Rombach, & Ommer, 2021]. Others have recently built on this work for text- or class-driven image manipulation [Crowson, n.d.; C. Wu et al., 2021; Esser, Rombach, Blattmann, & Ommer, 2021], whereas we consider image-driven editing. [Z. Zhang et al., 2021] shows a (single) qualitative result for image-based out-painting using BERT [Devlin et al., 2019a], where a small set of tokens in the output sequence is fixed (termed *preservation controls*), and paired training data is sourced from the same image. We do not solve the out-painting problem, but the targeted editing problem, with the added challenge of preserving context while mixing images with very different statistics.

3.3 Method

We wish to learn a model that can “naturally” blend a given source image x with a user-provided driver image y . Formally, we denote with $x, \hat{x} \in \mathbb{R}^{3 \times H \times W}$ the source and output (RGB) images and with $y \in \mathbb{R}^{3 \times H' \times W'}$ the driver image (where, usually, $H > H'$ and $W > W'$). Furthermore, we target the edit operation on a region $R \in \{0, 1\}^{H \times W}$, expressed as a binary mask. We cast the problem as

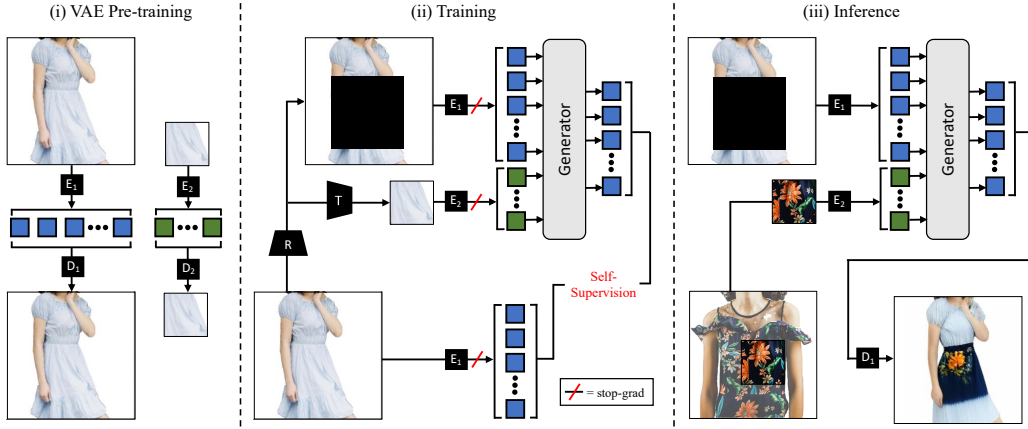


Figure 3.2: The E2EVE approach: (i) **VAE Pre-Training**: we train two quantized VAEs (one for whole images, and one for image patches), each consisting of an Encoder, E , and decoder, D . (ii) **Training**: Each data sample produces a masked source image (via the operation R), and driver image (via random transformation T on the masked region). Given these conditioning inputs, the model is self-supervised to predict the data sample. Following prior work, the VAEs are kept frozen while training the generator. (iii) **Inference**: E2EVE generates edited images when the source and driver are sampled independently from different images.

one of learning a conditional probability distribution $P(\hat{x}|x, y, R)$ and then sample the output image $\hat{x}$ conditioned on the source image x , the driver image y , and the edit region R (fig. 3.1).

Next, we discuss the advantages and requirements of this approach (section 3.3.1), propose a self-supervised learning formulation that does not require any manually-provided labels to train the model (section 3.3.2), and give the technical details of the neural network that learns the conditional distribution (section 3.3.3). An overview of our training and inference settings is shown in fig. 3.2.

3.3.1 End-to-End Conditional Generation

Our approach is to learn a conditional distribution $P(\hat{x}|x, y, R)$ *end-to-end*, which we do by means of an auto-regressive transformer network discussed in section 3.3.3. In order to train such a model, we require training quadruplets $(\hat{x}, x, y, R)$ sampled from the joint distribution $P(\hat{x}, x, y, R)$. Each of these quadruplets represents the outcome of a “creative” process, where a human artist combines images x and y to generate a new image $\hat{x}$. Because obtaining such training data would require the intervention of human artists, it would be very difficult to obtain a sufficiently large dataset to learn

the required conditional distribution. Hence, much of the research in image editing focuses on how to avoid this bottleneck and use instead data which is readily available.

A popular approach is to consider an *indirect* formulation and learn instead an unconditional image distribution $P(x)$, for instance expressed as a GAN generator $x = G(z)$. Then, the output image $\hat{x} = G(z^*)$ is “sampled” via an optimization process such as $z^* = \operatorname{argmin}_z d(G(z)|_R, y)$ where $d(\hat{x}|_R, y)$ measures compatibility between the region R of the generated image $\hat{x}$ and the driver image y . The advantage is that the model G can be learned from a collection $\mathcal{X}$ of unedited images $x \sim P(x)$, which is often easy to obtain at scale. The disadvantage is that this model is not optimized for the final task of image editing.

By contrast, in our approach we learn directly the model $P_\theta(\hat{x}|x, y, R)$ minimizing the standard negative log-likelihood loss:

$$\theta^* = \operatorname{argmin}_\theta \left(-\frac{1}{|\mathcal{T}|} \sum_{(\hat{x}, x, y, R) \in \mathcal{T}} \log P_\theta(\hat{x}|x, y, R) \right) \quad (3.1)$$

where $\mathcal{T}$ is a large collection of training quadruplets. Once learned, we can directly draw samples $\hat{x} \sim P_{\theta^*}(\hat{x}|x, y, R)$. The main challenge is how to obtain the training set $\mathcal{T}$. The key to our method is a way of constructing $\mathcal{T}$ from $\mathcal{X}$ in an automated fashion, at no extra cost. This is explained in the next section.

3.3.2 Synthesizing a Dataset of Meaningful Edits

Given a training set $\mathcal{X}$ of unedited images x , the goal is to create a dataset $\mathcal{T}$ of “edits” $(\hat{x}, x, y, R)$ consisting of the generated image $\hat{x}$, the source image x , the driver image y , and the edit region R . The difficulty is that these quadruplets should be representative of a “creative” process where the generated image $\hat{x}$ is a meaningful blend of the source and driver images, x and y , according to a human artist. Specifically, $\hat{x}$ should resemble x as much as possible except in the region R , where it should take the character of y , but without introducing unnatural artifacts (e.g., simply pasting y on top of x would not do).

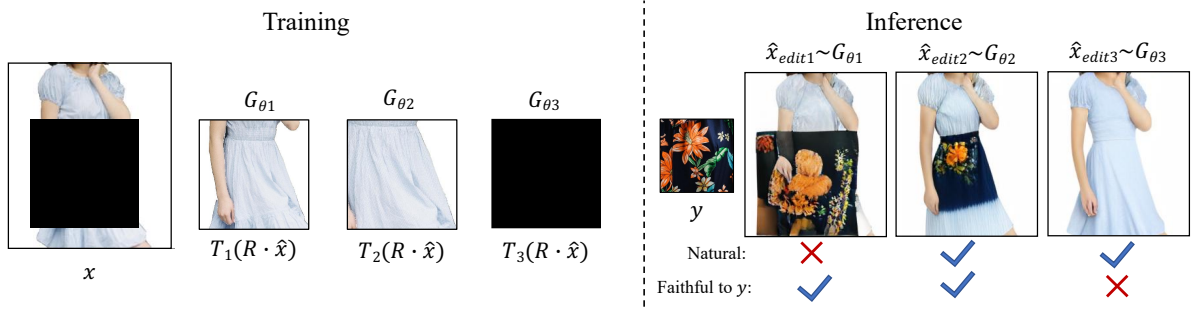


Figure 3.3: Intuition for augmenting training inputs. During training, the driver image y is computed by applying a random transformation T to the masked region of the source image x . Left: Three options for T that are used to train three different generators, G_{θ_n} . Right: Samples from the three trained models when conditioned on x and an independently sampled y . An optimal choice of T removes just enough information that generated images are both natural-looking, and are faithful to the driver image y .

We propose to build such quadruplets as follows (see also fig. 3.4). We sample an output image $\hat{x}$ from the unedited collection $\mathcal{X}$, thus pretending that the latter is, in fact, the result of an edit operation. Then, we define the source and driver images for this virtual edit as follows:

$$x = (1 - R) \odot \hat{x}, \quad y = T(R \odot \hat{x}), \quad (3.2)$$

where R is the mask of a random image region, $\odot$ is the element-wise product (where broadcasting is used as required), and $T : \mathbb{R}^{3 \times H \times W} \rightarrow \mathbb{R}^{3 \times H' \times W'}$ is a *random image transformation*, also known as an “augmentation”.

By optimizing the log-likelihood loss in eq. (3.1), the model $P_{\theta}(\hat{x}|x, y, R)$ learns to predict the full image $\hat{x}$ from x (which misses the region R), and y , which preserves some information about the missing region. Because $\hat{x}$ is originally an unedited image, the model learns to predict a natural-looking output.

The key design choice here is the random transformation T . For example, if we set $T = 1$ to be the identity function, then the output image can be reconstructed exactly as $\hat{x} = x + y$; in this case, the model $P_{\theta}(\hat{x}|x, y, R)$ learns to paste y onto x , which is uninteresting (see model G_{θ_1} in fig. 3.3). Inversely, if we set $T = 0$ to be the null function, then y does not provide any information; in this case,

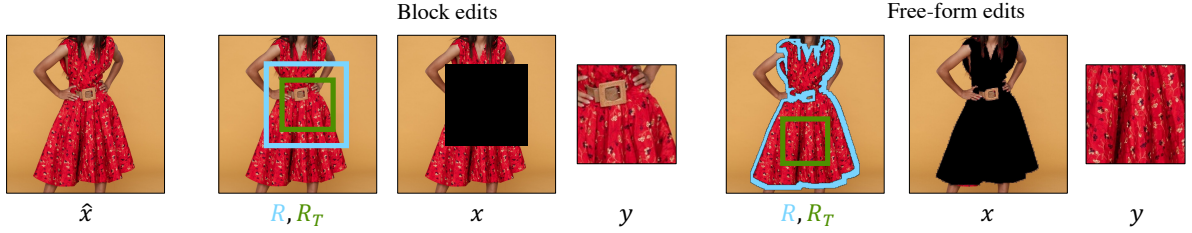


Figure 3.4: *Left: block edits.* We sample an output image $\hat{x}$ and generate the corresponding edit input (x, y, R) by sampling a square region R , a transformation T cutting a sub-region R_T from R , and extracting the source image x and a driver y from these two. *Right: free-form edits.* The same approach, but this time R is a pixel-accurate segmentation mask extracted manually or automatically (R_T remains a square sub-region).

the model $P_\theta(\hat{x}|x, y, R)$ learns to inpaint $\hat{x}$, filling in the missing region in a non-trivial manner, but ignoring the driver image y altogether (see model G_{θ_3} in fig. 3.3). The augmentations T should find a *sweet spot* and remove just the right amount of information from y (see model G_{θ_2} in fig. 3.3). While finding the optimal choice for the random transformations T is ultimately an empirical process, we describe next some important design criteria that were crucial for our results.

Decorrelating source and driver images. A difficulty with our approach is that, because both source image x and driver image y are derived from the same image $\hat{x}$, they are *not* independent but *paired*. This is a problem because the user should be free to choose almost *any* driver image y for editing, so the generator must work well for *unpaired* inputs x and y too. We can approximate this condition by making x and y as uncorrelated as possible during training.

While we cannot achieve this result exactly, we propose two methods to approximate it, *block* and *free-form*, both shown in fig. 3.4. **Block edits** are simple: we let the transformation $T(R \odot \hat{x})$ take a further sub-crop R_T of the crop R it is given as input, thus removing the most direct source of correlation between x and y : spatial continuity. Furthermore, we found empirically that if the sub-crop is always centered in R and of a fixed relative size w.r.t. R , then the model learns, as one would expect, to paste this crop in the middle. Instead, we let T further randomize the position (`pos_augment`) and size (`size_augment`) of the sub-crop relative to the edit region during training. We parameterize the sub-cropping operation via α , which defines the ratio of the sub-crop width, to the edit region width.

The `size_augment` operation allows α to vary during training. The model hence learns to find a meaningful placement for the patch y in the context of x , without assuming spatial continuity or a specific geometric arrangement. Optionally, we further decorrelate source and driver images via **free-form edits**. The difference is that we let R be the output of a semantic segmentation network, while T still takes a square sub-crop R_T from region R . Because y is fully contained in the edit region R and the latter separates a foreground object from the background, this significantly reduces the correlation between x and y . While this approach requires additional machinery (the segmentation network), empirically it can obtain impressive results (fig. 3.5).

Controlling the learned editor. In order to favour generalization, the augmentations above should remove as much information as possible from the crop y *except* for the information that the editor should transfer from the driver image y to the generated image $\hat{x}$. For example, it would be possible to consider further augmentations such as color jitter, but this would cause the editor to learn to ignore the color, which we usually wish to transfer. In general, by choosing different augmentations we can *control* what information the editor learns to transfer from the driver image to the generated one (e.g., style and colour), and what to ignore (e.g., the specific spatial arrangement).

3.3.3 Two-Stage Conditional Auto-Regressive Image Generation

In order to implement the conditional distribution $P_\theta(\hat{x}|x, y, R)$, we use an *auto-regressive (AR) model*. AR models have been shown to be highly expressive for image generation [Ramesh et al., 2021b; Esser, Rombach, & Ommer, 2021], they can be conditioned on multiple signals elegantly and without architectural changes, and, unlike GANs [Salimans et al., 2016], are mode-covering. In practice, this leads to more varied generation results and the ability to model datasets with more variation. We summarise next how this model is applied to our case and point the reader to the supp. mat. for additional details.

The goal is to model a conditional distribution $P(\hat{x}|c)$, where c lumps together all conditioning information. An AR model further decomposes $\hat{x} = (\hat{x}_1, \dots, \hat{x}_M)$ into M components and factorizes the distribution as the product $P(\hat{x}|c) = \prod_{m=1}^M P(\hat{x}_m|\hat{x}_1, \dots, \hat{x}_{m-1}, c)$. The model is trained by minimizing the negative log likelihood (3.1) (avoiding unstable adversarial techniques used in GANs). For

modelling images x , a challenge lies in finding a suitable decomposition, such that the individual factors $P(\hat{x}_m|\hat{x}_1, \dots, \hat{x}_{m-1}, c)$ can be implemented effectively. To this end, we build upon the two stage process of Esser *et al.* [Esser, Rombach, & Ommer, 2021] and use a transformer on top of a discrete autoencoder. Note that the focus of this work is on end-to-end targeted image editing and the training formulation; we describe the two-stage auto-regressive method for completeness and reproducibility.

Specifically, in the first stage we learn a compressed and discretized representation $z = \Phi(\hat{x}) \in \{1, \dots, K\}^M$ of the images, where here K is the size of the discrete encoding space and M the resulting number of discrete tokens. For this, we use the VQ-GAN [Esser, Rombach, & Ommer, 2021], achieving 16-fold image compression (we use separate encoders for x and y). Naturally, the encoder comes with a paired decoder $\hat{x} = \Psi(z)$ that allows to reconstruct the image from the code. This achieves two important goals: it allows (1) to scale the generator model to higher resolution images, and (2) to predict discrete distributions for the second stage.

The second stage uses a transformer to model the factors $P(\hat{x}_m|\hat{x}_1, \dots, \hat{x}_{m-1}, c)$. Specifically, the conditioning information $c = (x, y, R)$ consists of source image x , driver image y and region R . The sequence of tokens $S_m = (z_1, \dots, z_m) \oplus \Phi(x) \oplus \Phi(y)$ (where $\oplus$ denotes concatenation), comprising the partially-predicted output tokens along with the conditioning tokens, is fed to the transformer to output a K -dimensional histogram $P(z_m = \cdot | S_m)$. Spatial encodings are added to the tokens, but there is no need to explicitly encode the region R as it can be inferred from x because $x = (1 - R) \odot \hat{x}$ has a R -shaped ‘hole’. While this way of encoding for R may seem naïve, it is in fact simple and powerful: prior work such as EdiBERT [Issenhuth et al., 2021] use “occlusion tokens”, and hence lose the ability to express pixel-accurate edit regions, which we can do effortlessly.

In practice, we train a GPT-2 [Radford et al., 2019] style transformer. Because the factors $P(\hat{x}_m|\hat{x}_1, \dots, \hat{x}_{m-1}, c)$ require each predicted token to depend only on those prior to it in the sequence, GPT-2 uses causal masking allowing only unidirectional attention towards earlier tokens in the sequence. All factors are trained efficiently in parallel using *teacher forcing* [Williams & Zipser, 1989; A. Goyal et al., 2016]. During inference, only the conditioning information is provided so the target sequence is predicted iteratively, sampling one symbol z_m at a time from the corresponding histogram. In practice, inference is faster than for some GAN alternatives, as shown in the sup. mat.

3.4 Experiments

We compare our method to others that, given a source image x and a driver image y , produce one or more edits $\hat{x}$. Good edits have three properties: (1) *naturalness* (the edit $\hat{x}$ looks like a sample from the prior $P(x)$); (2) *locality* ($\hat{x}$ is close to x outside the edit region R — although a certain amount of slack is necessary to allow the edit to blend in naturally); and (3) *faithfulness* ($\hat{x}$ resembles y within the edit region R). Achieving only one of the three objectives is trivial (for example, setting $\hat{x} = x$ ignoring y is natural and local but unfaithful whereas copying y on top of x is local and faithful but unnatural) so a good model must seek for a trade-off between these properties.

Measuring these properties is not entirely trivial; for the quantitative analysis, we take the standard FID measure for naturalness [Heusel et al., 2017], the L^1 distance $|(1 - R) \odot (\hat{x} - x)|_1$ to measure locality, and a retrieval approach to measure faithfulness. For the latter, we consider a set $\mathcal{Y}_{\text{dstr}}$ of 100 distractor images of the same size as the driver image y , use the edited region $\hat{x}|_R$ as a query, and find its nearest neighbour $y^* = \operatorname{argmin}_{\hat{y} \in \{y\} \cup \mathcal{Y}_{\text{dstr}}} d(\hat{x}|_R, \hat{y})$, incurring the loss $\delta_{y^* \neq y}$. In this expression, $d(\cdot, \cdot)$ is the Inception v3 [Szegedy et al., 2016] feature distance (pre-trained on ImageNet [J. Deng et al., 2009]), which is the same encoder used for FID calculation.

Evaluation data. Recall that our goal is to evaluate the quality of automated editing algorithms. To feed such algorithms, we need triplets (x, y, R) consisting of a source image x , a driver image y and an edit region R . In section 3.3.2 we explained how to build such a dataset for the purpose of *training* our model — with the added complexity that, for training, we *also* need to know the result $\hat{x}$ of the edit process. We could use the same dataset for evaluation, but this would unfairly advantage our model. Instead, since knowing the output $\hat{x}$ is *not* required to measure naturalness, locality and faithfulness, we are free to choose *new* and less constrained triplets (x, y, R) for evaluation, resulting in more challenging edits and a fairer evaluation. However, we wish to avoid too many cases in which cohesive blending is impossible (e.g., where y is a patch of sky and $x|_R$ is a face). Hence, we build evaluation triplets as follows: given a sample image x and an edit region R , we define $y = x'|_R$ to be a crop taken at the same spatial location from a *different* image x' in the dataset. The effect is to (very) weakly constrain $x|_R$ and y to be compatible (e.g., both sky regions, or face regions) by exploiting the

photographer bias in the datasets we consider.

We conduct experiments on three datasets: (1) the private *Dresses-7M* dataset containing 7 million images mainly depicting a woman wearing a dress; (2) LSUN bedrooms [Yu et al., 2015] containing 3M images of bedrooms; and (3) FFHQ [Karras et al., 2019], containing 70K aligned faces. We sample x by considering 1024 images from the validation sets of *Dresses-7m* and FFHQ, and 256 for LSUN bedrooms (due to its small size). For each (x, y, R) , we consider 10 edit samples and obtain naturalness, locality and faithfulness by averaging over all images thus generated (totalling 10,240 and 2,560 samples, respectively). All images in this paper are from UnSplash² (dresses and bedrooms), or DFDC [Dolhansky et al., 2020] (faces).

Implementation Details. We use a transformer architecture with 24 layers, 16 head multi-head attention, embedding size 1024, and we train it using standard cross-entropy loss. The masked source image $(1 - R) \odot x$ and the driver image y are encoded using VQ-GANs with $16\times$ compression and 1024 codebook size. Source x and output $\hat{x}$ images have resolution 256×256 and y resolution 64×64 . The token sequences S_m (comprising the coded x , y and partial $\hat{x}$) has maximum length 516. Our final model uses `pos_augment` and `size_augment`—the latter forms y by taking a sub-crop in the region R , with α varying from 0.4 to 0.7. We use a batch size of 512, and the AdamW [Loshchilov & Hutter, 2019] optimizer with learning rate $4.5e-6$. During inference, for each input (x, y, R) , we generate 20 samples $\hat{x}$ and keep the 10 with highest similarity to y (a method termed *Filter*, in table 3.2). In order to focus the sampling on more realistic/likely outputs we use *nucleus sampling* [Holtzman et al., 2020] with a p-value of 0.9. Following prior work [Esser, Rombach, & Ommer, 2021; Ramesh et al., 2021b], the VQ-GANs are kept frozen when training the transformer.

Baselines. We compare our method against the following image composition baselines: (1) *Copy-paste* generates $\hat{x}$ by pasting y onto x at the specified location R ; (2) *Inpaint* ablates our method by removing the tokens y from the input, thus generating $\hat{x}$ by inpainting the region R unconditionally while disregarding y ; (3) *GAN inv*, inspired by [Issenhuth et al., 2021; Xu et al., 2021; J. Zhu et al., 2020], takes the copy-paste output and uses the StyleGANv2 [Karras, Laine, et al., 2020] or

²www.unsplash.com

Table 3.1: Results for *block-* and *free-form edits*. Naturalness is computed over the whole image (Image), and just the edit-region (Edit-R) using FID. Faithfulness is computed via retrieval, where R@K measures whether or not the sample is retrieved in the top-k instances. Locality is measured using the L^1 distance outside of the edit-region

		Naturalness ($\downarrow$)		Faithfulness ($\uparrow$)			Locality ($\downarrow$)
		Image	Edit-R	R@1	R@5	R@20	(L1)
Dresses-7m (<i>block-edits</i>)	Baseline: Copy-Paste	21.457	35.924	1.000	1.000	1.000	0.000
	Baseline: Inpaint	15.797	25.769	0.071	0.214	0.515	0.095
	EdiBERT [Issenhuth et al., 2021]	17.193	32.621	0.554	0.837	0.963	0.052
	(ours) E2EVE	14.411	24.743	0.797	0.937	0.978	0.056
FFHQ (<i>block-edits</i>)	Baseline: Copy-Paste	33.330	25.811	1.000	1.000	1.000	0.000
	Baseline: Inpaint	18.328	12.665	0.421	0.704	0.895	0.139
	GAN inv [Abdal et al., 2020]: StyleGANv2	26.583	16.223	0.590	0.823	0.948	0.198
	GAN inv [Abdal et al., 2020]: StyleGANv2-Ada	26.657	16.290	0.593	0.821	0.949	0.199
	In-domain [J. Zhu et al., 2020]	19.880	14.270	0.539	0.800	0.938	0.178
	EdiBERT [Issenhuth et al., 2021]	13.192	12.230	0.718	0.925	0.983	0.093
	(ours) E2EVE	12.770	10.574	0.853	0.970	0.994	0.106
LSUN Bedrooms (<i>block-edits</i>)	Baseline: Copy-Paste	24.402	28.828	1.000	1.000	1.000	0.000
	Baseline: Inpaint	15.080	21.493	0.113	0.297	0.596	0.161
	GAN inv [Abdal et al., 2020]: StyleGANv2	23.735	33.530	0.405	0.689	0.866	0.259
	In-domain [J. Zhu et al., 2020]	32.333	43.544	0.171	0.363	0.608	0.208
	EdiBERT [Issenhuth et al., 2021]	16.518	27.528	0.537	0.816	0.946	0.111
	(ours) E2EVE	14.107	22.187	0.789	0.923	0.981	0.119
Dresses-7m (<i>free-form edits</i>)	Baseline: Copy-Paste	23.107	58.259	0.581	0.700	0.817	0.000
	Baseline: Inpaint	13.718	24.516	0.193	0.385	0.659	0.103
	EdiBERT [Issenhuth et al., 2021]	15.277	27.359	0.650	0.843	0.937	0.079
	(ours) E2EVE	14.000	25.973	0.814	0.920	0.951	0.072

StyleGANv2-ADA [Karras, Aittala, et al., 2020] networks to re-encode and thus denoise the resulting image via GAN inversion [Abdal et al., 2020], “blending” the edit; (4) *EdiBERT* [Issenhuth et al., 2021] is a related transformer-based approach, which iteratively refines the output of copy-paste output using BERT [Devlin et al., 2019a] (for fairness, we use the same VQ-GAN and sample filtering by similarity to driver as for our method); (5) *In-Domain GAN* [J. Zhu et al., 2020] uses a regularised form of GAN inversion to blend source and driver images. Pre-trained models are available for all test datasets except *Dresses-7M*; unfortunately, we were unable to successfully train the GAN-based models on the latter (possibly due to the significant diversity of this data), so in this case we limit the other baselines. Some off-the-shelf models are trained on the validation sets that we use for testing, which disadvantages our approach in the comparison. For more details on experimental settings, please see supp. mat.

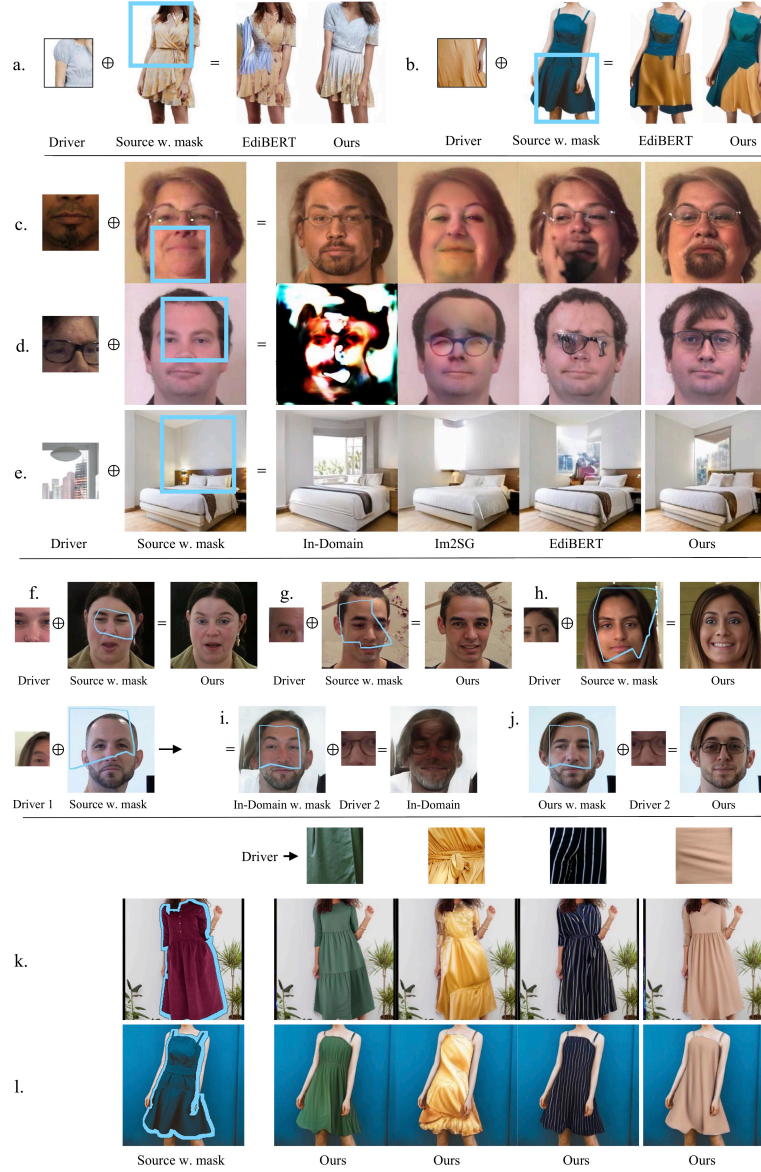


Figure 3.5: Qualitative results from E2EVE. Rows a,b: Block edits from E2EVE trained on *Dresses-7M*. Row c,d: Comparisons to prior work trained on FFHQ. Row e: Comparisons to prior work trained on LSUN-Bedrooms. Rows f,g,h: Edits from E2EVE on FFHQ using random masks of increasing size. Rows i,j: one edit after another (sequential) on the same source image using prior work, and E2EVE, respectively. Rows k,l: *free-form* edits from E2EVE trained on *Dresses-7M*. Please zoom in for details. In each case, the masked region in the source image is that contained within the blue line.

3.4.1 Quantitative Evaluation

Block edits. Table 3.1 reports the evaluation metrics for all baselines and datasets. Our approach significantly outperforms others in **naturalness**: because our method is trained explicitly with the goal of blending source and driver images, it works even for cases where the images are poorly aligned, where prior works based on fitting priors on unedited images fail (see fig. 3.5).

The copy-paste baseline outperforms other methods on the **faithfulness** metric but has very poor naturalness — this is expected as the edited image contains a 1-to-1 copy of the driver image. The opposite is true for the inpainting baselines, which attain good naturalness but very poor faithfulness as they ignore the driver image altogether. Our method is second only to copy-paste in faithfulness while also scoring best in naturalness. Other baselines sit somewhere in between, but generally do not fair very well in faithfulness because, by projecting the composite image to the prior manifold, they distort the cue too much.

As for **locality**, copy-paste is also optimal, as it does not change the context region at all. Compared to non-trivial baselines, our method is first or second best in this metric, affecting the context region much less than GAN methods. However, EdiBERT is also very competitive as it is designed to leave the context nearly exactly unchanged (via *periodic collage* of the output with the context). However, relaxing locality is often necessary to obtain a more reasonable blending effect — an intuitive fact that we show qualitatively in fig. 3.5f.

Finally, we also conduct a **human-study** on the *Dresses-7M* dataset using Amazon Mechanical Turk. We showed 256 edit samples using our method and EdiBERT to 3 human assessors each, asking two questions: which of the two outputs is more realistic, and which is more faithful to the driver image. The results show that by majority vote, human annotators think that our samples are more natural 83.2% of the time, and more faithful 80.5% of the time. The key to the performance improvements are that E2EVE is trained *end-to-end* for targeted visual editing, whereas prior work are not. Furthermore, we outperform EdiBERT, which uses the same VQ-GAN with a 50% larger transformer.

Free-form edits. Here, we use a semantic segmentation network to extract the edit region R from image x , while we define y as a crop taken from within a semantic region of a different image x' (see sec-



Figure 3.6: E2EVE generalises surprisingly well to out-of-domain driver images (e.g. images of weather and nature). Three examples using the same source image from *block-edit Dresses-7M* model. As shown, E2EVE generates cohesive and varied samples.

tion 3.3.2) We conduct experiments on *Dresses-7M* and compare to EdiBERT, with results in table 3.1. Our approach again outperforms the previous methods in terms of **naturalness** and **faithfulness** and, this time, **locality** too, because we can better capture irregular edit regions compared to EdiBERT which results in blocky artifacts. Additional experiments with random free-form masks can be found in the supp. mat.

3.4.2 Qualitative Evaluation

We show qualitative comparisons against prior work in fig. 3.5. Our edits combine naturalness, faithfulness and locality, whereas others fail at achieving all three goals as well as we do. Due to the augmentations in our training edits, our method is better able to cope with uncorrelated driver images y than other approaches that only rely on a pre-learned unconditional prior distribution $P(x)$. For example, in fig. 3.5c,d our approach can successfully mix images coming from faces with different gender or pose, showing better naturalness and faithfulness. As for locality, while EdiBERT is highly competitive in table 3.1, this comes at a cost: in fig. 3.5a,d our method achieves better naturalness by coloring both sleeves in the same way and by completing the glasses even though part of them lie outside of the edit region, whereas EdiBERT cannot. fig. 3.5k,l shows free-form edits where the entire clothing item is masked. Although structural details of the dress are hidden by the mask, E2EVE generates natural and varied structure that is different to the source and faithful to the driver image. In fig. 3.5e E2EVE generates more natural looking samples than prior work that edit with respect to the spatial geometry of the room. We also see that E2EVE generalises surprisingly well to out-of-domain driver images, as shown in fig. 3.6 for the block-edit *Dresses-7M* model. In fig. 3.5f,g,h, E2EVE generates impressive

samples with small, medium or large random free-form masks. In fig. 3.5j E2EVE makes sequential edits with random R s while maintaining naturalness – empirically this fails with prior works (fig. 3.5i). See the supp. mat. for additional results.

3.4.3 Ablations

In table 3.2 we analyse and ablate design choices in E2EVE. We report additional metrics: negative log likelihood (NLL) on the validation set and sample diversity, computed pairwise between samples from the same inputs using LPIPS [R. Zhang et al., 2018].

Starting from the construction of the training edits $\mathcal{T}$ (section 3.3.2), reducing α (rows a-d) means removing more of the image $\hat{x}$ from the crop y . As predicted in section 3.3.2, removing information from y increases naturalness (lower FID) at the expense of weaker faithfulness (lower R). $\alpha = 0.6$ provides a balance. `pos_augment` (row e vs. b) increases naturalness by preventing the model from simply pasting the driver image in the centre of the edit region. `size_augment` (row h) randomizes the choice of α in a range during training, so that the editor learns to automatically resize the driver image as needed. This significantly improves naturalness at the cost of a reduction of faithfulness (row h vs. e). In part, this is likely due to limitations of the retrieval model used to measure faithfulness, which struggles to cope with geometric deformations even when they preserve the style of the driver. α and augmentation have no effect on the locality (rows a-h). A benefit is that our final model generates more diverse samples (row h vs g) by learning to place the driver image at different positions and sizes. Interestingly, NLL is also minimised by the final model despite the fact that y is *less* correlated to $\hat{x}$ than in other cases: this is likely because additional augmentations reduce overfitting to the training data. For further discussion, please see supp. mat.

3.5 Conclusions, Limitations and Future work

We present E2EVE, a new approach for targeted visual image editing. The key innovation is an effective method for self-supervising the model end-to-end, based on only an unlabelled collection of natural images. E2EVE learns a conditional image generator that responds well to diverse user inputs,

Table 3.2: Model ablations and sweeps for *block edits*. Key: NLL: Negative log likelihood. α , pos-aug, size-aug: the parameters used to define the sub-cropping transformation T . Filter: filtering E2EVE samples by visual similarity to the driver image. 2VQ: using two VQ-GANs rather than one. Datasets: D : *Dresses-7m*, B : *Bedrooms*, F : *FFHQ*.

	α	pos-aug	size-aug	Filter	2VQ	Data	Naturalness ($\downarrow$)		Faithfulness ($\uparrow$)			Locality ($\downarrow$)	NLL ($\downarrow$)	Diversity ($\uparrow$)	
							Image	Edit-R	R@1	R@5	R@20	(L1)		Image	Edit-R
a.	0.8	$\times$	$\times$	$\checkmark$	$\checkmark$	D	17.241	30.076	0.882	0.980	0.996	0.056	2.181	0.135	0.309
b.	0.6	$\times$	$\times$	$\checkmark$	$\checkmark$	D	15.593	29.364	0.920	0.986	0.997	0.056	1.704	0.137	0.315
c.	0.4	$\times$	$\times$	$\checkmark$	$\checkmark$	D	13.967	26.975	0.811	0.954	0.988	0.056	1.594	0.139	0.327
d.	0.0	$\times$	$\times$	$\checkmark$	$\checkmark$	D	15.797	25.769	0.071	0.214	0.515	0.095	1.537	0.190	0.419
e.	0.6	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	D	15.605	26.513	0.887	0.980	0.997	0.056	1.518	0.142	0.338
f.	0.5-0.6	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	D	14.951	26.186	0.856	0.968	0.992	0.056	1.460	0.143	0.344
g.	0.4-0.7	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	D	14.589	27.824	0.864	0.970	0.992	0.056	1.494	0.139	0.328
h.	0.4-0.7	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	D	14.411	24.743	0.797	0.937	0.960	0.056	1.448	0.143	0.344
i.	0.4-0.7	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	D	13.913	24.583	0.611	0.817	0.929	0.056	1.448	0.145	0.351
j.	0.4-0.7	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	B	14.347	22.998	0.636	0.831	0.929	0.119	2.942	0.287	0.460
k.	0.4-0.7	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	F	12.636	10.699	0.723	0.899	0.976	0.106	2.392	0.203	0.321
l.	EdiBERT [Issenhuht et al., 2021]			$\times$		B	16.643	29.775	0.356	0.627	0.823	0.111	-	0.291	0.575
m.	EdiBERT [Issenhuht et al., 2021]			$\times$		F	13.036	12.891	0.536	0.778	0.925	0.093	-	0.181	0.423
n.	0.4-0.7	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	D	14.107	23.916	0.720	0.891	0.963	0.056	1.454	0.144	0.347

significantly outperforming prior work qualitatively and quantitatively without any manual supervision. Limitations remain: our data generation technique might be difficult to extend to text-based edits and some proposed edits are unreasonable (see sup. mat.) because the model lacks a full understanding of the semantic content of images. Furthermore, as our model is unsupervised and data-driven, it may contain surprising unwanted biases. Next steps include extending E2EVE beyond images to spatial editing in 3D scenes and spatio-temporal video editing.

Acknowledgements. We are grateful to the advice and support of Yanping Xie, Antoine Toisoul, Thomas Hayes, and the EdiBERT authors.

Appendices

Appendices can be accessed online.³

Statement of Authorship

A statement of authorship for this work can be found in Appendix A.

³<https://arxiv.org/pdf/2205.01668.pdf>

Part II

Audio-Visual Human-Centric Video Understanding

4 | Playing a Part: Speaker Verification at the Movies

Andrew Brown^{*1} Jaesung Huh^{*1} Arsha Nagrani^{*1}

Joon Son Chung² Andrew Zisserman¹

¹VGG, Oxford ²Naver Corporation

^{*}Equal Contribution

Abstract

The goal of this work is to investigate the performance of popular speaker recognition models on speech segments from movies, where often actors intentionally disguise their voice to play a character. We make the following three contributions: (i) We collect a novel, challenging speaker recognition dataset called `VoxMovies`, with speech for 856 identities from almost 4000 movie clips. `VoxMovies` contains utterances with varying emotion, accents and background noise, and therefore comprises an entirely different domain to the interview-style, emotionally calm utterances in current speaker recognition datasets such as `VoxCeleb`; (ii) We provide a number of domain adaptation evaluation sets, and benchmark the performance of state-of-the-art speaker recognition models on these evaluation pairs. We demonstrate that both speaker verification and identification performance drops steeply on this new data, showing the challenge in transferring models across domains; and finally (iii) We show that simple domain adaptation paradigms improve performance, but there is still large room for improvement.

Published in the Proceedings of the International Conference on Acoustics Speech and Signal Processing, 2021.

4.1 Introduction

All the world's a stage, and all the men and women merely players; They have their exits and their entrances, and one man in his time plays many parts.

- W. Shakespeare, *As You Like It*

After hearing the actor Steve Martin's smooth American accent on the Ellen show, is it possible to recognise that he is the voice behind 'Inspector Jacques Clouseau', the comical French character in the movie 'Pink Panther', without access to his visual appearance? Or recognise Anne Hathaway's voice in the movie 'Les Misérables', where she plays 'Fantine', singing through tears about her sadness and desperation? In this paper we investigate the challenging task of speaker recognition for actors from two different domains – the first being when they are speaking '*naturally*' in interviews, and second while *playing a part* in a movie, where they may be intentionally modifying their voice in order to play the role of a character or show emotion.

While recent years have shown great successes in speaker recognition [Snyder et al., 2018; Garcia-Romero et al., 2020; Wan et al., 2018; Chung, Huh, Mun, Lee, et al., 2020a], these successes have been reliant on the collection of large, labelled datasets such as VoxCeleb [Nagrani et al., 2017; Chung et al., 2018; Nagrani, Chung, Xie, & Zisserman, 2020] and others [McLaren et al., 2016; Fan et al., 2020]. The VoxCeleb datasets, while valuable, have been collected *entirely from interviews* of celebrities in YouTube videos and are limited in terms of linguistic content (celebrities mostly speak about their professions [Nagrani, Chung, Albanie, & Zisserman, 2020]), emotion, and background noise. In contrast, movies contain speech covering emotions such as anger, sadness, assertiveness, and fright, and varied background conditions – imagine the shouting in a violent scene from an action movie, or a romantic scene of reconciliation in a romcom. As we show in this paper, models trained on VoxCeleb, when applied to a novel domain such as speech in movies, suffer from significant degradation in performance. In order to accurately measure this, there is a compelling need for real-world datasets and evaluation sets across these domains. Collecting and annotating datasets for every new domain encountered in the real-world, however can be an extremely expensive and time-consuming process. We introduce a scalable method to automatically generate data in a new domain (movies), and investigate the performance

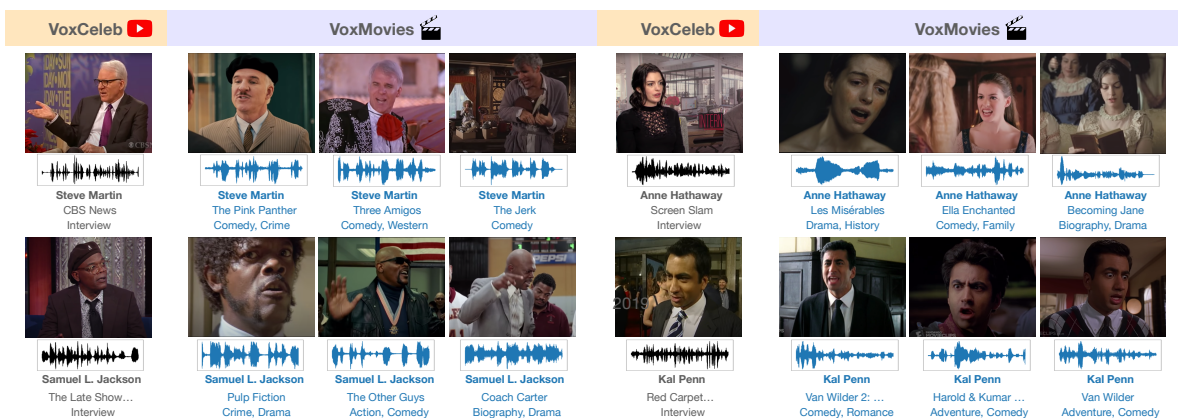


Figure 4.1: Domain Gap between VoxCeleb and VoxMovies: Unlike VoxCeleb (left in each panel), which consists of utterances from interviews, VoxMovies is sourced from movies of different genres (right in each panel). While VoxCeleb features utterances largely with calm, unvaried emotions, VoxMovies has challenging emotion and background noise. We show only a single frame for each utterance, below which is the name and genre of the film/video. VoxMovies contains 24 utterances from 5 movies for Anne Hathaway, including singing in the musical, ‘Les Misérables’, reading in an English accent in ‘Becoming Jane’, and arguing heatedly in an American accent in ‘Ella Enchanted’. Samuel L. Jackson has 59 utterances from 14 movies, including his famous, assertive speech in ‘Pulp Fiction’ and bragging as an arrogant policeman in ‘The Other Guys’.

of state-of-the-art speaker recognition models on this data, where actors are intentionally disguising their voice. Being able to detect human identity under such conditions of spoofing is valuable for security and authentication [W. Cai et al., 2017; Z. Chen et al., 2017; Das et al., 2019], and as shown by psychology studies [Reich & Duke, 1979; Hirson & Duckworth, 1993], is a challenging task even for humans.

VoxMovies has been used to create a challenging test set for the VoxCeleb Speaker Recognition Challenge [Nagrani, Chung, Huh, et al., 2020a] (VoxSRC2020)¹. Data: <https://www.robots.ox.ac.uk/~vgg/data/voxmovies/>.

In order to encourage research in domain adaptation for speaker recognition, we make the following three contributions: (i) We collect a novel speaker recognition dataset called VoxMovies, from 3,792 popular movie clips uploaded to YouTube. Our dataset consists of almost 9,000 utterances from 856 identities that appear in the VoxCeleb datasets, and contains challenging emotional, linguistic and channel variation (Fig. 4.1); (ii) We provide a number of domain adaptation evaluation sets, and bench-

¹<http://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2020.html>

mark the performance of state-of-the-art speaker recognition models on these evaluation pairs. We demonstrate that performance drops steeply on this new data for both speaker verification *and* identification, showing the challenge in transferring models across domains (from interviews to movies). We also investigate performance on positive pairs sampled across different movies, and reveal further performance drops; and finally (iii) We demonstrate that domain adaptation approaches added on top of already trained models improve performance, but there is still a severe degradation.

4.2 Related Work

Due to the reliance of machine learning on large, labelled datasets, domain adaptation has become popular in fields such as computer vision [Tzeng et al., 2017; Hoffman et al., 2018], text classification [P. Liu et al., 2017; Y. Zhang et al., 2017], speech enhancement [Liao et al., 2018; Meng et al., 2018] and speaker verification [Garcia-Romero & McCree, 2014; Garcia-Romero et al., n.d., 2014]. Recent interest in domain adaptation for speaker recognition has largely focused on boosting the performance on datasets such as NIST-SRE16 or Speakers in the Wild (SITW) [McLaren et al., 2015], which does not have a large training set. In this case, most methods train on VoxCeleb [Nagrani et al., 2017; Chung et al., 2018] and evaluate on these evaluation sets.

Recent methods focus on adversarial training techniques, with [Q. Wang et al., 2018] introducing domain adversarial training that exploits domain prediction, and [Luu et al., 2020] proposing a channel adversarial training method by adding a channel discriminator to mitigate the domain mismatch problem by using the video labels in VoxCeleb, with both works [Luu et al., 2020; Q. Wang et al., 2018] using gradient reversal layers. [Rohdin et al., 2019] proposes an end-to-end domain adversarial training method adopting the architecture of Wasserstein Generative Adversarial Network (GAN) [Arjovsky et al., 2017] with adversarial domain loss and cross-entropy loss. It significantly improves the performance on language adaptation. [Bhattacharya, Alam, & Kenny, 2019] also introduces a domain adaptation technique to new language or recording conditions by using a Domain Adversarial Neural Speaker Embeddings (DANSE) model which contains a 1-dimensional self-attentive residual block. [Bhattacharya, Monteiro, et al., 2019] explores various GANs to make the domain discriminator unable to distinguish whether embeddings are from the source or target domain. [Kang et al., 2020]

exploits Model-Agnostic Meta-Learning (MAML), projecting speaker representations to a generalized embedding space and achieves better results on CNCeleb dataset [Fan et al., 2020]. Other research methods train channel-invariant, noise-robust speaker recognition models which can help improve the overall performance of the model in diverse domains. [Zhou et al., 2019] shows a multi-task learning method that trains both a speaker recognition network and a discriminator that distinguishes types of noise in speaker representations. [Chung, Huh, & Mun, 2020] also tackles a similar problem by using an environment network and a KL-divergence based confusion loss to learn speaker-discriminative, environment-invariant representations.

Unlike existing works, we provide novel domain data for the *same identities as in VoxCeleb*, allowing us to investigate both cross-domain speaker verification, where pairs have one segment from either domain, and cross-domain identification, where speaker identification models are tested on a new, previously unseen domain. Both have not been previously possible. We focus on the domain adaptation from interviews to movies for male and female actors (explored for faces in [Nagrani & Zisserman, 2018]). We propose and benchmark on such evaluations conditions, and additionally on several within-domain verification tasks.

4.3 Cross-Domain Data

Our goal in this work is to investigate the effects of cross-domain speaker verification in movies. The domain change we focus on here is from YouTube interviews (domain $D-I$), to speech in different genres of movies (domain $D-M$). The data for the two different domains are sourced as follows:

$D-I$: Interviews from VoxCeleb [Nagrani et al., 2017; Chung et al., 2018] These datasets are sourced solely from interviews uploaded to YouTube. These are mainly in studio, outdoor or red-carpet locations. In turn, and due to the often *professional* context, voices are mainly calm and rarely show any strong emotion. These utterances are degraded with real world noise that would be expected from these environments, such as background chatter or laughter.

$D-M$: Movies from CMD [Bain et al., 2020] (**VoxMovies**) For the second domain, we curate a dataset of speech from movies, called `VoxMovies`, which consists of 8,905 utterances for 856 different identities, sourced from 3,792 video clips from 1,452 movies. These movies cover a range of genres (see Fig-

ure 4.2). The utterances in `VoxMovies` are sourced from the Condensed Movies dataset (CMD) [Bain et al., 2020], which covers the *key scenes* from movies. The distinctive change of domain can be seen in the following characteristics of `VoxMovies`:

(1) Emotion: In line with different movie genres, the utterances cover emotions such as anger, sadness, assertiveness, and fright. Furthermore, the videos in CMD represent scenes that are integral to the story-line and the different character developments, such as a fight between two main characters, or when they make up later in the film. Hence the utterances in the dataset often capture the most emotional parts of each movie.

(2) Background noise: Each key scene in the CMD dataset covers many different settings, from a loud basketball stadium in *Coach Carter*, to an 18th century gathering in *Becoming Jane* (see Figure 3.1). This represents a far more varied set of degradation for each of the utterances. Also, there is often background music.

Importantly, in `VoxMovies` this variety of emotion and background noise is seen both within and across different identities. Firstly this is because on average each identity has utterances from 2.7 different movies (see Table 4.1), and these movies are likely to be of different genres. Secondly, the videos in this dataset show the important character arcs of these identities within each movie, where they show different emotions at different points in the story-line. Examples and further details can be found in Figure 3.1.

Note that the utterances in `VoxMovies` are all from identities that are represented in `VoxCeleb1` and `VoxCeleb2`. We create an evaluation set, `VoxMovies-(Test)`, featuring identities from `VoxCeleb1`. We also provide a small amount of `VoxMovies` data for training domain adaptation methods, `VoxMovies-(Train)`, using a subset of the identities in the `VoxCeleb2` dev set. There is no identity overlap between these partitions. Statistics are shown in Table 4.1.

4.4 Dataset Collection Pipeline

Our dataset collection pipeline is similar to the one used to collect the `VoxCeleb` datasets, albeit applied to YouTube clips of movie scenes from the Condensed Movies Dataset (CMD) [Bain et al., 2020], described below.

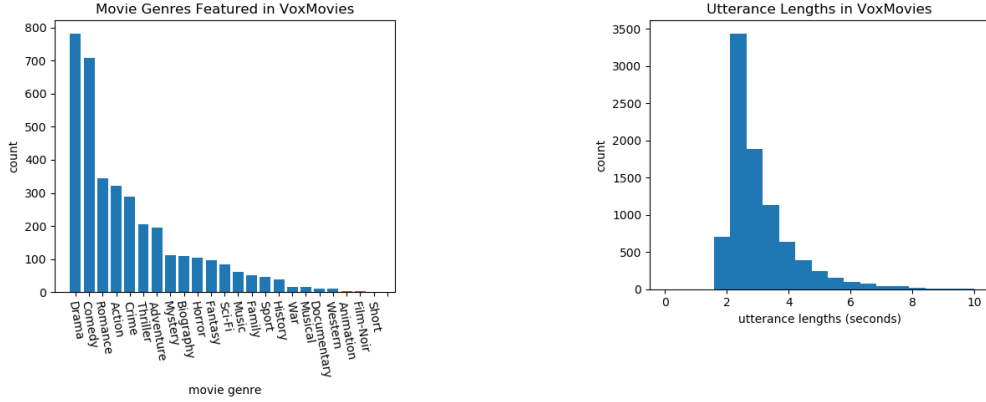


Figure 4.2: **Statistics of the VoxMovies dataset.** (a) The different movie genres that the utterances in VoxMovies are sourced from, and (b) the distribution of utterance lengths. The minimum length is 2 seconds by design choice.

Partition	ID Source	IDs	# Utter.	Clips / ID	Movies / ID
VoxMovies-(Test)	VoxCeleb1	485	4,943	5.1	2.7
VoxMovies-(Train)	VoxCeleb2	371	3,962	5.0	2.5

Table 4.1: **Dataset Statistics for VoxMovies.** The utterances are sourced from the Condensed Movies Dataset [Bain et al., 2020], which contains clips of key scenes from movies. Identities (IDs) overlap with the VoxCeleb1 (test) and VoxCeleb2 (dev) datasets.

Condensed Movies Dataset. CMD [Bain et al., 2020] consists of key scenes from over 3K movies, totalling 1,270 hours. Provided alongside the dataset are cast lists for each of the featured movies, face-tracks for each of the clips, and face embeddings for discriminating identity for each of these face-tracks. The cast lists give the names of people who are likely to appear in the clip. Our three stage method for collecting VoxMovies is as follows:

Stage 1. Sourcing candidate names. We compute the intersection between the VoxCeleb1 and VoxCeleb2 identities and the CMD cast lists.

Stage 2. Face verification. In this stage, face-tracks from the CMD dataset are classified as to whether they depict any of the candidate names from the previous stage using a three step process. (1) Example face images for each of the candidate names for the train and test set are sourced from the VGGFace2 [Q. Cao et al., 2018b] and VGGFace1 [Parkhi, Vedaldi, & Zisserman, 2015] datasets (these face datasets contain the same identities as the VoxCeleb2 and VoxCeleb1 datasets respectively). (2)

256D Embeddings are taken from the final hidden layer of a SE-Net50 architecture CNN for each of the face images, trained for discriminating face-identity using the VGGFace2 dataset [Q. Cao et al., 2018b]. For each identity, these embeddings are average pooled across all example face-images and L2 normalised, leaving one embedding per identity. (3) Verification is performed by computing the cosine similarity between the identity embeddings and the face-track embeddings from CMD. Any face-track with a similarity to an identity embedding above a high threshold is assigned that name.

Stage 3. Active speaker verification. Here we identify which face-tracks from the previous stage are speaking. This is performed using an audio-visual synchronisation network [Chung et al., 2019], which predicts the correlation between the audio track and the motion of the mouth, and outputs a synchronisation confidence score for each 5 frame window. A window with confidence above a high threshold is classified as speaking. Face-tracks with a minimum of 2 seconds of consecutive speech are kept, and the audio track is clipped to the speaking period. See Figure 4.2 for a histogram of segment lengths.

Discussion. A manual check of VoxMovies-(Test) reveals that the automated process has a precision of $> 99\%$. The $< 1\%$ false positive face verifications, or false positive active speaker verifications are manually removed. The high thresholds for stages 2 and 3 are chosen as 0.55 and 0.13, respectively. These values (both cosine similarity) achieved high precision on a manually labelled validation set. As shown in Table 4.1, VoxMovies-(Test) has slightly more utterances than VoxMovies-(Train). This is due to the fact that a larger proportion of the identities in VoxCeleb1 are well-known actors than in VoxCeleb2, and hence they appeared in more of the movies in CMD. Impressively, on average, each identity has utterances from 2.7 different movies, with Robert DeNiro appearing in 25 movies.

4.5 Experiments

4.5.1 Evaluation Tasks

Our goal is to determine the performance of state-of-the-art models trained on the VoxCeleb2 dev set for speaker recognition performance on data from the movie domain ($D-M$).

Eval. set	Positives Source	Negatives Source	# Utter.	# Positive Pairs	# Negative Pairs
E-1	<i>D-M (same)</i>	<i>D-M</i>	20,572	10,286	10,286
E-2	<i>D-I, D-M</i>	<i>D-I, D-M</i>	46,578	23,289	23,289
E-3	<i>D-I, D-M</i>	<i>D-I</i>	46,804	23,402	23,402
E-4	<i>D-I, D-M</i>	<i>D-M</i>	46,866	23,433	23,433
E-5	<i>D-M (diff)</i>	<i>D-M</i>	41,090	20,545	20,545

Table 4.2: **Statistics for the different evaluation sets:** Our evaluation sets are sourced from two different domains, interview material (*D-I*) and movie material (*D-M*). Key: **Utter.:** Utterances., *D-M (same)*: Segments in a pair are sourced from the same movie, *D-M (diff)*: Segments sourced from different movies.

Verification: We use the VoxCeleb1 and VoxMovies-(Test) datasets for evaluation, as these sets have no overlap with the identities in the VoxCeleb2 dev set (which is used to train all baselines). Given the two domains (*D-I* and *D-M*), we note that there are three different ways that pairs can be sourced for evaluation: pairs can be sourced entirely from *D-I*, entirely from *D-M*, or from both, where one utterance is from *D-I* and one is from *D-M*. For positive pairs sourced from *D-M*, we can add the further constraint that both utterances must come from the same movie (*D-M (same)*) or from different movies (*D-M (diff)*). We use these options to create 5 challenging evaluation sets E1-E5 (Table 4.2), of increasing difficulty. More details are provided in Section 4.6.

Identification: We also demonstrate domain mismatch from VoxMovies and VoxCeleb, with speaker identification. Here we add a single linear layer on baseline models trained on VoxCeleb2 dev set for verification. The task is to fine-tune this layer with the cross-entropy loss using utterances from 485 speakers in the VoxCeleb1 dataset (using only the training data from the identification split, see Table 5 in [Nagrani et al., 2017]). We then test performance using the VoxCeleb1 identification test data (in-domain test set) and the utterances from VoxMovies (out of domain test set).

Evaluation Metrics: For verification, we report equal error rate (**EER**- %) and minimum detection cost (**minDCF**) with $C_{miss} = 1$, $C_{fa} = 1$ and $P_{target} = 0.05$. The formula of minimum detection cost function is the same as the one used in NIST SRE [“NIST 2018 Speaker Recognition Evaluation Plan”, 2018 (accessed 31 July 2020)] and the VoxSRC2020 evaluations. For identification, we report top 1 and top 5 % accuracy.

Eval set	I-vec. [Dehak et al., 2010]		X-vec. [Snyder et al., 2018]		Thin-R34 [W. Cai et al., 2018]		Thick-R34 [Heo et al., 2020]	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF
VoxCb1†	5.53	0.336	3.30	0.220	2.05	0.166	1.05	0.084
VoxCb1-H†	9.13	0.467	5.82	0.352	4.37	0.283	2.39	0.154
VoxCb1-E†	5.55	0.338	3.35	0.221	2.27	0.164	1.22	0.086
VoxSRC20*	11.07	0.566	8.22	0.450	6.35	0.374	3.79	0.213
E-1	16.6	0.727	12.92	0.665	9.72	0.562	6.09	0.365
E-2	17.91	0.822	14.75	0.712	10.58	0.610	7.40	0.423
E-3	18.91	0.917	13.58	0.806	10.58	0.666	7.50	0.484
E-4	19.83	0.872	21.56	0.845	12.52	0.737	9.23	0.579
E-5	21.5	0.913	17.97	0.861	14.11	0.760	10.47	0.574

Table 4.3: **Baseline results on various VoxMovies test evaluation pairs.** We show the performance of 4 popular state-of-the-art speaker recognition models. † Cleaned versions of these evaluation pairs from the VoxCeleb1 dataset. *VoxSRC2020 validation set¹.

4.5.2 Baseline models

We compare four baseline models to investigate the performance on our dataset. All models are trained on the VoxCeleb2 dev set.

- 1. I-vector** [Dehak et al., 2010]: Following the Kaldi [Povey et al., 2011] VoxCeleb recipe v1, we extract 400D features, followed by Probabilistic LDA (PLDA) scoring.
- 2. X-vector** [Snyder et al., 2018]: Following the Kaldi [Povey et al., 2011] VoxCeleb recipe v2, we train a x-vector model with a PLDA back-end to extract 512D features.
- 3. Thin ResNet-34** [W. Cai et al., 2018]: Consists of 34 residual blocks (one-fourth of the channel dimensions of original ResNet-34 [K. He et al., 2016b]), with self-attentive pooling. [Chung, Huh, Mun, Lee, et al., 2020a] trains this network with an angular variant of the prototypical loss. We use the pre-trained model which is publicly available [Chung, Huh, Mun, Lee, et al., 2020a].
- 4. Thick ResNet-34** [Heo et al., 2020]: This has double the number of channels in Thin ResNet-34 and uses attentive statistical pooling to model higher order statistics such as standard deviation. The model is trained with both angular prototypical loss and vanilla softmax loss to improve performance. We use the pre-trained model which is publicly available [Chung, Huh, Mun, Lee, et al., 2020a]. This model currently represents the state-of-the-art on the VoxCeleb1 test sets. 512D features are extracted

Evaluation set	Top1 accuracy	Top5 accuracy
VoxCeleb1-test†	89.47%	97.38%
VoxMovies-(Test)	52.23%	73.31%

Table 4.4: Identification results for 485 identities on the VoxMovies-(Test) set (out of domain) and on the VoxCeleb1 test set (same domain). Note how performance drops steeply on the out of domain test set. † VoxCeleb1 test set for identification.

Eval set	Baseline	FT	S-norm	FT + S-norm
E-1	6.09	5.76	5.89	5.66
E-2	7.40	7.10	7.18	7.03
E-3	7.50	8.38	8.16	8.48
E-4	9.23	7.37	8.03	7.19
E-5	10.47	9.55	10.15	9.35
E-3a	1.15	1.53	1.29	1.58
E-3b	0.87	0.97	0.68	0.98
E-3c	7.72	9.90	9.64	10.23

Table 4.5: **Domain transfer results for Thick ResNet-34.** EER(%) is reported. **FT:** Fine-tuning on the VoxMovies-(Train) set. **S-norm:** Score-normalisation.

for both thick and thin ResNet-34 architectures.

4.5.3 Domain Transfer

In this section, we implement two common domain transfer methods using the *small* amount of data provided in the VoxMovies-(Train) set.

Fine-tuning on a small amount of target domain data. We fine-tune the pretrained Thick-ResNet34 with data from the VoxMovies-(Train) set and the VoxCeleb2 dev set (overlapping speakers only). To decrease the domain gap between the datasets, we always pick one utterance from VoxMovies and another from VoxCeleb2 to form positive pairs in each mini-batch. This forces the model to decrease the distance between embeddings from the same speaker’s utterances, hence reducing the domain gap during training. The network is trained with angular prototypical loss [Chung, Huh, Mun, Lee, et al., 2020a], for 500 epochs using Adam with learning rate of 1e-5. Only the last fully-connected layer is fine-tuned while weights of other layers are fixed.

Score Normalisation. [Matejka et al., 2017] introduces various score normalisation techniques for test conditions with diverse domains. A *cohort* is used to estimate the amount of shift and scale for

normalisation, allowing robust threshold setting. We experiment with the Z-norm, T-norm and S-norm, and find the best performance to be using the S-norm (Sec.2 in [Matejka et al., 2017]). We use the VoxMovies-(Train) set as the cohort (speakers in the VoxMovies-(Train) and (Test) sets are disjoint, which fits the assumption of cohort selection).

4.6 Results

Verification using Baseline Models. The results for the baseline models on the different verification tasks are given in Table 4.3. The change of domain in the VoxMovies evaluation sets offers a significant challenge – the Thick ResNet-34 which achieves an impressive 1.05 EER on the VoxCeleb1 test set can only achieve 6.09 EER on the least challenging set, E-1. When comparing eval sets that share a positives or negatives source (see Table 4.2), several conclusions are made: (1) Verifying the same speaker with cross domain utterances (*D-I*, *D-M* - E-4) is harder than with utterances from the same movie (*D-M (same)* - E-1). Interestingly, this shows that the change in an actor’s voice from a calm interview setting in *D-I* to strong emotions, accents and different background noise in *D-M*, is more challenging for speaker verification than the differing emotions in an actor’s voice within the same movie at different points in the story-line. Furthermore, the same speaker from different movies (*D-M (diff)* - E-5) is harder still, showing that an actor’s voice changes most between different movies. (2) Negative pair verification is hardest when the negatives are both taken from the unseen domain, *D-M* (E-4). This is more difficult than when they are taken from *D-I* (E-3) or *D-I*, *D-M* (E-2), which are of roughly equal difficulty to the Thick ResNet-34. This is largely because the baseline models were trained on *D-I*, so any source of negatives or positives exclusively from that domain will be less challenging.

Identification. We use the Thin ResNet-34 model for identification. Table 4.4 shows identification accuracy on both domains. As expected, the identification accuracy drops significantly from VoxCeleb1-test to VoxMovies-(Test) by 37.24% (top1% acc.). Anne Hathaway is one of the hardest to identify (top1 acc. 29.17% - down from 90% in VoxCeleb1-test), whereas Samuel L. Jackson is easier (acc. 100% drops to 72.88% in movies). This may be due to Hathaway’s multiple accents in her movies (Fig. 3.1).

Domain Transfer. We report the results in Table 4.5. Fine-tuning with VoxMovies-(Train) reduces the EER on most of the evaluation sets. The largest improvement is seen in evaluation sets that showed the worst performance in baseline experiments, namely by 2.04% in E-4 and 1.12% in E-5. Both E-4 and E-5 source negatives from *D-M*, showing that fine-tuning and score normalisation work well when transferring existing models from *D-I* to *D-M*. E-3 on the other hand, which sources negatives from *D-I* shows performance degradation. We conclude that fine-tuning on *D-M* degrades the performance on negative pairs only from *D-I*. To verify this conclusion, we introduce three new evaluation sets, E-3a (positives: *D-I*, negatives: *D-I*), E-3b (positives: *D-I*, negatives: *D-I*, *D-M*), E-3c (positives: *D-M*, negatives: *D-I*) in Table 4.5. The fine-tuned model becomes worse at verification in *D-I* (as shown by the degradation of E-3a). Performance on negatives from *D-I* contribute most to this (as shown by the degradation of E-3c, relative to E-3b).

4.7 Conclusion

In this paper, we provide a novel speaker recognition dataset from movies called VoxMovies which contains almost 9,000 utterances with diverse emotion, accents and background conditions. We demonstrate that state-of-the-art models trained on interview data from VoxCeleb degrade significantly on cross-domain evaluation sets from VoxMovies and while simple domain adaptation techniques boost performance, there is still large room for improvement. We hence encourage the research community to develop new methods and systems for this challenging new domain.

Acknowledgements. We are grateful to Max Bain for his help with the CMD dataset. AB is funded by an EPSRC DTA Studentship, JH by a Global Korea Scholarship, and AN by a Google PhD Fellowship. This work is supported by the EPSRC Programme Grant Seebibyte EP/M013774/1.

Statement of Authorship

A statement of authorship for this work can be found in Appendix A.

5 | Automated Video Labelling: Identifying Faces by Corroborative Evidence

Andrew Brown Ernesto Coto Andrew Zisserman

VGG, Oxford

Abstract

We present a method for automatically labelling all faces in video archives, such as TV broadcasts, by combining multiple evidence sources and multiple modalities (visual and audio). We target the problem of ever-growing online video archives, where an effective, *scalable* indexing solution cannot require a user to provide manual annotation or supervision. To this end, we make three key contributions: (1) We provide a novel, simple, method for determining if a person is famous or not using image-search engines. In turn this enables a face-identity model to be built reliably and robustly, and used for high precision automatic labelling; (2) We show that even for *less-famous people*, image-search engines can then be used for *corroborative evidence* to accurately label faces that are named in the scene or the speech; (3) Finally, we quantitatively demonstrate the benefits of our approach on different video domains and test settings, such as TV shows and news broadcasts. Our method works across three disparate datasets without any explicit domain adaptation, and sets new state-of-the-art results on all the public benchmarks.

Published in the Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval, 2020.

5.1 Introduction

There has been an exponential growth in the volume of video content (in the form of TV and film material) being produced and released online. Such content is rich with useful information for researchers, historians and the general public. However, the sheer scale of the data, coupled with a lack of relevant metadata, makes indexing, analysing and navigating this content an increasingly difficult task. Relying on additional, manual human annotation is no longer feasible, and without an effective way to navigate these videos, this bank of knowledge is largely inaccessible.

Interestingly, most video indexing and analysis is *human-centric*. For example, we might want to navigate to a scene where two particular people interact, or where a group of people first appear together. This is partly because many videos in online archives are centred on humans, but also because of our natural interest in human actions and interactions. The focus of this paper is therefore the labelling of all faces in videos, in a way that does not require any additional manual annotation from a user, be it in the form of provided transcripts [Everingham et al., 2006; Everingham, Sivic, & Zisserman, 2009; Parkhi et al., 2020; Sivic et al., 2009; Bauml et al., 2013], or a list of appearing people known *a-priori* [Parkhi et al., 2012; Nagrani & Zisserman, 2017b]. Our approach is hence scalable to large video archives where collecting manual annotations is infeasible.

To solve this task, we take a human-inspired approach. Imagine that you are watching a video and encounter a new person. In order to confidently identify them, you would first look for clues of their name either in the video such as text on the screen, their name being mentioned in speech, or in a list of cast members from an internet archive. You might then find some evidence to verify that this name is correct, by searching for the person online. In this work we follow this approach by harnessing the freely-available weak annotations on the internet, such as IMDB names lists and image-search engine rankings, to provide evidence for recognising faces automatically and with confidence. Consequently, this method is applicable to identities for whom there are images online.

We denote people with many images of themselves online as *famous*, and introduce a novel approach for automatically identifying if an identity is *famous* without any additional manual supervision or annotation. For identities that are *less-famous* according to our approach, we present a novel method



Figure 5.1: Modern, large, unlabelled online video archives, such as TV broadcasts, are growing at an exponential rate. These important resources are inaccessible due to the lack of annotations. Our proposed method for automated video labelling is scalable to these large archives, as it does not require additional manual supervision to label faces in videos. Examples from the BBC Videos and Sherlock dataset are shown above. The method is able to label a wide range of people, over different domains, lighting conditions and in extreme poses.

for using other identifying clues in the video together with image-search engine results as *corroborating evidence* to provide confident person labels.

Beyond visual content, videos also contain identity information in the audio track, as humans can be recognised from the sound of their voice. Occasionally when used independently, neither face-appearance or voice can provide a confident identity label, be it due to a slightly obscured face, or a noisy or brief audio signal [Nagrani & Zisserman, 2017b]. By fusing the two modalities and using a separate query expansion step, we show faces can be labelled with confidence, such that the recall of the automated labelling is improved without sacrificing precision. Figure 5.1 gives example results of our automated labelling method.

In summary, the task of this paper is to assign identity labels (‘tags’) to all people in a set of test videos, without the use of any additional manual annotation beyond what is freely and automatically available on the internet. The base unit for labelling used in this paper is the face-track *i.e.* face-detections from consecutive frames of the same identity that are linked together within shots. The method consists of

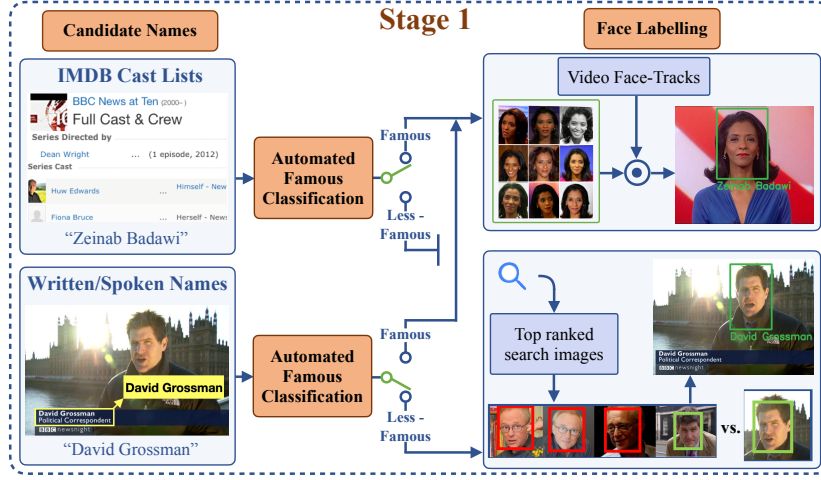


Figure 5.2: The first stage of our method automatically finds candidate names and labels faces. Candidate names are automatically sourced from IMDB name lists, displayed text (*written names*) and spoken words. Each person is classified as *famous* or not. For the *famous* people, face-identity models are automatically assembled and they are labelled throughout the videos. For the *less-famous* people who were found in written or spoken names, the temporal occurrence combined with *corroborating evidence* from image-search engines is used to provide labels.

two key stages:

Stage 1 – Using image-search engines as sole/corroborating evidence. This stage automatically identifies candidate names that may appear in the test videos and classifies them as *famous* or not. Image-search engines are used as the sole source of evidence if they are *famous*, or as *corroborating evidence* if they are not in order to tag the faces. This is presented in Section 5.3 and shown in Figure 5.2.

Stage 2 – Boosting the number of tags. This increases the number and variety of face tags across the test videos, using two techniques: (i) fusing the information from the modalities of face-appearance and voice; and (ii) query expansion. These techniques improve label recall, while maintaining very high precision (see Section 5.4).

Our method for person labelling can proceed completely automatically starting from just the names of the programmes being labelled (and will obtain lists of appearing names in the process). We can

also test specific modules of the method on standard benchmarks where cast lists are provided. We quantitatively demonstrate the benefits of our approach on several different video domains and test settings, such as TV shows and news broadcasts, as described in Section 5.5. The results in Section 5.6 show that our method works across these disparate datasets without any explicit domain adaptation, and sets new state-of-the-art results on all the public benchmarks. Further details can be found at https://www.robots.ox.ac.uk/~vgg/research/person_id_in_video/.

5.2 Related Work

Labelling People in Videos: This task has been well studied [Nagrani & Zisserman, 2017b; Parkhi et al., 2012; Everingham et al., 2006; Everingham, Sivic, & Zisserman, 2009; Parkhi et al., 2020; Bauml et al., 2013; Haurilet et al., 2016], due to its uses for story understanding [Bain et al., 2020] and archive indexing. In previous works, different levels of prior information are assumed to be available: Everingham *et al.* [Everingham et al., 2006] make use of transcripts aligned with subtitles to provide weak supervision, with many other works following suit [Cinbis et al., 2011a; Cour et al., 2010; Parkhi, Rahtu, & Zisserman, 2015; Sivic et al., 2009; Tapaswi et al., 2012; Everingham, Sivic, & Zisserman, 2009; Parkhi et al., 2020; Bauml et al., 2013]. Often the task is posed as one of Multiple Instance Learning (MIL) [Haurilet et al., 2016; Parkhi et al., 2020; Köstinger et al., 2011; Wohlhart et al., 2011]. Nagrani and Zisserman [Nagrani & Zisserman, 2017b] instead presume the existence of cleaned web-downloaded images for actor-level supervision. Many of these methods are tasked with labelling only a small number of known, main characters, and either require transcripts or some additional manual annotation in the pipeline. Our method, on the other hand, does not require either, and so crosses the domain gap to real-world large video archive scenarios.

The labelling of people in news videos is also a well studied and challenging task, due to its open-ended nature, where a list of appearing people or transcripts is not available [Satoh & Kanade, 1997; N. Le et al., 2017; J. Yang et al., 2005]. Both Canseco *et al.* [Canseco et al., n.d.] and Mauclair *et al.* [Jousse et al., 2009; Mauclair et al., 2006] establish correspondences between spoken names and speech-turns to label people. Several works use the co-occurrence of overlaid names in a scene with speech-turns to make person labels [Poignant et al., 2014; Gay et al., 2014; Bendris et al., 2013; Tian et al., 2019],

with much of this area of research accelerated by the MediaEval “Person Discovery in Broadcast TV” challenges [Poignant et al., 2017]. However, many of these techniques rely upon heuristics based on TV-broadcast structure to make confident labels. These heuristics do not generalise well to other domains. Our method on the other hand does not rely on such heuristics, and can hence be applied to movies, TV material and broadcast news alike.

Using Image-Search Engines for Supervision: Although widely used in the Vision Community for object recognition [Chatfield & Zisserman, 2012; Fergus et al., 2005; Schroff et al., 2007; Berg & Forsyth, 2006; W.-H. Lin et al., 2003], and more recently for face recognition [Nagrani & Zisserman, 2017b; Parkhi et al., 2012; Parkhi, Vedaldi, & Zisserman, 2015; D.-D. Le & Satoh, 2008; Holub et al., 2008; Pham et al., 2010; Z. Chen et al., 2019; A. Brown et al., 2020], the problem faced is that retrieved results have varying precision. Previous person-identification work has either relied on manually removing false positives from retrieved images [Nagrani & Zisserman, 2017b], or has focused on automatically improving the precision [Parkhi, Vedaldi, & Zisserman, 2015; D.-D. Le & Satoh, 2008; Holub et al., 2008], starting with a pre-determined list of well-known people. Both are infeasible in real-world video archive scenarios. In this work, to our knowledge, we present the first method for automatically determining the usefulness of a set of search-results, before automatically removing outliers from the retrieved results of those deemed useful, resulting in a completely automatic, scalable and high-precision method.

5.3 Stage 1 – Using image-search engines as sole/corroborating evidence

In this section, we describe our approach for using image-search engines to obtain sole or *corroborating evidence* for labelling people in videos. As a precursor, there are three cases for the usefulness of image-search engines depending on the candidate name: if the person is *famous*, many images of them will be found online (retrieved results will have high precision); for the *less-famous* people, there may be one or two images of them returned; and for the *never-famous* people there will be none (retrieved results have zero recall).

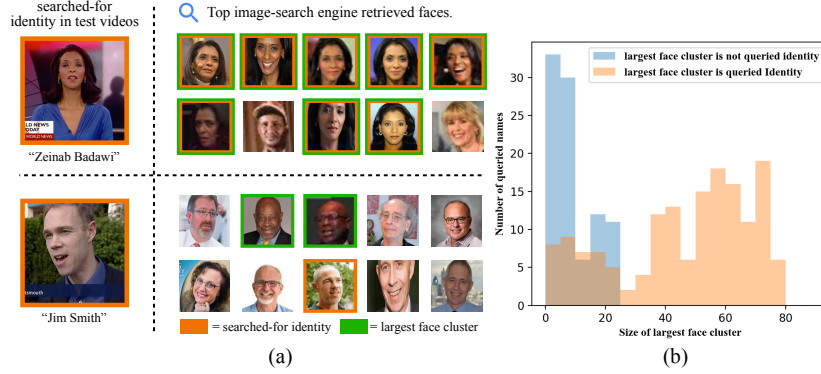


Figure 5.3: (a) Examples of the top-10 retrieved image-search engine results for two different candidate names. For the “Zeinab Badawi” query (top row), the large face cluster (shown in green) has 8 faces, and all depict the queried person (shown in orange). The actual cluster for “Zeinab Badawi” in the top-100 retrieved images has 76 faces. For the “Jim Smith” query (bottom row), the largest face cluster is far smaller with 2 images, and does not depict the searched-for person. The observation is made that if the largest face cluster has many images, then it will likely contain the searched-for person. (b) A plot of whether the largest face cluster contains the searched-for identity for 350 randomly chosen candidate names found for the BBC Videos dataset. Clearly, if the largest face cluster has many faces in it (*e.g.* more than 30 in the top-100) then it is very likely to contain the searched-for person.

This stage of the method proceeds in two steps. The first is **obtaining candidate names**. The names of appearing people are not known *a-priori* and therefore must be first automatically gathered before any labelling can commence. For video material like TV news broadcasts or movies, we use three different sources: (i) IMDB names lists (ii) Scene text: names appearing in text displayed on screen during the programmes, such as overlaid banners. (iii) Speech: names appearing in the speech of the audio track. This step results in a set of candidate names for each episode. Details of the fully automatic processing to obtain these candidate names are given in Section 5.5.1 and the Supplementary Material.

The second step is **labelling famous and less-famous people**. Given a candidate name, we first determine if the person is *famous* or not using a novel method explained below based on downloaded images from an image-search engine. We then proceed in one of two ways: (1) For each *famous* name, a face model is built from the downloaded images and used as the sole evidence to label that person throughout the video; or (2) For the *less-famous* people, we use the temporal occurrence of their spoken or written (displayed) name in a video as primary evidence, and any single occurrence of them in

retrieved image-search engine results as *corroborating evidence*. There is no labelling method for the *never-famous* people, as image-search engines provide no examples of their appearance. The following describes these methods in detail.

Sourcing Candidate Names from IMDB: The name lists are freely and automatically obtained starting from just the name of the programme (which often can be found automatically in video metadata). These lists do not constitute curated cast lists as they often contain thousands of names of briefly appearing characters. This differs from previous non-scalable work on automated face labelling, e.g. [Nagrani & Zisserman, 2017b; Parkhi, Vedaldi, & Zisserman, 2015; Parkhi et al., 2012], which use a curated list of *famous* appearing names.

Classifying Candidate Names as *famous*: When a candidate name is queried in the image-search engine, we use the key observation that if many of the top-ranked retrieved results correspond to the same person, then this person is *famous*. This observation is illustrated in Figure 5.3. In detail, faces are detected in the top 100 ranked results, and clustered using Agglomerative-Clustering [Jain & Dubes, 1988] on their L2-normalised face embeddings [Q. Cao et al., 2018b] (using a cosine-distance threshold of 0.7). If the largest face-image cluster has more than α faces (in this work we use $\alpha = 30$, learnt on a validation set as described in Section 5.6.1, and also illustrated in Figure 5.3), then the identity is classified as *famous*.

Building a Face-Identity Model For the *famous* People: For the candidate names that are classified as *famous*, we simply build a biometric model for that identity and use it to label any face-tracks depicting that person in all test videos, as shown in Figure 5.2. We take the largest cluster of face-embeddings from the downloaded images and average-pool and L2-normalise them into a single embedding. This single embedding is now a face-identity model that can be used for labelling. Taking only the embeddings from the largest cluster serves the purpose of removing false positives from the downloaded images. Face-tracks in the test videos are then labelled by measuring the cosine similarity between their embeddings and the face-identity embedding. If a face-track embedding has a similarity score higher than a threshold learnt on a validation set, then it is labelled with the famous name.

Finding Corroborating Evidence for *less-famous* People: When a candidate name is not classified as *famous*, there may still be a few images of the person in the downloaded images (e.g. the bottom

example in Figure 5.3). These low-precision images cannot serve as the sole evidence as was the case for the *famous* people, but can serve as *corroborating evidence*. Hence, for *less-famous* people we use the temporal occurrence of their spoken or written name as primary evidence of their appearance, and then a single correct retrieved face from image-search engines as the necessary *corroborating evidence* to label. In detail, the *corroborating evidence* is that at least one of the 20 top ranked faces from an image-search engine matches the face-track that appeared in the test video when the name was found. The face-track is then labelled with that name. We are here using 1-to-1 face verification. This is less accurate than the template-based (many-to-one) face verification that we use for the *famous* names, however seeing as the evidence is supported by the presence of the name in the scene or audio track, it is sufficient here.

5.4 Stage 2 – Boosting the Number of Tags

In this section, we describe the two methods used for boosting the number and variety of tags in the test videos. This includes fusing the evidence sources of face-appearance and speech, as well as query expansion.

5.4.1 Fusing face-appearance and voice

This section explains how we use additional information from the speech modality as corroborative evidence to label further face-tracks when the face-appearance alone is not enough to make a confident tag. For each tagged face-track, we use Active Speaker Detection [Chung & Zisserman, 2016; Chung et al., 2019] (ASD), to classify whether the face is speaking. For the speaking faces of each tagged person, we extract temporally aggregated speaker embeddings [Chung, Huh, Mun, Lee, et al., 2020b] using the overlapping audio segments, which after average pooling form a speaker model for that identity. For the remaining un-labelled speaking face tracks in the video, we compute the similarity score between the speaker embedding and the speaker ID models (voice score), and the similarity score between face and the face ID models (face score). We then simply average the two scores (fusion score), and label the face-track if it is above a given threshold. We find empirically that the simple rule of averaging the

two scores is highly effective. For any speaking face-track to be incorrectly classified with the fusion score, both the voice and the face score needs to be high for the same, incorrect identity. We see during experiments that this is very rarely the case, due to the lack of coupling between the modalities. In the supplementary material we present experiments with more complex architectures and show that this very simple rule achieves comparable performance. In the next sections we refer to this method as “Stage 2 Fusion”.

5.4.2 Query Expansion

Query expansion (QE) is a popular re-ranking method [Chum et al., 2007; Arandjelović & Zisserman, 2012; Buckley et al., 1995; Salton & Buckley, 1999]. Methods assume top ranking instances to be from the same class as the query, and use these to supplement the original query to create a new, superior ranking. Nagrani and Zisserman [Nagrani & Zisserman, 2017b] perform QE by training a new classifier for each identity with their top ranked test-video tags, and show it to be helpful for crossing the domain gap between the search engine face-images and the TV-material face-tracks (similarly explored in [A. Brown, Huh, et al., 2021] for voice). We adopt the same technique in this work, except we do not train any additional parameters at this stage, but instead just average-pool all tagged face-image embeddings to form a new face-identity model, which is then used to make further tags. In the next sections we refer to this method as “Stage 2 QE”.

5.5 Datasets, Evaluations and Implementation

In this section we first describe the datasets and the evaluations used for assessing the method, and then give some details on implementation. The video labelling method can proceed completely automatically given only the programme names. This ‘plug and play’ automation is assessed in experiments on two datasets: BBC Videos and MediaEval. We also test out Stage 2 of the method on the standard person identification benchmark, Sherlock, where cast lists and corresponding face-images are provided. Statistics of the three datasets used are given in Table 5.1. Different test protocols are used, when either testing the whole method, or just Stage 2, so that previous published methods can be compared to.

Table 5.1: Dataset Statistics and information on which parts of the method are tested on each of the datasets. Annotations are either provided at the face-track level (BBC Videos, Sherlock) or at the scene level (MediaEval).

Dataset	BBC Videos	MediaEval [Poignant et al., 2017]	Sherlock [Nagrani & Zisserman, 2017b]
No. Identities	66	1,971	31
No. Annotations	1,971	6,889	5,246
No. Videos	5	79	3
No. Hours	2.2	49.1	4
Test Stage 1	✓	✓	
Test Stage 2 - Fusion	✓	✓	✓
Test Stage 2 - QE	✓	✓	✓

BBC Videos: This dataset consists of five episodes of different BBC television programmes (BBC news, BBC World News, Newsnight, Question Time). The challenging dataset for identification provides annotations for all *human-identifiable* characters, from road-side interviewees to well-known politicians. These are the people whose names are alluded to somewhere in the episode, or who are well-known *i.e.* which a human-annotator with access to the internet could annotate. This includes most people barring audience members, pedestrians, etc. Only the names of the programmes are provided to the method. The first episode constitutes the validation set.

MediaEval [Poignant et al., 2017]: This dataset featured in the MediaEval 2015 challenge “Multi-modal person discovery in broadcast TV” [Poignant et al., 2017]. The test set [Bredin, 2015] consists of 79 episodes from the French TV news show, “*Le 20 heures*”, with a total of 6,889 speaking faces annotated at the scene level. Only the name of the programme is provided to the method.

Sherlock [Nagrani & Zisserman, 2017b]: This dataset consists of three episodes of the crime drama show “Sherlock”, where the face-tracks depicting main characters have been annotated. Each episode is approximately 80 minute long. A cast list and face-identity models (in the form of image-search engine images) are provided for each of the annotated characters, and the task is to classify each face-track by identity. The fast-paced show contains many visually challenging scenes (*e.g.* dark, quick camera movement), making it a difficult task for person labelling. For fair comparison to previous works, the face-tracks provided with the Sherlock dataset are used in the experiments.

Evaluation Metrics: When starting from the programme name alone, the BBC Videos and MediaEval experiments constitute open-set retrieval tasks, and so are evaluated using retrieval metrics: Precision,

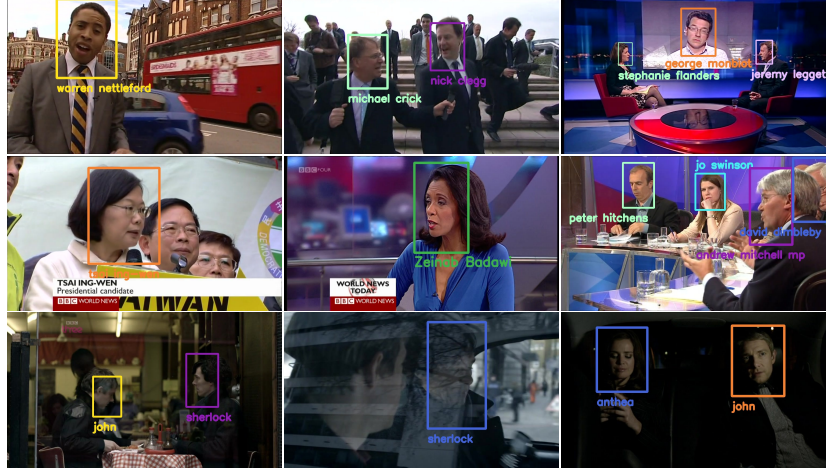


Figure 5.4: Correctly labelled faces from BBC Videos (top, middle row) Sherlock (bottom row). These visually disparate datasets set challenging scenarios for person labelling, such as low resolution, lighting and extreme poses. BBC Videos: (left) Face labels obtained using *corroborating evidence* from image-search engines. (middle) Talking faces labelled by the Fusion step. (right) The QE step labels small, occluded faces, and extreme poses. Sherlock: From left to right, extreme poses, occlusion from glare, and dark faces.

Recall, mAP, class recall – a measure of how many of the total classes have had 1 instance correctly retrieved. The dataset testing Stage 2 alone (Sherlock) provides a closed-set classification task, and so is evaluated using classification accuracy.

5.5.1 Implementation Details

For automatic preparation of each dataset, we compute face-tracks (from face-detections [Chung & Zisserman, 2016]) and speech-turns (linked to faces using Active-Speaker Detection (ASD) [Chung & Zisserman, 2016] for each test-video. Additionally, we extract approximate transcripts using Automatic Speech Recognition [Afouras, Chung, & Zisserman, 2020] (ASR) and any scene-detected-text is found using Optical Character Recognition (OCR) techniques [D. Deng et al., 2018; Y. Liu et al., 2018]. Full details on all video pre-processing method are given in the suppl. material.

Face and speech embeddings. For face-tracks and speech-turns we use pre-trained embeddings to perform face [Q. Cao et al., 2018b] or speaker [Chung, Huh, Mun, Lee, et al., 2020b] verification, respectively. For face-tracks, embeddings from each face detection are average pooled and L2-normalised

into a single embedding. For speech turns, temporal average pooling of the features along the time domain produces a single utterance-level embedding.

5.6 Results

In this section, we first investigate the automatic method of classifying whether a name is *famous* or not, and then evaluate either both stages, or just Stage 2, on the three datasets.

5.6.1 Determining if people are *famous* or not

Here we investigate the choice of the *famous* classification parameter α on candidate names for the BBC Videos dataset. A high α (> 25) means that it is likely that the faces in the largest cluster of downloaded images correspond to the searched-for person. This results in very confident face-identity models, and subsequent perfect face-track precision levels (Figure 5.5a) when these models make correct tags. However, a high α also means that the *famous* classification is limited to very well-known people. This results in low face-track recall (Figure 5.5b), as many of the candidate names are then not classified as *famous* and so are not tagged. A low α (< 25) leads to many names being classified as *famous*. This leads to a high face-track recall (> 0.85 for Stage 1, and > 0.93 for Stage 2) because face-identity models are built for many people in the videos. However, Figure 5.3 shows how at low α , the largest cluster does not always depict the searched-for person, so face-identity models become polluted with false positives. This leads to poor face-track precision as incorrect tags are made with bad face-identity models. Stage 2 QE (blue in Figure 5.5a/b) then creates new face models from incorrect tags, worsening the problem (precision < 0.6 for $\alpha < 10$).

The chosen value of $\alpha = 30$ reflects the balance between achieving high face-track recall, whilst ensuring high precision. This value selects 1,967 *famous* names from a total of 2,906 sourced from IMDB. The IMDB names lists are not curated cast lists, and so can contain hundreds of names irrelevant to this task. For written and spoken names, 129 are classified as *famous* and 170 as *less-famous*. α is not influenced by the video being labelled and so remains constant.

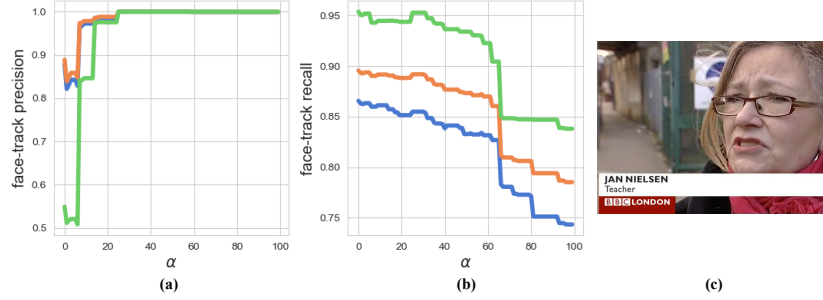


Figure 5.5: Analysis of the *famous* threshold α on the BBC Videos dataset. (a) the precision of the face-track tags as α increases. (b) the face-track recall as α increases. The colors are: blue for Stage 1, orange for Stage 2 Fusion only, and green for Stage 2 Fusion + QE. (c) A missed tag from the BBC Videos dataset (see Section 5.6.2). This figure is optimally viewed in colour.

Table 5.2: BBC Videos dataset results.

BBC Videos			
Method	Face-Tracks		Class Recall
	Precision	Recall	
Stage 1	1.0	0.860	0.91
Stage 1 + (Stage 2 Fusion only)	1.0	0.894	0.91
Stage 1 + (Stage 2 QE only)	1.0	0.934	0.91
Stage 1 + (Stage 2 Fusion + QE)	1.0	0.953	0.91

5.6.2 BBC Videos

The BBC Videos dataset is used to test the full automated pipeline (Stages 1 and 2). This dataset was annotated exclusively for this research. This means that we cannot compare to prior work. Instead we use the dataset to perform a quantitative analysis of the different stages of the method. Results are shown in Table 5.2. For each face-track either the correct label is assigned, the incorrect label is assigned, or we refuse to predict a label as no model has sufficient confidence.

Stage 1: The intended design choice is to only present correctly labelled face-tracks, so the classifier threshold (for tagging faces in the test videos) is chosen on the validation set such that we make no labelling mistakes. This results in a precision of 1.0 across all episodes, meaning that no people were incorrectly tagged, whilst achieving a high face-track recall of 0.86 . A class recall of 0.91 indicates

that 61 of the 66 people in the dataset were correctly tagged at least once (the missed classes are those written or spoken names for whom no *corroborating evidence* could be found on search engines). These results are impressive given that the only information provided was the programme name. Most image-search engine images are frontal faces (see Figure 5.3), and this is reflected in the face tags made by this stage (see Figure 5.4, left column, top/middle row). As no further evidence of new, unlabelled identities is found after Stage 1, the Class Recall does not increase further. Further details are given in the suppl. material.

Stage 2: The fusion method improves face-track recall by 3.4%. Here, the voice modality is harnessed to confidently tag extreme face poses that Stage 1 could not (see Figure 5.4, central column, top/middle row). The QE step (Stage 1 + (Stage 2 QE only)) leads to a comparatively larger 6.4% improvement over Stage 1. QE here is able to bridge the domain gap from image-search engine images to the test video, and increase the number of correct tags. For the combined experiment (Stage 1 + (Stage 2 Fusion + QE), QE builds new identity models from the tags made by both Stage 1 and the fusion step. When combined, these offer a rich variety of poses, that are representative of the within-class variations. This therefore leads to the largest improvement to 0.953 face-track recall, where faces in a range of poses are tagged throughout the videos (see Figure 5.4, right column, top/middle row).

Figure 5.5c shows an example of a missed tag. This person was not labelled even though their name is displayed, as no *corroborating evidence* could be found on search engines of their appearance. This problem is non-trivial, as a displayed name does not always correspond to the displayed person. Human annotators used the fact that this person is introduced in a prior scene where they were not present. For an automated process, this tag requires complex, longer term reasoning capabilities. This opens possibilities for future work.

5.6.3 MediaEval

Our results on the MediaEval dataset are shown in Table 5.3a. We experiment with both the original MediaEval 2015 challenge rules [Poignant et al., 2017] (no external biometric models may be used, so no image-search engines), and with the combined Stages 1 and 2 of the proposed method. Challenge participants use the strong prior that a written name is very likely to belong to the co-occurring

Table 5.3: (a) Results on the MediaEval dataset. (b) Results on the Sherlock dataset - values are the per-character classification accuracy for the main characters in each of the three episodes. Our method improves over the state-of-the-art for both datasets. Key: ME: MediaEval.

MediaEval [Poignant et al., 2017]		Sherlock [Nagrani & Zisserman, 2017b]			
Method	mAP (%)	Method	E01	E02	E03
Baseline [Bredin, 2015]	74.89	Nagrani & Zisserman[Nagrani & Zisserman, 2017b]	0.92	0.90	0.88
SOTA [Poignant et al., 2017]	82.80	Ours (face)	0.88	0.81	0.86
Ours (ME 2015 rules)	85.22	Ours (face + Stage 2)	0.95	0.93	0.94
Ours (Stage 1 + Stage 2)	89.10				

speaking-and-visible face in the scene. Impressively our MediaEval 2015 rules method gains a 10% improvement on the original baseline method [Bredin, 2015], and a significant 2.2% improvement on the state-of-the-art [Poignant et al., 2014], while using just pre-trained out-of-the-box features. Using our full method (Stages 1 and 2), results in an impressive further 3.88% improvement in mAP. This gain is seen because our method is able to provide labels regardless of the proximity of a face to their corresponding written name, and is able to use stronger biometric models through the assistance of image-search engines. Our MediaEval 2015 rules method, as well as the previous baselines, fails to identify very well known people if their name is never spoken or found written. Without any manual supervision, our proposed method is able to correctly identify these people.

5.6.4 Sherlock

The results for testing out Stage 2 of our method on the Sherlock dataset are shown in Table 5.3b. The dataset provides face images from image-search engines for each identity in the test set, so it is not necessary to run the complete Stage 1 of our method. Instead we obtain the Stage 1 labels as follows: face embeddings are extracted for each of the provided images, and average pooled for each identity to give a face model; then the cosine similarity between each of the identity models and the test video face-tracks is computed; finally, each track is labelled with the identity and score of the model that has the maximum cosine similarity. This gives a preliminary set of face labels (‘Face’ result in Table 5.3b), from which Stage 2 can now boost tag numbers (‘Face + Stage 2’ result in Table 5.3b). Our method

surpasses the previous state-of-the-art [Nagrani & Zisserman, 2017b] considerably by a margin of 3-6% on all three episodes. This is particularly impressive, as the original work uses extra parameters for a SVM multi-way classifier, whereas our work simply uses a nearest neighbour classifier from aggregated identity features, which has no extra parameters and requires no extra training. The improvements are due to superior face embeddings, and also to the evidence fusion, which is able to classify characters where the original work, which treated modalities independently, could not. Figure 5.4 show examples of tagged faces on this challenging dataset.

5.7 Conclusions

In this paper, we propose a novel method for the automated labelling of people in videos through the use of *corroborating evidence*, both from image-search engines, and from different information modalities. The method performs impressively over a set of visually disparate domains both when starting from just the programme name, and also when testing out certain stages, setting new state-of-the-art results in the process. This method therefore provides a robust and reliable technique for labelling large video archives.

Acknowledgment

This work is supported by an EPSRC DTA Studentship, and the EPSRC programme grants EP/M013774/1: Seebibyte, and EP/T028572/1: Visual AI. We are grateful to Arhsa Nagrani, Shaya Ghadimi, and Maya Gulieva for proof reading, and the reviewers for their helpful feedback.

Statement of Authorship

A statement of authorship for this work can be found in Appendix A.

6 | Face, Body, Voice: Video Person-Clustering with Multiple Modalities

Andrew Brown¹ Vicky Kalogeiton² Andrew Zisserman¹

¹VGG, Oxford ²LIX, Ecole Polytechnique

Abstract

The objective of this work is person-clustering in videos – grouping characters according to their identity. Previous methods focus on the narrower task of face-clustering, and for the most part ignore other cues such as the person’s voice, their overall appearance (hair, clothes, posture), and the editing structure of the videos. Similarly, most current datasets evaluate only the task of face-clustering, rather than person-clustering. This limits their applicability to downstream applications such as story understanding which require person-level, rather than only face-level, reasoning.

In this paper we make contributions to address both these deficiencies: first, we introduce a *Multi-Modal High-Precision Clustering algorithm* for person-clustering in videos using cues from several modalities (face, body, and voice). Second, we introduce a *Video Person-Clustering dataset*, for evaluating multi-modal person-clustering. It contains body-tracks for each annotated character, face-tracks when visible, and voice-tracks when speaking, with their associated features. The dataset is by far the largest of its kind, and covers films and TV-shows representing a wide range of demographics. Finally, we show the effectiveness of using multiple modalities for person-clustering, explore the use of this new broad task for story understanding through character co-occurrences, and achieve a new state of the art on all available datasets for face and person-clustering.

Published in the Proceedings of the International Conference on Computer Vision Workshops, 2021.



Figure 6.1: **Video Person-Clustering – an essential step towards story understanding.** Imagine trying to understand the story in the scenes above, given only the *non-greyed-out* parts. Face-level understanding (left) omits important information, such as characters with their backs turned. This work addresses the new task of video person-clustering, which develops *person-level* understanding (right) in a scene by clustering all people, regardless of if their faces are showing or not. This is in contrast to the more limited, established, task of face-clustering. Person-level understanding is essential for downstream applications of grouping-by-identity such as story understanding, and cannot be achieved by face-clustering alone.

6.1 Introduction

Clustering people by identity in videos is an appealing and much-visited topic in computer vision [Cinbis et al., 2011b; Fitzgibbon & Zisserman, 2002; Tapaswi, Parkhi, et al., 2014; B. Wu, Zhang, et al., 2013; Jin et al., 2017; Kalogeiton & Zisserman, 2020; Tapaswi et al., 2019]. It has several real-world applications, such as enabling person-specific browsing, organisation of video collections, character based fast-forwards, automatic cast listing; and story understanding, all without requiring any explicit identity labeling. A successful person-clustering framework can therefore alleviate the tremendous annotation cost that is otherwise necessary for such applications.

However, methods for clustering by identity are almost always limited to only using information from faces. Such methods have two significant drawbacks: First, they ignore many available, informative cues that a human would use to solve the task: (i) the person’s voice available from the audio track; (ii) the person’s overall appearance (from their hair, clothes, posture); and, (iii) the editing structure (in edited material) – such as the co-occurrence of characters in nearby shots and within a scene. Second, they limit the utility of clustering for downstream applications such as story understanding. Understanding the story-line in a scene requires knowledge of *all* the characters present in a scene,

not just those whose faces are visible, *i.e.* *person-level* not *face-level* reasoning. This is illustrated in Figure 6.1.

Our objective in this paper is to cluster people (or more precisely person-tracks, which depict an entire body in any pose) by identity in movies and TV-material, as a first step towards *story-level understanding*. We cluster people, rather than just faces, and use all cues (face, voice, body appearance, editing structure), including tracks of people from behind without a visible face.

To see the value and necessity of this multi-modal approach, consider the problem of determining if two poor resolution faces depict the same person or not – the voice can discriminatively resolve this ambiguity. Similarly, consider the problem of determining if a person seen speaking to camera in one shot, is the same as the person seen from behind in a following shot – the hair and clothes can provide the link. In Figure 6.1, for example, how would the people seen from behind be identified other than by clustering their hair, clothes or voice with instances in neighboring shots?

More generally, modalities arising from the same person are both *redundant* and *complementary*, and can be used to address two fundamental problems in clustering: how to obtain *pure* clusters (*i.e.* containing tracks from a single person); and, how to *merge* clusters without violating their purity (*i.e.* by contaminating them with tracks from another person). They can be used to obtain very pure clusters by requiring agreement (*e.g.* on both face and voice) in order for tracks to be grouped together; and can be used to merge clusters which could not otherwise be confidently merged with a single modality, *e.g.* by using the common voice to merge a frontal with a profile face cluster (where the face descriptors of each cluster may be different). In this way, multiple modalities provide a *bridge* between otherwise unmergeable clusters. Methods that merge clusters using a single modality inevitably sacrifice purity.

In this paper, we introduce a new method for the task of video person-clustering, *Multi-Modal High-Precision Clustering (MuHPC)*, that uses multiple modalities – face, voice, and body appearance. It builds on recent methods that use first nearest neighbour [Sarfraz et al., 2019; Kalogeiton & Zisserman, 2020; Jarvis & Patrick, 1973] clustering algorithms, and is designed to take advantage of the redundancy and complementarity of the modalities, as discussed above, and to incorporate lessons from the face-clustering literature, such as cannot-link constraints and using the video editing structure [Bauml

et al., 2013; Cinbis et al., 2011b] (Section 6.3).

To evaluate the multi-modal person-clustering task, we require a dataset with person-level annotations. However, there are very few such datasets due to the previous emphasis on face-clustering and moreover, most face-clustering and labelling datasets, such as Buffy [Everingham et al., 2006] and TBBT [Roy et al., 2014], are based on TV material with limited diversity in skin color. For these reasons, we introduce a new *Video Person-Clustering Dataset (VPCD)* where we: (i) re-purpose multiple existing face datasets by adding person-level multi-modal annotations (*e.g.* all person-tracks and voice utterances); and (ii) include different TV shows and films (hereby referred to under the unified term *program sets*) to address this lack of diversity. *VPCD* consists of visually disparate program sets, and includes body-tracks, face-tracks when visible; and voice utterances when speaking, for all annotated characters. We provide features so that future clustering algorithms can be compared easily and fairly (Section 6.4).

We show the effectiveness of multi-modality and outperform strong baselines for person-clustering on *VPCD* (Section 6.5.1), and explore this new expansive task for story understanding (Section 6.5.4). Our method also significantly outperforms the face-clustering state of the art on both TBBT and Buffy by over 10% NMI (Section 6.5.3). Note that our goal is multi-modal clustering and not representation learning. Thus, we do not propose a new architecture or train a network for better features. Instead, we use features from pre-trained networks (for face and speaker recognition) and only train a network where it is necessary for body Re-ID. A broader impact statement is included in the appendix.

6.2 Related Work

In this work, we focus on multi-modal person-clustering in videos. Similar works target the more limited task of face-clustering or labelling, person Re-ID, or person search. We describe them, and also discuss similar datasets to hello.

Face-Clustering. A well-studied task for both images [Otto et al., 2017; Ho et al., 2003; Berg et al., 2004; Rothlingshofer et al., 2019] and videos [Kalogeiton & Zisserman, 2020; Tapaswi, Parkhi, et al., 2014; Tapaswi et al., 2019; Z. Zhang et al., 2016], with difficulty arising from the variation of pose,

lighting, and emotion [W.-A. Lin et al., 2018; Y. He et al., 2018] in faces of the same identity. Most video approaches exploit the spatio-temporal continuity and find must-link and cannot-link clustering constraints [Wagstaff et al., 2001; de Amorim, 2012; Bauml et al., 2013; Cinbis et al., 2011b; Kalogeiton & Zisserman, 2020; V. Sharma et al., 2019a, 2020; Tapaswi, Parkhi, et al., 2014; B. Wu, Zhang, et al., 2013; S. Xiao et al., 2014], or additional constraints from the structure of videos [Tapaswi, Parkhi, et al., 2014]. Most works approach face-clustering with metric or representation learning [B. Wu, Lyu, et al., 2013; B. Wu, Zhang, et al., 2013; Cinbis et al., 2011b; Fitzgibbon & Zisserman, 2002; V. Sharma et al., 2019a, 2020, 2019b; Tapaswi et al., 2019]. For instance, [Tapaswi et al., 2019] map features from the same identity to a fixed-radius sphere, while Sharma *et al.* use supervision from video constraints [V. Sharma et al., 2019a,b] or weak clustering labels [V. Sharma et al., 2020]. These methods, however, are limited by the relatively small training sets available from particular TV-shows. For this reason, some recent approaches focus on simply clustering pre-trained features that have been learnt on very large-scale face datasets. [Sarfraz et al., 2019] propose a simple first nearest neighbour clustering method upon pre-trained features, FINCH, and show impressive results. More recently, [Kalogeiton & Zisserman, 2020] combines [Sarfraz et al., 2019] with spatio-temporal constraints and improves performance. *All* the above works focus on the limited task of clustering faces (Figure 6.1 - left), whereas our focus is multi-modal person-clustering *i.e.* clustering every appearance of characters, regardless of whether their face is visible (Figure 6.1 - right), and using multiple modalities.

Face-Labelling. The task of classifying faces by identity - most works address this by using face-appearance with supervision from transcripts aligned to subtitles [Everingham et al., 2006; Cour et al., 2010; Parkhi, Rahtu, & Zisserman, 2015; Sivic et al., 2009; Everingham, Sivic, & Zisserman, 2009; Bauml et al., 2013; ?; Ramanathan et al., 2014], for example by using Multiple Instance Learning [Hau-rilet et al., 2016; Köstinger et al., 2011; Wohlfart et al., 2011]. Some exploit cues other than faces from videos: [Ramanan et al., 2007] use clothing to match faces in TV-shows across shot boundaries, while [Nagrani & Zisserman, 2017b; A. Brown, Coto, & Zisserman, 2021] use face and voice to label faces. [Poignant et al., 2017] use face and voice to retrieve a list of shots containing a named person, by searching for their name in subtitles and displayed text. These works focus only on visible faces and although some are multi-modal (face, voice and/or text supervision), the text is typically obtained from external sources (*i.e.* transcripts). Our task is different, as we cluster rather than label, and thus do

not require character-classifiers or ID supervision or extra annotation, and we use all available cues *i.e.* editing structure and multi-modality.

Person Re-ID. The task of re-identifying pedestrians in CCTV - typically [L. Zheng et al., 2015; Z. Zheng et al., 2017; W. Li et al., 2014; Wei et al., 2018], each body is fully visible and walking, the clothing remains constant for each identity, and the images are low resolution. This differs substantially from person-clustering in TV and film material, where there is large pose variation (*e.g.* sitting, standing, lying down), occlusion, and the clothing frequently changes for each identity. A full literature review is out of scope. Closer to our task are works on person-retrieval in photo albums [N. Zhang et al., 2015; Joon Oh et al., 2015; Sivic et al., 2006] or person-search from portraits in videos [Huang et al., 2018; J. Xia et al., 2020]. [Huang et al., 2018; Sivic et al., 2006] use face and body features, while [J. Xia et al., 2020] use audio. The TRECVID Instance Search challenges [Awad et al., 2019] involved retrieving a list of shots that contain an identity, given a query video for that identity. In contrast, we cluster all characters at the track-level in videos without requiring search queries.

Related Datasets. Various face-clustering datasets have been proposed [Everingham et al., 2006; Roy et al., 2014; Kalogeiton & Zisserman, 2020; Ozerov et al., 2013; Ghaleb et al., 2015; Cour et al., 2009]. These follow some similar trends: (a) are limited in size, consisting of a movie or some TV show episodes; (b) under-represent most demographic groups; and (c) contain only face annotations, so cannot be used for the broader multi-modal person-clustering task. Several story understanding [Huang et al., 2020; Bain et al., 2020] or person-search [Huang et al., 2018] datasets with face and/or body annotations exist. These cannot be used for our task, as they lack audio [Huang et al., 2018, 2020] or contain only partial annotations such as keyframes [Huang et al., 2020] or for a subset of tracks [Bain et al., 2020]. Furthermore none contain labelled voice utterances. Instead, *VPCD* contains 6 different TV-shows and movies, representing a more diverse range of characters, and containing *multi-modal annotations* for all annotated characters.

Story Understanding. This targets automatic understanding of human-centred story-lines in videos. It has been formulated in several ways, *e.g.* grouping scenes by story threads [Ercolessi et al., 2012; A. Rao et al., 2020], learning character interactions [Tapaswi, Bauml, & Stiefelwagen, 2014; Marin-Jimenez et al., 2019] or relationships [Kukleva et al., 2020], creating movie graphs [Vicol et al., 2018];

or text-to-video retrieval from narrating captions [Bain et al., 2020], with several datasets [Bain et al., 2020; Huang et al., 2020] introduced. Many works [Bain et al., 2020; Vicol et al., 2018; Kukleva et al., 2020] highlight the importance of knowing who is present in a scene for understanding the story. This is the focus of our work.

6.3 Method

Here, we describe the *Multi-Modal High-Precision Clustering (MuHPC)* method for person-clustering in videos. It is a single hierarchical agglomerative clustering [Reddy & Vinzamuri, 2018] (HAC) approach that groups person-tracks by identity using similarities of modality features, together with constraints arising from the video structure. *MuHPC* uses pre-computed features, and hence does not require any training outside of simply learning optimal hyper-parameters, and can then run out of the box for any video dataset. In this work, we use three modalities (face, voice, and body appearance) but *MuHPC* can easily scale to any number of modalities.

Overview. *MuHPC* consists of three stages (Figure 6.2). **Stage 1** creates high-precision clusters using a single modality, here face. We group person-tracks that share a first nearest neighbour (NN) using multiple iterations of HAC, as in [Sarfranz et al., 2019; Kalogeiton & Zisserman, 2020; Jarvis & Patrick, 1973]. We follow this trend subject to two additional *constraints*: a cannot-link constraint for concurrent tracks (as in [Kalogeiton & Zisserman, 2020] based on [Bauml et al., 2013; Cinbis et al., 2011b]), and a conservative threshold on the maximum NN distance. This results in K_1 clusters (Section 6.3.1). **Stage 2** exploits multi-modality to *bridge* clusters that were otherwise unmergeable by the single face modality with a conservative threshold; in particular, by requiring that different modalities (*i.e.* face and voice) concur on the merge (Section 6.3.2). **Stage 3** clusters tracks without visible faces, and hence that are not yet clustered by the first two stages. Constraints from the editing structure (neighboring shots) and a conservative threshold on body features (so that they depict the same person with the same clothing) are used to link face-less person-tracks to clusters with faces (Section 6.3.3). Here, we describe the stages, algorithm design choices, and how the hyper-parameters are learnt.

Notation. Given a dataset with person-tracks and C characters, where x_i is a single person-track, the

goal is to cluster all x_i by identity into C clusters (C is unknown). Each person-track x_i is represented by one feature vector per available modality, *i.e.* $x = \{x_f, x_v, x_b\}$, with x_f, x_v, x_b the face, voice and body-track features, respectively. The availability of each feature vector is dependant upon the part of the person that is visible (face and/or body), and if they are speaking. For each person, at least one of x_f, x_b are available. Let $d(x_i, x_j)$ be the distance between two track features of the same modality, and d_f, d_v and d_b the distances between two face, voice or body-tracks, respectively; the lower the value, the more likely the tracks depict the same identity. NN is nearest neighbor; $n_{x_i}^1$ is the first NN track of track x_i . The set of video frames that x_i is present in is denoted by T_i .

6.3.1 Stage 1: High-Precision Clustering

Stage 1 creates high-precision clusters, each containing tracks of the same identity. It uses only the face modality as this is the most discriminant of the three (face, voice and body), and thus is least likely to group different identities in the same cluster. Here, we use a NN clustering method [Sarfranz et al., 2019; Kalogeiton & Zisserman, 2020], subject to two clustering constraints.

Clustering Constraints. A NN is only considered valid if the resulting merge satisfies: (1) *A Spatio-Temporal Cannot-link Constraint*: Tracks that have (partial) temporal overlap cannot be grouped together, since they must represent different characters as they appear together in at least in one frame (introduced by [Kalogeiton & Zisserman, 2020]); and (2) *A NN Distance Constraint*: the distance $d_f(x_i, n_{x_i}^1)$ between a track x_i and its first NN $n_{x_i}^1$ is less than a strict threshold τ_f^{tight} for Stage 1.

Clustering process. At every iteration (cluster partition Γ), each cluster is grouped with its NN cluster, *i.e.* the closest. Specifically, the first partition groups tracks into clusters through first NN relations, while following partitions group the clusters formed in the previous partition; each cluster is represented by the average of the features it contains. Following the notation of [Sarfranz et al., 2019], at each partition Γ , the method forms K_Γ clusters by merging tracks that are either first NN (mutually or one is the first NN of the other) or have a common NN $n_{x_i}^1$, as described by the adjacency matrix:

$$A(x_i, x_j) = \begin{cases} 1 & \text{if } (x_j = n_{x_i}^1 \text{ or } n_{x_j}^1 = x_i \text{ or } n_{x_i}^1 = n_{x_j}^1) \text{ and } T_i \cap T_j = \emptyset, d_f(x_i, n_{x_i}^1) \leq \tau_f^{\text{tight}} \\ 0 & \text{otherwise.} \end{cases} \quad (6.1)$$

Discussion. In standard HAC the clustering continues until all clusters merge to one. Including the constraints introduces strict stopping criteria, and therefore the clustering stops when either the clusters are all more than a distance τ_f^{tight} apart, or they are separated by a cannot-link constraint. This results in K_1 high-precision clusters, where we expect $K_1 \geq C$. The very simple addition of a distance threshold leads to a significant improvement in clustering results over [Kalogeiton & Zisserman, 2020; Sarfraz et al., 2019; Tapaswi et al., 2019] (Section 6.5.3). Without this constraint, little prevents an incorrect merging of clusters of different identities and the subsequent creation of low-precision clusters.

6.3.2 Stage 2: Multi-modal Cluster Bridging

Combining a discriminative modality with the constraints results in *high-precision* clusters. However, a single modality alone cannot continue making confident merges without sacrificing purity. Thus, Stage 2 merges these clusters by exploiting multiple modalities *i.e.* face and voice.

Modality-pair merges. To further merge clusters, we demand that two modalities agree that the clusters contain the same identity. Therefore, we require that the distances for the face and voice are both below new thresholds, *i.e.* $d_f < \tau_f^{\text{loose}}$ and $d_v < \tau_v^{\text{loose}}$. Note, here we use features taken from tracks within clusters, rather than averaged cluster features. τ_f^{tight} is raised by just a small margin, δ , *i.e.* $\tau_f^{\text{loose}} = \tau_f^{\text{tight}} + \delta$, due to the concurrent agreement from the voice.

Discussion. This stage results in K_2 clusters with high-precision, where $K_2 \leq K_1$. Here, we use face and voice as they have been shown to be coupled [Nagrani et al., 2018b,a] and to contain redundant, identity discriminating information. An alternative is to require that the voice modality alone provides a confident (*i.e.* tight threshold) match, *e.g.* two person-tracks with the same voice. We find however that voice alone cannot reliably join clusters of the same identity. This can be because two identities with the same emotion in their voice (*e.g.* shouting, crying) can appear similar to the less discriminative

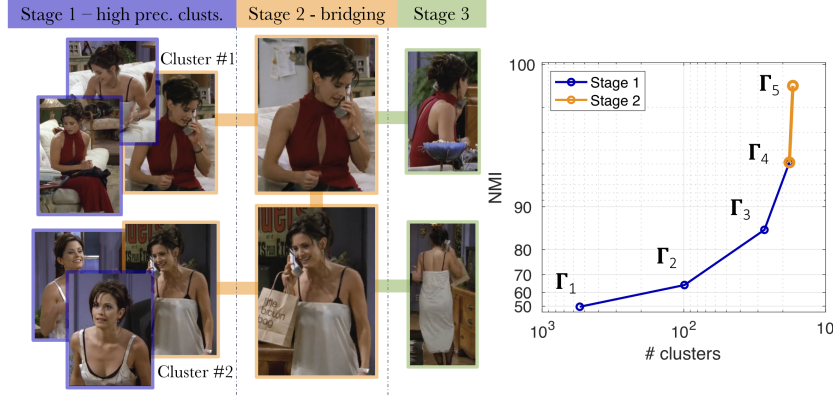


Figure 6.2: **The clustering process of MuHPC.** (Left) Example person-tracks at each stage of MuHPC. Two high-precision clusters from Stage 1 depicting the same character. One contains near-frontal faces (below) and one profiles (top), hence the single face modality cannot confidently merge the two. Stage 2 uses a talking person-track from each cluster to form a bridge, by demanding the agreement of both face and voice modalities that these contain the same identity. Stage 3 merges face-less bodies into the formed cluster. (right) The NMI and number of clusters at each partition, Γ , of stages 1 and 2 on an example video from VPCD. At each partition the number of clusters decreases, while the normalised mutual information increases. At Γ_4 Stage 1 clustering stops. Stage 2 progresses to Γ_5 by bridging clusters. Stage 3 does not affect the number of clusters.

voice embedding (more in the appendix).

6.3.3 Stage 3: Clustering backs

Stages 1 and 2 result in high-precision clusters. Nevertheless, they do not account for person-tracks with no visible face, for instance when viewed from behind, *i.e.* a *face-less* person. The goal of Stage 3 is to add the face-less person-tracks into their respective high-precision clusters using the modality of body-appearance. Here, we use the editing structure of the videos, given that the appearance of the same character can change dramatically between scenes. As discussed above, body features may not be discriminative for identifying if characters are wearing very similar clothing. We determine such body-tracks using the simple ratio-test introduced in [Lowe, 2004]. Specifically, for each body-track we compute the first and second NN distances, d_{b,x_i}^1 and d_{b,x_i}^2 . If the ratio, $d_{b,x_i}^1/d_{b,x_i}^2$ is higher than a threshold ρ then the body-track is classified as non-distinctive and is ignored.

For assigning face-less people to clusters, we find the NN body-track (that has a face and hence is already clustered) that does not violate the ratio-test in a neighbouring shot, and assign the face-less

person to this cluster. Given that the same person is most likely wearing the same outfit in the same or neighbouring shots, we only examine the distance between body-tracks from these shots. At this stage, some backs cannot be clustered with high confidence, either because they are not similar to any nearby body or because they fail the ratio test for being a non-distinctive feature. Our design choice is to ignore these backs, *i.e.*, we ignore any back for which the NN distance is more than a threshold τ_b^{back} . Note, this stage keeps the number of clusters to K_2 .

Required Number of Clusters. Suppose we know the number of characters C , and hence the number of clusters. Our goal is to reduce K_2 to the desired C (typically $K_2 \geq C$). Previous methods [Tapaswi et al., 2019] employ HAC; however, this suffers from reliance on features that can no longer confidently discriminate between clusters of the same person. Instead, we employ a cluster prior: there is no identity overlap amongst the largest clusters *i.e.* they contain unique identities, and conversely there is likely an identity overlap between a small and large cluster. Our intuition is that big clusters contain ample information about an identity, and consequently if two large clusters contained the same identity, then they would have been merged. Therefore, we iteratively merge the smallest with the largest cluster until there are C clusters. In practice, we observe that small clusters contain blurry or low-resolution tracks, and so could not confidently be merged at earlier stages.

Discussion. Most methods [Tapaswi et al., 2019; V. Sharma et al., 2019a, 2020] fine-tune character features on a video dataset. Instead, *MuHPC* operates on pre-trained features, thus reducing the computational burden and leading to increased generalisation capabilities. An extension would be to replace the constraints with a cost function optimisation approach, allowing a cannot-link to be correctly broken for a person’s reflection in a mirror.

6.3.4 Learning Hyper-Parameters

The hyper-parameters for *MuHPC* are learnt on the validation partition of *VPCD*. The visually disparate program sets in the test partition are disjoint from those in the validation, yet these parameters are kept constant. For the hyper-parameter associated with the face modality (τ_f^{loose}) this is possible as the face features are trained on millions of faces [Q. Cao et al., 2018b], and therefore are highly discriminative and universal (generalise well across different program sets). However, voice identity features are less

universal than face features, and hence there is not a single good choice for τ_v^{loose} that would generalise across the audibly disparate program sets. Instead, we learn a unique value *automatically* for each. Our goal is to choose τ_v^{loose} to be lower than the minimum distance between voices from different people. The cannot-link constraints automatically provide face-track pairs of different identities. We measure the distances between different people’s voices. In practice, there are too few constraints between speaking faces to provide an accurate representation of the negative distances, as rarely two face-tracks speak in the same shot. We combine the cannot-link speaking face-tracks with clusters from Stage 1 to provide more examples. This leads to many negative distances and an accurate representation of their distribution. We select τ_v^{loose} as the lower 99.9 percentile of these distances. This provides a robust automatic threshold measure. For program sets with similar sounding characters, this process gives a low τ_v^{loose} (e.g. Buffy – many similar sounding teenagers).

6.4 Video Person-Clustering Dataset

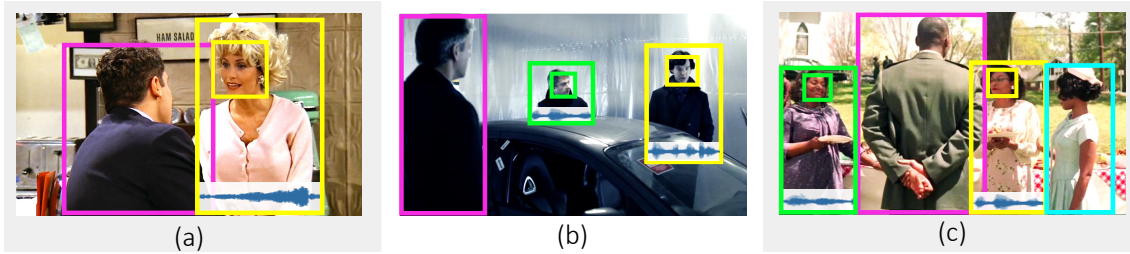


Figure 6.3: VPCD dataset. It consists of different and diverse TV shows and movies; here, we display a subset of them: (a) Friends, (b) Sherlock, (c) Hidden Figures. VPCD contains face, body and voice tracks annotated for many characters. Here, we display such examples. Each face-body pair is displayed with a unique color. A more representative range of characters are captured in a variety of scenes (e.g. dark (b)), viewpoints (e.g.(c)); and poses, including backs of bodies (magenta, cyan). When speaking, we also include a voice-track (blue signal below body-tracks).

In this section, we describe the dataset (Section 6.4.1), the annotation (Section 6.4.2), and the feature extraction processes (Section 6.4.3). The dataset is built on top of existing video datasets that have face-level annotations (labeled face-tracks) by adding and annotating body-tracks, and annotating voice utterances. This is for three reasons: first, it enriches the existing dataset by raising them to have person-level annotations; second, it enables comparisons on face-level clustering with prior work on

Dataset	#eps	length	#IDs	Gender	#Tracks		
				F/M	body	face	voice
TBBT [Roy et al., 2014]	6	2h 6m	103	53/50	4,276	3,908	1,047
Buffy [Everingham et al., 2006]	6	4h 9m	109	37/70	7,561	5,832	1,835
Sherlock [Nagrani & Zisserman, 2017b]	3	4h 30m	31	16/15	6,232	6,247	1,615
Friends [Kalogeiton & Zisserman, 2020]	25	9h 22m	49	23/26	18,360	17,333	3,961
ALN [Somandepalli et al., 2020]	1	1h 40m	10	4/6	1,932	1,614	404
HF [Somandepalli et al., 2020]	1	2h 7m	24	11/13	1,416	1,463	303
VPCD		23h 54m	326		39,777	35,396	9,165

Table 6.1: **Video Person-Clustering Dataset statistics.** For each program set in *VPCD* we detail video and annotation statistics. #eps: number of episodes; #IDs: number of unique characters; TBBT: The Big Bang Theory; (movies) ALN: About Last Night; HF: Hidden Figures. We cite the first published work that used each respective program set for face-clustering, but we provide additional full multi-modal annotations for each.

these datasets; and third, it means that the video material is already publicly available and we need only release the new annotations (and features).

6.4.1 VPCD content

VPCD contains *full multi-modal annotations* for primary and secondary characters for a range of diverse and visually disparate TV-shows and movies (statistics in Table 6.1, examples in Figure 6.3). *VPCD* contains annotations for 39,777 body-tracks, 35,396 face-tracks for whenever the face is visible, and 9,165 manually annotated voice-tracks for whenever each of them are speaking. Identity discriminating features (embeddings from deep networks) are provided for all modalities. A total of 23 hours of video cover a range of genres and styles such as Hollywood Drama (Hidden Figures, 2016), Romance (About Last Night, 2014), fast-paced Action/Mystery (Sherlock, Buffy) and live studio-audience sitcoms (Friends, TBBT). A large variety of characters are annotated, ranging from small casts shown over many episodes (*e.g.* Friends) to program sets with a long-tailed distribution with many secondary/background characters (*e.g.* Buffy). *VPCD* is by far the largest dataset of its kind. The program sets were chosen such that *VPCD* is representative of the diversity of people’s appearance in the real world. There is a validation set and a test set - these are disjoint. The validation set is the first five episodes of Friends.

6.4.2 Annotation Process

Here, we describe the annotation process for the face, body, and voice tracks in *VPCD*. For all component program sets, the face annotations already exist, and define the characters of interest for that video. Our goal is to annotate their body and voice-tracks. Very often in videos, a character is seen facing from behind (Figure 6.3). This means that the existing face-tracks cannot be used to trivially annotate the body-tracks by spatial overlap (since there will be no face-track). We therefore combine automatic and manual annotation methods (more details in the appendix).

Face. We use the same face bounding-box/track annotations and ID labels as were provided with the original datasets so that we can compare to previous works on face-clustering.

Body. We detect bodies with a Cascade R-CNN [Z. Cai & Vasconcelos, 2018] trained on MovieNet [Huang et al., 2020] and form tracks with an IOU tracker. When a body-track clearly corresponds to a face-track (*i.e.* no significant IOU with any other face-track), the body-track is automatically annotated with the character name of that face-track. We manually annotate the remainder as well as the body-tracks corresponding to characters from behind.

Voice. We manually segment the audio-track into the speaking parts for all annotated characters. To ensure the correctness of the segmentation, the audio track was first segmented by one human annotator, and then verified by different ones.

#	Modality			TBBT			#C _s = 130			Buffy			#C _s = 165			Sherlock			#C _s = 50			Friends			#C _s = 239			Hidden Figures			#C _s = 10			About Last Night			#C _s = 24			Average			#C _s = 618		
	F	B	V	WCP	NMI	CP	CR	WCP	NMI	CP	CR	WCP	NMI	CP	CR	WCP	NMI	CP	CR	WCP	NMI	CP	CR	WCP	NMI	CP	CR	WCP	NMI	CP	CR	WCP	NMI	CP	CR	WCP	NMI	CP	CR						
B-ReID				80.5	69.7	49.6	55.0	65.0	60.9	52.7	46.8	61.2	28.9	43.6	44.3	70.9	60.4	71.0	56.3	32.6	23.4	36.8	19.6	41.0	14.1	37.4	32.6	58.5	42.9	48.5	42.4														
B-C1C	✓	✓		87.7	69.2	39.4	50.6	73.6	58.2	34.6	41.6	77.7	41.6	29.3	43.6	85.3	77.1	69.5	70.8	76.2	69.8	55.2	50.3	94.4	85.8	68.0	76.8	82.5	67.0	49.3	55.6														
MuHPC-✓				93.5	84.6	76.4	77.6	80.0	66.7	63.8	65.2	83.8	52.3	51.2	58.4	85.7	73.7	81.3	79.0	77.6	70.4	59.1	52.1	95.7	89.7	98.2	86.3	86.1	72.9	71.7	69.8														
MuHPC _v ✓			✓	93.5	84.6	76.4	77.6	80.1	67.2	64.2	64.7	84.5	59.3	54.9	57.3	86.9	75.3	84.0	82.8	77.6	70.4	59.1	52.1	96.0	90.5	98.3	86.4	86.4	73.5	72.3	69.7														
MuHPC _b ✓			✓	96.9	92.8	80.4	79.6	85.7	75.6	68.1	67.9	84.1	52.9	51.7	54.3	89.5	81.3	84.6	82.4	77.6	70.3	59.0	52.0	95.7	89.4	98.2	86.3	88.2	77.1	74.0	70.0														
MuHPC ✓			✓	96.9	92.8	80.4	79.6	85.8	76.4	68.4	67.2	84.8	60.0	55.2	57.2	90.8	83.1	87.7	86.6	77.6	70.3	59.0	52.0	96.0	90.2	98.3	86.4	88.6	78.8	74.8	71.5														

Table 6.2: **Person-Clustering Results on VPCD.** For each program set, each metric is averaged across all episodes. AT protocol. The ‘Average’ column reports averaged metrics across all six program sets. #C_s is the sum of ground truth clusters across each episode in each program set. We report two strong baselines (B-ReID, B-C1C, Section 6.5.1) and an ablation on the modalities used. Keys: F-face, B-body, V-voice. *Modality*: used modalities.

6.4.3 Feature Extraction

Face. We use L2-normalised 256D features, extracted from an SENet-50 [Hu et al., 2018b] pre-trained on MS-Celeb-1M [?], and fine-tuned on VGGFace2 [Q. Cao et al., 2018b] (same as [Roy et al., 2014; Everingham et al., 2006; Nagrani & Zisserman, 2017b; Kalogeiton & Zisserman, 2020]).

Body. For all body detections, we extract 256D features with ResNet50 [K. He et al., 2016a] trained on CSM [Huang et al., 2018]. We average the features across each body-track, and then L2-normalise them.

Voice. Following [Chung, Huh, Mun, Lee, et al., 2020b], we extract a single, L2-normalised 512D speaker embedding from each voice segment using a thin-ResNet-34 [W. Xie et al., 2019; K. He et al., 2016a] trained on VoxCeleb2 [Chung et al., 2018].

6.5 Experiments

Here, we evaluate *MuHPC*. We first give experimental details, followed by person-clustering results on *VPCD* and provide ablations. We compare to previous face-clustering works and finally examine the advantages of person-clustering for story understanding. Further ablations and experiments on clustering all characters in all videos simultaneously are included in the appendix.

Implementation details. We use the face, body and voice track annotations and features from *VPCD* (Sections 6.4.1, 6.4.3). For all modalities, feature distances d_f, d_b, d_v are computed using (1 - cosine similarity). As described in Section 6.3.4, parameters are learnt on the *VPCD* val. set. The values are: $\tau_f^{\text{tight}}=0.48$, $\delta=0.025$, $\rho=0.9$ and $\tau_b^{\text{back}}=0.4$. These parameters are fixed for all experiments, and only have to be re-learnt if the features change. Details on the automatically selected τ_v^{tight} values are in the appendix.

Metrics. For each dataset in *VPCD*, we measure each metric at the episode level and average over all episodes. Following [Kalogeiton & Zisserman, 2020; Tapaswi et al., 2019], we use Weighted Cluster Purity (**WCP**) and Normalized Mutual Information (**NMI**). WCP weights the purity of a cluster by the number of tracks belonging in it. NMI [Manning et al., 2008b] measures the trade-off between cluster-

ing quality and number of resulting clusters. **Character Precision and Recall (CP, CR)** are computed using the number of ground truth identities. Each identity is uniquely assigned to a cluster. CP is the proportion of tracks in a cluster that belong to its assigned character, while CR is the proportion of that character’s total tracks that appear in the cluster. They are averaged across all characters, thus weighting each equally.

Test protocol. We evaluate: (i) automatic termination (AT), *i.e.* unknown number of clusters, and (ii) oracle cluster (OC), when known. AT is realistic for applications, while OC offers a fair comparison to the state of the art.

6.5.1 Person-Clustering

Baselines. To evaluate person-clustering, we compare to two strong baselines stemming from the best existing face-clustering algorithm, C1C [Kalogeiton & Zisserman, 2020]. The first, B-ReID, is inspired by person Re-ID [L. Zheng et al., 2015; Z. Zheng et al., 2017; W. Li et al., 2014] and uses C1C to cluster body rather than face features. It ignores person-tracks without bodies (<2% of person-tracks). For the second, B-C1C, we use regular C1C to cluster faces, with the addition of Stage 3 of *MuHPC* for clustering face-less bodies.

Results and analysis. Table 6.2 reports person-clustering results when testing on *VPCD*. For all metrics, *MuHPC* (full method) significantly outperforms the strongest baseline by on average 6.1% in WCP and 11.8% in NMI. B-ReID is poor due to frequent clothing changes. *MuHPC* outperforms B-C1C thanks to (1) the NN distance threshold that prevents incorrect merges and subsequent low-precision clusters, and (2) the multi-modal bridges that merge clusters which face alone cannot. This validates that using all available video cues, such as multi-modality and editing structure aids video person-clustering substantially. The clustering process for a character in *VPCD* is visualised qualitatively and quantitatively in Figure 6.2. *MuHPC* improves most upon the baselines on the more unconstrained program sets with many secondary characters and long-tailed character distributions (*e.g.* TBBT, Buffy, Friends, Sherlock). Here, *MuHPC* uses the NN distance threshold to keep the clusters of the many characters separated, and then merges any repeated clusters of main-characters via talking person-tracks. The *MuHPC* clustering process is visualised in Figure 6.2.

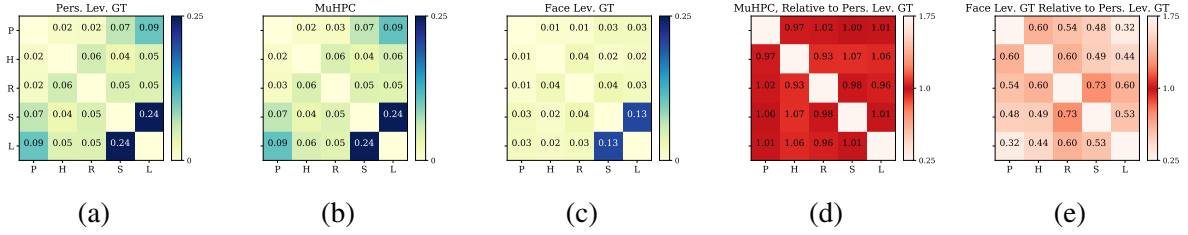


Figure 6.4: **Character co-occurrences for story understanding** between the 5 main characters in the 6 episodes of The Big Bang Theory in *VPCD*. (a) The ground truth co-occurrences as a proportion of the combined temporal length of the 6 videos. It is generated from *VPCD* which annotates each character whenever they are visible. Higher indicates more co-occurrence. (b,c): *MuHPC* and face-level clustering co-occurrences, as a proportion of the 6 videos; (d,e): *MuHPC* and face-level co-occurrences, relative to the ground truth (a). (d,e) are obtained by dividing (b,c) by (a), respectively. 1.0 indicates that the prediction is the same as the ground truth. Key: P: Penny, H: Howard, R: Raj, S: Sheldon, L: Leonard.

6.5.2 Ablation

Here, we perform ablations on the different modalities in *MuHPC*. Detailed results and parameter sweeps can be found in the appendix. Table 6.2 includes an ablation of the multi-modality, *i.e.* using voice (*MuHPC_v* – Stage 2) or body (*MuHPC_b* – Stage 3) modalities or both (*MuHPC*). Experiments without the body modality do not use Stage 3, and instead cluster each face-less body to the temporally-closest (Temporal-NN) body with a face in a nearby shot. Due to the threading structure [Hoai & Zisserman, 2014] of edited videos, there is a strong prior that the Temporal-NN is correct.

Adding either the voice or body offers a benefit over *MuHPC*–, due to the increased discriminative capabilities from an additional modality. *MuHPC_b* outperforms *MuHPC_v*, as there are many face-less bodies in *VPCD*, and the body modality allows for these to be clustered correctly. Using the voice in conjunction with the body (*MuHPC*) performs best, as their benefits are compounded, and the multi-modal bridges connect clusters with higher purity. The voice gives a higher boost when used alongside the body modality, as otherwise the multi-modal bridges are merging lower precision clusters. The voice adds significant benefit in NMI on multiple program-sets. This is impressive as the tight voice thresholds were found *automatically*. Sometimes the voice does not lead to an improvement, due to the absence of speaking person-tracks in merge-able clusters (*e.g.* TBBT). Additionally, the body offers little improvement in the two movies (Hidden Figures, About Last Night) that have many dark scenes and non-distinctive clothing. Here temporal-NN is able to assign face-less bodies to clusters well. Note,

NMI increases more than WCP when adding the voice modality, because bridging two high-precision clusters will not greatly effect the purity; however, it leads to increased NMI as there is less identity overlap between the resulting clusters.

MuHPC requires manually diarised speech segments. Preliminary results show that automatic diarisation methods lead to smaller improvements from the voice modality than when manually diarised voice is used, but we leave this to future work. We highlight that with 24 hours of manually diarised audio, VPCD provides a unique test bed for future research on moving beyond requiring manual diarisation.

6.5.3 Face-Clustering

Here, we compare to previous works by experimenting only on face-tracks, excluding person-tracks without faces. We compare to FINCH [Sarfraz et al., 2019] (evaluated at the required number of clusters, from [Kalogeiton & Zisserman, 2020]), BCL [Tapaswi et al., 2019] and C1C [Kalogeiton & Zisserman, 2020]. For TBBT and Buffy, the face annotations are the same as [Tapaswi et al., 2019; Kalogeiton & Zisserman, 2020]. Here, we do not compare to works that use the less challenging [Kalogeiton & Zisserman, 2020] subset of the annotations [V. Sharma et al., 2020; Rothlingshofer et al., 2019]. For our method, we present: (i) *MuHPC*– uses only face-tracks, *i.e.* exactly the *same* information and features as other methods, hence results are directly comparable; and (ii) *MuHPC_v* uses face-tracks with multi-modal bridges (*i.e.* voice). Following [Kalogeiton & Zisserman, 2020; Tapaswi et al., 2019], performance is evaluated at frame level.

Table 6.3 reports face-clustering results. For both AT and OC protocols, *MuHPC*– significantly outperforms the state of the art in all metrics, as it avoids incorrect merges, hence maintaining cluster purity. For instance, NMI, CP and CR boost by +10-14% for Buffy and TBBT for OC, and by over 10% for WCP averaged across all datasets for AT. *MuHPC_v* also leads to a boost over *MuHPC*– in most datasets. We observe that the more challenging the dataset, the higher the boosts by multi-modality, *e.g.* +3.8% in CR for Friends and +7.4% in NMI for Sherlock. We note that on NMI, WCP, the performance on TBBT is now almost saturated. A full discussion of results is given in the appendix.

Method	protocol	TBBT				#C _s = 130	Buffy				#C _s = 165
		WCP	NMI	CP	CR	#C _p	WCP	NMI	CP	CR	#C _p
BCL [Tapaswi et al., 2019]	AT	90.8	85.7	-	-	83	85.0	78.8	-	-	121
C1C [Kalogeiton & Zisserman, 2020]	AT	89.2	87.4	29.1	40.9	41	66.3	68.8	14.9	27.1	40
<i>MuHPC</i> –	AT	99.4	97.8	87.8	88.6	168	96.1	92.8	85.6	85.5	223
<i>MuHPC_v</i>	AT	99.4	97.8	87.8	88.6	168	96.1	93.7	85.9	84.8	221
Finch [Sarfraz et al., 2019]	OC	90.8	80.5	46.1	44.2		82.9	75.3	49.6	41.0	
BCL [Tapaswi et al., 2019]	OC	94.0	85.0	-	-		86.5	77.6	-	-	
C1C [Kalogeiton & Zisserman, 2020]	OC	95.3	84.5	54.9	57.3		88.1	79.1	58.1	55.4	
<i>MuHPC</i> –	OC	99.1	97.4	79.3	83.0		95.6	92.2	72.3	73.8	
<i>MuHPC_v</i>	OC	99.1	97.4	79.3	83.0		95.6	93.1	71.5	73.2	
		Friends				#C _s = 239	Sherlock				#C _s = 50
C1C [Kalogeiton & Zisserman, 2020]	AT	88.2	89.8	62.4	73.2	185	76.3	50.3	20.2	41.0	25
<i>MuHPC</i> –	AT	98.7	94.9	98.1	94.0	543	86.7	60.3	79.1	71.2	96
<i>MuHPC_v</i>	AT	98.4	95.9	97.7	95.3	522	86.3	66.0	78.4	74.5	86
Finch [Sarfraz et al., 2019]	OC	92.2	89.9	85.2	85.6		81.6	58.6	59.8	56.8	
C1C [Kalogeiton & Zisserman, 2020]	OC	94.3	93.2	79.1	85.5		81.6	53.8	40.5	51.7	
<i>MuHPC</i> –	OC	96.3	92.7	89.0	88.8		84.0	56.5	55.4	59.9	
<i>MuHPC_v</i>	OC	97.1	94.6	92.3	92.6		85.1	63.9	59.6	62.9	

Table 6.3: **Face-Clustering Results.** Comparisons to previous state of the art on four program sets using only face-tracks with unknown (AT), and known (OC) number of clusters. We report metrics averaged over each episode in each program set, and the number of predicted clusters, summed over each episode ($\#C_p$). *MuHPC*– uses only face; *MuHPC_v* uses the multi-modal bridges from voice and face. Where not reported in respective publications, numbers are computed using official implementations. Finch has no stopping criterion so results for AT are not reported.

6.5.4 Enabling Story Understanding

Here, we explore how close we have come to enabling story understanding. Clustering people (rather than faces) indicates who is present in a scene (Figure 6.1) – an essential and necessary step for predicting character co-occurrences, and hence their interactions [Kukleva et al., 2020; Marin-Jimenez et al., 2020] that make up a story. Specifically, we ask two questions: (1) Can clustering on the face-level predict the co-occurrence of two characters correctly? (2) How close is *MuHPC* to correctly predicting co-occurrences? To answer these, we experiment with the five main characters from the six episodes of TBBT in *VPCD*. Figure 6.4a shows the ground truth (Pers. Lev. GT) co-occurrence of character pairs as a proportion of all frames in the show, and hence is a measure of their interaction, e.g. Sheldon and Leonard co-occur for 24% of all frames.

For the first question, we visualise the co-occurrences of characters according to the face-track annotations (Face Lev. GT) in *VPCD* (shown in absolute terms in Figure 6.4c, and relative to the GT in

Figure 6.4e). The face-level co-occurrences are poor – with an average error from GT of 48%. For instance, Penny and Leonard, whose romance is a main story-line, are shown to co-occur in only 3% of the videos vs the GT 9%. Furthermore, the GT shows that this is the second most commonly occurring pair; nevertheless, the face-level annotations fail to pick up that it is significant relative to other pairs. This is expected as often one or more characters do not show their face when appearing together (Figure 6.1). Hence, any co-occurrences predicted from the face-level are a limited foundation for story understanding.

For the second question, we cluster with *MuHPC* and assign each cluster to the character that appears most within it (Figures 6.4b, 6.4d). We observe that these predictions are very close to the GT, with an average error of just 3% (Figure 6.4d). This impressive result shows that the presence of each character, their co-occurrence and hence their possible interactions can be found completely automatically and accurately using our proposed method. This provides a rich and informative foundation for story understanding.

The assignment of character names to clusters can be automated by combining *MuHPC* with methods that focus on the automated labelling of face-tracks with names [Everingham et al., 2006; A. Brown, Coto, & Zisserman, 2021].

6.6 Conclusions

In this work we propose *MuHPC*, a novel method for multi-modal person-clustering in videos. For evaluation we introduced *VPCD*, the largest and most diverse dataset of its kind. We showed that using all available video cues is essential for person-clustering, leading to significant improvements on *VPCD*, and to state-of-the-art performance for face-clustering. Importantly, we demonstrated that *MuHPC* allows each character appearance and co-occurrence to be predicted completely automatically and accurately. We hope this can support downstream story understanding tasks such as the learning of relationships [Kukleva et al., 2020]. *MuHPC* has intriguing benefits for creative video editing/understanding, such as the automated collation of character-based “highlight reels” *i.e.* scenes containing a certain two character’s interactions, or one character’s story-line. **Acknowledgments:** This

work is supported by an EPSRC DTA Studentship, and the EPSRC programme grant EP/T028572/1: Visual AI.

Appendices

Appendices can be accessed online.¹

Statement of Authorship

A statement of authorship for this work can be found in Appendix [A](#).

¹https://openaccess.thecvf.com/content/ICCV2021W/CVEU/supplemental/Brown_Face_Body_Voice_ICCVW_2021_supplemental.pdf

7 | Discussion

In this chapter we discuss the main achievements and impact of the research contributions presented in this thesis (Section 7.1), and also suggest directions for future work (Section 7.2).

7.1 Achievements and Impact

In this thesis, we have present methods for enabling end-to-end learning of deep neural networks (Chapters 2 and 3), and for audio-visual human-centric video understanding (Chapters 4 to 6). At the time of writing in October 2022, the work in this thesis has been cited over 190 times according to Google Scholar. In this Section, we discuss the achievements and impact of that work.

Enabling End-to-End Learning. In Chapter 2 we introduce a method for training models end-to-end for image retrieval by directly optimising *Average Precision*, via a new objective function termed *Smooth-AP*, which optimises a smoothed version of the ranking metric. We present an analysis as to why optimising *Smooth-AP* is a closer approximation to Average Precision than prior work, and demonstrate significant performance improvements on large-scale real-world datasets when training a model with this new objective.

Our proposed objective, *Smooth-AP* has been shown to be a versatile and reliable choice of objective function when training deep neural networks for Average Precision, as demonstrated by the wide variety of published research projects where *Smooth-AP* has been used. For example, Smooth-AP has been used while training models for person re-identification [Y. Wu et al., 2021], aerial image retrieval [Gominski et al., 2021, 2022], cross-modal image-text retrieval [Cheng et al., 2022], and word-spotting [Riba et al., 2021]. Additionally, researchers from KU Leuven and ETH Zurich used *Smooth-AP* as the objective function for self-supervised image representation learning [Varamesh et al., 2020]. In this case, the Smooth-AP objective was used to rank different views/augmentations of the same instance above images of different instances in terms of similarity during training, and was shown to outperform state-of-the-art approaches such as SwAV [Caron et al., 2020] and SimCLR [T. Chen et al., 2020] in certain training settings. The authors focus on the use of real-world, non object-centric

training datasets. In this case, standard augmentation techniques for generating different views of the same instance may result in *noisy positives*, *i.e.* two images depicting very different content. A key characteristic of Smooth-AP is that the ranking of *hard positives* is not considered as important as the ranking of less hard positives; a core characteristic of the Average Precision metric. The authors in [Varamesh et al., 2020] argue that this characteristic allows the model to ignore very hard *noisy* positives, resulting in faster convergence on real-world datasets.

The introduction of *Smooth-AP* has motivated a great number of research projects focused on directly optimising Average Precision. These projects include works focusing on the direct improvement of Smooth-AP through analysis of gradients during training [Ramzi et al., 2021; Z. Li et al., 2021], and also those that take different approaches to direct AP optimisation [G. Wang et al., 2021; Shu et al., 2022; Tao et al., 2021], including the offer of convergence guarantees [Qi et al., 2021].

The greatest impact of the research in Chapter 2 has been the motivation of other works which demonstrate the benefits of end-to-end training in other target settings with different non-differentiable evaluation metrics. Such works include those focused on the direct optimisation of the recall@K metric [Patel et al., 2022], the False/True Acceptance Rate metrics [J. Liu et al., 2022], the area under the ROC curve [Tang et al., 2022], or the Kendall’s Tau metric [H. Gu et al., 2022]. In all of these published research projects, the technique popularised in Chapter 2 of smoothing non-differentiable metrics by replacing Heaviside step functions with a smooth sigmoid was used.

In Chapter 2 we argue that some of the datasets used for comparing image retrieval models are too small to reliably differentiate the performance of different methods. To this end, we introduce a new challenging large-scale image retrieval evaluation set constructed from the INaturalist dataset [Van Horn et al., 2018]. A further impact of this work has been the adoption of this evaluation set in the image retrieval community as a standard benchmark for comparing models [Xing et al., 2021; Z. Yang et al., 2022].

In Chapter 3 we present a novel method for enabling end-to-end training for the task of targeted visual editing by using self-supervision to synthesise the target data (example edits) which does not otherwise exist in sufficient quantity. The resulting trained model significantly outperforms all prior work, which instead circumvent the problem of a lack of example edits by learning task-agnostic priors that are not

necessarily optimal for the editing task. A further achievement of Chapter 3 is to present the advantages of *visual* editing. A large focus of prior work [Bau et al., 2021; Shi et al., 2021] is to consider the image editing task, but while using textual descriptions for specifying the desired change. Instead, we argue that specifying the edits visually offers a far more fine-grained and explicit level of control, which ultimately results in a more useful editor.

VoxMovies. In Chapter 4 we introduce a challenging dataset of speech segments from movies, *VoxMovies*, for analysing domain adaptation for speaker verification models. We publicly released the dataset along with several challenging evaluation sets for analysing how well models could adapt to this setting where actors are often intentionally trying to disguise their voice. A significant achievement of this work was to demonstrate that current state-of-the-art speaker verification models, which achieve near perfect performance on existing benchmarks, struggle to adapt to this new target domain. As a result, *VoxMovies* has been widely used in the literature for quantifying if trained speaker verification models are robust to changes in domain [X. Liu, Sahidullah, & Kinnunen, 2021; X. Liu et al., 2022].

Furthermore, *VoxMovies* was used as a key component of the VoxSRC 2020 workshop challenge [Nagrani, Chung, Huh, et al., 2020b]. One goal of this annual challenge is to test state-of-the-art speaker verification models on challenging, in-the-wild data, via the release of difficult hidden test sets which participants compete for the best results on. The focus of the 2020 verification challenge was on how well models could adapt to *VoxMovies* data, which presented a significant challenge to the highly optimised state-of-the-art competing systems.

Person Labelling for Human-Centric Video-Understanding. In this thesis, we present methods for understanding *who* is present in a human-centric video, as a key component of video understanding. In Chapter 5 we present a method for automatically labelling faces in unlabelled video archives, while using significantly less supervision than prior work. Our proposed method takes a human-inspired approach for labelling the people throughout the videos, by collecting evidence from audio-visual biometric signals, image-search engines, speech, and text found on screen. Seeing as it does not rely on any human supervision beyond simply the name of the TV-program or movie, our proposed method constitutes a feasible automatic annotation solution for the contemporary problem of

large unlabelled video archives.

The work presented in Chapter 5 was awarded the Best Student Paper Award at the IEEE conference on Multimedia Information Processing and Retrieval 2021, and the Runner-Up Research prize at the annual British Library Labs awards in 2020. Our proposed system can be used to automatically annotate identities in Computer Vision datasets based upon human-centric videos. One such example is the Condensed Movies Dataset [Bain et al., 2020], where our system automatically annotated 8,375 different characters in 25,760 (78%) of the constituent videos. The Condensed Movies dataset targets the difficult story-understanding task by training models to retrieve video clips using high level captions that describe story-lines. Utilising the annotations of *who* was present in each scene as an additional source of supervision resulted in significant performance improvements from the models. Furthermore, in this work we presented a method for robustly collecting evidence of a person’s identity from image search engines by automatically determining if a person was *famous* or not and removing outlier images. This idea of automatically determining the main subject identity in a set of images was built upon in [Franc & Yermakov, 2021].

In Chapter 6 we focus on the task of *person-clustering* in TV-shows and movies using multi-modal signals from a person’s face, voice and body appearance. Our main contribution was to build a system that exploited the complementarity between multi-modal signals arising from the same stimulus in order to cluster people by identity. In doing so, our proposed method has *person-level* rather than just *face-level* reasoning; a factor that we demonstrate to be advantageous for downstream tasks such as story understanding through character co-occurrence. We additionally release the *Video Person-Clustering Dataset*; the largest dataset of its kind, containing annotated body-tracks, face-tracks and voice-tracks for a wide range of characters. Beyond multi-modal person clustering, researchers have utilised the annotations in *Video Person-Clustering Dataset* for understanding the representation and depiction of background characters in TV-broadcasts [R. Sharma & Narayanan, 2022a], and for audio-visual speaker diarisation [R. Sharma & Narayanan, 2022b].

7.2 What Comes Next?

In this Section, we present some open-challenges that still exist in the areas covered in this Thesis.

7.2.1 Multi-Task Multi-Modal End-to-End Learning

In this thesis, we present methods for the end-to-end training of task-specific models, which are experts at performing specific tasks such as image retrieval (Chapter 2), and targeted visual editing (Chapter 3), often using domain-specific architectures, such as the 2D CNN [K. He et al., 2016b] for image inputs. In contrast, humans are able to harness overlapping skills in order to perform a large number of tasks without having to *re-train*. Training task-specific models for all possible tasks is clearly not feasible. An exciting future research direction is hence the end-to-end training of models that can perform any number of tasks, and using a wide variety of modalities.

In the NLP community, progress was made towards this goal by using increasingly flexible and task-agnostic ways of pre-training language representations, which could then be fine-tuned for a large number of downstream tasks [Radford & Narasimhan, 2018; Devlin et al., 2019a]. More recently, researchers found that *scale* (in terms of data and architecture) was key for unlocking *few-shot* adaptability to a wide range of tasks [T. B. Brown et al., 2020; Alayrac et al., 2022]. These works showed that large-scale models could rapidly and effectively adapt to a new task, given just a few demonstrations in the form of a prompt, without any parameter updates.

A challenge exists in how to train such models to use any number of input modalities, with promising directions existing in the form of architectures that can handle multiple modality inputs such as the transformer [Vaswani et al., 2017] and perceiver IO [Jaegle et al., 2022]. Currently, exciting progress has been made with models such as Flamingo [Alayrac et al., 2022], which demonstrates few-shot adaptation capabilities using vision and language, and Gato [Reed et al., 2022], a generalised agent that is able to perform a wide number of multi-modal tasks without re-training. However, a key issue is that, at the time of writing in October 2022, it is exceptionally computationally expensive to explore this research direction, making it largely unfeasible for anyone except the largest of industry labs to make progress in this research space. A further problem exists in how to compare these models, due to the

reliance on internet-scale pre-training datasets that are largely not shared between labs (an exception is the emergence of some open, freely accessible large-scale datasets such as LAION [Schuhmann et al., 2021]).

Multi-Modality and Temporal Reasoning for Video Understanding. The goal of the second part of this thesis was to improve human-centric video understanding via the use of uni-modal representations from audio and visual modalities. There is however much more progress to be made, and in order to build robust human-like video understanding systems, it is essential that we train models to harness both end-to-end trained multi-modal representations using all modalities that are available to human perception systems, and also temporal reasoning. Prior works have learnt multi-modal representations for video understanding [Y. Liu et al., 2019; Aytar et al., 2017], but challenges must still be overcome in dealing with the varying granularities of different modalities [Alayrac et al., 2020], as well as constructing large enough multi-modal datasets .

When solving video understanding tasks, humans exhibit long-term temporal reasoning. These skills are essential for understanding motivations, intent, or for giving explanations for a certain series of events. Datasets have been proposed for evaluating the long-range temporal reasoning capabilities of models, including long-range video retrieval [Bain et al., 2020], and video question answering datasets [Tapaswi et al., 2016; Zellers et al., 2019]. However, challenges remain in how to aggregate information temporally, and how to avoid reliance on biases present in datasets that make long-term reasoning unnecessary for solving the task [Sevilla-Lara et al., 2021; S. Xie et al., 2018].

Self-Supervision for Human-Centric Story-Line Understanding. The goal of Chapters 5 and 6 was the training of models for understanding *who* is present in a scene, as an essential enabling technology for the task of human-centric story-line understanding. Many open challenges exist for training models for this task, which is focused on understanding the motivations, intentions and actions of human characters in videos. For example, a key issue is the availability of sufficient amounts of labelled data for supervision. This is largely due to the cost of annotating such long-range, subjective labels. An exciting future research direction is hence the exploration of self-supervised approaches to human-centric story-line understanding.

Current proxy-tasks for self-supervised video representation learning take creative approaches such as future frame prediction [Han et al., 2019], determining the correct ordering of a sequence of frames [Misra et al., 2016], and temporal cycle-consistency [Dwibedi et al., 2019]. Understandably, in these cases, the unit of prediction in the proxy-tasks is a frame-level feature representation. However, it is not clear that the resulting learnt representations would be useful for the downstream task of story-line understanding. For extending these approaches to story-line understanding, an exciting research direction is to instead move to using higher-level abstracted concepts as the unit of prediction. For example, proxy tasks could include predicting future interactions, given current character motivations, or predicting the correct ordering of a sequence of clips that are key to the story-line. For this, challenges would need to be overcome in how to learn at this increased level of abstraction without any manual annotation.

Ethical Considerations and Broader Impacts. Finally, it is important to note that as training datasets grow to *internet-scale* [T. B. Brown et al., 2020; P. Goyal et al., 2022], and the potential real-world applications of trained models increase in number, that an essential direction of future work is on the increased ethical and broader impact considerations of such models. It is vital that we consider the broader impact on the academic and social world of our research going forward [Sim et al., 2021; Bender et al., 2021], and that ethical discussions are a key part of the exciting future of artificial machine learning research.

7.3 Conclusion

In this thesis we explore and overcome a series of challenges as related to the training of deep neural networks. First, we explore the enabling of end-to-end learning, via the direct optimisation of non-differentiable objective functions (Chapter 2), and the self-supervised synthesis of target data (Chapter 3). Second, we develop techniques for audio-visual human-centric video understanding, where we harnessed both the redundancy and complementarity between the two modalities for this task. Overcoming such challenges in training deep neural networks will provide the means for building artificial machine learning systems with human-level perception capabilities.

Bibliography

- Abdal, R., Qin, Y., & Wonka, P. (2019). Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*.
- Abdal, R., Qin, Y., & Wonka, P. (2020). Image2stylegan++: How to edit the embedded images? In *CVPR*.
- Abdal, R., Zhu, P., Femiani, J., Mitra, N. J., & Wonka, P. (2021). Clip2stylegan: Unsupervised extraction of stylegan edit directions. *arXiv:2112.05219 [cs.CV]*.
- Afouras, T., Chung, J. S., & Zisserman, A. (2020). Asr is all you need: Cross-modal distillation for lip reading. In *Proc. ICASSP*.
- Afouras, T., Owens, A., Chung, J. S., & Zisserman, A. (2020). Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*.
- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... Simonyan, K. (2022). *Flamingo: a Visual Language Model for Few-Shot Learning*. arXiv. Retrieved from <https://arxiv.org/abs/2204.14198> doi: 10.48550/ARXIV.2204.14198
- Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., Fauw, J. D., ... Zisserman, A. (2020). Self-supervised multimodal versatile networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.
- Albanie, S., Nagrani, A., Vedaldi, A., & Zisserman, A. (2018, October). Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia (ACM MM)*.
- Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Arandjelović, R., & Zisserman, A. (2012). Three things everyone should know to improve object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Arandjelovic, R., & Zisserman, A. (2013). All about vlad. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Arandjelović, R., & Zisserman, A. (2018). Objects that sound. *ArXiv, abs/1712.06651*.

- Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.
- Awad, G., Butt, A., Curtis, K., Lee, Y., Fiscus, J., Godil, A., . . . Quenot, G. (2019). Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. *ArXiv, abs/2009.09984*.
- Aytar, Y., Vondrick, C., & Torralba, A. (2017). *See, Hear, and Read: Deep Aligned Representations*.
- Babenko, A., Slesarev, A., Chigorin, A., & Lempitsky, V. (2014). Neural codes for image retrieval. In *Proceedings of the European Conference on Computer Vision*.
- Bain, M., Nagrani, A., Brown, A., & Zisserman, A. (2020). Condensed movies: Story based retrieval with contextual embeddings. In *Asian Conference on Computer Vision (ACCV)*.
- Barlow, H. B. (1989). Unsupervised learning. *Neural computation*, 1(3), 295–311.
- Bau, D., Andonian, A., Cui, A., Park, Y., Jahanian, A., Oliva, A., & Torralba, A. (2021). Paint by word. *arXiv:2103.10951 [cs.CV]*.
- Bau, D., Strobel, H., Peebles, W., Wulff, J., Zhou, B., Zhu, J.-Y., & Torralba, A. (2019). Semantic photo manipulation with a generative image prior. *ACM Trans. Graph.*.
- Bau, D., Zhu, J.-Y., Strobel, H., Lapedriza, A., Zhou, B., & Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*.
- Bau, D., Zhu, J.-Y., Wulff, J., Peebles, W., Strobel, H., Zhou, B., & Torralba, A. (2019). Inverting layers of a large generator. In *ICLR 2019 Debugging Machine Learning Models Workshop*.
- Baumel, M., Tapaswi, M., & Stiefel, R. (2013). Semi-supervised learning with constraints for person identification in multimedia data. In *Proc. CVPR*.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
- Bendris, M., Favre, B., Charlet, D., Damnati, G., Senay, G., Auguste, R., & Martinet, J. (2013). Unsupervised face identification in tv content using audio-visual sources. In *2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI)* (pp. 243–249).

- Berg, T. L., Berg, A. C., Edwards, J., Maire, M., White, R., Teh, Y.-W., . . . Forsyth, D. A. (2004). Names and faces in the news. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Berg, T. L., & Forsyth, D. A. (2006). Animals on the web. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York*.
- Berrada, L., Zisserman, A., & Kumar, M. P. (2018). Smooth loss functions for deep top-k classification. *International Conference on Learning Representations*.
- Bhattacharya, G., Alam, J., & Kenny, P. (2019). Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Bhattacharya, G., Monteiro, J., Alam, J., & Kenny, P. (2019). Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Bredin, H. (2015). *Mediaeval Person Discovery task, Evaluation Package 2015*. <https://github.com/MediaevalPersonDiscoveryTask/EvaluationPackage2015>. GitHub.
- Brown, A., Coto, E., & Zisserman, A. (2021). Automated video labelling: Identifying faces by corroborative evidence. In *MIPR*.
- Brown, A., Huh, J., Nagrani, A., Chung, J. S., & Zisserman, A. (2021). Playing a part: Speaker verification at the movies. In *Proc. ICASSP*.
- Brown, A., Xie, W., Kalogeiton, V., & Zisserman, A. (2020). Smooth-AP: Smoothing the path towards large-scale image retrieval. In *European Conference on Computer Vision (ECCV)*.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Amodei, D. (2020). Language models are few-shot learners. *CoRR*, *abs/2005.14165*. Retrieved from <https://arxiv.org/abs/2005.14165>.
- Buckley, C., Salton, G., Allan, J., & Singhal, A. (1995). Automatic query expansion using SMART. In *TREC-3 Proc.*
- Burges, C. J., Ragno, R., & Le, Q. V. (2007). Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems*.

- Cai, W., Cai, D., Liu, W., Li, G., & Li, M. (2017). Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion. In *INTERSPEECH*.
- Cai, W., Chen, J., & Li, M. (2018). Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. *Speaker Odyssey*.
- Cai, Z., & Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. In *Proc. CVPR*.
- Cakir, F., He, K., Xia, X., Kulis, B., & Sclaroff, S. (2019a). Deep metric learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Cakir, F., He, K., Xia, X., Kulis, B., & Sclaroff, S. (2019b). *FastAP: Deep Metric Learning to Rank*. <https://github.com/kunhe/Deep-Metric-Learning-Baselines>. GitHub.
- Canseco, L., Lamel, L., & Gauvain, J.-L. (n.d.). A comparative study using manual and automatic transcriptions for diarization. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005*. (pp. 415–419).
- Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018a). Vggface2: A dataset for recognising faces across pose and age. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*.
- Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018b). VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*.
- Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., & Li, H. (2007a). Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the International Conference on Machine Learning* (pp. 129–136).
- Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., & Li, H. (2007b). Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the International Conference on Machine Learning*.
- Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9912–9924). Curran Associates, Inc. Retrieved from <https://proceedings.neurips.cc/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf>

- Chai, L., Wulff, J., & Isola, P. (2021a). Using latent space regression to analyze and leverage compositionality in {gan}s. In *International Conference on Learning Representations*. Retrieved from <https://openreview.net/forum?id=sjuuTm4vj0>
- Chai, L., Wulff, J., & Isola, P. (2021b). Using latent space regression to analyze and leverage compositionality in gans. In *ICLR*.
- Chapelle, O., Le, Q., & Smola, A. (2007). Large margin optimization of ranking measures. In *Advances in Neural Information Processing Systems*.
- Chatfield, K., & Zisserman, A. (2012). Visor: Towards on-the-fly large-scale object category retrieval. In *Proceedings of the Asian Conference on Computer Vision*. Springer.
- Chen, H., Xie, W., Vedaldi, A., & Zisserman, A. (2019). Autocorrect: Deep inductive alignment of noisy geometric annotations. In *Proceedings of the British Machine Vision Conference*.
- Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. In *ICML*.
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.
- Chen, Z., Xie, Z., Zhang, W., & Xu, X. (2017). Resnet and model fusion for automatic spoofing detection. In *INTERSPEECH*.
- Chen, Z., Zhang, W., Deng, B., Xie, H., & Gu, X. (2019). Name-face association with web facial image supervision. *Multimedia Systems*.
- Cheng, Q., Tan, Z., Wen, K., Chen, C., & Gu, X. (2022). Semantic pre-alignment and ranking learning with unified framework for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 1-1. doi: 10.1109/TCSVT.2022.3182549
- Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*.
- Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Chum, O., Mikulik, A., Perd'och, M., & Matas, J. (2011). Total recall II: Query expansion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

- Chum, O., Philbin, J., Sivic, J., Isard, M., & Zisserman, A. (2007). Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proceedings of the 11th International Conference on Computer Vision, Rio de Janeiro, Brazil*.
- Chung, J. S., Huh, J., & Mun, S. (2020). Delving into voxceleb: environment invariant speaker recognition. *Speaker Odyssey Workshop*.
- Chung, J. S., Huh, J., Mun, S., Lee, M., Heo, H. S., Choe, S., . . . Han, I. (2020a). In defence of metric learning for speaker recognition. In *INTERSPEECH*.
- Chung, J. S., Huh, J., Mun, S., Lee, M., Heo, H. S., Choe, S., . . . Han, I. (2020b). In defence of metric learning for speaker recognition. In *INTERSPEECH*.
- Chung, J. S., Nagrani, A., & Zisserman, A. (2018). VoxCeleb2: Deep speaker recognition. In *INTERSPEECH*.
- Chung, J. S., & Zisserman, A. (2016). Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*.
- Chung, S., Chung, J. S., & Kang, H. (2019). Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (p. 3965-3969).
- Chung, S.-W., Chung, J. S., & Kang, H.-G. (2019). Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Cinbis, R. G., Verbeek, J., & Schmid, C. (2011b). Unsupervised metric learning for face identification in tv video. In *Proceedings of the International Conference on Computer Vision*.
- Cinbis, R. G., Verbeek, J. J., & Schmid, C. (2011a). Unsupervised metric learning for face identification in TV video. In *Proceedings of the International Conference on Computer Vision* (p. 1559-1566).
- Cour, T., Sapp, B., Jordan, C., & Taskar, B. (2009). Learning from ambiguously labeled images. In *Proc. CVPR*.
- Cour, T., Sapp, B., Nagle, A., & Taskar, B. (2010). Talking pictures: Temporal grouping and dialog-supervised person recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Crowson, K. (n.d.). *VQGAN-CLIP*. (<https://github.com/nerdyrodent/VQGAN-CLIP>, 2021)
- Dai, B., & Wipf, D. (2019). Diagnosing and enhancing VAE models. In *ICLR*.

- Damen, D., Doughty, H., Farinella, G. M., Furnari, A., Kazakos, E., Ma, J., ... Wray, M. (2020). Rescaling egocentric vision. *CoRR*, *abs/2006.13256*. Retrieved from <https://arxiv.org/abs/2006.13256>
- Das, R. K., Yang, J., & Li, H. (2019). Long range acoustic features for spoofed speech detection. In *INTER-SPEECH*.
- de Amorim, R. C. (2012). Constrained clustering with minkowski weighted k-means. In *CINTI*.
- Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*.
- Deng, D., Liu, H., Li, X., & Cai, D. (2018). Pixellink: Detecting scene text via instance segmentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019a). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019b, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/N19-1423> doi: 10.18653/v1/N19-1423
- Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., ... Tang, J. (2021). Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*.
- Doersch, C., Gupta, A., & Efros, A. A. (2015, dec). Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision (ICCV)* (p. 1422-1430). Los Alamitos, CA, USA: IEEE Computer Society. Retrieved from <https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.167> doi: 10.1109/ICCV.2015.167

- Dolhansky, B., Bitton, J., Pfau, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). *The DeepFake Detection Challenge Dataset*.
- Duan, Y., Zheng, W., Lin, X., Lu, J., & Zhou, J. (2018). Deep adversarial metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Dwivedi, D., Aytar, Y., Tompson, J., Sermanet, P., & Zisserman, A. (2019, June). Temporal cycle-consistency learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018, July). HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 31–36). Melbourne, Australia: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/P18-2006> doi: 10.18653/v1/P18-2006
- Edelman, G. M. (1987). *Neural Darwinism: The theory of neuronal group selection*. Basic books.
- Engilberge, M., Chevallier, L., Pérez, P., & Cord, M. (2019). Sodeep: a sorting deep net to learn ranking loss surrogates. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Ercolessi, P., Bredin, H., & Sénac, C. (2012). Stoviz: story visualization of tv series. In *Proc. ACMMM*.
- Esser, P., Rombach, R., Blattmann, A., & Ommer, B. (2021). Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. In *NeurIPS*.
- Esser, P., Rombach, R., & Ommer, B. (2020). A disentangling invertible interpretation network for explaining latent representations. In *CVPR*.
- Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *CVPR*.
- Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J., & Zisserman, A. (2009). *The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results*. <http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html>.
- Everingham, M., Sivic, J., & Zisserman, A. (2006). “Hello! My name is... Buffy” – automatic naming of characters in TV video. In *Proceedings of the 17th British Machine Vision Conference, Edinburgh*.
- Everingham, M., Sivic, J., & Zisserman, A. (2009). Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5).

- Fan, Y., Kang, J., Li, L., Li, K., Chen, H., Cheng, S., ... Wang, D. (2020). Cn-celeb: a challenging chinese speaker recognition dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Fauw, J. D., Dieleman, S., & Simonyan, K. (2019). Hierarchical autoregressive image models with auxiliary decoders. *arXiv:1903.04933 [cs.CV]*.
- Fergus, R., Fei-Fei, L., Perona, P., & Zisserman, A. (2005). Learning object categories from Google's image search. In *Proceedings of the 10th International Conference on Computer Vision, Beijing, China*.
- Fitzgibbon, A. W., & Zisserman, A. (2002). On affine invariant clustering and automatic cast listing in movies. In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark* (Vol. 3, pp. 304–320). Springer-Verlag.
- Franc, V., & Yermakov, A. (2021). Dominant subject recognition by bayesian learning. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)* (pp. 01–05).
- Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., & Taigman, Y. (2022). *Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors*. arXiv.
- Galatolo, F., Cimino, M., & Vaglini, G. (2021). Generating images from caption and vice versa via clip-guided generative latent space search. *Proceedings of the International Conference on Image Processing and Vision Engineering*.
- Garcia-Romero, D., & McCree, A. (2014). Supervised domain adaptation for i-vector based speaker recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Garcia-Romero, D., McCree, A., Shum, S., & Vaquero, C. (n.d.). Unsupervised domain adaptation for i-vector speaker recognition..
- Garcia-Romero, D., McCree, A., Snyder, D., & Sell, G. (2020). Jhu-hltcoe system for the voxsrc speaker recognition challenge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Garcia-Romero, D., Zhang, X., McCree, A., & Povey, D. (2014). Improving speaker recognition performance in the domain adaptation challenge using deep neural networks. In *SLT*.

- Gay, P., Dupuy, G., Lailier, C., Odobez, J.-M., Meignier, S., & Deléglise, P. (2014). Comparison of two methods for unsupervised person identification in tv shows. In *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)* (pp. 1–6).
- Ge, W., Huang, W., Dong, D., & Scott, M. (2018). Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision*.
- Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, 23(3), 183-209. Retrieved from <https://www.sciencedirect.com/science/article/pii/001002778690034X> doi: [https://doi.org/10.1016/0010-0277\(86\)90034-X](https://doi.org/10.1016/0010-0277(86)90034-X)
- Ghaleb, E., Tapaswi, M., Al-Halah, Z., Ekenel, H. K., & Stiefelhagen, R. (2015). Accio: A data set for face track retrieval in movies across age. In *Proc. ICMR*.
- Ghosh, P., Zietlow, D., Black, M. J., Davis, L. S., & Hu, X. (2021). Invgan: Invertible gans. *arXiv:2112.04598 [cs.CV]*.
- Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*. Retrieved from <https://openreview.net/forum?id=Slv4N2l0->
- Gominski, D., Gouet-Brunet, V., & Chen, L. (2021). Connecting images through sources: Exploring low-data, heterogeneous instance retrieval. *Remote Sensing*.
- Gominski, D., Gouet-Brunet, V., & Chen, L. (2022, June). Cross-dataset learning for generalizable land use scene classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (p. 1382-1391).
- Gordo, A., Almazan, J., Revaud, J., & Larlus, D. (2016). Deep image retrieval: Learning global representations for image search. In *Proceedings of the European Conference on Computer Vision*.
- Goyal, A., Lamb, A., Zhang, Y., Zhang, S., Courville, A., & Bengio, Y. (2016). Professor forcing: A new algorithm for training recurrent networks. In *NeurIPS*.
- Goyal, P., Duval, Q., Seessel, I., Caron, M., Singh, M., Misra, I., ... Bojanowski, P. (2022). Vision models are more robust and fair when pretrained on uncured images without supervision. *arXiv preprint arXiv:2202.08360*.

- Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., . . . Malik, J. (2018, June). Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Gu, H., Li, J., Fu, G., Wong, C., Chen, X., & Zhu, J. (2022, June). Autoloss-gms: Searching generalized margin-based softmax loss function for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 4744-4753).
- Guan, S., Tai, Y., Ni, B., Zhu, F., Huang, F., & Yang, X. (2020). Collaborative learning for faster stylegan embedding. *arXiv:2007.01758 [cs.CV]*.
- Guiver, J., & Snelson, E. (2008). Learning to rank with sofrank and gaussian processes. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Han, T., Xie, W., & Zisserman, A. (2019). Video representation learning by dense predictive coding. In *Workshop on Large Scale Holistic Video Understanding, ICCV*.
- Han, T., Xie, W., & Zisserman, A. (2020). Self-supervised co-training for video representation learning. In *Neurips*.
- Härkönen, E., Hertzmann, A., Lehtinen, J., & Paris, S. (2020). Ganspace: Discovering interpretable gan controls. *arXiv:2004.02546 [cs.CV]*.
- Harwood, B., Kumar, B., Carneiro, G., Reid, I., & Drummond, T. (2017). Smart mining for deep metric learning. In *Proceedings of the International Conference on Computer Vision*.
- Haurilet, M.-L., Tapaswi, M., Al-Halah, Z., & Stiefelwagen, R. (2016). Naming tv characters by watching and analyzing dialogs. In *Proc. WACV*.
- He, K., Cakir, F., Adel Bargal, S., & Sclaroff, S. (2018). Hashing as tie-aware learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- He, K., Lu, Y., & Sclaroff, S. (2018). Local descriptors optimized for average precision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Deep residual learning for image recognition. In *Proc. CVPR*.

- He, Y., Cao, K., Li, C., & Loy, C. C. (2018). Merge or not? learning to group faces via imitation learning. In *AAAI*.
- Heilbron, F. C., Escorcia, V., Ghanem, B., & Niebles, J. C. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 961-970). doi: 10.1109/CVPR.2015.7298698
- Henderson, P., & Ferrari, V. (2016). End-to-end training of object class detectors for mean average precision. In *Proceedings of the Asian Conference on Computer Vision*.
- Heo, H. S., Lee, B.-J., Huh, J., & Chung, J. S. (2020). Clova baseline system for the voxceleb speaker recognition challenge 2020. *arXiv preprint arXiv:2009.14153*.
- Hermans, A., Beyer, L., & Leibe, B. (2017). In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*.
- Hirson, A., & Duckworth, M. (1993). Glottal fry and voice disguise: a case study in forensic phonetics. *Journal of biomedical engineering*.
- Ho, J., Yang, M.-H., Lim, J., Lee, K.-C., & Kriegman, D. (2003). Clustering appearances of objects under varying illumination conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Hoai, M., & Zisserman, A. (2014). Thread-safe: Towards recognizing human actions across shot boundaries. In *Proceedings of the Asian Conference on Computer Vision*.
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735-1780. Retrieved from <https://doi.org/10.1162/neco.1997.9.8.1735> doi: 10.1162/neco.1997.9.8.1735
- Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., . . . Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the International Conference on Machine Learning*.
- Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *ICLR*.

- Holub, A., Moreels, P., & Perona, P. (2008). Unsupervised clustering for google searches of celebrity images. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*
- Hu, J., Shen, L., & Sun, G. (2018a). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Hu, J., Shen, L., & Sun, G. (2018b). Squeeze-and-excitation networks. In *Proc. CVPR*.
- Huang, Q., Liu, W., & Lin, D. (2018). Person search in videos with one portrait through visual and temporal links. In *Proceedings of the European Conference on Computer Vision*.
- Huang, Q., Xiong, Y., Rao, A., Wang, J., & Lin, D. (2020). Movienet: A holistic dataset for movie understanding. In *Proc. ECCV*.
- Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2017, jul). Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4).
- Isola, P., & Liu, C. (2013). Scene collaging: Analysis and synthesis of natural images with semantic layers. In *ICCV*.
- Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *CVPR*.
- Issenhuth, T., Tanielian, U., Mary, J., & Picard, D. (2021). Edibert, a generative model for image editing. *arXiv:2111.15264 [cs.CV]*.
- Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., ... Carreira, J. (2022). Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*. Retrieved from <https://openreview.net/forum?id=fILj7WpI-g>
- Jahani, A., Chai, L., & Isola, P. (2020). On the "steerability" of generative adversarial networks. In *ICLR*.
- Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.
- Jarvis, R. A., & Patrick, E. A. (1973). Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Computers*.
- Jégou, H., Douze, M., & Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the European Conference on Computer Vision*.

- Jégou, H., Douze, M., & Schmid, C. (2011). Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., & Schmid, C. (2011). Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Jin, S., Su, H., Stauffer, C., & Learned-Miller, E. (2017). End-to-end face detection and cast grouping in movies using erdos-renyi clustering. In *Proceedings of the International Conference on Computer Vision*.
- Joon Oh, S., Benenson, R., Fritz, M., & Schiele, B. (2015). Person recognition in personal photo collections. In *Proc. ICCV*.
- Jousse, V., Petit-Renaud, S., Meignier, S., Esteve, Y., & Jacquin, C. (2009). Automatic named identification of speakers using diarization and asr systems. In *Proc. ICASSP*.
- Kalogeiton, V., & Zisserman, A. (2020). Constrained video face clustering using 1nn relations. In *Proceedings of the British Machine Vision Conference*.
- Kang, J., Liu, R., Li, L., Cai, Y., Wang, D., & Zheng, T. F. (2020). Domain-invariant speaker vector projection by model-agnostic meta-learning. *arXiv preprint arXiv:2005.11900*.
- Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 3128-3137). doi: 10.1109/CVPR.2015.7298932
- Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data. In *NeurIPS*.
- Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *CVPR*.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In *CVPR*.
- Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... Zisserman, A. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

- Ke, Y., Sukthankar, R., & Hebert, M. (2005). Efficient visual event detection using volumetric features. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* (Vol. 1, p. 166-173 Vol. 1). doi: 10.1109/ICCV.2005.85
- Kim, D., Cho, D., Yoo, D., & Kweon, I. (2018, mar). Learning image representations by completing damaged jigsaw puzzles. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (p. 793-802). Los Alamitos, CA, USA: IEEE Computer Society. Retrieved from <https://doi.ieeecomputersociety.org/10.1109/WACV.2018.00092> doi: 10.1109/WACV.2018.00092
- Kim, H., Choi, Y., Kim, J., Yoo, S., & Uh, Y. (2021). Exploiting spatial dimensions of latent in gan for real-time image editing. In *CVPR*.
- Kim, W., Goyal, B., Chawla, K., Lee, J., & Kwon, K. (2018). Attention-based ensemble for deep metric learning. In *Proceedings of the European Conference on Computer Vision*.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*.
- Köstinger, M., Wohlhart, P., Roth, P., & Bischof, H. (2011). Learning to recognize faces from videos and weakly related information cues. In *avss*.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25). Curran Associates, Inc. Retrieved from <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>
- Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). Hmdb: A large video database for human motion recognition. In *2011 International Conference on Computer Vision* (p. 2556-2563). doi: 10.1109/ICCV.2011.6126543
- Kukleva, A., Tapaswi, M., & Laptev, I. (2020). Learning interactions and relationships between movie characters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Lai, Z., & Xie, W. (2019). Self-supervised learning for video correspondence flow. In *Proceedings of the British Machine Vision Conference*.
- Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3(3), 299-321. Retrieved from <https://www.sciencedirect.com/science/article/pii/0885201488900147> doi: [https://doi.org/10.1016/0885-2014\(88\)90014-7](https://doi.org/10.1016/0885-2014(88)90014-7)

- Larsson, G., Maire, M., & Shakhnarovich, G. (2016). Learning representations for automatic colorization. In *European Conference on Computer Vision (ECCV)*.
- Law, M. T., Urtasun, R., & Zemel, R. S. (2017). Deep spectral clustering learning. In *Proceedings of the International Conference on Machine Learning*.
- Le, D.-D., & Satoh, S. (2008). Unsupervised face annotation by mining the web. In *ICDM*.
- Le, N., Bredin, H., Sargent, G., India, M., Lopez-Otero, P., Barras, C., ... Freire, I. L. (2017). Towards large scale multimedia indexing: A case study on person discovery in broadcast news. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing* (p. 18).
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4), 541–551.
- Lei, J., Yu, L., Bansal, M., & Berg, T. L. (2018). Tvqa: Localized, compositional video question answering. In *EMNLP*.
- Li, K., Huang, Z., Cheng, Y.-C., & Lee, C.-H. (2014). A maximal figure-of-merit learning approach to maximizing mean average precision with deep neural network based classifiers. In *Proc. ICASSP*.
- Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proc. CVPR*.
- Li, Z., Min, W., Song, J., Zhu, Y., & Jiang, S. (2021). Rethinking ranking-based loss functions: Only penalizing negative instances before positive ones is enough. *CoRR*, abs/2102.04640. Retrieved from <https://arxiv.org/abs/2102.04640>
- Liao, C.-F., Tsao, Y., Lee, H.-Y., & Wang, H.-M. (2018). Noise adaptive speech enhancement using domain adversarial training. *arXiv preprint arXiv:1807.07501*.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Dollar, P. (2014). Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*.
- Lin, W.-A., Chen, J.-C., Castillo, C. D., & Chellappa, R. (2018). Deep density clustering of unconstrained faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Lin, W.-H., Jin, R., & Hauptmann, A. (2003). Web image retrieval re-ranking with relevance model. In *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*.

- Lipton, Z. C., & Tripathi, S. (2017). Precise recovery of latent vectors from generative adversarial networks. *arXiv:1702.04782 [cs.LG]*.
- Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In *ECCV*.
- Liu, J., Qin, H., Wu, Y., & Liang, D. (2022). Anchorface: Boosting tar@far for practical face recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Liu, P., Qiu, X., & Huang, X. (2017). Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.
- Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Liu, X., Park, D. H., Azadi, S., Zhang, G., Chopikyan, A., Hu, Y., ... Darrell, T. (2021). More control for free! image synthesis with semantic diffusion guidance. *arXiv:2112.05744 [cs.CV]*.
- Liu, X., Sahidullah, M., & Kinnunen, T. (2021). Optimized power normalized cepstral coefficients towards robust deep speaker verification. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 185–190).
- Liu, X., Sahidullah, M., & Kinnunen, T. (2022). Learnable nonlinear compression for robust speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7962–7966).
- Liu, Y., Albanie, S., Nagrani, A., & Zisserman, A. (2019). *Use What You Have: Video Retrieval Using Representations From Collaborative Experts*.
- Liu, Y., Wang, Z., Jin, H., & Wassell, I. (2018). Synthetically supervised feature learning for scene text recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 435–451).
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *ICLR*.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*.
- Lu, J., Xu, C., Zhang, W., Duan, L.-Y., & Mei, T. (2019). Sampling wisely: Deep image embedding by top-k precision optimization. In *Proceedings of the International Conference on Computer Vision*.

- Luu, C., Bell, P., & Renals, S. (2020). Channel adversarial training for speaker verification and diarization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Lv, J., Liu, W., Zhou, L., Wu, B., & Ma, H. (2018). Multi-stream fusion model for social relation recognition from videos. In *MMM*.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008a). *Introduction to Information Retrieval*. Cambridge University Press.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008b). *Introduction to information retrieval*. Cambridge university press.
- Mansimov, E., Parisotto, E., Ba, J., & Salakhutdinov, R. (2016). Generating images from captions with attention. In *ICLR*.
- Marin-Jimenez, M. J., Kalogeiton, V., Medina-Suarez, P., & Zisserman, A. (2019). LAEO-Net: revisiting people Looking At Each Other in videos. In *Proc. CVPR*.
- Marin-Jimenez, M. J., Kalogeiton, V., Medina-Suarez, P., & Zisserman, A. (2020). LAEO-Net++: revisiting people Looking At Each Other in videos. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Matejka, P., Novotný, O., Plchot, O., & Burget, L. (2017). Analysis of score normalization in multilingual speaker recognition..
- Mauclair, J., Meignier, S., & Estève, Y. (2006). Speaker diarization: about whom the speaker is talking? In *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop* (pp. 1–6).
- Maze, B., Adams, J., Duncan, J. A., Kalka, N., Miller, T., Otto, C., . . . Cheney, J. (2018). Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*.
- McFee, B., & Lanckriet, G. R. (2010). Metric learning to rank. In *Proceedings of the International Conference on Machine Learning*.
- McLaren, M., Ferrer, L., Castan, D., & Lawson, A. (2016). The speakers in the wild (sitw) speaker recognition database..
- McLaren, M., Lawson, A., Ferrer, L., Castan, D., & Graciarena, M. (2015). The speakers in the wild speaker recognition challenge plan. *Interspeech 2016 Special Session, San Francisco*.
- Meng, Z., Li, J., & Gong, Y. (2018). Adversarial feature-mapping for speech enhancement. In *INTERSPEECH*.

- Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.
- Misra, I., Zitnick, C. L., & Hebert, M. (2016). Shuffle and learn: Unsupervised learning using temporal order verification. In *ECCV*.
- Mohapatra, P., Rolinek, M., Jawahar, C., Kolmogorov, V., & Pawan, K. (2018). Efficient optimization for rank-based loss functions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Mokady, R., Benaim, S., Wolf, L., & Bermano, A. (2018). Mask based unsupervised content transfer. *arXiv:1906.06558 [cs.CV]*.
- Movshovitz-Attias, Y., Toshev, A., Leung, T. K., Ioffe, S., & Singh, S. (2017). No fuss distance metric learning using proxies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Nagrani, A., Albanie, S., & Zisserman, A. (2018a). Learnable pins: Cross-modal embeddings for person identity. In *Proceedings of the European Conference on Computer Vision*.
- Nagrani, A., Albanie, S., & Zisserman, A. (2018b). Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Nagrani, A., Chung, J. S., Albanie, S., & Zisserman, A. (2020). Disentangled speech embeddings using cross-modal self-supervision. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Nagrani, A., Chung, J. S., Huh, J., Brown, A., Coto, E., Xie, W., . . . Zisserman, A. (2020a). *VoxSRC 2020: The Second VoxCeleb Speaker Recognition Challenge*.
- Nagrani, A., Chung, J. S., Huh, J., Brown, A., Coto, E., Xie, W., . . . Zisserman, A. (2020b). *VoxSRC 2020: The Second VoxCeleb Speaker Recognition Challenge*. arXiv. Retrieved from <https://arxiv.org/abs/2012.06867> doi: 10.48550/ARXIV.2012.06867
- Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2020). Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60, 101027.
- Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: a large-scale speaker identification dataset. In *INTERSPEECH*.
- Nagrani, A., & Zisserman, A. (2017a). From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script. In *BMVC*.

- Nagrani, A., & Zisserman, A. (2017b). From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script. In *Proceedings of the British Machine Vision Conference*.
- Nagrani, A., & Zisserman, A. (2018). From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script. *arXiv preprint arXiv:1801.10442*.
- Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... Chen, M. (2021). *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*.
- Nist 2018 speaker recognition evaluation plan [Computer software manual]. (2018 (accessed 31 July 2020)). (https://www.nist.gov/system/files/documents/2018/08/17/sre18_eval_plan_2018-05-31_v6.pdf, See Section 3.1)
- Noroozi, M., & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*.
- Oh Song, H., Jegelka, S., Rathod, V., & Murphy, K. (2017). Deep metric learning via facility location. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Oh Song, H., Xiang, Y., Jegelka, S., & Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Opitz, M., Waltner, G., Possegger, H., & Bischof, H. (2017). BIER: Boosting Independent Embeddings Robustly. In *Proceedings of the International Conference on Computer Vision*.
- Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., & Yokoya, N. (2016). Learning joint representations of videos and sentences with web image search. In *ECCV Workshops*.
- Otto, C., Wang, D., & Jain, A. K. (2017). Clustering millions of faces by identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Ozerov, A., Vigouroux, J.-R., Chevallier, L., & Pérez, P. (2013). On evaluating face tracks in movies. In *Proceedings of the IEEE International Conference on Image Processing*.
- Park, T., Liu, M.-Y., Wang, T.-C., & Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *CVPR*.

- Parkhi, O. M., Rahtu, E., Cao, Q., & Zisserman, A. (2020). Automated video face labelling for films and tv material. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Parkhi, O. M., Rahtu, E., Cao, Q., & Zisserman, A. (2020). Automated video face labelling for films and tv material. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Parkhi, O. M., Rahtu, E., & Zisserman, A. (2015). It’s in the bag: Stronger supervision for automated face labelling. In *ICCV Workshop: Describing and Understanding Video & The Large Scale Movie Description Challenge*.
- Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2012). On-the-fly specific person retrieval. In *International Workshop on Image Analysis for Multimedia Interactive Services*. IEEE.
- Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proceedings of the British Machine Vision Conference*.
- Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., & Lischinski, D. (2021). Styleclip: Text-driven manipulation of stylegan imagery. *arXiv:2103.17249 [cs.CV]*.
- Patel, Y., Tolias, G., & Matas, J. (2022). Recall@k surrogate loss with large batches and similarity mixup. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7502–7511).
- Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., & Efros, A. (2016). Context encoders: Feature learning by inpainting. In *CVPR*.
- Peebles, W., Peebles, J., Zhu, J.-Y., Efros, A. A., & Torralba, A. (2020). The hessian penalty: A weak prior for unsupervised disentanglement. In *ECCV*.
- Perronnin, F., Liu, Y., Sánchez, J., & Poirier, H. (2010). Large-scale image retrieval with compressed fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Pham, P. T., Moens, M.-F., & Tuytelaars, T. (2010). Naming persons in news video with label propagation. In *2010 IEEE International Conference on Multimedia and Expo* (pp. 1528–1533).
- Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis*.
- Poignant, J., Besacier, L., & Quénot, G. (2014). Unsupervised speaker identification in tv broadcast based on written names. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

- Poignant, J., Bredin, H., & Barras, C. (2017). Multimodal person discovery in broadcast tv: lessons learned from mediaeval 2015. *Multimedia Tools and Applications*.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Schwarz, P. (2011). The kald speech recognition toolkit. In *IEEE workshop on automatic speech recognition and understanding*.
- Press, O., Galanti, T., Benaim, S., & Wolf, L. (2019). Emerging disentanglement in auto-encoder based unsupervised image content transfer. In *ICLR*.
- Qi, Q., Luo, Y., Xu, Z., Ji, S., & Yang, T. (2021). Stochastic optimization of areas under precision-recall curves with provable convergence. In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems*. Retrieved from https://openreview.net/forum?id=Q_64PF6XNut
- Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., & Jin, R. (2019). Softtriple loss: Deep metric learning without triplet sampling. In *Proc. ICCV*.
- Qin, T., Liu, T.-Y., & Li, H. (2010). A general approximation framework for direct optimization of information retrieval measures. *Information Retrieval*.
- Radenovic, F., Tolias, G., & Chum, O. (2016). CNN image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Proceedings of the European Conference on Computer Vision*.
- Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434 [cs.LG]*.
- Radford, A., & Narasimhan, K. (2018). Improving language understanding by generative pre-training..
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.
- Ramanan, D., Baker, S., & Kakade, S. (2007). Leveraging archival video for building face datasets. In *Proceedings of the International Conference on Computer Vision*.
- Ramanathan, V., Joulin, A., Liang, P., & Fei-Fei, L. (2014). Linking people in videos with their names using coreference resolution. In *Proc. ECCV*.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents*. arXiv.

- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... Sutskever, I. (2021a). *Zero-Shot Text-to-Image Generation*. arXiv. Retrieved from <https://arxiv.org/abs/2102.12092> doi: 10.48550/ARXIV.2102.12092
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... Sutskever, I. (2021b). Zero-shot text-to-image generation. In *ICML*.
- Ramzi, E., THOME, N., Rambour, C., Audebert, N., & Bitot, X. (2021). Robust and decomposable average precision for image retrieval. In *Thirty-Fifth Conference on Neural Information Processing Systems*. Retrieved from <https://openreview.net/forum?id=VjQw3v3FpJx>
- Rao, A., Xu, L., Xiong, Y., Xu, G., Huang, Q., Zhou, B., & Lin, D. (2020). A local-to-global approach to multi-modal movie scene segmentation. In *Proc. CVPR*.
- Rao, Y., Lin, D., Lu, J., & Zhou, J. (2018). Learning globally optimized object detector via policy gradient. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Razavi, A., van den Oord, A., & Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*.
- Reddy, C., & Vinzamuri, B. (2018). A survey of partitional and hierarchical clustering algorithms..
- Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016, 20–22 Jun). Generative adversarial text to image synthesis. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of The 33rd International Conference on Machine Learning* (Vol. 48, pp. 1060–1069). New York, New York, USA: PMLR.
- Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., ... de Freitas, N. (2022). *A Generalist Agent*. arXiv. Retrieved from <https://arxiv.org/abs/2205.06175> doi: 10.48550/ARXIV.2205.06175
- Reich, A. R., & Duke, J. E. (1979). Effects of selected vocal disguises upon speaker identification by listening. *The Journal of the ASA*.
- Revaud, J., Almazá, J., Rezende, R. S., & de Souza, C. R. (2019). Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the International Conference on Computer Vision*.
- Riba, P., Molina, A., Gomez, L., Terrades, O. R., & Lladós, J. (2021). Learning to rank words: Optimizing ranking metrics for word spotting. *CoRR*, abs/2106.05144. Retrieved from <https://arxiv.org/abs/2106.05144>

- Rohdin, J., Stafylakis, T., Silnova, A., Zeinali, H., Burget, L., & Plchot, O. (2019). Speaker verification using end-to-end adversarial language adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- RolÁnek, M., Musil, V., Paulus, A., Vlastelica, M., Michaelis, C., & Martius, G. (2020). Optimizing rank-based metrics with blackbox differentiation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Roth, K., & Brattoli, B. (2019). *Deep Metric Learning Baselines*. <https://github.com/Confusezius/Deep-Metric-Learning-Baselines>. GitHub.
- Roth, K., Brattoli, B., & Ommer, B. (2019). Mic: Mining interclass characteristics for improved metric learning. In *Proceedings of the International Conference on Computer Vision*.
- Rothlingshofer, V., Sharma, V., & Stiefelhausen, R. (2019). Self-supervised face-grouping on graphs. In *Proc. ACMMM*.
- Roy, A., Guinaudeau, C., Bredin, H., & Barras, C. (2014). Tvd: a reproducible and multiply aligned tv series dataset. In *LREC*.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*.
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. *arXiv:1606.03498 [cs.LG]*.
- Salton, G., & Buckley, C. (1999). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4), 288–297.
- Sanakoyeu, A., Tschernezki, V., Buchler, U., & Ommer, B. (2019). Divide and conquer the embedding space for metric learning. In *Proc. CVPR*.
- Sarfraz, S., Sharma, V., & Stiefelhausen, R. (2019). Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Satoh, S., & Kanade, T. (1997). Name-it: Association of face and name in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Schaldenbrand, P., Liu, Z., & Oh, J. (2021). Styleclipdraw: Coupling content and style in text-to-drawing synthesis. *arXiv:2111.03133 [cs.CV]*.

- Schroff, F., Criminisi, A., & Zisserman, A. (2007). Harvesting image databases from the web. In *Proceedings of the 11th International Conference on Computer Vision, Rio de Janeiro, Brazil*.
- Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., . . . Komatsuzaki, A. (2021). LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR, abs/2111.02114*. Retrieved from <https://arxiv.org/abs/2111.02114>
- Schwettmann, S., Hernandez, E., Bau, D., Klein, S., Andreas, J., & Torralba, A. (2021). Toward a visual concept vocabulary for gan latent space. In *ICCV*.
- Sevilla-Lara, L., Zha, S., Yan, Z., Goswami, V., Feiszli, M., & Torresani, L. (2021). Only time can tell: Discovering temporal data for temporal modeling. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 535-544.
- Sharma, R., & Narayanan, S. (2022a). Audio visual character profiles for detecting background characters in entertainment media. *arXiv preprint arXiv:2203.11368*.
- Sharma, R., & Narayanan, S. (2022b). Using active speaker faces for diarization in tv shows. *arXiv preprint arXiv:2203.15961*.
- Sharma, V., Tapaswi, M., Sarfraz, M. S., & Stiefelhagen, R. (2019a). Self-supervised learning of face representations for video face clustering. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*.
- Sharma, V., Tapaswi, M., Sarfraz, M. S., & Stiefelhagen, R. (2019b). Video face clustering with self-supervised representation learning. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.
- Sharma, V., Tapaswi, M., Sarfraz, M. S., & Stiefelhagen, R. (2020). Clustering based contrastive learning for improving face representations. *Proc. Int. Conf. Autom. Face and Gesture Recog.*.
- Shen, Y., Gu, J., Tang, X., & Zhou, B. (2020). Interpreting the latent space of gans for semantic face editing. In *CVPR*.
- Shi, J., Xu, N., Zheng, H., Smith, A., Luo, J., & Xu, C. (2021). Spaceedit: Learning a unified editing space for open-domain image editing. *arXiv:2112.00180 [cs.CV]*.

- Shocher, A., Gandelsman, Y., Mosseri, I., Yarom, M., Irani, M., Freeman, W. T., & Dekel, T. (2020). Semantic pyramid for image generation. In *CVPR*.
- Shu, X., Wang, X., Zhang, S., Zhang, X., Chen, Y., Li, G., & Tian, Q. (2022). Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 32, 4390-4403.
- Sim, K., Brown, A., & Hassoun, A. (2021). Thinking through and writing about research ethics beyond "broader impact". *CoRR*.
- Sivic, J., Everingham, M., & Zisserman, A. (2009). "Who are you?" – learning person specific classifiers from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*.
- Sivic, J., Zitnick, C. L., & Szeliski, R. (2006). Finding people in repeated shots of the same scene. In *Proceedings of the 17th British Machine Vision Conference, Edinburgh*.
- Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Somandepalli, K., Hebbar, R., & Narayanan, S. (2020). Multi-face: Self-supervised multiview adaptation for robust face clustering in videos. *arXiv preprint arXiv:2008.11289*.
- Song, H. O., Xiang, Y., Jegelka, S., & Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23, 828-841.
- Suh, Y., Han, B., Kim, W., & Lee, K. M. (2019). Stochastic class-based hard example mining for deep metric learning. In *Proc. CVPR*.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *CVPR*.
- Tang, S., Luo, F., & Wu, J. (2022). Smooth-auc: Smoothing the path towards rank-based ctr prediction. In

- Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.*
- Tao, C., Li, Z., Zhu, X., Huang, G., Liu, Y., & Dai, J. (2021). Searching parameterized AP loss for object detection. In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems*. Retrieved from <https://openreview.net/forum?id=hLTZCN7f3M->
- Tapaswi, M., Baeuml, M., & Stiefelbogen, R. (2012). “knock! knock! who is it?” probabilistic person identification in tv series. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Tapaswi, M., Bauml, M., & Stiefelbogen, R. (2014). Storygraphs: visualizing character interactions as a timeline. In *Proc. CVPR*.
- Tapaswi, M., Law, M. T., & Fidler, S. (2019). Video face clustering with unknown number of clusters. In *Proceedings of the International Conference on Computer Vision*.
- Tapaswi, M., Parkhi, O. M., Rahtu, E., Sommerlade, E., Stiefelbogen, R., & Zisserman, A. (2014). Total cluster: A person agnostic clustering method for broadcast videos. In *Proc. ICVGIP*.
- Tapaswi, M., Zhu, Y., Stiefelbogen, R., Torralba, A., Urtasun, R., & Fidler, S. (2016). Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4631–4640).
- Taylor, M., Guiver, J., Robertson, S., & Minka, T. (2008). Softrank: Optimising non-smooth rank metrics. In *WSDM*.
- Tian, Y., Zhou, L., Zhang, Y., Zhang, T., & Fan, W. (2019). Deep cross-modal face naming for people news retrieval. *Proc. TKDE*.
- Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., & Cohen-Or, D. (2021). Designing an encoder for stylegan image manipulation. *arXiv:2102.02766 [cs.CV]*.
- Tsai, Y.-H., Shen, X., Lin, Z. L., Sunkavalli, K., Lu, X., & Yang, M.-H. (2017). Deep image harmonization. In *CVPR*.
- Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proc. CVPR*.
- Ustinova, E., & Lempitsky, V. (2016). Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems*.

- van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. *arXiv:1711.00937 [cs.LG]*.
- Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., ... Belongie, S. (2018). The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Varamesh, A., Diba, A., Tuytelaars, T., & Gool, L. V. (2020). Self-supervised ranking for representation learning. *CoRR*, abs/2010.07258. Retrieved from <https://arxiv.org/abs/2010.07258>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*.
- Vicol, P., Tapaswi, M., Castrejon, L., & Fidler, S. (2018). Moviegraphs: Towards understanding human-centric situations from videos. In *Proc. CVPR*.
- Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015, jun). Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 3156-3164). Los Alamitos, CA, USA: IEEE Computer Society. Retrieved from <https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298935> doi: 10.1109/CVPR.2015.7298935
- Vlastelica, M., Paulus, A., Musil, V., Martius, G., & RolÁnek, M. (2020). *Differentiation of Blackbox Combinatorial Solvers*. <https://github.com/martius-lab/blackbox-backprop>. GitHub.
- Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., & Murphy, K. (2018, September). Tracking emerges by colorizing videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Voynov, A., & Babenko, A. (2020). Unsupervised discovery of interpretable directions in the gan latent space. In *ICML*.
- Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained k-means clustering with background knowledge. In *Proceedings of the International Conference on Machine Learning*.
- Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset..
- Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018). Generalized end-to-end loss for speaker verification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

- Wang, G., Yang, M., Zhang, L., & Yang, T. (2021). Momentum accelerates the convergence of stochastic AUPRC maximization. *CoRR*, *abs/2107.01173*. Retrieved from <https://arxiv.org/abs/2107.01173>
- Wang, J., song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., ... Wu, Y. (2014). Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Wang, J., Zhou, F., Wen, S., Liu, X., & Lin, Y. (2017). Deep metric learning with angular loss. In *Proceedings of the International Conference on Computer Vision*.
- Wang, Q., Rao, W., Sun, S., Xie, L., Chng, E. S., & Li, H. (2018). Unsupervised domain adaptation via domain adversarial training for speaker recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*.
- Wang, X., Han, X., Huang, W., Dong, D., & Scott, M. R. (2019). Multi-similarity loss with general pair weighting for deep metric learning. In *Proc. CVPR*.
- Wang, X., Hua, Y., Kodirov, E., Hu, G., Garnier, R., & Robertson, N. (2019). Ranked list loss for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Wang, X., Jabri, A., & Efros, A. A. (2019). Learning correspondence from the cycle-consistency of time. In *CVPR*.
- Wang, X., Zhang, H., Huang, W., & Scott, M. R. (2020). Cross-batch memory for embedding learning. In *Proc. CVPR*.
- Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer gan to bridge domain gap for person re-identification. In *Proc. CVPR*.
- Weinberger, K., Blitzer, J., & Saul, L. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*.
- Weinberger, K. Q., & Saul, L. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, *10*, 207–244.
- Williams, R. J., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*

- Wohlhart, P., Köstinger, M., Roth, P. M., & Bischof, H. (2011). Multiple instance boosting for face recognition in videos. In *DAGM-Symposium*.
- Wu, B., Lyu, S., Hu, B.-G., & Ji, Q. (2013). Simultaneous clustering and tracklet linking for multi-face tracking in videos. In *Proceedings of the International Conference on Computer Vision*.
- Wu, B., Zhang, Y., Hu, B.-G., & Ji, Q. (2013). Constrained clustering and its application to face clustering in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., & Duan, N. (2021). N\ " uwa: Visual synthesis pre-training for neural visual world creation. *arXiv:2111.12417 [cs.CV]*.
- Wu, C.-Y., Manmatha, R., Smola, A. J., & Krahenbuhl, P. (2017). Sampling matters in deep embedding learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Wu, Y., Wu, X., Li, X., & Tian, J. (2021). Mgh: Metadata guided hypergraph modeling for unsupervised person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*.
- Wu, Z., Lischinski, D., & Shechtman, E. (2020). Stylespace analysis: Disentangled controls for stylegan image generation. *arXiv:2011.12799 [cs.CV]*.
- Xia, J., Rao, A., Huang, Q., Xu, L., Wen, J., & Lin, D. (2020). Online multi-modal person search in videos. In *Proc. ECCV*.
- Xia, W., Yang, Y., Xue, J.-H., & Wu, B. (2021). Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*.
- Xiao, S., Tan, M., & Xu, D. (2014). Weighted block-sparse low rank representation for face clustering in videos. In *Proceedings of the European Conference on Computer Vision*.
- Xiao, Z., Yan, Q., Chen, Y.-A., & Amit, Y. (2019). Generative latent flow. *arXiv:1905.10485 [cs.CV]*.
- Xie, S., Sun, C., Huang, J., Tu, Z., & Murphy, K. P. (2018). Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*.
- Xie, W., Nagrani, A., Chung, J. S., & Zisserman, A. (2019). Utterance-level aggregation for speaker recognition in the wild. In *International Conference on Acoustics, Speech, and Signal Processing*.
- Xing, Y., He, T., Xiao, T., Wang, Y., Xiong, Y., Xia, W., ... Soatto, S. (2021). Learning hierarchical graph neural networks for image clustering. In *ICCV 2021*. Retrieved from

<https://www.amazon.science/publications/learning-hierarchical-graph-neural-networks-for-image-clustering>

- Xu, Y., Shen, Y., Zhu, J., Yang, C., & Zhou, B. (2021). Generative hierarchical features from synthesizing images. In *CVPR*.
- Xuan, H., Souvenir, R., & Pless, R. (2018). Deep randomized ensembles for metric learning. In *Proceedings of the European Conference on Computer Vision*.
- Yang, C., Shen, Y., & Zhou, B. (2020). Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision*.
- Yang, H.-F., Lin, K., & Chen, C.-S. (2016). Cross-batch reference learning for deep classification and retrieval. In *Proc. ACMMM*.
- Yang, J., Yan, R., & Hauptmann, A. G. (2005). Multiple instance learning for labeling faces in broadcasting news video. In *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 31–40).
- Yang, Z., Bastan, M., Zhu, X., Gray, D., & Samaras, D. (2022, January). Hierarchical proxy-based loss for deep metric learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (p. 1859-1868).
- Yu, F., Zhang, Y., Song, S., Seff, A., & Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv:1506.03365 [cs.CV]*.
- Yuan, Y., Yang, K., & Zhang, C. (2017). Hard-aware deeply cascaded embedding. In *Proceedings of the International Conference on Computer Vision*.
- Yue, Y., Finley, T., Radlinski, F., & Joachims, T. (2007). A support vector method for optimizing average precision. In *SIGIR*.
- Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 6713-6724). doi: 10.1109/CVPR.2019.00688
- Zhang, N., Paluri, M., Taigman, Y., Fergus, R., & Bourdev, L. (2015). Beyond frontal faces: Improving person recognition using multiple cues. In *Proc. CVPR*.
- Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *ECCV*.

- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.
- Zhang, Y., Barzilay, R., & Jaakkola, T. (2017). Aspect-augmented adversarial networks for domain adaptation. *Transactions of the Association for Computational Linguistics*.
- Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2016). Joint face representation adaptation and clustering in videos. In *Proceedings of the European Conference on Computer Vision*.
- Zhang, Z., Ma, J., Zhou, C., Men, R., Li, Z., Ding, M., . . . Yang, H. (2021). UFC-BERT: Unifying multi-modal controls for conditional image synthesis. In *NeurIPS*.
- Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E. I., & Xu, Y. (2021). Large scale image completion via co-modulated generative adversarial networks. In *ICLR*.
- Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proc. ICCV*.
- Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proc. ICCV*.
- Zhou, J., Jiang, T., Li, L., Hong, Q., Wang, Z., & Xia, B. (2019). Training multi-task adversarial network for extracting noise-robust speaker embedding. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Zhu, J., Shen, Y., Zhao, D., & Zhou, B. (2020). In-domain gan inversion for real image editing. In *ECCV*.
- Zhu, J.-Y., Krähenbühl, P., Shechtman, E., & Efros, A. A. (2016). Generative visual manipulation on the natural image manifold. In *ECCV*.