Spatial Audio and Spatial Audio-Visual Learning

Yuhang He

St Hugh's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2024

This thesis is dedicated to my dearest parents
for their altruistic giving and vigorous support.

Acknowledgements

I am tremendously grateful for the outpouring support I have received during my DPhil study. I must admit that I have seriously underestimated the challenge and difficulty that I have had to face as a DPhil candidate, encompassing both academic and non-academic staff. Despite years of industrial research experience I have had before joining University of Oxford which endowed me the essential capability to address the foreseen life obstacles, new challenges and difficulties kept emerging during my whole DPhil journey and sometimes they inevitably became much more severe due to the pandemic. In spite of thesis crisis, numerous people have generously offered me generous support including but not limited to their time, lending resources, encouragement and instant academic feedback. I firmly believe that my DPhil experience would be totally different without their phenomenal support.

Fist and foremost, I owe a debt of gratitude to Professor Andrew Markham and Professor Niki Trigoni, who collaboratively guided me as co-supervisors throughout my DPhil journey. Their visionary research approach and deep insight into technically challenging yet overlooked reals with current research community have profoundly influenced my scholarly path. Their mentorship not only steered my research trajectory but also positioned me at an advantageous position as I move to the next phase of my academic postgraduate journey. Their encouragement for me to go into spatial audio learning during my initial DPhil research stage, starting from the very theoretical signal processing techniques to the advanced modern deep learning supported spatial audio learning, directed me to an underexplored yet highly promising research path marked by occasional frustration and loneliness. The flexibility and freedom they generously gave me enabled me to passionately explore diverse areas of knowledge without the imposition of rigid publication goals. I am truly grateful for the precious and unparalleled freedom for the exploration and research

that I have been fortunate to experience, an opportunity that is, regrettably, becoming increasingly rare for many computer science DPhil students at current time. Reflecting on my early publications, I find it hard to believe how they would have evolved and taken shape without their generous support.

University of Oxford as a world-renowned research hub has been phenomenally supportive for DPhil students to conduct interdisciplinary research. During my DPhil journey, I have been fortunate enough to have various opportunities to discuss (brainstorm or even debate) with DPhil (or MSc.) students from different backgrounds. The memories of those perspiration-filled moments still linger, haunting me as if they happened just yesterday. Specifically, I can easily blurt out numerous professional words including "rocking system", "plastic deformation" (by talking with Yixiong Jing, Zhengyou Zhang in civil engineering), "string theory", "black hole" (by talking with Sirui Ning in theory physics), "DNA repair" (by talking with Hannan Xu in Biochemistry), "ammonoids", "earth quake" (by talking with Dr. Qi Ou in earth science). I must admit that my research interest and passion have been tremendously propelled and catalysed from such interdisciplinary and aimless discussion. I would like to express my deepest gratitude to all of them. Moreover, I want express sincerest gratitude to Hanyu Hu from Said business school, Zhan Huang from Anthropology, Jiaxing Zhong, Kai Lu, Sangyun Shin, Madhu Vankadari, Jialu Wang in CPS lab.

The global pandemic (Covid-19) commenced shortly after I started my DPhil life. During its inception, nobody could imagine how deep and everlasting the pandemic would change the whole world and everyone's normal life. On retrospection, I do not think I can survive those painful and grinding times without the generous support from the my friends inside and outside Oxford (especially Prof. Chen Feng in NYU, Dr. Anoop Cherian, Dr. Moitreya Chatterjee in MERL), family members, especially my parents, younger sister.

Abstract

As humans, we extensively depend on multimodal signals to perceive, interact with, and analyze our surrounding 3D spatial environment, so as to accomplish various complex tasks. Amongst all our multimodal senses, sound and vision are the two most ubiquitous signals in real world scenarios. Equipping machines with such spatial audio-visual multimodal inferring and learning capability, or human-level audio-visual intelligence, is a vital yet challenging task. Although the widely accepted importance for spatial audio-visual intelligence, current research community's focus has been heavily biased towards vision, with far less attention paid to spatial audio. Research in spatial audio-visual learning typically takes spatial audio as an auxiliary input to assist vision-centred tasks, ignoring the intricate and multifaceted interactions between audio and vision.

These observations and limitations discussed above motivate my two main DPhil research focus: spatial audio learning and spatial audio-visual learning. While the first one serves as the preparatory exploration aims to fill up the gap between the well-explored vision learning and the under-explored spatial audio learning, the second one raises new challenges in modality issue and exhibits strong applicability in real-scenarios. In summary, my DPhil research is driven by the following three main questions:

1. In the modern deep neural network era, how can classic signal processing based sound waveform feature extraction methods benefit from the powerful expressiveness of deep neural network?
2. Is the widely-adopted process of spectrogram extraction which treats sound as ordinary 2D image an optimal representation? If not, can we design novel neural networks specifically tailored for sound?
3. How can we design robust audio-visual multimodal learning framework in embodied settings where sound and vision are weakly associated?

Exploring these questions requires treating sound as being equally important as vision, yet fundamentally distinct. To accomplish these aims, **First**, we first systematically explore the fundamental knowledge behind classical acoustic and visual signal processing, and further study the modality difference between acoustic signals and visual signals. **Second**, based on the previous exploration, we further investigate how these fundamentals behind acoustic and visual signals can guide deep neural network design for sound raw waveform learning. **Third**, we show how sound can be efficiently and effectively processed together with vision to provide an initial attempt towards spatial audio-visual intelligence under sound-vision weak-correlation condition.

Contents

Chapter 1

Introduction

Humans and animals perceive the world through a variety of senses. Out of the five main senses (touch, smell, taste, sound, sight), sound and sight enable detailed spatial information to be garnered about the surrounding environment [1] . As predominantly visual creatures, it is thus unsurprising that research in processing visual information (*e.g.*, from cameras) has largely dominated the field of machine learning for rich spatial awareness. Machine learning for audio has somewhat lagged behind, with many approaches treating sound as an image (*e.g.*, by converting it into a time-frequency spectrogram). Although this allows audio to benefit from the burgeoning body of work in computer vision, it naturally begs the question of whether this is the only, and indeed the best, representation. This becomes particularly apparent when considering *spatial audio e.g.*, audio which is detected by two or more synchronized microphones. This is because spatial information is conveyed through subtle time-delays, which manifest equivalently as phase-shifts of frequency components. The aim of this thesis is to investigate the relative merits of machine learning approaches that treat audio as a first-class citizen, not merely a poor cousin of computer vision.

1.1 Spatial Audio

Spatial audio can potentially relate to a wide range of technologies, such as 3D sound field, spatial sound creation and reproduction. In this thesis, we constrain the scope of "spatial audio learning" to spatial multi-channel (microphone-array based) acoustic signal processing from deep neural network perceptive, especially for tasks like the sound source detection and spatial acoustic effect prediction.

[1] Some animals can perceive a rich spatial olfactory environment, but smell is not a well honed spatial sense for humans.

Spatial audio has become increasingly ubiquitous across various domains, ranging from the natural physical world that we may not always consciously perceive to the digital environment that requires immersive acoustic spatial effect. Spatial audio captures the changes in the audio signal as it propagates from a source position to a set of receivers. It is capable of informing us of a sense of depth, bearing angle, and the environment layout. As humans, we unconsciously rely on Psychoacoustics [117] to acoustically perceive the 3D physical world. In most cases, we can infer the number and type of sound sources in the environment, what direction and distance (i.e. their position) they come from based on our binaural auditory system, and further infer the overall scene properties and rough spatial size. For example, bird song and natural audio are likely associated with a nature reserve, the sounds of cars and traffic may be associated with a public open street area, and highly reverberant audio is associated with a spacious and empty indoor space.

Spatial realism is also vital in artificially created digital environments such as virtual reality (VR) and augmented reality (AR). Here, the main task of spatial audio in the digital environment is to create an immersive spatial audio experience informed by, and grounded in, the digital environment such that exploring the digital environment resembles exploring a realistic physical world. It is safe to say we anticipate the spatial audio to become increasingly present in our technological interactions of the future.

In summary, spatial audio presents in our daily life, either in the physical world we are living in or artificial digital platform (*e.g.*, AR/VR, film industry and music industry). Its ubiquity will continue to grow, shaping the way we perceive and interact with audio content in our increasingly digital and interconnected world.

1.1.1 Spatial Audio Sources

Spatial audio can be generated by specific sources that emit audio by themselves (*e.g.*, an alarm, or telephone ringing). As the sound waves propagate through space, they travel at the speed of sound 340m/s. If two or more receivers (i.e. microphones) are placed in the environment at spatially distinct locations, the different path lengths cause the waveforms to be delayed in time, which is equivalent to undergoing a frequency-dependent shift in phase angle. By examining the time delay or phase shift between the channels, the difference in path length can be estimated, which allows the angle of arrival to be inferred. Given multiple sound sources (*e.g.*, polyphony), the phase shift can be used as an informative cue to separate out different sources.

In addition to this unitary object source spatial audio generation paradigm, spatial audio can be generated by inter-object interaction and such spatial audio is called impact audio [35]. Impact audio is ubiquitous in a dynamic space and can be used to infer the spatial dynamics (*e.g.* who are involved and what are happening) and crossmodal information. For example, given the continuous footstep spatial audio generated by pedestrian's interaction with the ground, it naturally informs us a pedestrian or multiple pedestrians are nearby and approaching towards or away from us. Evolving both temporally and spatially, impact audio usually provides us with ample cues to inversely characterise the acoustic scene so as to further foster deeper understanding and engagement with complex concepts and scenarios.

1.1.2 Applications of Spatial Audio

Spatial audio has a wide range of applications, some of which are outlined here.

Spatial audio technology is playing an increasingly pivotal role in wildlife acoustic surveillance [158] and smart city surveillance [34], revolutionizing the way we monitor and analyze audio data in natural and urban environments. In wildlife acoustic surveillance, spatial audio enables researchers and conservationists to accurately localize and identify animal species based on their vocalizations, facilitating biodiversity monitoring, habitat assessment, and wildlife management efforts. By deploying spatial audio sensors in remote and challenging environments, such as forests, wetlands, and marine ecosystems, researchers can capture rich audio recordings that capture the spatial distribution of animal calls and behaviors, providing valuable insights into ecosystem health and biodiversity dynamics. Similarly, in smart city surveillance applications, spatial audio technology enhances the situational awareness and monitoring capabilities of urban surveillance systems by enabling the precise localization and tracking of sound events in real-time. By integrating spatial audio sensors with existing video surveillance networks and IoT (Internet of Things) infrastructure, city planners and law enforcement agencies can detect and analyze a wide range of acoustic events, including traffic accidents, emergencies, and public disturbances, improving public safety, traffic management, and environmental monitoring in urban areas. As spatial audio technology continues to advance, it holds the potential to transform wildlife acoustic surveillance and smart city surveillance into more effective, efficient, and sustainable monitoring solutions, with far-reaching implications for conservation, urban planning, and public safety.

In the realm of entertainment, spatial audio revolutionizes the way we consume and interact with audiovisual content. With the emergence of AR/VR technologies [82, 13, 14], spatial audio adds depth, dimension, and realism to immersive experiences, transporting users to virtual worlds where sound behaves just like it does in the real world. Whether it's experiencing the roar of a crowd at a live concert, the rustle of leaves in a serene forest, or the thunderous footsteps of a towering monster in a video game, spatial audio creates a sense of presence and immersion that traditional audio formats cannot replicate. This heightened level of realism not only elevates entertainment experiences but also opens up new avenues for storytelling, artistic expression, and creative exploration in film, music, theater, and interactive media.

Spatial audio can be artificially created to revolutionize the music industry, offering artists and producers new creative tools to enhance the immersive quality of their compositions and recordings. By harnessing spatial audio techniques such as surround sound, ambisonics, and binaural recording, musicians can create multi-dimensional (or 360°) sonic experiences that transport listeners into the heart of the music. Spatial audio allows the precise placement of instruments and sound elements within a three-dimensional soundstage, enabling artists to craft intricate spatial arrangements that envelop the listener and evoke a heightened sense of presence and immersion. From live performances captured in immersive 360° audio to studio recordings mixed and mastered in spatial audio formats, the possibilities for spatial audio in music production are virtually limitless. Moreover, spatial audio technologies enable new forms of musical expression and experimentation, blurring the lines between traditional stereo recordings and immersive sonic experiences. As spatial audio continues to gain attention in the music industry, it promises to redefine the way we experience and interact with music, opening up exciting new avenues for artistic innovation and sonic storytelling.

In gaming, spatial audio is an indispensable tool for creating immersive and engaging game play experiences [15]. By accurately positioning sound sources in three-dimensional space, spatial audio enhances situational awareness, spatial navigation, and environmental immersion for players. From detecting enemy footsteps and gunfire in first-person shooters to locating hidden treasures and clues in adventure games, spatial audio adds a layer of depth and realism that enhances gameplay mechanics and player engagement. Moreover, spatial audio enables innovative gameplay mechanics such as spatialized voice chat, dynamic sound propagation, and interactive audio en-

vironments, paving the way for more immersive and interactive gaming experiences across all genres and platforms.

In the field of communication and collaboration [160], spatial audio facilitates natural and immersive remote interactions, bridging the gap between physical and virtual spaces. With the rise of remote work, virtual meetings, and teleconferencing, spatial audio technologies enable users to engage in more immersive and productive conversations, regardless of their physical position and orientation. By spatially rendering participants' voices and ambient sounds in virtual meeting spaces, spatial audio recreates the sense of presence and proximity that is crucial for effective communication and collaboration. This enables more naturalistic group discussions, presentations, and brainstorming sessions, fostering better teamwork, creativity, and decision-making in distributed teams and remote work environments.

In summary, spatial audio is a versatile and transformative technology with vast applications across a wide range of industries and domains. From entertainment and gaming to communication, education, healthcare, and beyond, spatial audio enhances user experiences, enables novel interactions, and drives technological innovation. As spatial audio continues to evolve and proliferate, it holds the potential to reshape the way we perceive, interact with, and experience audio in our increasingly digital and interconnected world.

1.1.3 Common Tasks in Spatial Audio

The most prominent research problem in spatial audio is that of sound event detection and localization, which is sometimes termed SELD [2], in which the goal is to jointly spatially localize each sound source and classify its semantic label. Here sound source localization means localizing the sound source both temporally and spatially: finding a sound event's start time and end time along the temporal axis, pinpointing its spatial position simultaneously, either through directional of arrival (DoA) or actual physical position ($[x, y, z]$ coordinates). This problem encompasses several challenges, including the accurate estimation of sound source positions and trajectories in ambient noisy, reverberant and polyphonic acoustic environments. Addressing this research problem requires the combination of advanced signal processing methods, machine learning especially modern deep neural network and various advanced sensor technologies. It also requires interdisciplinary collaborations between researchers in audio engineering, computer science, psychology and physics. Based on this problem, numerous further tasks can be carried out including source separation, acoustic scene analysis [184]. Chapter 3 focuses on this research problem by verifying the

robustness and efficacy of proposed learnable filters for both sound event direction of arrival (DoA) task and physical position estimation task ($[x, y, z]$).

Another core research problem is to accurately model spatial audio propagation processes in reverberant environments. Spatial audio propagation from the source position to the target position in an enclosed 3D space is a complex process that involves changes like absorption, diffraction and reflection. Accurately measuring this propagation process is a challenging task in real-world settings, involving exhaustive pairwise measurements and decorrelation. In an enclosed indoor 3D space that is linear time invariant (LTI), the propagation process is measured by a room impulse response (RIR), which is lengthy in data points due to the complex propagation behaviour. Classic RIR data collection processes are highly inefficient and interpolating between spatial locations is not trivial. In the deep learning era, especially given the advancement of neural implicit fields in recent years, a promising alternative is instead to utilize a deep neural network to learn a spatially continuous field to implicitly model audio propagation process from arbitrary source position to arbitrary receiver position from sparsely sampled RIR data. Chapter 4 focuses on implicit audio rendering.

Apart from the aforementioned research problems, other relevant research problems based on spatial audio learning include acoustic scene classification [177], acoustic anomaly detection [59] and scene synthesis [138]. All of these research problems have been discussed from various perspectives.

1.1.4 Motivation in Spatial Audio Learning

In the realm of audio signal processing, a prevalent technique involves transforming the one-dimensional audio waveform in the time domain into a detailed two-dimensional time-frequency representation, often achieved through methods such as the short-time Fourier transform (STFT) [21, 20, 27, 26, 169]. This transformation unveils the frequency components concealed within the temporal domain of the sound waveform by explicitly presenting them in the form of a 2D time-frequency map. An inherent advantage of this conversion lies in the ability to depict spatial audio in the format of conventional 2D images. This affords us the opportunity to leverage well-established image-based deep neural networks, seamlessly applying them to a myriad of tasks related to spatial audio. However, the ease of this transition to a visual representation also brings with it a noteworthy consequence – typically the loss of phase and interchannel information. This tendency to overlook the specific exploration of deep neural network architectures tailored to the unique characteristics

and intricacies of spatial audio is a limitation of current work. By treating spatial audio merely as 2D images, there emerges a gap in the development of deep learning models that possess an internal understanding of the wave-like nature inherent in spatial audio. This absence of specialized exploration inhibits the realization of more nuanced and sophisticated deep neural networks that can unlock the full potential of spatial audio processing (*e.g.*, that are phase-aware). Hence, it becomes imperative to delve into the design of deep neural networks explicitly centered around the wave nature of spatial audio, ensuring a more comprehensive and tailored approach to its analysis and synthesis.

1.2 Spatial Audio-Visual Signal

Serving as one of the primary sensing modalities, vision plays a crucial role in shaping our perception of the world. It allows us to process information from the electromagnetic spectrum, discerning shapes, colors, and spatial relationships between objects. Vision not only aids in recognizing and analyzing objects situated in the physical world, but also contributes significantly to reasoning and understanding the overall context of a scene. Specifically, we can estimate a visual object's semantic label, spatial position (*e.g.*, via stereo-vision), and motion from visual signals. It is natural to thus explore how we can extend spatial audio to incorporate visual information for scene perception and understanding.

In particular, we note that through the intricate orchestration of audio-visual sensory perception, humans seamlessly navigate and comprehend their surroundings. Moreover, in the dynamic physical world, acoustic and visual signals are often entangled, with many visual signals being accompanied with acoustic signals. For example, a crying baby is both visually and acoustically observable (in this case, audio and vision signals indicate the same entity); a person walking creates spatial audio signifying the position and motion of the visually observable pedestrian (in this case, the acoustic signal is a byproduct of the pedestrian-floor interaction). The integration of these senses allows a holistic understanding of the world, enabling adaptive responses and enriching the human experience.

1.2.1 Challenges in Audio-Visual Research

Although it has been long ago recognized the importance and necessity of audio-visual multimodal perception, the corresponding progress and development that have achieved is not that promising, when comparing the progress that has been already

achieved in their individual modalities. In the acoustic signal processing domain, the complete and systematic development of the Fourier transform and Wavelet transform "kingdoms" have laid solid mathematical and physical foundation of various acoustic signal processing techniques; In the visual signal processing (either the digital signal processing or modern computer vision), we have witnessed the clear and complete evolution from mature classic computer vision fundamentals (*e.g.* multiview geometry, 3D-2D projection and camera pose estimation and epipolar line constraints) to modern deep neural network supported computer vision techniques (examplar works include ResNet [61], ViT [87], FasterRCNN [146]). For many years, the two domains have developed relatively independently, and the modern deep neural network based spatial audio learning development is far lagged behind than its counterpart visual signals.

We conjecture that the reason for the mismatched research attention that has been assigned to audio-visual spatial multimodal learning compared with its importance derives from three main aspects: modality difference, dataset scarcity and audio-visual heterogeneous correction.

Modality Difference. Being constrained in a narrow electromagnetic radiation wavelength range (from 380 nanometers to 700 nanometers), vision relies on the reception of electromagnetic waves within the visible spectrum. It captures information about shapes, colors, and spatial arrangements of objects. Visual signals originate from light reflecting off surfaces and entering the eyes. The brain processes these signals to create a visual representation of the environment. It mainly provides a detailed, dense and high-resolution 2D spatial representation. On the contrary, spatial audio derives from vibrations of air molecules, creating variations in air pressure that reach the ears or microphones. Rather than directly providing dense 2D spatial representation, spatial audio gives sparse 2D representation for only those objects that are audible. These objects' semantic information and mutual relationship implicitly lie in captured spatial audio's temporal effect, including reverberation, time-frequency representation and variation and inter-channel phase difference. These modality differences inevitably make audio-visual spatial multimodal learning an intricate and difficult problem.

Dataset Scarcity. Despite the omnipresence of audio-visual data in the real physical world, the task of capturing such data with comprehensive audio and visual labels, essential for the effective training of machine learning algorithms, proves to be exceptionally challenging. This difficulty amplifies as scenes become more intricate and sophisticated, making the alignment and coordination between audio and visual

labels increasingly elusive. Beyond the escalating challenge of visual labeling, understanding the behavior of spatial audio during propagation from the source position to the target position becomes exceptionally intricate in complex scenes [26, 154]. In addition to the increased complexity in visual labeling, modeling the intricate dynamics of spatial audio propagation in more complex scenes poses significant challenges. This complexity is further underscored in realistic spatio-temporally evolving 3D audio-visual environments, where the creation of appropriate audio-visual spatial data becomes an even more formidable task. The evolving nature of such environments introduces dynamic elements such as sound source movement, object motion, and the emergence of spatially impactful audio resulting from inter-object interactions. Addressing the scarcity of datasets suitable for training models to navigate these complexities is crucial. Current limitations in available data hinder the development of robust machine learning algorithms capable of handling the intricacies inherent in realistic audio-visual environments. As we venture into more sophisticated scenarios, efforts must be directed toward the generation and curation of datasets that encompass the multifaceted dynamics of both audio and visual components, fostering advancements in the understanding and application of machine learning in audio-visual processing.

Heterogeneous Correlation. In the real physical world, audio and vision can be either entangled and disentangled depending on level of audio-visual correlation. Most existing audio-visual multimodal learning approaches assume vision-audio are strongly correlated, which means the sound source and visual entity can be one-to-one associated and both are discernible in their own respective modality. For example, given that a sound source comes from particular visually observable objects such as a church bell, a running train, and a clock [170, 164, 74], the object of interest can be detected from either modality. In addition to the widely discussed audio-visual strong-correlation situation, audio and vision can be also weakly-correlated or even non-correlated. The sound source can be either too small to be visually observed (weak-correlation), or out-of-view and even exhibit no visual entities (no-correlation). Such free and multi-faceted audio-vision correlation poses new research challenges that have been largely ignored by the current research community.

The aforementioned difficulties lead to under-exploration of audio-visual spatial multimodal learning. Existing work on audio-visual multimodal learning failed to consider the 3D spatial information jointly revealed by audio-visual cues. Most of them have simply focused on constrained 2D image area and depended on the widely

accessible video dataset for tasks where monoaural audio and vision are tightly coupled. For example, the "guitar playing" audio comes from visually observable guitar.

1.2.2 Motivation in Audio-Visual Learning

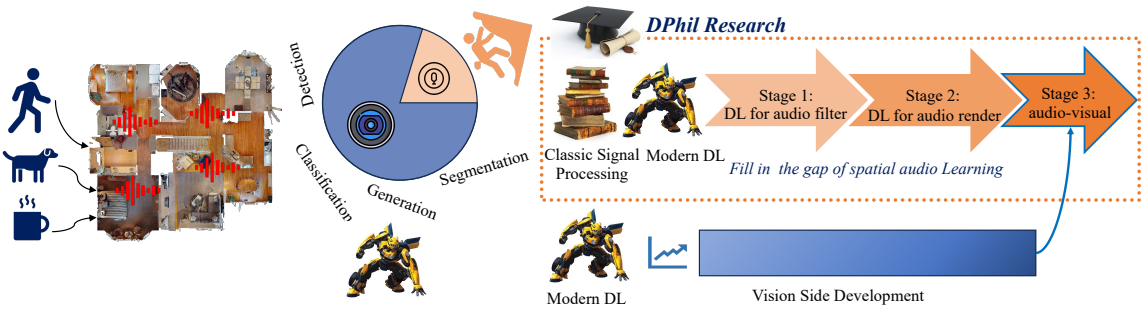
Although the aforementioned challenges and difficulties have constrained the research attention in audio-visual learning, it is still desirable to work on this research direction especially when audio-visual simulators have been developed and become publicly available in recent years [27]. Those audio-visual simulators have reduced the "dataset scarcity" challenge. If we set some constraints to the audio-visual scene, we can conduct various preliminary researches on top of those audio-visual simulators. For example, we can simply assume the audio sources are at fixed positions, just one clean audio source at a time that is constantly emitting sound [26] or that the audio signal is correlated with visual signal [25]. However, it is clear that there is significant room for exploring more challenging spatial audio-visual scenarios.

1.3 DPhil Research Question and Motivation

Due to the imbalanced research attention between vision and spatial audio, and the ubiquity of spatial audio-visual signal in our daily lives (as presented in previous sections), the ultimate DPhil research ambition is to explore spatial audio-visual learning. However, given the current state of the art of spatial audio processing in both standalone spatial audio learning [55, 20, 121] and audio-visual spatial learning methods [27, 26, 139], it is clear that significant attention needs to be paid to improving machine learning for spatial audio alone. The thesis can be divided into three main parts: Spatial Audio Filter Learning, Spatial Audio Neural Rendering and Spatial Audio-Visual Learning.

1. **Spatial Audio Filter Learning**. Existing research typically treats acoustic information as pre-converted time-frequency 2D maps. This relies on defined filter lengths and equally spaced frequency bands, with a fundamental trade-off between time and frequency resolution. Instead of relying on this image-centric approach, we instead pose the following questions:

- Can we exploit the potential of modern deep neural networks to design learnable spatially-aware filters that are capable of directly learning from raw sound waveform?



A. Our living environment is audio-visual multimodal. B. Spatial audio learning is far lagged behind. C. My DPhil research first fills in the gap of spatial audio learning and then goes into audio-visual spatial learning.

Figure 1.1: DPhil research question motivation illustration. Our living space is filled with acoustic and visual signals. For example, in the indoor living space shown in sub-figure A, the daily activities from either residents or the pet dog create a dynamic environment where acoustic and visual signal co-exist. However, the current research community has heavily focused on vision-only research and mostly ignored the associated acoustic signals, resulting in the far-lagged behind research attention on spatial audio (sub-figure B). My DPhil research thus aims to first fill in the gap of spatial audio learning before delving into audio-visual spatial learning.

- Can we build on classical signal processing theories to inform the design of the learnable filters?

2. **Spatial Audio Neural Rendering**. Key to the realism of spatial audio is the complexity of the interaction of the sources with their environment e.g. the reverberation and echo of a corridor. Traditionally, this is achieved through detailed room impulse response (RIR) measurements [95]. Since accurately measuring RIR is difficult and challenging in practice, a research question that naturally arises is:

- Can we adopt deep neural networks to implicitly learn a neural rendering field to capture the acoustic propagation process?

3. **Audio-Visual Multimodal Spatial Learning**. Given the focus of existing audio-visual multimodal learning which predominantly assume that audio and vision are strongly correlated and only exist in the 2D image domain,ignoring spatial effects, We extend audio-visual multimodal learning to the 3D world-based spatial setting, where the audio-visual crossmodal signals are learned in 3D spatial setting and acoustic and visual signals enjoy flexible correlations. Specifically, the two following questions are explored:

- How can we learn from spatial audio-visual signals where such signals are weakly correlated or entirely uncorrelated?

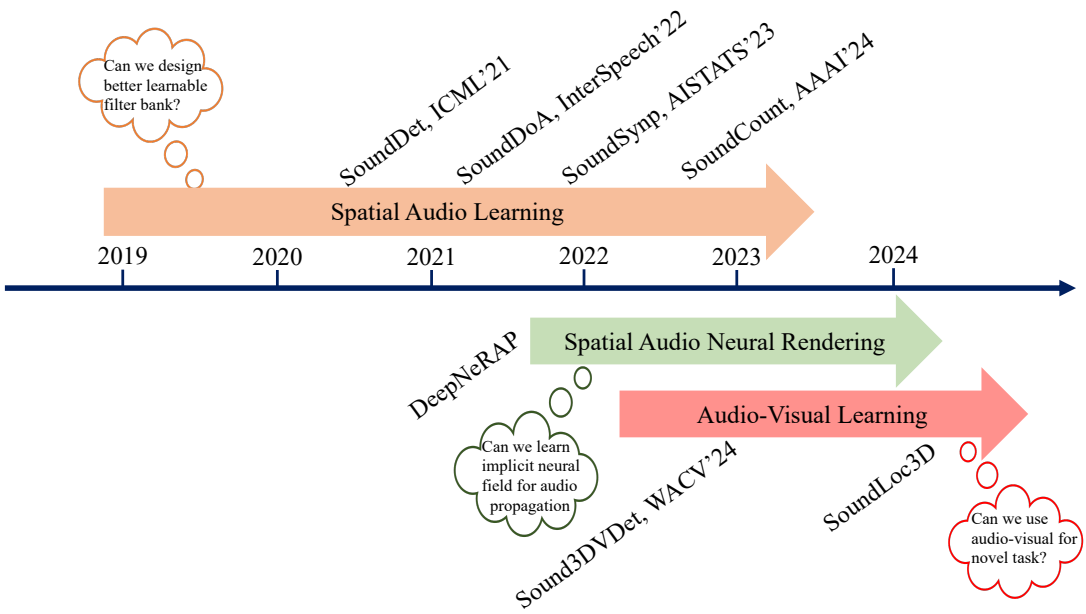


Figure 1.2: DPhil research roadmap visualization. My DPhil research journey can be divided into three main parts, each of which has been driven by a different research question. Chronologically, we have explored "spatial audio learning", "spatial audio neural rendering" and "audio-visual learning".

- How can audio-visual spatial multimodal learning be applied to an embodied AI task to improve the embodied agent's capability in understanding the 3D environment?

Overall, the DPhil research questions are visualized in Fig. 1.1: Most living spaces for our daily lives are highly audio-visual. Acoustic signals co-exist with visual signals where they cooperatively form a dynamically evolving audio-visual ecosystem. Although there is a wide ubiquity of audio-visual signals, currently the research community has heavily focused on the vision side, while the acoustic counterpart has been less researched. While tremendous progress has been made in various vision tasks in modern deep learning era, including detection, classification, generation and segmentation, the development of spatial audio learning is relatively isolated. In this DPhil research, we aim to bridge the gap between vision and spatial audio research.

1.4 Contribution and Research Roadmap

Based on aforementioned discussion, my DPhil research chronologically spanned to three main areas: Spatial Audio Learning, Spatial Audio Neural Rendering and

Audio-Visual Learning. My research roadmap is visualized in Fig. 1.2, and is further structured in this thesis in the following organization format:

- **In Chapter 2**, we systematically present the background knowledge centers at the fundamentals of spatial acoustic signal and visual signals. We discuss the physical and mathematical principles governing the two modal signals' generation and propagation, and further present the common way to represent the two modal signals.

  Contribution: Readers will be able to have a panoramic understanding of the fundamentals of audio and visual signals. They will serve the knowledge base for reading the following chapters.

- **In Chapter 3**, we present three works on designing learnable filter bank to extract more representative and robust time-frequency from either monochannel or multichannel raw sound waveform. These learnable filter design process roots traditional signal processing theories such as Fourier transform kingdom and Wavelet transform kingdom, enabling the designed filter bank to be physically and mathematically meaningful. Specifically, we focused on two main tasks: sound source detection from multichannel sound waveform (SoundDet [68], SoundSynp [66]) and sound event count from monochannel sound waveform (SoundCount [62]).

  Contribution: Readers will learn the technical details and motivation behind designing learnable filter bank for processing sound raw waveform. We show designing novel neural network to directly learn from sound raw waveform usually leads to better performance than learning from pre-computed time-frequency map. This chapter is based on three publications:

  [1] **Yuhang He**, Niki Trigoni, and Andrew Markham. "SoundDet: Polyphonic Moving Sound Event Detection and Localization from Raw Waveform." In International Conference on Machine Learning (ICML), 2021.

  [2] **Yuhang He**, Andrew Markham. "SoundSynp: Sound Source Detection from Raw Waveforms with Multi-Scale Synperiodic Filterbanks." International Conference on Artificial Intelligence and Statistics (AISTATS), 2023.

  [3] **Yuhang He**, Zhuangzhuang Dai, Long Chen, Niki Trigoni, Andrew Markham. "SoundCount: Sound Counting from Raw Audio with Dyadic Decomposition Neural Network." The 38th Annual AAAI Conference on Artificial Intelligence (AAAI), 2024.

- **In Chapter 4**, we present how deep neural network can learn neural room acoustic field, an implicit neural representation to model sound propagation in an enclosed 3D space. We present the common physical principles sound propagation process should obey (such as reciprocity, superposition) and further show how guide the whole neural network design.

  Contribution: Readers will learn the way spatial audio propagates in an enclosed 3D space, and how deep neural network can be utilized to implicitly model the process. This work is based on the following paper [69].

  [4] **Yuhang He**, Anoop Cherian, Gordon Wichern, Andrew Markham. "Deep Neural Room Acoustics Primitive." International Conference on Machine Learning (ICML), 2024.

- **In Chapter 5**, we present audio-visual multimodal learn for novel tasks that can not be easily solved by each single modality alone. We present to detect 3D sound source (localize the sound source $[x, y, z]$ coordinates and classify its semantic labels) from multiview RGB and MicArray acoustic signal data. While the 3D sound source exhibits no visual entity, we show incorporating multiview RGB images can help this task.

  Contribution: Based on our previous exploration on spatial audio learning, we further explored how spatial audio and vision can be combined for novel tasks that can not be easily solved by each single modality. While it seems to be a natural step forward to combine audio and vision, we move further to consider novel audio-vision correlation situation that has been rarely discussed in previous work. This chapter is mainly based on the following publication:

  [5] **Yuhang He**, Sangyun Shin, Anoop Cherian, Niki Trigoni, Andrew Markham. "Sound3DVDet: 3D Sound Source Detection Using Multiview Microphone Array and RGB Images." IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024.

Chapter 2

Background Knowledge

This chapter provides the literature review across four main aspects: 1. Audio, 2. Spatial Audio, 3. Spatial Audio-Visual. The goal of this chapter is to provide panoramic review of the research status quo up to date, helping to better understand the motivation underneath my DPhil research.

2.1 Audio Fundamentals

Audio signals refer to variations in air pressure that propagate as waves through a medium such as air, liquid and solid medium. These acoustic waveforms are generated from the oscillation of air particles and can be characterized by their frequency, amplitude, and wave nature. From the scientific research perspective, the science of acoustics studies the production, transmission and biological and psychological effects, governed by the physical and mathematical principles. The mathematical analysis of acoustic signals often involves concepts from wave mechanics and fluid dynamics.

From the mathematical analysis perspective, acoustic signals can be represented as waveforms, with primary parameters,

1. Frequency f: The number of oscillations per unit of time, measured in Hertz (Hz). Human auditory system is more sensitive to acoustic signals of lower frequency than higher frequency.

2. Amplitude A: Amplitude indicates the "size" of an acoustic waveform or the maximum displacement air particles from their equilibrium position. Amplitude equivalently determines the acoustic signal's loudness. Sound intensity is also relevant to amplitude, and often known as the "volume" of a sound wave or the rate at which a wave moves energy per unit of an area.

3. Wavelength λ: The distance between successive peaks of a wave. The mathematical representation of an acoustic signal often involves the use of Fourier analysis to decompose complex signals into their constituent frequencies. This allows for a more detailed understanding of the frequency content and spectral characteristics of the signal.

Audio signals play a crucial role in communication, both in human and animal contexts. In humans, speech is a complex acoustic signal produced by the articulation of the vocal tract. Understanding and interpreting acoustic signals involve the auditory system, where the ear transforms incoming sound waves into electrical signals that the brain interprets as sound. Specifically, acoustic wave is first converted into neuronal signal at the eardrums, which is then further transported to the inferior colliculus and cochlear nucleus in the brain stem for further process. Subsequently, the processed acoustic signal is fed to the primary auditory cortex for final encoding [52].

Beyond human communication, audio signals are integral in various fields, including music, environmental monitoring, and industrial applications. Analysis of acoustic signals is essential for tasks such as speech recognition, audio processing, and the study of sound propagation in different environments.

2.2 Spatial Audio Fundamentals

Spatial audio refers to a technology that reproduces or simulates sound in a way that creates a three-dimensional auditory experience for the listener. Unlike traditional audio systems that present sound from fixed points, spatial audio aims to replicate the way we perceive sound in the real world by considering directionality, distance, and elevation. This immersive audio technology provides the listener with a sense of sound sources coming from different directions and distances, contributing to a more realistic and enveloping auditory experience. Spatial audio is commonly used in applications such as virtual reality, gaming, film production, and live events to enhance immersion and create a more lifelike and engaging audio environment.

2.2.1 Spatial Audio Propagation Mathematical Equation

In time domain, the propagation progress of the sound waveform is governed by the wave equation that can be expressed as,

$$\frac{\partial^2 p(x,t)}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 p(x,t)}{\partial t^2} = 0 \tag{2.1}$$

where $p(x,t)$ is the measured sound pressure. Equation 2.1 describes a wave varying characteristics in space and time. In spatial audio, we are concerned about particle displacement $u(x,t)$ which is usually proportional to the sound pressure $p(x,t)$. For simplicity, we can assume $p(x,t) = u(x,t)$. The solution $u(x,t)$ for a 1D propagating spatial wave can be represented as

$$u(x,t) = Ae^{\pm\alpha x}e^{i(\omega t \pm \kappa x)} \tag{2.2}$$

where ω is the angular frequency, $\alpha(\omega)$ is the frequency-dependent attenuation coefficient (also called damping coefficient), κ is the wave number. Together they make the *wave propagation coefficient* $\gamma(\omega) = \alpha(\omega) + i\kappa(\omega)$. $\alpha(\omega)$ is a positive even function, and $\kappa(\omega)$ is an odd function and positive for $\omega > 0$. The $\pm$ sign indicates the wave propagating in two directions. Energy dissipation is induced when wave propagating in, for example, viscoelastic medium, which is captured by the attenuation term $e^{\pm\alpha x}$. Phase changes are decided by the term $e^{i(\omega t \pm \kappa x)}$. In acoustics, we also have the relation

$$|\kappa(\omega)| = \frac{2\pi}{\lambda(\omega)} = \frac{\omega}{c(\omega)} \tag{2.3}$$

where c is the wave speed. Note that, α, κ, c are both frequency-dependent and material-dependent.

In most cases, we apply Fourier transform [11] or Wavelet transform [109] to the time domain wave equation to get the time-frequency representation,

$$\frac{\partial^2}{\partial x^2}\tilde{P}(x,\omega) + \frac{\omega^2}{c^2}\tilde{P}(x,\omega) = 0 \tag{2.4}$$

and the frequency domain solution

$$\tilde{P}(x,\omega) = \tilde{P}_0(\omega)e^{-\gamma x} + \tilde{P}_0'(\omega)e^{\gamma x} \tag{2.5}$$

where $\tilde{P}_0(\omega)$ and $\tilde{P}_0'(\omega)$ are the magnitudes for each frequency component propagating in left and right directions (they are analogous to the initial time domain magnitude A in Eq.2.2). They can be determined by initial conditions.

Substituting the wave solution Eqn. 2.5 into the wave equation Eqn. 2.4, we have

$$(\alpha^2 + i2\alpha\kappa)\tilde{P}_0(\omega)e^{-(\alpha+i\kappa)x} + (\alpha^2 + i2\alpha\kappa)\tilde{P}_0'(\omega)e^{(\alpha+i\kappa)x} = 0 \tag{2.6}$$

2.2.2 Spatial Audio in Room Acoustics

Room acoustics [96] is the study of how sound behaves within an enclosed 3D space, such as an indoor environment, theatre and auditorium. It models the interaction of the propagating spatial audio waveform with enclosed 3D environment during its propagation path, typical interaction include absorption, reflection and diffraction. The received spatial audio at the receiver position from the source position contains reverberation effect that results from the interaction between propagating spatial audio and room environment. The main task of room acoustics is to model the reverberation effect caused from the source position to the receiver position. This reverberation effect is often represented as 1D room impulse response (RIR) feature, which is lengthy in data points (usually more than 20k data points) and chaotic in data distribution. In most cases, it is comprised of three main components: direct sound, early reflection and late reverberation.

There are two main ways to model room acoustics: wave-based modelling [9, 133, 86, 10] and geometry-based modelling (also called geometrical acoustics) [153]. The wave-based modelling utilizes sound wave nature to model sound wave propagation, whereas geometry-based modelling treats sound propagation as optic rays. Typical geometry-based modelling methods include ray tracing [92], image source method (ISM [6]), beam tracing [48] and acoustic radiosity [73, 122]. Here we introduce two main modelling methods: geometry-based modelling and wave-based modelling.

Geometry-based Modelling The most fundamental idea of geometry-based modelling is assuming propagating spatial audio act as rays [153]. In the simplest situation, propagating spatial audio is modeled as straight travelling lines. Whenever the ray hits an obstacle (*e.g. wall, furniture*), the propagating ray changes its direction and continue to propagate along a new direction. Once treating spatial audio as a ray, there is a strong correspondence between geometry-based modelling and the light propagation modelling in computer graphics whose goal is to produce photo-realistic images. The main difference between them is the "ray speed": while the light propagates at infinite speed so that the light can reach everywhere immediately, the acoustic-ray travels at finite speed (around 340 m/s), resulting in room impulse response to represent the reverberation effect. Since the wave nature of the spatial audio gets neglected by modelling the audio as rays, wave-related phenomena are missing so the resulting behavior does not fully reflect the actual propagating procedure. The diffraction effect, the propagation behaviour difference regarding to the spatial audio frequency difference fail to be modelled. For example, when the

audio source position and receiver position is blocked by a small barrier so that there is line-of-sight between the two positions. In reality, however, most of the spatial audio is capable of reaching to the receiver position due to edge diffraction, under the effect of which only the high frequency part is blocked.

Among all geometry-based modelling methods, image-source method [6] is the most-widely used. First proposed by Cremer in 1948 [36] and Mintzer in 1950 [115] for simple rectangular room and then extended to polyhedra room environment by Borish in 1984 [81], image-source method has been fully researched and developed and explored from both theoretical and applicable aspects. The fundamental idea behind is that the specular reflection on a rigid surface can be equivalently represented by a direct path from source mirrored against the reflecting wall. The mirrored source emits the same audio raw waveform. When the reflected spatial audio wave reaches to another rigid surface, the newly reflected path can be further obtained by mirroring the current mirrored source (first order) to the new mirrored source position (second order). As the propagating audio waveform constantly hits the rigid surface more and more times, more and more mirrored virtual sources are accordingly created (1-st order, 2-nd order, $\cdots$, n-th order).

Wave-based Modelling. Wave-based audio modeling methods revolutionize the simulation of audio propagation by leveraging principles from wave physics to create highly accurate and immersive audio environments. Unlike traditional geometry-based methods, which often simplify acoustic interactions within a scene, wave-based approaches delve into the intricacies of wave behavior. These methods, such as Finite Difference Time Domain (FDTD) simulations or Wave Field Synthesis (WFS), model audio propagation by solving the wave equation, considering factors like diffraction, reflection, and interference. The mathematical foundation lies in the wave equation, a partial differential equation that characterizes the spatial and temporal evolution of acoustic waves. By solving this equation, wave-based methods capture the dynamic nature of sound, providing realistic simulations of how audio waves interact with the environment. This stands in contrast to geometry-based methods, which often rely on ray tracing and geometric approximations, overlooking the nuanced effects of diffraction and wave interference. In essence, wave-based audio modeling offers a more comprehensive and physically accurate representation of audio propagation, contributing to the creation of immersive audio experiences in applications like virtual reality, gaming, and architectural acoustics.

2.2.3 Room Impulse Response

The Room Impulse Response (RIR) [132] serves as a cornerstone in the realm of acoustics and audio signal processing, providing an intricate depiction of the acoustic interactions within an enclosed space. Essentially, the RIR characterizes the way sound waves behave within a room, encompassing reflection, absorption, and diffraction that occur during the propagation path. This temporal and spectral representation of acoustic phenomena offers a comprehensive insight into the unique acoustic fingerprint of a given space, revealing the intricacies of how sound evolves over time within the room's geometry. Specifically, the room impulse response can be represented as 1D data points $h(t)$ in time domain, it encodes all the changes (reflection, absorption, diffraction, *etc.*) along the propagation path from the source position to thee receiver position. Let's represent the source audio $x(t)$ at position p_s, the received recorded audio $y(t)$ at position p_r, the corresponding room impulse response $h(t)_{p_s \to p_r}$, the receiver recorded audio can be mathematically expressed by convolving $h(t)_{p_s \to p_r}$ with the source audio $x(t)$,

$$y(t) = x(t) \circledast h(t)_{p_s \to p_r} + n(t) \tag{2.7}$$

where $\circledast$ indicates the 1D convolution operation, $n(t)$ is the associated independent noise. Analyzing the Room Impulse Response proves invaluable in various applications. In room acoustics, the RIR is pivotal for understanding the reverberation characteristics and identifying potential issues such as echo, flutter echoes, or unwanted resonances. Additionally, in audio engineering and sound system design, the RIR is used to optimize audio reproduction by accounting for the room's impact on sound quality. Moreover, in the burgeoning field of virtual acoustics, the RIR is crucial for creating realistic simulations of acoustic environments, enabling the synthesis of convincing spatial audio experiences.

In essence, the Room Impulse Response stands as a key tool for researchers, audio engineers, and acousticians, providing a nuanced understanding of the acoustic behavior within enclosed spaces. Its applications extend beyond traditional room acoustics, influencing the design and refinement of technologies that deliver immersive audio experiences in virtual environments and contribute to the continuous pursuit of high-fidelity audio reproduction.

2.2.4 Single-Channel Audio Representation

In audio signal processing, the representation of a signal in the time domain serves as the foundation for understanding its frequency content. Let $x(t)$ denote the audio signal, a function of time t. The Fourier Transform, a pivotal tool in signal processing, transforms this signal into the frequency domain, revealing its spectral components. The Fourier Transform of $x(t)$ is given by the equation,

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft}\, dt \tag{2.8}$$

where $X(f)$ represents the frequency content of $x(t)$. The result is a complex function, often represented as magnitude and phase map (or the real and imaginary map, respectively). In practice, it is often represented as a time-frequency representation map by squaring the real and imaginary map, and one typical example is shown in Fig. 2.2 (right-most sug-figure). The Fourier Transform provides valuable insights into the frequency distribution of the audio signal, allowing identification of dominant frequencies. Wavelet Transform [109], another powerful tool, refines this analysis by capturing not only frequency but also temporal information. The Continuous Wavelet Transform (CWT) of $x(t)$ is given by:

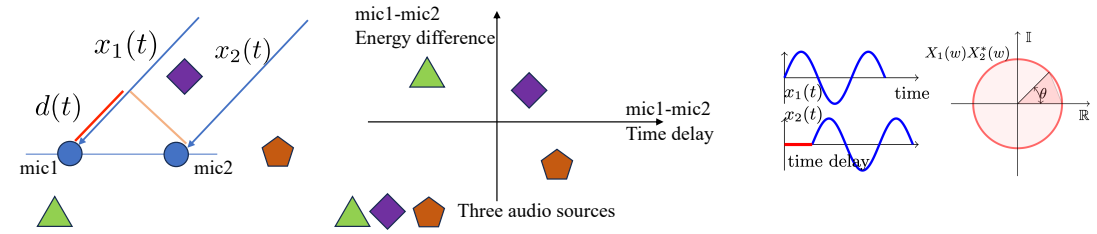
$$W(a, b) = \int_{-\infty}^{\infty} x(t) \Psi^* \left(\frac{t - b}{a} \right) dt \tag{2.9}$$

Where a represents the scale parameter, b is the translation parameter, and $\Psi(t)$ is the analyzing wavelet. Wavelet Transform provides a time-scale representation, offering a localized view of the signal in both time and frequency domains simultaneously.

2.2.5 Multi-Channel Audio Representation

Single channel spatial audio does not explicitly carry audio source spatial position information. The widely adopted way to jointly encode audio source semantic information (arise from single-channel audio) or spatial position information is to use microphone-array device to record the audio source by multiple microphones that are carefully configured into a microphone-array device.

Common microphone array configurations include linear, planar, and spherical arrays, each with its unique advantages and applications. Linear arrays consist of microphones arranged in a straight line, offering directional sensitivity along a single axis. Planar arrays feature microphones arranged in a two-dimensional plane, providing directional sensitivity in both horizontal and vertical dimensions. Spherical arrays



A. Time-delay visualization for two mics.

B. Different audio source exhibits time-delay and energy difference on two mics

C. Time delay visualization in time and frequency domain

Figure 2.1: We use two-microphone setting to illustrate the time-delay cue. A. given two microphones mic1 and mic2, the arrival time of the audio waveform $x_1(t)$ and $x_2(t)$ from the same audio source to mic1 and mic2, respectively is different, resulting in time-delay $d(t)$. Alongside the time-delay, the received audio energy by mic1 and mic2 is also different. Such time-delay and audio energy difference can be exploited to differentiate audio source position. For example, the three audio sources (triangle, rhombus and pentagon) can be successfully separated in sub-figure B. We further visualize the time-delay in both time and frequency domain for clear understanding in sub-figure C.

employ microphones distributed over the surface of a sphere, enabling omnidirectional sensitivity to capture sound from all directions.

The spacing between microphones in an array depends on various factors such as the desired spatial resolution, frequency range of interest, and intended application. In general, smaller microphone spacing results in higher spatial resolution but may also introduce spatial aliasing effects. Larger spacing provides greater robustness to spatial aliasing but may sacrifice spatial resolution.

For linear arrays, the spacing between microphones is typically uniform, with distances ranging from a few centimeters to several meters depending on the desired beam width and directivity. Planar arrays may have uniform spacing in both horizontal and vertical dimensions or vary depending on the array geometry and desired coverage pattern. Spherical arrays often have non-uniform spacing to achieve uniform sensitivity across all directions, with smaller spacing near the poles and larger spacing near the equator.

For microphone-array based device, the recorded spatial audio is multi-channel. The audio source spatial position cue arises from the inter-channel audio arrival time delay (or phase difference): for a given audio source, its arrival time to different microphones varies. Exploiting the time-delay (or phase difference) among the microphone-array audio can help inversely infer the audio source spatial position. Accompanying the time-delay, the received spatial audio energies on different microphones also vary

because the audio energy gradually decays along the propagation path. We visualize the inter-channel time-delay in Fig. 2.1

The four-channel sound waveforms provide enough cues to estimate a sound source's 3D spatial position and semantic label, in which the time-frequency representation obtained by short time Fourier transform (STFT) reveals the semantic label and inter-channel phase difference encodes its spatial position. Given one MicArray $\mathcal{A}_i = [a_{i,1}, a_{i,2}, a_{i,3}, a_{i,4}]$, we follow the practice in [182, 20, 57] to jointly encode the time-frequency representation in log-mel scale for each single channel waveform and generalized cross-correlation phase transform (GCC-Phat [12]) between each channel pair. GCC-Phat has been widely used to encode inter-channel phase difference [20, 2, 178, 21]. Given two channel sound waveforms $a_{i,k}$ and $a_{i,l}$ in $\mathcal{A}_i$, the GCC-Phat $f^{k,l}_{\text{gccphat},i}$ is expressed as,

$$f^{k,l}_{\text{gccphat},i} = \text{ifft}(\frac{F(a_{i,k}) \cdot F^*(a_{i,l})}{|F(a_{i,k})| \cdot |F^*(a_{i,l})|}), \quad k \neq l \tag{2.10}$$

where $\text{ifft}(\cdot)$ indicates inverse short time Fourier transform, $F(\cdot)$ indicates short-time Fourier transform (afterwards transformed to Log-mel scale), $(\cdot)^*$ indicates the conjugate. $k, l = \{1, 2, 3, 4\}$. Given the four-channel sound waveforms, we can extract 6 such GCC-Phat feature maps.

2.2.6 Literature Review on Audio Feature Representation

Based on the presentation in Sec. 2.2.4 and Sec. 2.2.5, we can transform the monochannel spatial audio into time-frequency representation, and further compute the inter-channel time-delay 2D feature map from multichannel spatial audio. Moreover, all these transformation can be achieved by hand-crafted signal processing methods with fixed and empirically chosen parameters (*e.g.*, hop length, window size). Most existing mature image-based neural network can be directly adopted to learn on top of the pre-computed 2D images such as convolutional neural networks [182] and recurrent neural networks [33, 72].

Existing approaches [55, 130, 168, 182, 121, 169] on SELD task unanimously follow this process pipeline. To obtain a single-channel audio time-frequency representation, existing methods compute different time-frequency representation, including short-time Fourier Transforms (STFT) [89, 91], Mel-Spectrograms [17, 60] and Log-Mel Spectrograms [1, 188, 20, 55, 166]. To obtain the inter-channel time-delay 2D map, GCC-Phat [12] has been widely adopted [55, 130, 121, 169]. We summarise those existing methods' time-frequency representation with hand-crafted transform

Table 2.1: Literature review on existing SELD-task based work's pre-processing of input spatial audio.

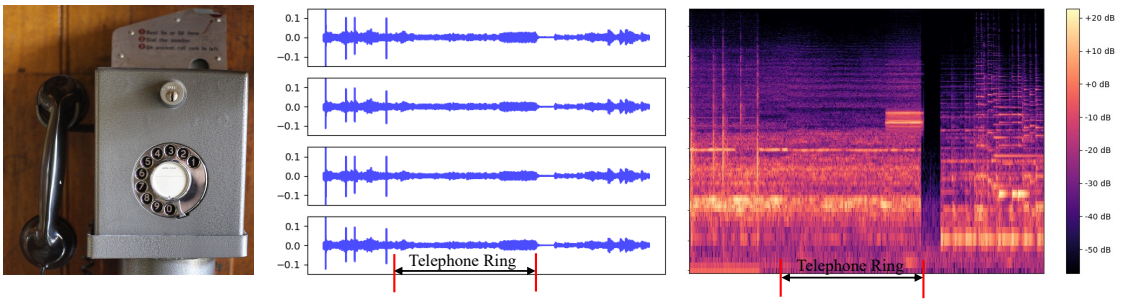
Paper	Year	STFT	MFCC	Mel-scale	LogMel-scale
Kothinti *et. al.* [89]	2019	✓	✗	✗	✗
Krause *et. al.* [91]	2019	✓	✗	✗	✗
Mesaros *et. al.* [126]	2010	✗	✓	✗	✗
Cakir *et. al.* [17]	2016	✗	✗	✓	✗
Hayashi *et. al.* [60]	2017	✗	✗	✓	✗
Pham*et. al.* [130]	2018	✗	✗	✓	✗
Adavanne *et. al.* [1]	2016	✗	✗	✗	✓
Xia *et. al.* [188]	2019	✗	✗	✗	✓
Cao *et. al.* [20]	2021	✗	✗	✗	✓
Grondin *et. al.* [55]	2019	✗	✗	✗	✓
Thi Ngoc *et. al.* [166]	2021	✗	✗	✗	✓
Tho Nguyen *et. al.* [121]	2022	✗	✗	✗	✓
Tho Nguyen *et. al.* [169]	2022	✗	✗	✗	✓

in Table 2.1, from which we can see their unanimous emphasis on extracting time-frequency feature with traditional hand-crafted signal processing methods. This observation serves as the fundamental motivation for us to design novel learnable filter bank to directly learn from spatial audio raw waveform. In the meantime, we also observed some preliminary work on designing learnable filters for acoustic data (especially speech) [155, 123, 79, 150, 194, 143, 195]. However, designing the corresponding filters for spatial audio and multi-channel microphone-array based spatial audio still remains as an uncharted territory.

2.3 Audio-Visual Signal Fundamentals

Sight and sound play a vital role in humans' perception and understanding of the surrounding 3D physical environment. They dominate the signals humans frequently depend on to interact with the physical world and communicate with each other. The abundance of various acoustic and visual signals in the physical world and the fast development of digital visual and acoustic signal recording technologies enable us to easily acquire massive acoustic and visual dataset for various research purpose, although for a long time the research on vision and research on audio have been developed in parallel and independently.

Sight and sound are integral components that shape the way humans perceive and comprehend their three-dimensional physical environment. The two sensory modalities not only dominate the signals individuals rely on for interacting with the world



A. Telephone B. Sound Waveform C. STFT Spectrum

Figure 2.2: The comparison of object of interest representation in vision modality and sound modality. A. In vision modality, the telephone is represented by coherent and contiguous area sharing the similar geometric primitives such as lines, texture and semantic context (image courtesy to ImageNet [39]). B. Four-channel microphone array recorded sound waveforms containing telephone ring visualization (sound courtesy to DCASE challenge [134]). C. The top channel sound waveform's STFT time-frequency map visualization.

but also serve as primary venue for communication among individuals. The richness of diverse acoustic and visual signals in the physical world, coupled with the rapid advancements in digital recording technologies for visual and acoustic signals, facilitates the effortless acquisition of extensive datasets for various research purposes. Despite the wealth of available data, it is noteworthy that, for a considerable duration, research in vision and audio has progressed along parallel trajectories, operating independently of each other.

In the face of abundant acoustic and visual information, the parallel development of vision and audio research has somewhat obscured the potential synergies that could arise from a more integrated approach. The convergence of these two realms holds immense promise for a deeper and more holistic understanding of our environment. Bridging the gap between visual and audio research opens avenues for cross-modal exploration, where the intricate interplay between sight and sound can be harnessed to bring new insights and enrich our comprehension of the complex dynamics of the physical world. As interdisciplinary collaboration becomes increasingly vital, the opportunity to synchronize and couple vision and audio research promises not only a more comprehensive exploration of sensory perception but also the potential for innovative applications across various domains.

2.3.1 Fundamentals of Visual Signal

The most fundamental source of visual information is light. Light is defined as the electromagnetic radiation with wavelengths visible to human eyes (between 380 and 750 nm). Electromagnetic radiation, such as light, is generated by changes in movement (vibration) of electrically charged particles. Light is characterized by its dual nature, behaving both as waves and particles (photons). It travels in straight lines and can interact with matter through reflection, refraction, diffraction, and absorption. In the context of visual signals, light serves as the stimulus that allows the visual system to perceive objects, colors, and the surrounding environment [124, 70], it carries the details of the visual world to the eyes. When light interacts with objects, it can be reflected, refracted, or absorbed, influencing the visual information that reaches the eyes. Our eyes, in turn, capture and focus light onto the retina, where photoreceptor cells convert it into electrical signals. These signals are then transmitted to the brain via the optic nerve, where complex neural processing occurs to form a visual perception. In essence, light is the medium through which visual information is conveyed, and vision is the result of the brain's interpretation of that information. When the light reaches to human eyes, it activates numerous photoreceptors (around 260 million) on the retina, the output is further is sent to and processed by 2 million ganglion cells to form high-level visual information [140]. Visual signals are often measured and characterized using various parameters and metrics depending on the specific context and goals of the analysis. Here are some commonly adopted measurements and metrics:

1. **Luminance**: Luminance measures the intensity of light emitted or reflected from a surface and is an important factor in understanding the brightness of visual stimuli.

2. **Contrast**: Contrast measures the difference in luminance between different parts of an image. High contrast enhances visibility and can be important in various applications, such as image quality assessment.

3. **Colorimetry**: Colorimetry involves the measurement of color properties, including hue, saturation, and brightness. Various color spaces, such as RGB (Red, Green, Blue) or CIE XYZ, are used for quantifying and analyzing colors.

4. **Brightness**: Brightness is a subjective perception of the overall lightness or darkness of an image. It is influenced by luminance but also by the characteristics of the display device.

5. **Gamma Correction**: Gamma correction is a nonlinear adjustment applied to the brightness values in an image to compensate for the nonlinear response of human vision.

6. **Visual Acuity**: Visual acuity measures the ability to distinguish fine details. It is often assessed using tests like the Snellen chart, and it is crucial in fields such as optometry and ophthalmology.

7. **Signal-to-Noise Ratio (SNR)**: SNR measures the ratio of the signal strength to the noise level in an image. A higher SNR indicates better image quality.

8. **Color Temperature**: Color temperature characterizes the color appearance of light sources and is often expressed in Kelvin (K). It is crucial in applications where accurate color representation is essential, such as in photography or video production.

9. **Dynamic Range**: Dynamic range measures the range of luminance values that a system can capture or display. A higher dynamic range allows for the representation of a broader range of light intensities.

The 2D images are typically used with light as the primary medium by computer vision research community. Within the 2D images, the object of interest usually corresponds to a localized contiguous area with similar color texture or relevant context (see Fig. 2.2 left, we use telephone as an example).

2.3.2 Audio-Visual Signal Integration

Audio-visual signal lies at the intersection of auditory and visual perception, encompassing the principles underlying the representation, processing, and interpretation of signals that combine both auditory and visual information. While audio signal is often characterized by sound waveform amplitude, frequency and phase and propagates through the medium such as air, wall and water, visual signal consists of light waves that is often characterised by brightness, color and dynamic range. One direct comparison of cross-modal signals representing an object interest (telephone) is shown in Fig. 2.2. The way to integrate the two modal signals should be task-specific and fully consider how the task can benefit from each single modal separately and the crossmodal signal (*e.g.*, audio-visual signal) jointly.

Due to the heterogeneous correlation between audio and visual signals, audio-visual signal integration should reflect the specific correlation between the two modalities and further design the corresponding integration algorithm accordingly. If the audio-visual signals are strongly-correlated, for example, the object of interest can be independently detected from each single modality separately, the integration then becomes much flexible and relatively simple. One can simply combine the two methods working for each single modal signal together to achieve the final task. If the audio-visual signal is weakly-corrected (*e.g.*, our work presented in Chapter 5), a deep investigation how the object of interest is revealed from each single modal signal and how the crossmodal audio-visual signal jointly enhances the detection of the object of interest is thus required. This investigation will later guide the audio-visual signal integration framework.

Given the ubiquity of audio-visual signal in real-scenarios, the development of audio-visual signal integration framework should reflect the real-scenario case. In fact, the audio-visual signal correlates in heterogeneous ways, ranging from strong-correlation, weak-correlation to no-correlation in a commonly seen dynamic environment. Seeking a unified and robust audio-visual signal integration framework that resembles the heterogeneous real-scenario remains as a long-term research goal.

2.3.3 Literature Review on Audio-Visual Learning

The widespread availability of online video data containing both images and audio has attracted significant research attention on audio-visual joint learning in recent years [199]. The audio-visual separation [49, 44, 4, 107, 118, 56, 200], as well as localization and navigation [50, 156, 171, 137, 198, 197, 157, 149, 127, 80]. Those video-based audio-visual learning work are constrained into the video image frames, and they assume audio and vision are strongly-correlated so that the object of interest can be independently informed by either audio signal or visual signal. For example, the guitar playing audio corresponds to guitar player who can be directly detected from the image.

Some recent work [27, 26, 88, 25] in embodied AI research proposed AudioGoal navigation task, in which the goal is to navigate to the audio goal by taking the ego-centric image and spatial audio as input. In this case, the spatial audio can be either uncorrelated with the vision (which means the audio goal is virtually placed in the environment), or strongly correlated with the vision [25] (*e.g.*, the audio goal is a telephone ring that can be detected from the vision).

In my DPhil research, we choose to focus on a new audio-vision weak-correlation case: the audio source exhibits no visual entity but lies on the object's physical surface. It reflects some real-scenarios. For example, gas leaking generates audio that the audio source lies on the surface of the gas pipe but exhibits no visual entity. We will present one audio-visual learning work that focuses on audio-visual weak-correlation situation.

Chapter 3

Spatial Audio Learning

3.1 Motivation Discussion

The motivation of this spatial audio learning is based on the finding that the majority of existing work [21, 20, 27, 26, 121, 169] on processing spatial audio that,

1. unanimously convert each input spatial audio channel into 2D spectrogram image and start learning on top of the spectrogram image, which can be easily obtained by applying classic Fourier transform (*e.g.*, short time Fourier transform, STFT).

2. heavily rely on the mature RGB image based neural network to learn from the 2D spectrogram image.

3. have simply tested on relatively simple acoustic scenarios where limited audio sources are available and the whole acoustic scene is clean (absence of noise).

While being able to achieve promising performance, the "treating spatial audio as 2D image" strategy naturally raises the research questions that:

1. Can we instead design learnable filter bank that can directly ingest multi-channel audio raw waveform without any hand-crafted pre-processing? It lends us the advantage of designing the whole neural network to be end-to-end trainable, which is a common practice in various vision based deep neural network design.

2. What is the downside in classic hand-crafted spectrogram map, when comparing with deep neural network learned spectrogram?

3. How can we handle the much more complex acoustic scene where multiple audio sources co-exist in the environment (high polyphonicity)?

To address these research questions, we dedicated ourselves to designing a novel learnable filter bank. The goal was to replace conventional hand-crafted preprocessing methods with learnable filter bank processing, without introducing much additional computational load while enhancing performance. Specifically, we present three main work: SoundDet [68], SoundSynp [66] and SoundCount [62].

SoundDet [68] (ICML'21): a learnable filter bank that is of constant length in time domain but vary in frequency domain. It jointly learns the time-frequency representation from each single channel and the inter-channel time-delay during one sweep. To encode the inter-channel time-delay, a learnable time-delay parameter is further associated with each single learnable filter. To validate the efficacy, we test our method on the sound event detection and localization (SELD) task (DoA estimation) [134].

SoundSynp [66] (AISTATS'23): a learnable multi-scale filter bank motivated by uncertainty principle. The learnable filterbank varies in both time frequency domain. The whole design in based on the theory that lower-frequency filter requires longer length in time domain, higher-frequency filter instead requires shorter length in order to achieve discriminative time-frequency resolution. To validate the efficacy, we test our method on the sound event detection and localization (SELD) task, but on both DoA estimation [134] and audio source 3D physical position estimation [58].

SoundCount [62] (AAAI'24): Existing dataset for SELD task just contain relatively simple acoustic scenario where a maximum of three audio sources can co-exist at the same time. Such simplified data may not reflect the real-scenario and can not fully test the potential of the designed learnable filter bank. How to design learnable filter bank to cope with highly polyphonic acoustic scene? To this end, we propose a SELD-related new task: count the sound event number arising from mono-channel audio snippet where multiple audio sources may overlap temporally. We further propose a dyadic decomposition framework to learn the time-frequency representation in a hierarchical and multi-stage manner. Unlike SoundDet and SoundSynp which obtain the time-frequency representation in one-step process, SoundCount learns the time-frequency in the coarse-to-fine dyadic decomposition manner, we show such hierarchical decomposition strategy is capable of learning more representative time-frequency representation in highly-polyphonic situation.

3.2 SoundDet: MaxCorr Multichannel Filter Bank

We present a new framework SoundDet, which is an end-to-end trainable and light-weight framework, for polyphonic moving sound event detection and localization. Prior methods typically approach this problem by preprocessing raw waveform into time-frequency representations, which is more amenable to process with well-established image processing pipelines. Prior methods also detect in segment-wise manner, leading to incomplete and partial detections. SoundDet takes a novel approach and directly consumes the raw, multichannel waveform and treats the spatio-temporal sound event as a complete "sound-object" to be detected. Specifically, SoundDet consists of a backbone neural network and two parallel heads for temporal detection and spatial localization, respectively. Given the large sampling rate of raw waveform, the backbone network first learns a set of phase-sensitive and frequency-selective bank of filters to explicitly retain direction-of-arrival information, whilst being highly computationally and parametrically efficient than standard 1D/2D convolution. A dense sound event proposal map is then constructed to handle the challenges of predicting events with large varying temporal duration. Accompanying the dense proposal map are a temporal overlapness map and a motion smoothness map that measure a proposal's confidence to be an event from temporal detection accuracy and movement consistency perspective. Involving the two maps guarantees SoundDet to be trained in a spatio-temporally unified manner. Experimental results on the public DCASE dataset show the advantage of SoundDet on both segment-based and our newly proposed event-based evaluation system.

3.2.1 Introduction

Acoustic source detection, classification and localization is a key task for a wide variety of applications, such as smart speakers/assistants, interactive robotics and scene visualization. Broadly, a microphone array (typically with four microphones) receives sound events from the environment, which are corrupted by background noise. The task is to firstly detect events, classify them, and finally estimate the physical location or direction-of-arrival angle (DoA).

Early work has demonstrated how deep learning techniques can be used to provide good performance for sound event detection and localization. However, the majority of these works treat event detection and localization as separate problems [2, 168]. For example, a framewise classifier and regressor are used separately [167, 54] to obtain event class and spatial location. In addition, they also rely heavily on handcrafted

pre-processing to convert the raw waveforms into time-frequency representations (*e.g.*, log-mel spectrograms, GCC-PHAT [12]) that are more amenable to process with mature 2-D CNN networks like ResNet [61] followed by LSTM [72] or GRU [33] network for handling temporal dependencies. Lastly, sound event detection is typically achieved following a segment-based approach (*e.g.*, cut the waveform into second-long snippet) which inevitably leads to incomplete or partial detection that span segment boundaries.

The relatively unexplored area becomes more challenging when it considers polyphonic event detection where events with different DoAs overlap temporally. In addition, rather than necessarily being stationary, these events undergo a motion. This increased realism leads to poor performance or even incapability of existing methods [111, 2] that make a strong assumption that only a single and stationary event exists at a time slot.

In this chapter, we rethink the polyphonic sound event detection and localization problem. We draw inspiration from the success of object detection in 2-D images and 3-D point clouds, and think of an event as being analogous to a "sound-object" that has a specific location (spatial and temporal onset/offset), size (duration) and class (semantic category).

Unlike computer vision where the primary challenge is occlusion of overlapping objects leading to partial views, the physics of sound are such that sound-objects superimpose and additively mix with one another. Sound objects thus have to rely on frequency-selective and phase-sensitive approaches so as to be distinguished from background or ambient noise. Rather than detecting in several separate and independent steps, a more desirable event detector needs to be end-to-end that directly consumes raw waveform and outputs predictions, whilst generates minimal computation cost. To this end, we propose a unified and light weight polyphonic sound event detection and localization framework, dubbed SoundDet, that naturally meets these requirements.

To the best of our knowledge, SoundDet is the first approach to directly consume the raw waveform and not rely on any spectral transformation pre-processing. To achieve this, we propose to learn a set of channel correlation aware and frequency sensitive bank of filters, these parametric filters naturally capture the phase difference between channels and frequency bandpass for different sound events. Comparing with standard 1D/2D convolution, learning such filter bank requires minimal computation cost and parameters, making it ideal for processing the raw waveform with large sampling rates. SoundDet takes a novel view and adopts a dense event proposal

strategy to directly estimate an event's location, size and category information, unlike current approaches that simply treat it as a segment-wise prediction problem. To better model a sound event's completeness and continuity, we introduce concept like motion smoothness, boundary sensitive temporal overlapness.

To comprehensively evaluate the performance, we further propose an event-based evaluation system. Unlike previous segment-based evaluation, event-based metric takes sound event's confidence score, length and motion consistency into account and accumulates performance under different score threshold. We embrace 2D image object detection metric [100] and propose to calculate average precision/recall, mean average precision (mAP), mean average recall (mAR) for each category. Experiment on 14 sound category DoA estimation DCASE dataset [3] shows the advantage and efficacy of SoundDet.

In summary, our main contributions are three fold: **First**, we propose the first framework for polyphonic moving sound event detection and localization that directly consumes raw waveform and estimate sound event from "event-object" perspective. **Second**, we propose a novel phase and frequency sensitive MaxCorr filter bank that is parameter and computation lightweight to process raw waveform. **Third**, an event-based evaluation system that avoids arbitrarily setting threshold is proposed to better evaluate the framework.

3.2.2 Problem Formulation

We have a multi-channel sound recording (aka sound waveform) W_c recorded by one microphone array station or multiple microphone array stations at a fixed sampling rate, W_c contains a sound event set [1] $E = \{e_i = (t_{s,i}, t_{e,i}, l_i, c_i)\}_{i=1}^{N}$, where $t_{s,i}$, $t_{e,i}$ are the i-th sound event's start time and end time, respectively, c_i indicates the sound event's semantic information (*e.g.*, category, baby cry, phone ring). l_i indicates the sound event's spatial location, it can either be a fixed location if the sound event is stationary or a spatial trajectory if it undergoes a motion. The task is to learn a model $\mathcal{F}$ to accurately recover these sound events directly from raw sound waveform,

$$E = \mathcal{F}(W_c) \tag{3.1}$$

These sound events are polyphonic, which means they may overlap in the time dimension.

[1] In this chapter, a sound event is restricted to be a temporally and spatially continuous and semantically meaningful. Two events with exactly the same temporal location belong to different class.

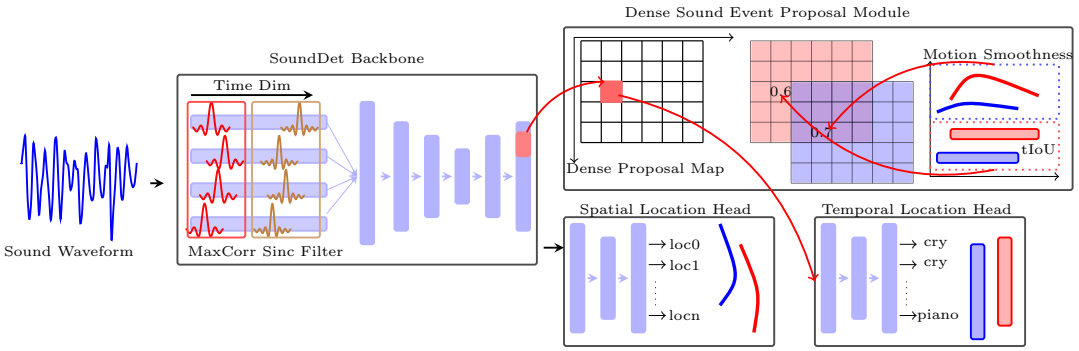


Figure 3.1: SoundDet pipeline: the input multi-channel waveform is fed to a backbone neural network to learn framewise representations. The backbone consists of MaxCorr filter bank and a encoder-decoder like network. A spatial location head and a dense event proposal generation head connect the backbone network in parallel. The spatial location head predicts per-class spatial location in a framewise manner. The dense proposal head gives all possible event proposals, each one associates with an event-wise feature, certain start/end time. The dense proposal head maintains two score maps, one map measures proposals' temporal overlap with group truth and the other represents proposals' motion smoothness, which directly comes from spatial location head. The two maps unifies spatial location head and temporal location head so that the whole framework can be jointly trained in a unified manner.

3.2.3 MaxCorr Band-Pass Filter Bank and Backbone

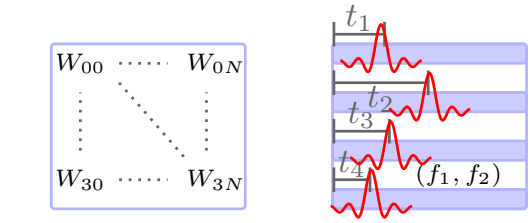
Sound waveform usually has high temporal resolution, a typical 1 minute recording of 30kHz has 1.8 million data points. This leads to unbearable computation burden (FLOPs) for standard 1D/2D convolution. To counteract the large data point size issue and further to enforce the neural network to learn meaningful sound event spatio-temporal representation, we propose to learn a set of rectangular band-pass filters in frequency domain (a Sinc filter in time domain equals to rectangular filters in frequency domain, see [143]), each of which contains a lower learnable frequency cuttoff f_1 and a higher learnable frequency cutoff f_2. After converting the rectangular band-pass filter to time domain, we obtain *sinc* convolution kernels k,

$$k[n, f_1, f_2] = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n) \tag{3.2}$$

where $sinc(x) = sin(x)/x$. Note that the band-pass filter is a parametric filter that the learnable parameter number is independent of the kernel length. Moreover, the kernel length is usually much larger than standard 1D/2D convolution kernel length. A typical band-pass filter kernel length can be set as a large number (in our case 481), whereas the standard 1D/2D convolution kernel size is usually 3 or 5. These

two properties guarantee these band-pass filters are highly efficient at both FLOPs computation and model parameter number. The SincNet filter has been successfully applied for speech recognition [143].

To further enforce the above designed filter banks to be able to dynamically decouple the sound event's temporal information and spatial information. We further expand the sinc filter to multiple channels to model the internal correlation between different channels. We draw inspiration from canonical time difference of arrival (TDoA) estimation to explicitly encode sound events' geometrical information recorded by a microphone array. TDoA is a well-established technique for sound source geolocalization. It exploits the arrival time difference among microphones to recover the sound event spatial information (*e.g.*, DoA or physical location) by computing the maximum time delay in either frequency domain or time domain [178]. We involve TDoA estimation into our band-pass filter bank design and propose a new filter bank $g(\cdot)$ that is capable of learning the correlation between different channels. Learning such filter bank helps the neural network to dynamically learn how to integrate between-channel correlation and event frequency band pass to best recover sound events, we thus call it MaxCorr filter bank,



Standard 1D/2D Conv MaxCorr Filter

Figure 3.2: Visual comparison of standard 1D/2D convolution with our proposed MaxCorr filter. While the learnable parameters of standard 1D/2D convolution equals to filter size (left side image), MaxCorr filter is parametric and introduces much fewer learnable parameters (right side image). For four channel waveform inputs, each MaxCorr filter merely has six learnable parameters (two frequencies and four time shifts).

$$g(n, f_1, f_2, t_1, \cdots, t_c) = [k[n + t_1], \cdots, k[n + t_c]] \tag{3.3}$$

where t_i is the i-th channel time shift parameter, f_1, f_2, $t_1, \cdots, t_c$ are learnable parameters and in our case $c = 4$ as there are four channels. To make the time shift parameter to be differentiable, we adopt sigmoid like soft truncation to approximate the round operation, like [47] do. Please see Fig. 3.2 for MaxCorr filter illustration and its comparison with standard convolution.

Following the MaxCorr filters, we add an encoder-decoder like 1D convolution network to learn framewise representations. In the encoder, we gradually reduce the

temporal length but increase the filter number, resulting in a compressed representation. In the decoder, it goes the opposite way and finally forms a "bottleneck"-like backbone neural network. To mitigate the model training degradation dilemma, we add skip connection between the encoder and decoder. The MaxCorr filter bank and encoder-decoder like neural network together constitute the Backbone neural network, which learns a framewise representation $[f_1, f_2, \cdots, f_N]$, where each representation's temporal duration equals to ground truth labelling resolution, such as 100 millisecond resolution.

We further present dense sound event proposal. Given the representation learned by the backbone, the next step is to generate sound event proposal. Potential sound events freely span in the temporal and spatial space, which means sound events are largely varying in their temporal length and spatial location. An efficient sound event proposal generation module thus should be: 1. able to handle sound events with large varying temporal length and spatial location. 2. computationally efficient enough (*e.g.*, generating potential event proposals at parallel). To this end, we propose a dense sound event proposal generation module, which organizes all potential sound events into a compact organization and all the potential events can be generated in parallelization.

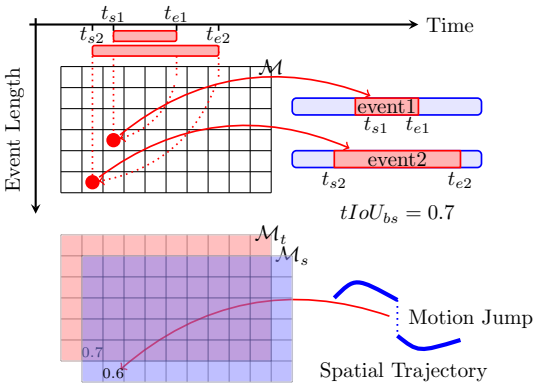


Figure 3.3: Dense sound event proposal. A matrix $\mathcal{M}$ is created to densely represent all potential sound events, each cell of the matrix corresponds to a sound event with certain start time and end time. For example $\mathcal{M}^{i,j}$ represents that event with start time t_j and end time t_{i+j}. The matrix maintains two score maps $\mathcal{M}_t$ and $\mathcal{M}_s$ of the same size. While $\mathcal{M}_t$ records proposals' temporal overlap with ground truth, $\mathcal{M}_s$ measures proposals' motion smoothness. The two score maps jointly represent a proposal's likeliness to be a true positive event. The motion jump (sound event's spatial position's sudden change), for example, largely violates the motion smoothness rule, so it is highly unlikely that the proposal is a true event.

Specifically, we compactly organize all potential sound events into a matrix-like representation $\mathcal{M}$, where the row indicates the sound event's temporal length and the column represents its start time. Each cell $C_{i,j}$ in $\mathcal{M}$ corresponds to a sound event with certain start time j and end time $i + j$. A careful selection of the size of matrix $\mathcal{M}$ guarantees all potential sound events have their

unique "cell" in $\mathcal{M}$. The matrix-like compact organization enables fast computation due to matrix parallelization computation supported by most GPU/CPU engines. Similar idea has been used for video-based activity recognition [99].

Accompanying the dense proposal map are two score maps of the same size: $\mathcal{M}_t$ measuring a sound event proposal's temporal overlap with ground truth, and $\mathcal{M}_s$ sound event proposal motion smoothness map. The two score maps jointly help to measure sound event proposals' confidence to be a true positive sound event. Specifically, the potential proposal indicated by cell $C_{i,j}$ is associated with a event wise feature representation f_{ij}, a temporal overlap confidence score $s^t_{i,j}$ from $\mathcal{M}_t$ and a motion smoothness score $s^s_{i,j}$ from $\mathcal{M}_s$. $f_{i,j}$ derives from the averaging features learned by Backbone network spanning its temporal duration $[t_j, t_{i+j}]$, $f_{i,j} = \frac{1}{i}\sum_{k=j}^{i+j} f_k$.

Temporal overlap score map $\mathcal{M}_t$ measures two sound events' temporal overlap degree. Here we adopt temporal intersection over union (tIoU). Given two sound events $s_1 = [t_{s1}, t_{e1}]$ at cell $\mathcal{M}_{i,j}$ and its ground truth temporal location $s_2 = [t_{s2}, t_{e2}]$, their tIoU at $\mathcal{M}^{i,j}_t$ is defined as,

$$tIoU = \frac{max\{0, min(t_{e1}, t_{e2}) - max(t_{s1}, t_{s2})\}}{max(t_{e1}, t_{e2}) - min(t_{s1}, t_{s2})} \tag{3.4}$$

We further introduce temporal overlapness score map. The higher of tIoU, the more likely the proposal at $C_{i,j}$ is a true positive proposal. Note that $\mathcal{M}_t$ is class-agnostic and its responsibility is to generate all potential event proposals (also can be thought as eventness proposal). In training, we train a $\mathcal{M}_t$ map regressor to learn the mapping from event wise feature representation to tIoU score. During test, we partially use the regressor to generate raw sound event proposals by choosing the large scores. To further enhance to regressor to fast localize a sound event, we propose a boundary sensitive tIoU $tIoU_{bs}$ which explicitly penalizes even a small temporal location alteration by multiplying an exponential term,

$$tIoU_{bs} = tIoU \cdot e^{-w\cdot(1-tIoU)} \tag{3.5}$$

where w is a decay weight controlling tIoU decay rate, which we set to 2. Our observation is that as a sound event temporal length increases, slight temporal alteration results in small tIoU score change. For example, given the ground truth sound event resides in frames [20,80] and the proposal is [22,82], the original tIoU will be very high (0.94 in this case) as there is a high degree of overlap. This does not give sufficient steering to the network to correctly align the event boundaries. Under our

new metric, however, $tIoU_{bs} = 0.83$, the misalignment is prominently magnified so that the network receives enough clue to improve its localization capability.

We then introduce spatial motion smoothness map. An independent sound event's spatial trajectory should be consistent and smooth. In other words, there should be no abrupt "jump" along the sound event's trajectory, regardless of its motion status. We call this property as motion smoothness. Specifically, we model motion smoothness as the maximum neighboring location displacement along the sound event's motion path. Mathematically, for the sound event $C_{i,j}$'s sequential spatial location $\{l_0, \cdots, l_i\}$, the motion smoothness is defined as the maximum neighboring spatial location displacement d.

The overall SoundDet training pipeline is shown in Fig. 3.1. The multi-channel raw waveform input is fed to the backbone neural network to efficiently learn framewise representation. The spatial location recovery head and dense event proposal generation head directly build on the learned framewise representation to learn framewise per-class sound event spatial locations and densely generate event proposals (per-event representation), respectively. The densely generated event proposals are fed to temporal recovery head to learn proposals' category label.

In practice, the temporal location head builds on event wise feature. It consists of three sub-heads: a multi-label classification head deciding the proposal's category label (*head*1); A binary classification head deciding whether a proposal is a foreground or background event (we can call it eventness prediction for better understanding). A tIoU regression head to learn to regress tIoU score based on the event wise feature input. The three heads consist of several full connection layers (FC). For multi-label classification and binary classification, we adopt the standard cross entropy loss. For the tIoU regression head, we adopt mean squared error (MSE) loss. The computation of $\mathcal{M}_s$ requires no specific learning process because it is directly derived from spatial location head, the MSE loss is adopted again to reinforce event's smoothness.

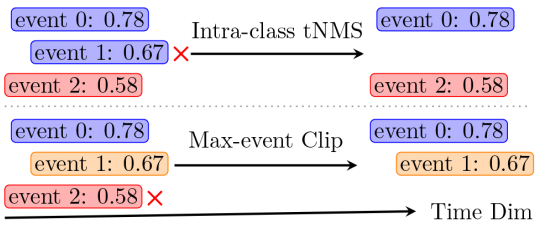


Figure 3.4: SoundDet postprocess illustration: we apply inra-class NMS to suppress sound events of the same classes along the temporal axis. Given multiple different temporally-overlapping sound events, max-event clip process is applied to clip untruthful events. Different colors indicate different sound event classes.

The dense sound event proposal map presented above covers all possible events, it inevitably leads to the data imbalance problem because most pre-generated proposals are negative and just very few proposals are positive. We adopt two strategies to mitigate this problem, 1) increase positive proposal number, in which we set a tIoU threshold t_d and any event with the tIoU larger than the threshold t_d is treated as true positive proposal. 2) negative proposal random dropout, with the remaining negative samples, we randomly drop part of them with a probability p_d. By carefully setting t_d and p_d value, we can adjust the number of positive and negative detection number (for example, higher t_d leads to fewer positive detections, smaller p_d leads to more negative detections), and further keep positive and negative ratio to be 1:1.

During inference, the calculated $\mathcal{M}_s$ and $\mathcal{M}_t$ are jointly used to generate sound event proposals. The sound event proposal at cell $C_{i,j}$ is treated as positive proposal only if it passes two test: 1) $\mathcal{M}_s^{i,j} > d_s$ so that it satisfies motion smoothness rule; 2) $\mathcal{M}_t^{i,j} > d_t$ so that it resides in the designated temporal location (t_j, t_{j+i}) with a high confidence (see Fig. 3.3 for illustration). d_s and d_t are predefined spatial motion smoothness and temporal overlapness threshold. Once it passes the two tests, the final event detection score derives from the multiplication of its motion smoothness score, temporal overlapness score and the corresponding multilabel classification score. The class label comes from the same multilabel classification head and the spatial location is derived from spatial location head.

The raw event candidates obtained above overlap in the temporal dimension. A post-processing is thus required to remove the redundancy (see Fig. 3.4). The post-processing consists of two main parts: intra-class temporal non-maximum suppression (tNMS) and max-event clip. In intra-class tNMS, we compare within the same event class, for two events of the same class with temporal overlap (tNMS) above a predefined threshold, we delete the event with smaller confidence score. In max-event clip, we restrain the maximum number events that can happen at the same time. Max-event clip is a class-agnostic operation, for a set of events that happen at the same time, we sort them according the confidence score s_i in descending order and only keep the top-M events. M is a predefined max-event threshold. The post-processing is an iterative process, after which we get the final K detected sound events from N raw proposals, where $K \leqslant N$.

3.2.4 Experiment

SoundDet is capable of estimating 2D/3D DoA, physical location, distance and motion of various sound events in polyphonic and moving scenario. In this experiment,

Table 3.1: Segment-based evaluation result. We report result for different event length: All (overall), Small (0-2s), Medium (2-7s) and large (> 7s.) SoundDet is more advantageous at predicting longer sound events than EIN [20]. For the LR_{CD} metric, please refer to the main paper. Note that SD indicates SoundDet, SD_bb indicates SoundDet_backbone.

Methods	$ER_{20^\circ}(\downarrow)$				$F_{20^\circ}(\uparrow)$				$LE_{CD}(\downarrow)(^\circ)$			
	All	Sma	Mid	Lar	All	Sma	Med	Lar	All	Sma	Med	Lar
SELDnet(foa)	0.63	0.64	0.63	0.79	0.46	0.48	0.48	0.42	23.1	22.8	23.2	20.5
SELDnet(mic)	0.66	0.68	0.66	0.79	0.43	0.44	0.45	0.42	24.2	23.9	23.6	22.8
EIN[20]	**0.25**	**0.30**	0.25	0.29	**0.82**	**0.80**	0.82	0.83	**8.0**	**8.1**	8.0	**8.5**
SD_bb(foa)	0.74	0.79	0.75	0.74	0.38	0.37	0.40	0.38	21.7	22.2	16.7	21.7
SD_bb(mic)	0.80	0.85	0.82	0.90	0.30	0.29	0.31	0.34	28.6	29.5	28.6	21.0
SD_nomcorr(foa)	0.90	0.91	0.90	0.90	0.05	0.05	0.06	0.06	79.4	79.3	79.8	79.0
SD_nomcorr(mic)	0.90	0.95	0.95	0.90	0.04	0.04	0.04	0.04	87.3	85.9	87.4	91.3
SD_nomots(foa)	0.31	0.37	0.38	0.35	0.77	0.72	0.74	0.72	9.0	10.1	9.8	7.4
SoundDet(foa)	**0.25**	0.31	**0.24**	**0.26**	0.81	0.79	**0.83**	**0.86**	8.3	8.5	**7.7**	8.1

we focus on indoor 3D DoA estimation tasks.

We evaluate SoundDet on TAU-NIGENS DCASE sound event detection and localization (SELD) [3] dataset. It is an indoor synthetic sound recording collected by an Eigenmike spherical microphone array and available in two data formats: first-order ambisonics (FOA) and tetrahedral microphone array (MIC). MIC indicates four microphones with different orientations, usually in spherical coordinates. FoA is known as B-format, consisting of omni-directional and x, y, z direction components. The possible azimuths of the sound recording span the whole range $[-180^\circ, 180^\circ)$ and the elevations lie in range $[-45^\circ, 45^\circ]$. Recorded sound events are either stationary or moving, and a maximum of two sound events overlap spatially and temporally. In total, 14 sound event categories are generated: alarm, crying baby, crash, barking dog, running engine, female scream, female speech, burning fire, footsteps, knocking on door, male scream, male speech, ringing phone and piano. Each sound recording is 1 minute long with sampling rate 24 kHz. For more details about this dataset, please refer to [3]. There are 8 folds recordings, each of which has 100 1-min *.wav* format recording. We follow the official splits and use 1-6 folds for train and the remaining 7-8 folds (200 1-min recordings) for test.

We incorporate two evaluation metrics:

Segment-based Metric is the standard evaluation metric for existing methods [2, 167, 111]. It compares prediction and ground truth within non-overlapping short temporal segments (one sec.) and evaluates each frame's prediction result by jointly considering its predicted class and spatial location. For event detection evaluation, it

calculates $F1$-score ($F_{\leq T^\circ}$) and Error Rate ($ER_{\leq T^\circ}$), where a true positive prediction has to be within a spatial distance threshold with the corresponding ground truth spatial location (typically with an angular threshold $T = 20^\circ$). For event localization evaluation, we compute class dependent localization error LE_{CD} which measures the average location distance between prediction and ground truth, and localization recall LR_{CD} measures the ratio of how many such localizations were estimated within a class. For more details, please refer to [111].

Event-based Metric draws inspiration from evaluation on 2D image object detection. It treats a sound event as an independent instance with specific start time, end time, framewise spatial location and confidence score belonging to one class (Rather than simply binarizing a detection as false or true detection with an arbitrary threshold, event-based metric comprehensively evaluates performance under various event confidence score and tIoU thresholds, calculating average precision (AP) and average recall (AR) for each class separately. Finally, mean average precision (mAP) and mean average recall (mAR) is accumulated by averaging AP and AR throughout classes respectively.

Specifically, given K detected sound events $E_p = \{t_{s,i}, t_{e,i}, l_i, c_i, s_i\}_{i=1}^{K}$ and N ground truth sound events $E_{gt} = \{t_{s,i}, t_{e,i}, l_i, c_i\}_{i=1}^{N}$ of the same class, we compute the average precision and average recall under tIoU in range $[0.1, 0.05, 1.0]$, where 0.05 is the stepsize. For a particular tIoU, the AP and AR is computed by averaging precision scores and recall scores obtained under various confidence score in range $[0.1, 0.05, 1.0]$. Please note that we do not take any predefined tIoU and confidence score threshold because arbitrarily chosen thresholds inevitably introduces human bias e.g. the choice of an entirely arbitrary angular threshold $T = 20^\circ$ in the segment based approach. Our proposed event-based metric instead provides a more objective and comprehensive metric.

We compare SoundDet with two existing methods: SELDNet [2] and EIN [20]. Since we focus on polyphonic and moving scenario, other relevant methods [168, 22] that merely work on stationary and monophonic sound events are not discussed here. SELDNet [2] jointly trains sound event detection and localization with CRNN network. It extracts Log-mel, GCC-PHAT, Intensity features as neural network input. Bidirectional GRU network is involved to model temporal dependency. SELDNet is treated as the baseline. EIN [20] is a very recent work which models sound event detection and localization with two identical but separate neural networks as a multi-task learning format. It uses a log-mel spectrogram as input for event detection, and

a GCC-PHAT approach for DoA estimation. The two parallel networks are independent but with soft parameter sharing. In addition, multi-head self-attention [179] is applied. EIN has large parameter size and is currently the state of the art approach under segment-based evaluation metric.

The above two methods generate framewise predictions, in which each frame is associated with class classification score and predicted DoA value. In order to generate a score that metrics event classification score and DoA closeness between predicted location and ground truth location, we propose to integrate mean classification score s_c and mean DoA Euclidean distance d_{DoA} together in the formula $s = s_c \cdot e^{-d_{DoA}}$. For fair comparison, we don't involve any data augmentation methods. We report the result for both FOA and MIC data format if possible.

Within SoundDet, we want to answer four main questions: 1. If our proposed MaxCorr band-pass filters is capable of learning useful representation for sound event spatio-temporal information recovery? 2. Event length impact on SoundDet and other methods. 3. Efficiency comparison in terms of parameters and inference time. 4. Quantitative comparison between SoundDet and other methods. To this end, we test four SoundDet variants: backbone only SoundDet with MaxCorr filter (SoundDet_backbone) and without MaxCorr filter (SoundDet_nomaxcorr). SoundDet without motion smoothness (SoundDet_nomots). Moreover, in addition to report the overall metric, we further divide the events into Small (0-2s), Medium (2s-7s) and Large (>7s), three categories and report result for them separately.

Please note that the network might need to be slightly changed to process waveform with different input length, sample rate and label interval length. H and W indicate dense proposal map height and width, respectively, in our experiment $H = 60$ and $W = 60$, they can be directly modified to fit other input waveform temporal length or ground truth labelling resolution.

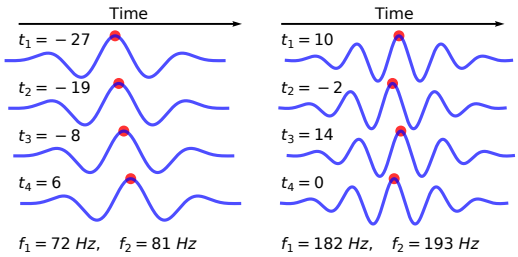


Figure 3.5: Two learned MaxCorr filters, they are controlled by different frequency cutoffs and four time shifts. The red dots are the filter symmetric point.

We adopt multi-stage training strategy. First, we train the SoundDet backbone network in a framewise DoA regression and multilabel classification manner, like SELDNet [2] and EIN [20] do. Training SoundDet backbone neural network at first provides two advantages: 1) it helps to test if our proposed MaxCorr filter bank

Table 3.2: Event-based evaluation result, model parameter number and input feature. We report mAP/mAR under different event temporal length threshold. The "Input" column labels are: 0. Raw waveform, 1. Log-Mel, 2. GCC-PHAT, 3. Intensity.

Methods	Params	Input	mAP(↑)				mAR(↑)			
			Ovall	Sma	Med	Larg	Ovall	Sma	Med	Large
SELDNet(foa)	0.5M	1,3	0.087	0.038	0.092	0.157	0.152	0.079	0.096	0.097
SELDNet(mic)	0.5M	1,2	0.079	0.035	0.081	0.143	0.140	0.067	0.099	0.086
EIN [20]	26.0M	1,2	0.134	0.088	0.187	0.183	0.256	0.175	0.186	0.257
SD_bb(foa)	10.0 M	0	0.040	0.025	0.055	0.090	0.096	0.066	0.051	0.033
SD_bb(mic)	10.0 M	0	0.035	0.021	0.091	0.090	0.094	0.065	0.049	0.027
SD_nomcorr(foa)	10.0 M	0	0.025	0.015	0.034	0.080	0.063	0.040	0.032	0.014
SD_nomcorr(mic)	10.0 M	0	0.017	0.059	0.016	0.044	0.041	0.017	0.079	0.016
SD_nomots(foa)	13.0 M	0	0.117	0.062	0.173	0.152	0.247	0.174	0.159	0.232
SoundDet(foa)	13.0 M	0	**0.197**	**0.098**	**0.201**	**0.216**	**0.294**	**0.189**	**0.223**	**0.289**

helps to learn essential representation for sound event temporal detection and spatial localization (as we compare in the experiment), 2) it guarantees to learn a reasonable framewise representation first so that the later joint training (combine the backbone and two heads) converges much faster. For the backbone training, we use SGD optimizer with an initial learning rate 0.5, the learning rate decays every 30 epochs with decay rate 0.7. With the pre-trained backbone network, we continue to train the whole SoundDet network with the same optimizer.

The segment-based evaluation result is shown in Table 3.1. We can see that, by involving MaxCorr filter banks, our backbone-only SoundDet achieves comparable performance comparing with SELDNet [2]. SELDNet extracts Log-mel, GCC-PHAT and Intensity features (see Table 3.2) as its neural network input, replacing these hand-crafted features with learnable MaxCorr filter bank helps to achieve similar performance, even with 1D convolution. If disabling SoundDet to learn between-channel correlation (SoundDet_nomaxcorr, no maximum correction), we can witness a huge performance drop in both FOA and MIC format, especially the high event detection error rate and large DoA error angle. This shows that our proposed MaxCorr filter bank is capable of learning essential representations for recovering temporal and spatial information. Actually, we observed a fast convergence of the between-channel time shift parameters (see Eqn. 3.3), SoundDet quickly updates time shift parameters in the very first couple of training iterations and swiftly achieves a relatively stable state. Two learned MaxCorr filters are shown in Fig. 3.5, from which we can observe that different MaxCorr filters have learned different time shifts and frequency cutoffs. These learned frequency-selective and phase-sensitive MaxCorr filter bank is of vital

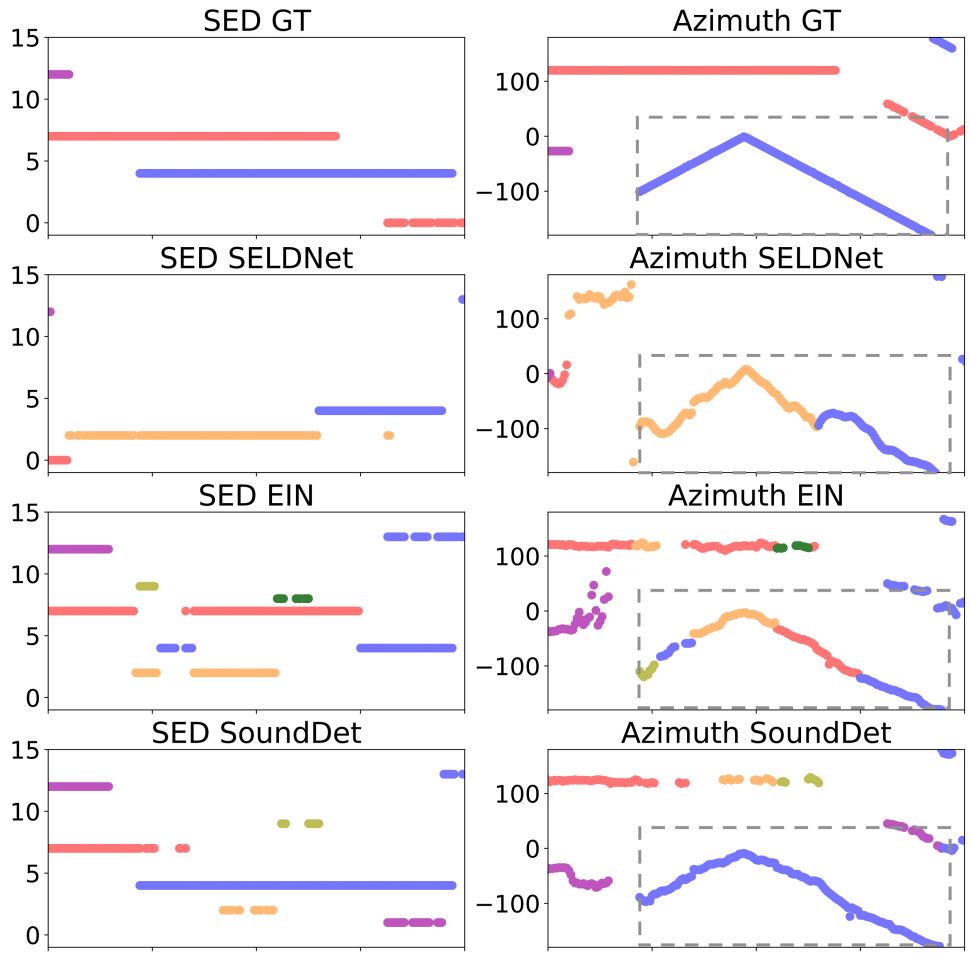


Figure 3.6: Qualitative comparison. We visualize detected sound events' temporal location (SED, left side) and Azimuth-only DoA (right side). While framewise based methods (SELDNet and EIN) produce discrete and segmented event, SoundDet maximally keeps an event's completeness and continuity. Pay attention to dotted grey box and different color indicates different sound event class.

importance for sound event representation learning.

Both SoundDet and SELDNet work better for FOA format than MIC format. Framewise based learning (both SoundDet, SELDNet and EIN) shows no obvious difference in estimating events with different temporal length, which is reasonable because of their framewise processing property. The whole SoundDet (combine backbone and two heads) far outperforms SELDNet baseline and achieves comparable performance with EIN with much less parameter number. Specifically, SoundDet shows advantage on longer sound event estimation. It thus attests the necessity of directly modelling "sound-object" when it comes to events with longer temporal length or more complex motion trajectory. Excluding motion smoothness largely reduces the

Table 3.3: Inference time on Intel(R) Core(TM) i9-7920X CPU. The waveform pre-processing time is contained for SELDNet and EIN.

SELDNet	EIN	SoundDet
1.20 s	2.20 s	1.25 s

performance, we observed involving motion smoothness map greatly helps the neural network to accurate localize sound events.

The event-based evaluation result is shown in Table 3.2, from which we can observe that SoundDet outperforms all existing methods in both the mAP and mAR metric. This shows that the segment-based metrics in current use do not comprehensively reflect an algorithm's performance. Rather than focusing on a local segment evaluation, the event-based metric highlights an event's completeness and continuity and SoundDet is naturally designed to meet these requirements. This advantage is echoed by the qualitative comparison in Fig. 3.6. We can clearly see that SELDNet and EIN inevitably contain many mixed sound events and frequently cut complete events into small disconnected segments. The situation becomes more serious when faced with overlapping situation. However, SoundDet better avoids this dilemma because by design it treats events as complete instances, rather than discrete segments.

We further report the average inference time of different methods to process a one-minute long audio in Table 3.3, showing that it is almost twice as fast as EIN, and comparable to SELDNet. In summary, SoundDet is capable of learning a sound event's spatial, temporal and class information from raw waveforms, achieving competing performance under existing segment-based metrics and leading results on the proposed event-based metrics.

3.2.5 Conclusion and Discussion

We have introduced a number of innovations in this chapter that attempt to unify the currently disparate fields of object detection in computer vision, with event detection and localization in audio settings. We also abandon prior approaches that resort to hand-crafted pre-processing and transformation of the multi-channel waveforms, and instead directly consume raw data. Experiment on DoA based SELD task verifies that: 1. we can design learnable filters to jointly encode per-channel time-frequency representation and inter-channel time-delay information in a unified manner. 2. Learning from audio raw waveform with appropriately designed filters in a data-driven way can lead to superior performance in DoA based SELD task.

3.3 SoundSynp: Learnable MultiScale Filterbank

3.3.1 Introduction

To detect sound sources, we often deploy a spatially-configured microphone array to record an acoustic environment. The recorded sound is a highly compressed 1D time series. Since different sound sources have different frequency properties, it is essential to convert 1D waveform into 2D time-frequency representation. This is often achieved by projecting the raw waveform onto various frequency bases. A sound source's spatial location clue lies in inter-channel difference among waveforms (*e.g.*, phase difference). It is essential to design a neural network that jointly encodes mono-channel time-frequency and inter-channel phase difference from the raw waveforms in a unified, parameter-frugal and computation-efficient manner. The learned representation should have expressive resolution in both time, frequency and space domains so that sound sources can be precisely detected.

However, learning such representation is a tough task. Challenges stem from both theoretical side and practical side. According to the Uncertainty Principle, we cannot achieve the optimal resolution in time and frequency domain at the same time, but instead keep a trade-off between them. Traditional hand-engineered sound feature [37, 20, 12] and some recently proposed learnable filter bank [143] empirically set the same length for all filters, resulting in human-biased, unadjustable time-frequency resolution map. Some other works [194, 68, 65] correlate filter frequency response and filter length by initializing in mel-scale, but it is neither scalable nor stable. Moreover, all of them process raw waveform with one-scale filter bank, we think such one-scale filter bank easily leads to non-optimal representation, especially when sound sources have spectrum overlap or undergo free spatial motion.

In this chapter, we first give theoretical analysis on the filter bank impact on time-frequency representation. Based on the analysis, we propose a simple yet effective synperiodic filter banks construction strategy, in which synperiodic means each filter's temporal length and its carried frequency response are synchronized by rotating periodicity such that each filter's length is inversely proportional to its frequency resolution. The resulting synperiodic filter banks (we call it filter banks as it contains multiple filter bank groups) thus internally maintain a better time-frequency resolution trade-off than traditional fixed-length filter bank. Coupling the filter length with its frequency response helps us to reduce human intervention in filter bank design. By simply alternating the periodicity term, we further construct a group of synperiodic filter banks, with which we achieve multi-scale perception in time domain. At the

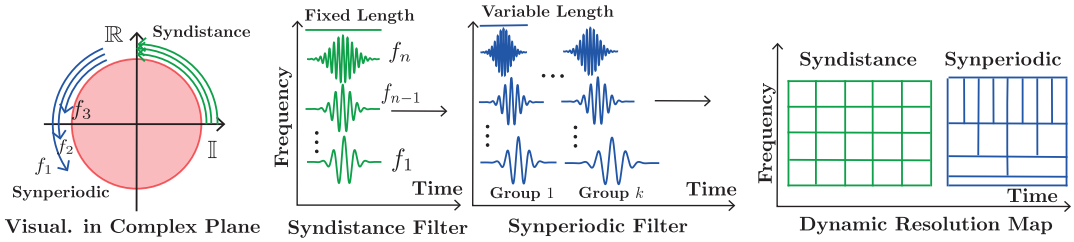


Figure 3.7: Synperiodic filter banks illustration: Syndistance filter bank (green color) rotates the same distance in complex plane and thus has the same kernel length, regardless of frequency it carries. Its time-frequency dynamic resolution map is thus rectangular. our proposed synperiodic filter banks (blue color) are generated by rotating the same periodicity number. So filter carrying lower frequency has larger kernel size than those with higher frequency response. As a result, synperiodic filter banks' time-frequency dynamic resolution map achieves better trade-off than traditional syndistance filter bank such as STFT and LogMel filter bank.

same time, by applying a synperiodic filter banks to process one raw waveform as well as its consecutively-downsampled versions, we achieve multi-scale perception in frequency domain. The multi-scale perception in both time and frequency domain of synperiodic filter banks enables the neural network to dynamically learn better representation for sound source detection in a data-driven way. It is worth noting that synperiodic filter banks parameter number is just linear to filter number (adds up to less than 1% of the whole parameters) and it can be efficiently implemented as a 1D convolution operator.

Following the learnable synperiodic filter banks, we further design backbone network (a small lite one and a large one) with two paralleling branches with layerwise soft-parameter sharing to learn sound source's semantic and spatial location related representation jointly. Experiment on both direction-of-arrival (DoA) task and physical location estimation task shows that our proposed framework outperforms comparing methods significantly. Replacing existing method's head with our proposed synperiodic filter banks also improves the performance significantly.

3.3.2 Background Knowledge Discussion

Sound source detection task aims at detecting each sound source's start/end time, the semantic identity and spatial location during its occurrence. Semantic identity cues mainly lie in mono-channel sound waveform time-frequency representation, and spatial location cues lie in inter-channel waveform difference (e.g. phase difference). To get the time-frequency representation for each mono-channel waveform, a frequency-

selective filter bank $\mathcal{F}$ is often used to project the waveform onto different frequency bases. A general filter bank $\boldsymbol{\mathcal{F}}$ of M filters can be mathematically represented as,

$$\boldsymbol{\mathcal{F}} = \{\mathcal{F}^i\}_{i=1}^{M},\ \mathcal{F}^i_{f_i,\sigma_i}(t) = \phi_{f_i}(t) \cdot \omega_{\sigma_i}(t) \tag{3.6}$$

Each filter $\mathcal{F}^i$ is a filter in time domain. It contains a frequency response f_i and filter length σ_i, that are independently controlled by frequency-selective filter $\phi_{f_i}(t)$ (e.g. [143]) and a window function $\omega_{\sigma_i}(t)$. An expressive filter bank should have good resolution capability in both time and frequency domains. Frequency resolution indicates the ability of discerning two adjacent frequency bins, time resolution corresponds to the capability of precisely localizing a sound source in time domain. According to the Uncertainty Principle, the frequency resolution Δ_f and time resolution Δ_t satisfies $\Delta_f \cdot \Delta_t \geq C$ (C is a constant, which means we cannot achieve the optimal resolution at time and frequency domains simultaneously, but instead maintain a trade-off between them. Therefore, it is essential to design a filter bank that maximally maintains a good time-frequency resolution.

Existing filter bank differs in their way of choosing $\phi_{f_i}(t)$ and $\omega_{\sigma_i}(t)$. Classic Fourier transform based filter banks, such as short-time Fourier transform (STFT), LogMel and MFCC [37], decide $\phi_{f_i}(t)$ and $\omega_{\sigma_i}(t)$ independently in arbitrary manner. They usually assign a fixed window length to all filters (where $\omega_{\sigma_i}(t)$ is a constant), so their extracted time-frequency representation resolution is fixed and unadjustable across all frequency bins. We call such filter bank **syndistance** filter bank to emphasize their equal length property across all frequency responses. Wavelet transform [109] inversely correlates window length with frequency response so that the filter with higher frequency response is naturally associated with shorter window. The resulting time-frequency map theoretically has better resolution than the one extracted by Fourier transform, but it is still fixed and heavily rely on empirical parameter tuning. Some recent work [68, 194] relax $\phi_{f_i}(t)$ to be trainable so that they can be further optimized in a data-driven way. They still involve much empirical parameter-tuning work (*e.g.*, $\omega_{\sigma_i}(t)$ selection). Moreover, they all process the raw waveform in one-scale manner, which often leads to non-optimal time-frequency representation.

Synperiodic filter banks address these issues from three perspectives: **First**, we inversely correlate $\phi_{f_i}(t)$ and $\omega_{\sigma_i}(t)$ by a *periodicity* term. Therefore, we do not have to explicitly set the window length ($\omega_{\sigma_i}(t)$) for each filter because it is internally decided by the filter's frequency response. In addition, inversely correlating $\phi_{f_i}(t)$ and $\omega_{\sigma_i}(t)$ naturally generates filter bank in which filters with high frequency responses

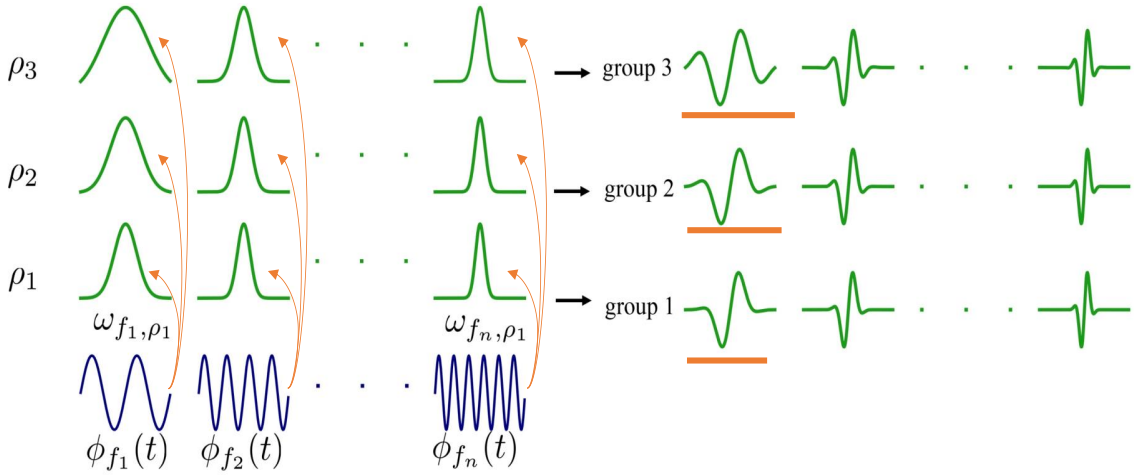


Figure 3.8: Synperiodic filter construction. Given a set of filters of infinite length $\phi_f(t)$ (dark blue) and predefined periodicity parameters $[\rho_1, \rho_2, \rho_3]$, windowing function $\omega_{f,\rho}$ takes the frequency f and periodicity ρ as input to create windows with locality property (green), and the window's active region length is inversely proportional to its corresponding filter's frequency response f. Multiplying the infinite filter and its associated window results in a group of Synperiodic filter bank. The filters of the same frequency response in different groups have different time length, helping to achieve multi-scale perception in time-scale.

are associated with shorter window length. Such filter bank naturally maintains a better time-frequency resolution. **Second**, by simply alternating the *periodicity* term, we create a group of filter banks that differ in their window length. Applying these filter banks to process the raw waveform helps achieve multi-scale perception in time domain. **Third**, applying the same filter banks to process recursively 2x downsampled waveforms helps achieve multi-scale perception in frequency domain.

3.3.3 Synperiodic Filter Banks Construction

In synperiodic filter banks, we inversely correlate ϕ_{f_i} and ω_{σ_i} by setting σ_i to be proportional to the filter's periodic term: rotating ρ periodic circles. Specifically, synperiodic filter banks $\boldsymbol{\mathcal{F}}_{synp} = \{\mathcal{F}_i^{synp}\}_{i=1}^{M}$ can be represented as,

$$\mathcal{F}_{i,f_i}^{synp}(t) = \phi_{f_i}(t) \cdot \omega_{f_i,\rho}(t) \tag{3.7}$$

where $\omega(f_i, \rho)$ indicates the periodic number the filter rotates. Since filters carrying higher-frequencies have shorter period length, requiring all filters to rotate the same period number naturally results in shorter filter length for high-frequency filters and wider filter length for low-frequency filters. Therefore, $\omega(f_i, \rho)$ defines the filter

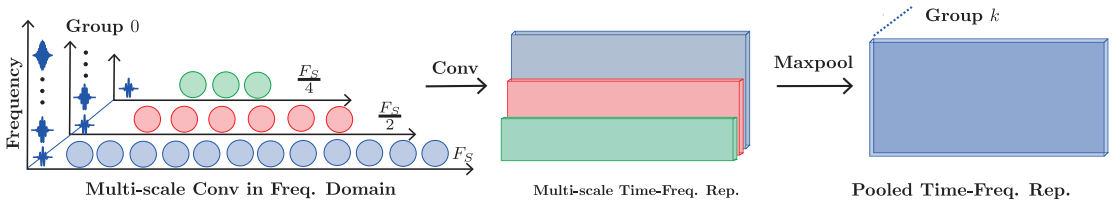


Figure 3.9: Multi-scale learning in frequency domain. Given the raw one channel sound waveform and pre-constructed synperiodic filter banks, we consecutively downsample the waveform by factor 2x, the newly downsampled waveform is processed by low-half filter bank from the proceeding filter bank. We can obtain time-frequency representation on each frequency scale. These time-frequency representations share the same time length by adjusting step size. The final time-frequency representation is obtained by max-pooling them together.

window length by constraining the period number it rotates. We call our filter banks synperioidc filter banks to emphasize that all filters' window lengths are automatically decided by the period number they rotate. To better understand the difference between syndistance and synperiodic filter banks, we visualize them in complex-valued plane (see Fig. 3.7, left-most), in which a filter is a complex exponential rotating in the complex plane counter-clockwisely, the rotating speed corresponds to the frequency it carries. In the complex-valued plane, all syndistance filters rotate to the same distance. Synperiodic filters, however, rotate to a predefined periodicity ρ, resulting in narrow window for high-frequency filters and wide window for low-frequency filters.

Synperiodic filter banks lend us three advantages: it **first** avoids us setting window length for each filter separately, which is empirical and heuristic; **second**, the constructed filter banks internally maintain a good time-frequency resolution trade-off; **third**, by simply varying the periodicity term ρ, we can easily obtain a group of synperiodic filter banks to process the raw waveforms in multi-scale manner. Our synperiodic construction strategy shares similar idea with Wavelet transform [109] where it adopts a time shift and "squeezing ratio" to achieve multi-scale perception. The difference is that we omit the time shift but instantiate the squeezing ratio with our proposed synperiodicity strategy. Moreover, synperiodic filter banks are multi-scale both time and frequency domain, and self-adjustable in a data-driven way.

There are many ways to instantiate $\omega(w_i, \rho)$, as long as we guarantee the window length gradually reduces as the frequency response increases. The simplest choice is to treat $\omega(w_i, \rho)$ as a constant, but we find it either results in too wide window for low-frequency filters or too narrow window for high-frequency filters. To mitigate this dilemma, we use logarithmic window function,

$$\omega(f_i, \rho) = 27 \cdot \log_{10}(f_i) - \rho, \ \ \rho = \{-6, -11, -16\} \tag{3.8}$$

We set ρ as $[-6, -11, -16]$ respectively to construct three synperiodic filter banks. The design of this window function is motivated by mel-scale frequency initialization strategy. By roughly setting a filter's bank width to be equal to the distance between its preceding and next frequency location in frequency domain, converting to time domain we can roughly get a logarithmic scale frequency-periodicity relationship.

Figure 3.8 visualizes synperiodic filter banks construction. In our implementation, synperiodic filter is created by multiplying a sinusoidal basis ($\phi(f_i)$ instantiation) with learnable frequency response initialized in mel-scale by a Gaussian window ($\omega(f_i, \rho)$ instantiation) with learnable width initialized through the windowing function by Eqn. (3.8). The initial synperiodic filter extracted features in different channels are in a complex format, we further encode cross-spectrum feature as spatial location relevant feature (see Sec. 1 in supp. material) and concatenate them with synperiodic filter extracted together as the overall sound source feature. Synperiodic filter banks are initialized with independent learnable frequencies and window length, they are independently updated during training stage.

For multi-scale perception in time domain, we use the previously constructed synperiodic filter banks to convolve with mono-channel sound waveform with the same step size and padding strategy, resulting in the same size output for each single synperiodic filter banks. Since different filter group has different window length, we achieve multi-scale learning in time domain (see the third figure in Fig. 3.7). It maximally avoids us empirically selecting one window scale ρ, but instead uses a group of filter banks to enforce the neural network to strike a better time-frequency resolution trade-off in a data-driven way.

Multi-scale perception in frequency domain is hierarchical: given a raw sound waveform with sampling frequency F_S, synperiodic filter banks' frequency is initialized within the range $[0, \frac{F_S}{2}]$ under Nyquist sampling theorem. If we downsample the sound waveform by a factor of 2, the resulting waveform can be processed by the lower-half filters in each group whose frequency response lies in $[0, \frac{F_S}{4}]$. Please note that merely using the lower-half filters to process the 2x-downsampled waveform helps us to avoid aliasing issue. This process that 2x-downsampling the waveform further process the downsampled waveform with filters with lower-half frequency response can potentially iterate a couple of times (in our case three times), resulting in multi-scale perception in frequency domain. Figure 3.9 illustrates how it works. Multi-scale learning in frequency domain brings us two extra benefits: 1) from data augmentation

perspective, 2x downsampling a waveform creates new low-quality waveform, equivalently we have extra dataset. 2) from the perception field perspective, the adjacent 2x-downsampling strategy leads to dilated convolution [43] for lower-frequency filters, because applying a filter to convolve with a downsampled waveform equals to convolve on the original waveform with dilated convolution (skip-2 connection). The resulting wider or dilated perception field for lower frequency filters enables to learn better sound representation along the time axis. In sum, by using learnable synperiodic filter bank group to process the raw waveforms in multi-scale manner, we achieve a dynamic time-frequency resolution that naturally maintains a better time-frequency resolution fitting for sound source detection in a data-driven way.

Synperiodic filter bank group introduces very few parameters (less than 1%) because they are parameterized filters. The trainable parameter number increases linearly w.r.t. synperiodic filter bank number. Their convolution with raw waveforms can also be efficiently implemented with 1D convolution.

Following the Synperiodic filter banks, we further design a backbone network to further learn sound source representation. Jointly learning sound source semantic label and spatial location representation is a multi-task problem [83, 116], we follow [20] to propose a backbone with two paralleling and identical branches to learn each sub-task separately. To enforce information communication, we add a layerwise information exchange module: for the intermediate semantic label feature f_s^i and spatial location feature f_g^i learned by the i-th block, a learnable weight W_i is introduced to linearly combine them together to get an updated f_s^i and f_g^i before feeding them to the next layer, $[f_s^i, f_g^i] = W_i \cdot [f_s^i, f_g^i]$. On top of the representation, we add trackwise permutation-invariant training (PIT) strategy to train the whole neural network in an end-to-end manner. We have designed two backbone versions: a lite version of $24M$ parameters and a large version of $60M$ parameters.

3.3.4 Experiments

We focus on two tasks: direction of arrival (DoA) and physical location estimation. For DoA estimation task, we use DCASE2020 sound event detection and localization dataset [134], in which the sampling rate is 24 k. It contains 14 sound sources with azimuth range $[-180°, 180°]$ and elevation range $[-45°, 45°]$. Two recording formats: FOA and MIC-array. More details are in [134]. For physical location, we use L3DAS22-SELD dataset [58].

For DoA, we compare with five latest methods:

Table 3.4: Result on DoA task. For Segment-based eval., we report detection error ER_{20°, F-measure F_{20° under DoA threshold 20°, and classification dependent localization error LE_{CD} and localization recall LR_{CD}. For event-based eval., we report mAP/mAR. The "Input" column labels are: 0. Raw waveforms, 1. Log-Mel, 2. GCC-Phat, 3. Intensity Vector. Top three performances are respectively highlighted by **red**, **green**, and **blue** color. Specifically, $ER_{20^\circ}(\downarrow)$, $F_{20^\circ}(\uparrow)$, LE(↓), LR(↑), mAP(↑), mAR(↑).

Methods	Input	Segment-Based Eval.				Event-based Eval.	
		ER_{20°	F_{20°	LE	LR	mAP	mAR
SELDNet(foa) [2]	1,3	0.63	0.46	23.1	0.69	0.087	0.152
SELDNet(mic) [2]	1,2	0.66	0.43	24.2	0.66	0.079	0.140
EIN-v2(foa) [20]	1,2	0.30	0.77	8.9	0.84	0.134	0.256
SoundDet(foa) [68]	0	0.25	0.81	8.3	0.82	0.197	0.294
SoundDoA(foa) [65]	0	0.23	0.85	7.9	0.87	0.220	0.301
UTSC-Iflytek(foa+mic) [183]	1,2,3	0.20	0.85	6.0	0.89	-	-
SoundSynp_lite(mic)	0	0.21	0.83	6.2	0.87	0.199	0.303
SoundSynp_large(mic)	0	0.19	0.86	5.5	0.91	0.210	0.313
SoundSynp_lite(foa)	0	0.20	0.85	5.6	0.89	0.205	0.309
SoundSynp_large(foa)	0	**0.15**	**0.89**	**4.3**	**0.94**	**0.232**	**0.327**

1. **SELDNet** [2], SELDNet is the baseline model and it jointly trains sound source's semantic label and spatial location with a convolutional recurrent neural network (CRNN) [33].

2. **EIN-v2** [20], EIN-v2 [20] is an improved version of EIN [21]. It adopts multi-heads self-attention [179] (MHSA) to model temporal dependency and track-wise permutation-invariant training to train the model.

3. **SoundDet** [68], SoundDet directly learns from raw waveforms with MaxCorr kernels, followed by an encoder-decoder neural network to learn frame-wise representation.

4. **SoundDoA** [65] SoundDoA also learns from raw waveform by a Gabor-like filter bank with an *Enhance* module. A backbone neural network is associated with the Gabor-like filter bank to learn time-frequency representation.

5. **Utsc-Iflytek** [183]. Utsc-Iflytek is ranked first in DECASE2020 challenge leaderboard [2], it combines MIC and FOA features and ensembles different models like ResNet [61] and Xception [31] to detect sound sources.

For sound source physical location estimation task, we additionally compare with Conf-EIN [76], which is based on EIN-v2 [20] and additionally contains conformer and dense blocks. We call our framework SoundSynp (the one with small backbone

[2] see link for leaderboard report.

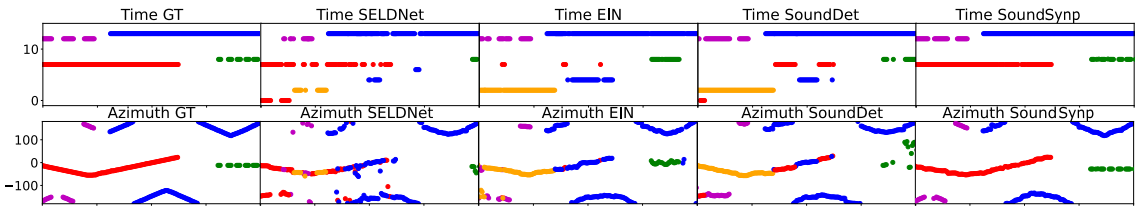


Figure 3.10: **DoA result visualization**. We show detected sound source temporal location (top row) and azimuth (bottom row). The horizontal axis is time, the vertical axis is semantic label (top) and azimuth in degree(bottom). Different color indicates different sound source class.

SoundSynp_lite, large backbone SoundSynp_large). All methods' comparison is in Table 3.7.

To train the neural network, we clip all 4s short snippets from the original one minute long four-channel raw waveforms so that we have the largest training dataset. The raw waveforms are normalized to $[-1, 1]$. We adopt Adam optimizer [85] with an initial learning rate 0.0002 in the first 100 epochs and 0.0001 in the following 50 epochs. The loss combination weight between classification head and regression head is $1 : 2$. During training, data augmentation method SpecAugment [128] is applied. For DoA task, we regress direction of arrival angle in Cartesian coordinates $[x, y, z]$. In synperiodic filter banks, the filter length is 1025, each group's filter number is 256 and the step size is 600. Particularly, we have observed the initialized learnable synperiodic filter banks update its parameters intensively during the early several epochs, and then gradually becomes stable. We train each model with Pytorch [129] five times independently and report the average score. The standard deviation is within 0.04 (for recall) and 0.17° for angle, 0.002 for mAP and mAR, we do not report the standard deviation in tables for succinct report.

We use two metrics. **Segment-based** metric is a widely adopted evaluation metric [2, 20, 65, 68], it couples semantic label and spatial location together: a semantic-correctly detected sound source needs to be spatially close enough to its ground truth location in order to be regarded as a true positive detection. **Event-based** based metric is newly proposed by [68] to comprehensively evaluate under different confidence scores. Like object detection from images [100], it computes mean average precision (mAP) and mean average recall (mAR).

The result is given in Table 3.4, from which we see that SoundSynp (both the lite and large versions) achieves the best performance over all comparing methods significantly. Both EIN-v2 and UTSC-Iflytek use pre-extracted hand-engineered sound features, such as Logmel, GCC-Phat and Intensity Vector. SELDNet [2] uses phase

Table 3.5: Existing Methods with Synperiodic Frontend.

Method	ER ↓	F ↑	LE↓	LR↑
SELDNet [2]	0.63	0.46	23.1	0.69
SELDNet_Synp	**0.50**	**0.53**	**21.0**	**0.78**
EIN-v2 [20]	0.30	0.77	8.9	0.84
EIN-v2_Synp	**0.22**	**0.84**	**6.7**	**0.89**
SoundDet [68]	0.25	0.81	8.3	0.82
SoundDet_Synp	**0.21**	**0.85**	**7.5**	**0.87**
SoundDoA [65]	0.23	0.85	7.9	0.87
SoundDoA_Synp	**0.21**	**0.87**	**7.2**	**0.90**

Table 3.6: Various SoundSynp Variants Results.

Variants	ER ↓	F ↑	LE↓	LR↑
SSynp_MSFreq	0.20	0.84	7.3	0.86
SSynp_MSTime	0.23	0.83	7.0	0.88
SSynp_Linear	0.22	0.83	8.4	0.84
SSynp_SScale	0.25	0.81	7.8	0.84
SSynp_Sinc	0.21	0.83	6.3	0.87
SSynp_LEAF	0.22	0.84	5.7	0.86
SoundSynp	**0.15**	**0.89**	**4.3**	**0.94**

Table 3.7: Network Architecture Highlight. MHSA: multi-head self-attention.

Variants	Network Blocks
SELDNet [2]	Conv2D, biGRU
EIN-v2 [20]	Conv2D, MHSA
SoundDet [68]	Conv1D LSTM
SoundDoA [65]	Conv1D/2D MHSA
SoundSynp	Conv1/2D MHSA

Table 3.8: Replace Synperiodic Learnable Frontend with traditional TF feature.

Method	ER ↓	F ↑	LE↓	LR↑
SSynp_MFCC	0.22	0.81	6.7	0.86
SSynp_LogMel	0.23	0.84	6.6	0.86
SSynp_STFT	0.23	0.82	7.0	0.84
SSynp_Gabor	0.22	0.82	7.3	0.85
SoundSynp	**0.15**	**0.89**	**4.3**	**0.94**

and spectrum. SoundDet [68], SoundDoA [65] and SoundSynp are the only three methods that directly learn from raw waveforms. At the same time, SoundSynp obtains better performance on FOA than MIC format, the same phenomena has been observed by all other methods. It thus shows FOA better fits for sound source detection than MIC format. It is worth noting that Utsc-Iflytek [183] ensembles different powerful image-based 2D models to detect sound sources. However, our proposed SoundSynp still outperforms Utsc-Iflytek by a large margin. SoundSynp_lite achieves comparable performance with Utsc-Iflytek with much smaller parameter size (24M). It thus shows our proposed synperiodic filter banks are capable of learning expressive representation for sound source detection. We do not report mAP/mAR value for Utsc-Iflytek because it is a complex system and no detail about their system is available.

We show one learned synperiodic filter in Fig. 3.11, in which the filter's temporal support region and frequency response is updated to achieve better time-frequency resolution.

We further run four ablation studies.

1. Existing Methods with Synperiodic Frontend. We replace either fixed TF feature extractor front-end of SELDNet, EIN-v2, or learnable TF extractor front-end of SoundDet [68] and SoundDoA [65] with our proposed Synperiodic filter banks front-end to test their performance. It removes the influence of the backbone neural network of different models and thus helps to get direct comparison of synperiodic filter banks with other fixed time-frequency features. The result is in Table 3.5, from which we can see using synperiodic filter banks as a replacement of existing filter bank dramatically improves the performance. Synperiodic filter banks can be used as a general plug-and-play front-end by existing methods.

2. Replace Synperiodic with Classic TF Feature. In SoundSynp, we replace the synperiodic filter bank group with MFCC [37] and Log-Mel (used by SELDNet [2] and EIN-v2 [20]), STFT and Wavelet [109] like filters (we use the typical Gabor filter bank, we call SSynp_Gabor). It helps us to understand the performance with/without synperiodic filter banks. The result is given in Table 3.8, from which we can see replacing SoundSynp's synperiodic filter banks with classic hand-engineered TF features inevitably reduces the performance under all evaluation metrics. It thus shows learning from either fixed or single-scale TF feature leads to worse performance than our proposed multi-scale synperiodic filter banks on sound source detection task.

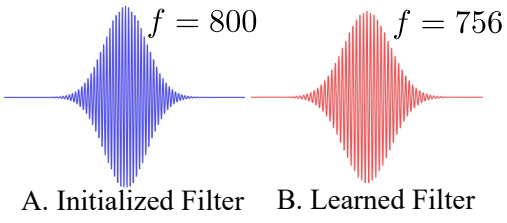


Figure 3.11: Learned Synperiodic Filterbank Visualization.

3. Necessity of Each SoundSynp Component. We internally test six synperiodic filter banks variants: (1) synperiodic filter banks with just multi-scale in frequency domain (SSynp_MSFreq); (2) just multi-scale in time domain (SSynp_MSTime); (3) Synperiodic filter banks with frequency responses linearly initialized in Nyquist frequency range (SSynp_Linear, compare with our mel-scale initialization); (4) just one synperiodic filter bank without multi-scale perception neither in time nor frequency domain (SSynp_SScale); (5) with rectangular band-pass frequency response initialization (SSynp_Sinc), like SincNet [143] does; (6) with LEAF [194] learnable filter bank (SSynp_LEAF). The result is given in Table 3.6. We can observe that the absence of multi-scale perception in either frequency domain or time domain inevitably reduces the performance. We find semantic label detection suffers more in single-scale perception in time domain than in frequency domain (see better performance

on ER_{20°, and F_{20° score), which shows frequency domain multi-scale perception is vital for semantic estimation. Similarly, we can observe that multi-scale perception in time domain is vital for sound source spatial location estimation (see better performance on LE and LR score). Linearly initialized filter bank frequency reduces the performance as well, which shows assigning more filters to the lower frequency range is important. But this conclusion might be data-biased because we find DCASE dataset contains many low-frequency sound events like burning fire. Moreover, reducing the synperiodic filter bank groups to one group with just single-scale perception leads to the worst performance, it thus shows multi-scale perception in both time and frequency domain is essential for DoA-based sound source detection. Moreover, SoundSynp_Sinc leads to slightly inferior performance than our used mel-scale initialization strategy, it attests synperiodic filter framework is general enough to be adopted by various initialization strategy.

One qualitative comparison is shown in Fig. 3.10. We can clearly see that SELDNet [2] generates mixed prediction at different time steps and DoA locations. SoundDet [68] and EIN-v2 [20] give non-existing sound sources (orange color). When multiple sound sources happen at the same time (polyphonicity), SoundDet and EIN are easily failed to predict the right spatial location (discretized blue and red color). Our method (SoundSynp_large) predicts more spatially and temporally consistent sound sources by maximally keeping sound source's continuity and completeness.

For physical location estimation, we run experiment on L3DAS22-SELD dataset [58], whose target is to predict sound source's 3D physical location $[x, y, z]$ in indoor room environment. The room is of size $6m \times 5m \times 3m$. It involves 14 seed sounds from FSD50K [46]. Up to 3 sources are co-emitting sound. The dataset contains 600/150/150 30s-clips for train/val/test. We report the result on validation set because the test set is held-out for the challenge and not publicly available. The evaluation metric used here is F-score [111, 58]. In addition to EIN-v2 [20], SoundDet [68] and SoundDoA [65] (SELDNet model does not converge in training), we also compare with the champion method Conf-EIN [76], it ranks the first in L3DAS22-SELD challenge. The result is in Table 3.9, we can see SoundSynp_large outperforms Conf-EIN [76] by a large margin with smaller parameter size and inference time (see Table 3.10). SoundSynp_lite achieves comparable performance with Conf-EIN [76]. It thus shows SoundSynp framework is capable of accurately detecting sound sources 3D physical location.

The inference time (Intel(R) Core(TM) i9-7920X CPU, 100 independent tests, report the average time) and model parameters of all methods are given in Table 3.10,

Table 3.9: Physical Location Detection Result.

Method	$F_{\leq 1m} \uparrow$	$F_{\leq 2m} \uparrow$
EIN-v2 [20]	0.621	0.636
SoundDet [68]	0.640	0.672
SoundDoA [65]	0.652	0.688
Conf-EIN [76]	0.685	0.715
SoundSynp_lite	0.644	0.681
SoundSynp_large	**0.722**	**0.733**

Table 3.10: Inference Time and Param. Size (M: million).

Method	Infer.	Param.
SELDNet [2]	1.20 s	0.5 M
EIN-v2 [20]	2.20 s	26 M
SoundDet [68]	1.25 s	13 M
SoundDoA [65]	2.10 s	27 M
Conf-EIN [76]	4.0 s	83 M
SoundSynp_lite	1.80 s	24 M
SoundSynp_large	3.10 s	60 M

from which we can see that SoundSynp_lite has comparable parameters and smaller inference time than EIN-v2 [20]. SoundSynp_large has fewer parameters and less inference time than Conf-EIN [76], and it outperforms Conf-EIN [76] in physical location based sound source detection.

3.3.5 Conclusion and Discussion

From the experiment on both DoA-based and physical position based SELD task, we show uncertainty principle inspired multiscale in both time-domain and frequency-domain learnable filter bank is capable of learning representative time-frequency representation for SELD task. Given the huge diversity of sound events in terms of event class and spatial position, learning a multi-scale time-frequency representation can help grasp granular time and frequency variability resulting from the huge diversity in sound events' spatial positions.

3.4 SoundCount: Learnable Dyadic Filter Bank

3.4.1 Introduction

Suppose you are in a party where lots of people are talking simultaneously. An interesting question that naturally arises is whether you can tell the number of speaking person around you merely from the sound you heard? Although a trivial example, this sound "crowd counting" problem has a number of important applications. For example, passive acoustic monitoring is widely used to record sounds in natural habitats, which provides measures of ecosystem diversity and density [5, 42, 32]. Sound counting helps to quantify and map sound pollution by counting the number of individual polluting events [7]. It can also be used in music content analysis [78], speech source count [29]. Despite its importance, research on sound counting has far lagged behind than its well-established crowd counting counterparts from either images [196, 181], video [98] or joint audio-visual [75].

We conjecture the lack of exploration stems from three main factors. First, sound counting has long been treated as an over-solved problem by sound event detection (SED) methods [113, 18, 2, 68], in which SED goes further to identify each sound event's (*e.g.* a bird call) start time, end time and semantic identity. Sound counting number then becomes easily accessible by simply adding up all detected events. Secondly, current SED only tags whether a class of sound event is present within a window, regardless of the number of concurrent sound sources of the same class like a series of baby crying or multiple bird calls [131]. Thirdly, labelling acoustic data is technically-harder and more time-consuming than labelling images, due to the overlap of concurrent and diverse sources. The lack of well-labelled sound data in crowded sound scenes naturally hampers research progress. Existing SED sound datasets [2, 71] capture simple acoustic scenarios with low polyphony and where the event variance is small. The simplified acoustic scenario in turn makes sound counting task by SED methods tackleable. But when the sound scene becomes more complex with highly concurrent sound events, SED methods soon lose their capability in discriminating different sound events [125, 18]. In the meantime, some researchers think sound counting is equivalent to sound source separation task [120, 173, 176, 162, 175], in which the sound is counted as the source number by isolating individual sound from sound mixture and assigning it to corresponding sound source. However, our proposed sound counting is different from source number counting, it directly counts the overlapping events number, regardless of if these events come from the same sound source. Therefore, a study specific for sound counting problem is desirable and overdue.

In this chapter, we study the general sound counting problem under highly polyphonic, cluttered and concurrent situations. Whilst the challenges of image-based crowd counting mainly lie in spatial density, occlusion and view perspective distortion, the sound counting challenges are two-fold. Firstly, acoustic scenes are additive mixtures of sound along both time and frequency axes, making counting overlapping sounds difficult (temporal concurrence and spectrum-overlap). Secondly, there is a large variance in event loudness due to spherical signal attenuation with distance.

To tackle these challenges, we propose a novel dyadic decomposition neural network to learn a sound density representation capable of estimating cardinality directly from raw sound waveform. Unlike existing sound waveform processing methods that all apply frequency-selective filters on the raw waveform in single stage [68, 20, 194, 65, 37], our network progressively decomposes raw sound waveform in a dyadic manner, where the intermediate waveform convolved by each parent filter is further processed by its two child filters. The two child filters evenly split the parent filter's frequency response, with one child filter encoding the waveform *approximation* (the one with the lower-half frequency response) and the other one encoding the waveform *details* (the one with the higher-half frequency response). To accommodate sound loudness variance, spectrum-overlap and time-concurrence, we further propose an energy gain normalization module to regularize each intermediate parent waveform before feeding it to two child filters for further processing. This hierarchical dyadic decomposition front-end enables the neural network to learn a robust TF representation in multi-stage coarse-to-fine manner, while introducing negligible extra computation cost. By setting each filter's frequency cutoff parameters to be learnable and self-adjustable during optimization in a data-driven way, the final learned TF representation can better characterize sound existence in time and frequency domain. Following the front-end, we add a backbone network to continue to learn a time framewise representation. Such representation can be used to derive the final sound count number by either directly regressing the count number, regressing density map (the one we choose) or following SED pipeline. Apart from the network, we further propose three polyphony-aware metrics to quantify sound counting task difficulty level: polyphony ratio, maximum polyphony and mean polyphony. We will give detailed discussion to show the feasibility of three metrics.

We run experiment on large amounts of sound datasets, including commonly heard bioacoustic, indoor and outdoor, real-world and synthetic sound. In the indoor environment, we run experiment with simulation [154] and we set small reverberation

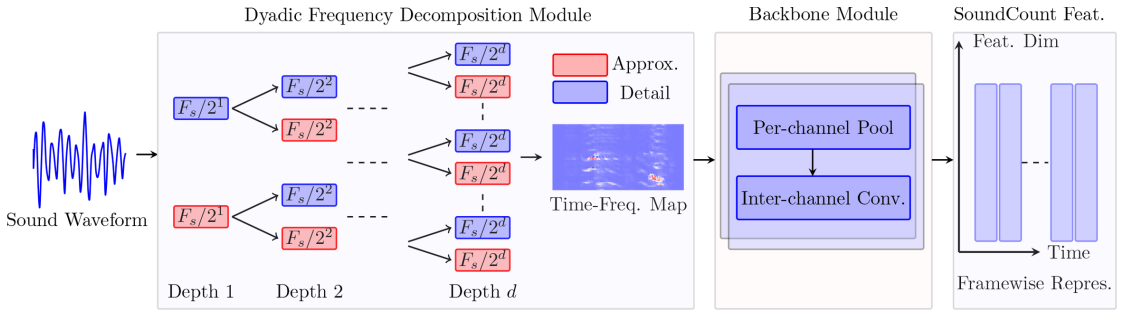


Figure 3.12: DyDecNet pipeline. We first feed the input raw sound waveform to the dyadic decomposition front-end to learn a time-frequency representation, which is further fed to a backbone neural network to continue to learn framewise representation. Such representation retains time information, so it is general enough to get count number by either regression or SED method. The dyadic decomposition front-end consists of a set of parameterized learnable band-pass filters. Each intermediate waveform processed by a parent filter is further processed by two child filters, with lower-half filter (red color) encoding *approximation* and higher-half filter (light-blue) encoding *details*.

parameters to minimize the reverberation impact. Comprehensive experimental results show the superiority of our proposed framework in counting under different challenging acoustic scenarios. We further show our proposed dyadic decomposition front-end can be used to tackle other acoustic task, like SELD [2, 68]. In summary, we make three main contributions: **First**, propose dyadic decomposition front-end to decompose the raw waveform in a multi-stage, coarse-to-fine manner, which better handles loudness variance, spectrum-overlap and time-concurrence. **Second**, propose a new set of polyphony-aware evaluation metrics to comprehensively and objectively quantify sound counting difficulty level. **Third**, show DyDecNet superiority on various counting datasets, and its potential to be used as a general learnable TF extraction front-end.

3.4.2 Dyadic Decomposition Neural Network

Different sound classes typically exhibit different spectral properties. A canonical way to process raw sound waveform is to apply a frequency-selective filter bank $\boldsymbol{\mathcal{F}}_f = \{f_i\}_{i=1}^k$ to project the raw sound waveform onto different frequency bins. Traditional Fourier transform [37] or Wavelet transform [109] construct fixed filter banks in which all filter-construction relevant hyperparameters are empirically chosen and thus may not be optimal for a particular task. Recent methods [68, 194] relax some hyperparameters to be trainable so that the filter bank can be optimized in a data-

driven way. A learnable filter bank often leads to better performance than fixed filters. However, all existing methods apply all filters, either learnable or fixed, on the raw waveform in a one-stage manner. Such shallow and one-stage processing may fail to learn powerful and robust representation for sound counting task where large loudness variance and heavy spectrum overlap exist. In our dyadic decomposition framework, we instead adopt a progressive pairwise decomposition strategy to obtain the time-frequency (TF) representation. It learns a TF representation from coarse to fine-grained granularity. Particularly, it consists of a dyadic frontend and a backbone.

In dyadic decomposition frontend, we construct a set of D hierarchical filter banks $\boldsymbol{\mathcal{F}}_{dyadic}^{D} = \{\boldsymbol{\mathcal{F}}_{2^1}^{1}, \boldsymbol{\mathcal{F}}_{2^2}^{2}, \cdots, \boldsymbol{\mathcal{F}}_{2^D}^{D}\}$. The d-th filter bank has 2^d filters, each filter is parameterized by a learnable high frequency-cutoff parameter and a low frequency-cutoff parameter. By cascading these filter banks, we consecutively decompose the raw waveform in frequency domain dyadically, leading to coarse-grained to fine-grained TF representation. Specifically, we denote the dyadic filter banks depth by D, in the depth d filter bank $\boldsymbol{\mathcal{F}}_{2^d}^{d}$, we have 2^d filters evenly divide the waveform sampling frequency F_s. Therefore, each single filter's frequency response length is $\frac{F_s}{2^d}$, the i-th filter f_i^d high frequency cutoff F_h and low frequency cutoff F_l are initialized as,

$$F_h(f_i^d) = \frac{F_s}{2^d} \cdot (i+1), \quad F_l(f_i^d) = \frac{F_s}{2^d} \cdot i \tag{3.9}$$

From Eqn. (3.9) we can see that dyadic decomposition frontend forms a complete binary-tree-like structure, in which the filter number doubles and each filter's frequency response length halves as the tree's depth increases by one. The intermediate waveform processed by a "parent" filter is just further processed by its two "children" filters. The frequency responses of the two children filters evenly split their parent filter's frequency response. The child filter carrying the higher half frequency response encode the parent's processed intermediate waveform's *detail* while the other one carrying the lower half frequency response instead encodes the *approximation*. For example, for the filter f_i^d in the d-th filter bank, its frequency response lies in $[\frac{F_s}{2^d} \cdot i, \frac{F_s}{2^d} \cdot (i+1)]$, its two children filters f_{2i}^{d+1} and f_{2i+1}^{d+1} in the depth $d+1$ evenly divide its frequency range, so f_{2i}^{d+1} carries $[\frac{F_s}{2^d} \cdot i, \frac{F_s}{2^d}(i+\frac{1}{2})]$. f_{2i+1}^{d+1} carries $[\frac{F_s}{2^d}(i+\frac{1}{2}), \frac{F_s}{2^d} \cdot (i+1)]$.

With the pre-constructed dyadic decomposition filter banks, we cascade them together to process the raw sound waveform, progressively learning the final TF representation. In our implementation, each filter in dyadic filter banks is a learnable band-pass filter. We adopt rectangular band-pass in frequency domain filter which comprises of a learnable high frequency cutoff parameter F_h and a learnable low frequency cutoff parameter F_l. Converting it to time domain through the inverse Fourier

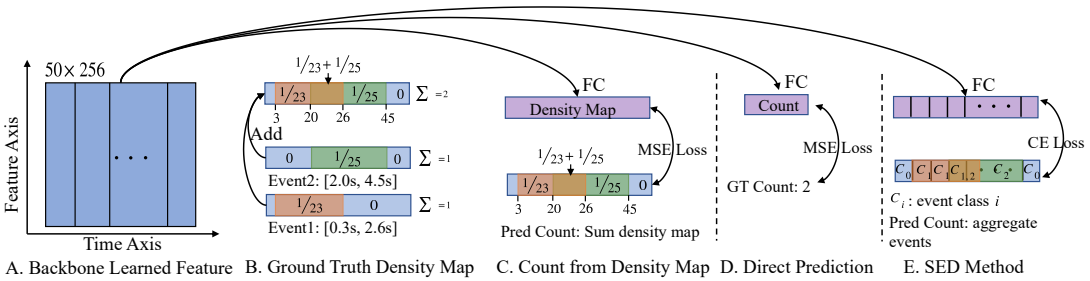


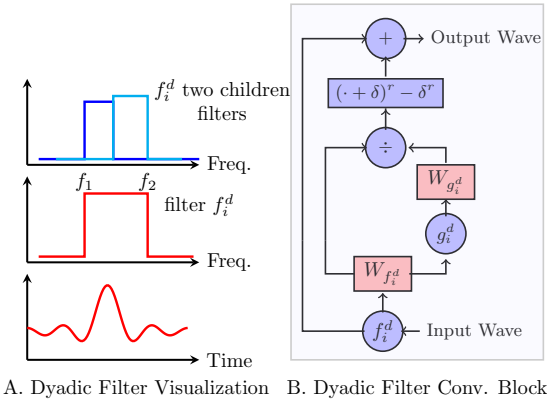
Figure 3.13: Three counting methods illustration. For density map (sub-fig. C), the sum (or integral) of the density map equals to the count number. We can also direct regress the final count number (sub-fig. D), or use SED method (sub-fig. E).

transform, we get $sinc(\cdot)$ function like filter that is used to convolve with the waveform. In our implementation, the learnable frequency cutoffs are initialized in Mel scale so that more filter banks are assigned to the lower frequency range. We further allow frequency overlap. Specifically, any two neighboring initialized filter banks overlap 20 Hz. For example, the filter f_i^d in Eqn. (3.9) is represented as,

$$f_i^d[t, F_h, F_l] = 2F_h sinc(2\pi F_h t) - 2F_l sinc(2\pi F_l t) \tag{3.10}$$

where $sinc(x) = sin(x)/x$, t indicates the filter's representation at time t. F_h and F_l are initialized according to Eqn. (3.9), but they can be further adjusted during the training process. $sinc(\cdot)$ filters have been successfully used in speech recognition [143] and sound event detection and localization [68]. In our dyadic decomposition frontend, each filter from different depth has separate and independent learnable parameters (high frequency cutoff and low frequency cutoff). Moreover, our constructed filter is much longer (1025 samples in our case) than traditional 1D/2D Conv filters (3 or 5, im mel-scale). Its wide length characteristic enables the filter to have a wide field-of-view on the raw waveform. Cascading them together allows the filters in later layers (larger depth) to have an even wider field-of-view on the input raw waveform. With this advantage, we do not have to model sound event temporal dependency explicitly with RNN network. As a result, the whole dyadic frequency decomposition frontend is fully convolutional and parametrically learnable, it is parameter-frugal and computationally efficient. In practice, the dyadic decomposition frontend depth is 8, so the output TF representation has 256 frequency bins. At the same time, we downsample the intermediate waveform by 2 before feeding it to its two children filters in the initial five stage dyadic filter banks to reduce the memory cost because dowmsampling by 2 reduces the waveform's temporal length.

We further design an energy gain normalization module to regularize each intermediate waveform before feeding them to the next dyadic filter bank. The motivation of introducing energy gain normalization is two-fold: first, to reduce sound event loudness variance led by sound events' different spatial locations; Second, to reinforce the frontend to learn to better tackle spectrum overlap challenge led by intra-class sound events in the sound scene. Specifically, for the intermediate waveform $W_{f_i^d}$ processed by a dyadic filter f_i^d, we first smooth it with a learnable 1D Gaussian kernel g_i^d parameterized by learnable width σ to get the corresponding smoothed waveform $W_{g_i^d}$ which just contains loudness. We then introduce a learnable automatic gain control parameter α to mitigate sound loudness impact. Furthermore, another two learnable compression parameters δ and γ are introduced to further compress $W_{f_i^d}$. The overall energy gain normalization can be represented as,



A. Dyadic Filter Visualization B. Dyadic Filter Conv. Block

Figure 3.14: Dyadic filter illustration. Left: In time domain, dyadic filter is a *sinc* function curve. In frequency domain, dyadic filter is a rectangular band-pass filter with learnable high frequency f_2 and low frequency cutoff f_1. The filter's two child filters (left-top) evenly splits the parent filter's frequency response. Right: dyadic filter convolution block. The input waveform is fed to an energy normalization module. Then a skip-connection is added.

$$W_{f_i^d} = (\frac{W_{f_i^d}}{(W_{g_i^d})^\alpha} + \delta)^\gamma - \delta^\gamma \tag{3.11}$$

where α, δ and γ are learnable parameters. As a result, the energy gain normalization *eg*-Norm (**e**nergy **g**ain normalization) is fully learnable and parameterized by four learnable parameters *eg*-Norm$(\sigma, \alpha, \delta, \gamma)$. Practically, each filter in dyadic filter banks is associated with an independent *eg*-Norm module. Similar energy normalization has been successfully used in tasks like keyword spotting [186, 105]. The difference lies in the fact that they apply exponential moving average operation to get smoothed waveform representation, so the computation is very slow because it iterates along the time axis to compute the averaged value step by step. Our proposed energy gain normalized strategy instead adopts a Gaussian kernel to get the smoothed waveform, in which it can be easily implemented as 1D convolution. The dyadic filter visualization and energy normalization module is shown in Fig. 3.14.

We add a lightweight backbone neural network to the frontend neural network to further learn a representation useful for call counting . The backbone network consists of two parts: per-channel pooling and inter-channel 1D convolution. Unlike existing methods [18, 2] that first convert 1D sound waveform into 2D map with fixed FFT-like transform, then learn from the 2D map with 2D Conv. operations, our method directly learns from sound raw waveform with learnable 1D Conv.. Specifically, we downsample each channel separately by assigning each channel with an independent frequency-sensitive learnable filter. We call such learnable downsampling per-channel pooling. It helps to learn sound event's frequency variance along the time axis individually. Moreover, we add normal 1D Conv. to achieve inter-channel communication, which enhances the neural network to learn concurrent sound events interaction. The backbone serves as the backend to learn framewise representation for counting.

The backbone network discussed above learns a framewise representation $[T_b, F_b]$, where T_b indicates the time steps and F_b indicates feature size. There are three potential ways to derive the final sound count number from the learned representation: 1. directly regress the count number; 2. SED method: detect sound events first and then aggregate results to get final count; 3. Predicting the density map. For a sound event with time location $[t_1, t_2]$, its density map is a 1D vector with value $\frac{1}{t_2 - t_1}$ during its occurrence time, otherwise it is 0. So the count number equals the vector integral. We show the regressing density map produces the best result (see Table 3.15). We thus adopt the mean squared error (**MSE**) loss during training to directly regress the density map. The comparison of three methods is shown in Fig. 3.13.

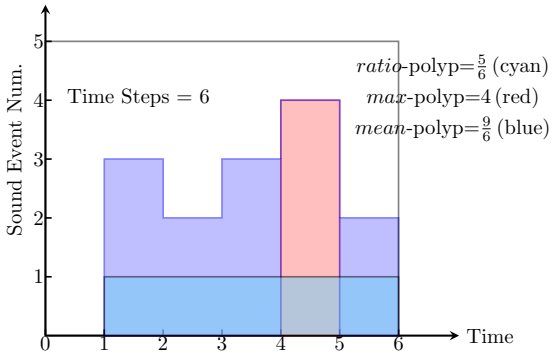


Figure 3.15: The max polyphony level in this sound clip is 4 (show in red, time step 5), so *max*-polyp=4. The *mean*-polyp indicates the purple area, averaging them over time gets $\frac{9}{6}$. *ratio*-polyp measures polyphony (no fewer than two concurrent sound events) existence ratio along the time axis (cyan), so it is $\frac{5}{6}$.

3.4.3 Counting Difficulty Quantification

Mean absolute error (**MAE**) and mean squared error (**MSE**) are two widely used metrics in crowd counting [106, 196]. Specifically, denote the ground truth count and predicted count by y_i and $\hat{y}_i$ respectively, for the i-th sound clip. MAE is defined

as MAE $= \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$, MSE is defined as MSE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$. We also involve accuracy rate (**AccuRate**) to show the ratio of accurately predicted count. We introduce a tolerance term p, where $p = 0$ means the predicted count number has to be exactly the same with ground truth number in order to be treated as an accurate counting; $p = 1$ relaxes the constraint so there can be one count mismatch for an accurate counting.

The aforementioned three general metrics do not reflect the impact of sound scene nature on algorithms. We introduce three polyphony-aware metrics to quantify the sound counting difficulty level reflected by the sound scene nature. The three metrics are time-window invariant so they can be used as general metrics to quantify difficulty level of sound scene of various lengths.

Polyphony Ratio (*ratio*-polyp) describes the ratio of polyphony (at least two sound events happen at the same time) over a period of time. It binarizes each time step as either polyphonic or non-polyphonic (monophoinc or silent) so the value lies between $[0, 1]$.

Maximum Polyphony (*max*-polyp) focuses on the maximum polyphony level over a time period. It is motivated by the fact that human's capability in discriminating different sound events reduces seriously when the number of temporal-overlapping sound event number increases. It is a positive integer and helps us to understand an algorithm's capability in tackling polyphony peak.

Mean Polyphony (*mean*-polyp) instead focuses on the averaging level of polyphony involved within a time period. It is designed to reflect an algorithm's capability in tackling the average polyphony level over an arbitrary time window.

Given T_n time steps sound vector $[p_1, p_2, \cdots, p_{T_n}]$, where $p_i \geq 0$ is the sound event number happening at time step T_i. The three metrics are defined as,

$$\begin{aligned} ratio\text{-polyp} &= \frac{\sum_{i=1}^{n}\mathbb{1}_2(p_i)}{n}; max\text{-polyp} = \max_{i=1,\cdots,n} p_i; \\ mean\text{-polyp} &= \frac{\sum_{i=1}^{n}\max(p_i - 1, 0)}{T_n} \end{aligned} \tag{3.12}$$

where $\mathbb{1}_2(p_i)$ is an indicator function, it is 1 if $p_i \geq 2$, otherwise 0. With the three metrics, we can report the general metrics (MAE, MSE) against various difficulty levels.

3.4.4 Experiment

We run experiments on five main category sound datasets that we commonly hear in everyday life.

1. Bioacoustic Sound. We focus on bird sound as bird sound is ubiquitous in most terrestrial environments with distinctive vocal acoustic properties. Specifically, we test three datasets: one real-world NorthEastUS [32] dataset and other two synthesized datasets: Polyphony4Birds (for heterophony test) and Polyphony1Bird (for homophony test). NorthEastUS data is recorded in nature reserve in northeastern United States. It encompasses 385 minutes of dawn chorus recordings collected in July 2018, with a total of 48 bird species. The average bird sound temporal length is very short (less than 1s) and the polyphony level (*max*-polyp and *mean*-polyp) is small. To test performance under highly polyphonic situations, we synthesize two bird sound datasets. Specifically, the first dataset contains four sounds: junco, American redhead, eagle, and rooster from copyright-free website `findsounds.com`. We call it Polyphony4Birds (heterophony test). The second dataset contains one sound: rooster. We call it Polyphony1Bird (homophony test).

2. Indoor Sound. We count telephone ring sound, the telephone ring seed sound comes from the same copyright-free website. We follow Polyphony1Bird synthesis procedure except that the room size is much smaller ($10m \times 10m \times 3m$) to minimize indoor reverberation effect (as smaller room exhibits smaller reverberation effects).

3. Outdoor Sound. We count car engine, as it is widely heard in outdoor scenario. The car engine seed sound comes from the same copyright-free website. We follow Polyphony1Bird synthesis procedure to create the dataset.

4. AudioSet [53] is a large temporally-strong labelled dataset with a wide range of sound event classes, including music, speech and water. The AudioSet data tests all methods' capability in counting under large different event classes scenario. Specifically, we train model on the training dataset which has 103,463 audio clips and 934,821 labels, and test the model on the evaluation which has 16,996 audio clips and 139,538 labels. In total there are 456 sound event categories.

5. Music Sound. We use OpenMic2018 dataset [78] to count musical instruments.

We compare DyDecNet with three main method categories: 1) traditional signal processing methods: Librosa-onset and Aubio-onset; 2) three SED-based methods. 3) one sound source separation method. **Librosa-onset** [110] provides an onset/offset detection method for music note detection. It measures the uplift or shift of spectral energy to decide the starting time of a note. We use its onset/offset detection ability

Table 3.11: MSE (↓) and MAE (↓) on the five main category sound (six in total) datasets. For more detailed result, please refer to the conference paper.

Method	AudioSet		NorthEastUS		Polyp4Birds		Polyp1Bird		TelepRing	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Librosa-onset	26.9	4.80	2.31	1.65	28.3	4.09	37.63	5.5	30.03	4.50
Aubio-onset	8.50	1.98	4.91	1.74	8.43	1.91	35.33	5.27	33.20	4.22
SELDNet [2]	0.93	1.28	1.35	1.79	0.92	1.41	0.89	1.19	0.97	1.30
CRNNNet [18]	0.92	1.07	1.33	1.77	0.74	1.10	0.87	1.16	0.92	1.31
DND-SED [43]	1.10	1.22	1.19	1.64	0.95	1.34	1.04	1.27	1.23	1.34
DPTasNet [120]	0.81	1.02	1.11	1.60	0.97	1.47	1.22	1.43	1.47	1.21
Our DyDecNet	**0.32**	**0.73**	**1.01**	**1.19**	**0.46**	**0.92**	**0.54**	**0.85**	**0.58**	**0.89**

to count sound events. **Aubio-onset** [16] achieves pitch tracking by aligning period and phase. We use its pitch tracking to count.

SED-based methods build on traditional fixed TF representation, such as short time Fourier transform (STFT) and LogMel. The TF representation is treated as a 2D image to be processed by a sequence of 2D Conv. operators. GRU [33] and LSTM [72] are often adopted to model temporal dependency. We compare three typical SED methods: 1) **CRNNNet** [18] consists of 2D Conv. to learn multiple compressed TF representations from the input TF map. Then it concatenates them together along the frequency dimension and further feeds it to LSTM [72] to learn framewise representation. 2) **DND-SED** [43] instead adopts depthwise 2D convolution and dilated convolution to avoid using RNN. 3) **SELDNet** [2] is originally used for joint sound event detection and localization. It adopts 2D Conv. to convolve the 2D TF map, and bidirectional GRU to model temporal dependency. The three comparing methods' network architectures are slightly adjusted to fit our dataset. For sound source separation method, we adopt **DPTasNet** [120], in which it trains a Dual-Path RNN (DPRNN) and TasNet to jointly separate each sound event and further count the event number. In this case, we treat each sound event as independent sound sources.

For all datasets, all input audios are segmented into 5 second long clips, with sampling rate 24 kHz. So the input waveform has 120,000 data points and is normalized into $[-1, 1]$. We train the models with Pytorch [129] on TITAN RTX GPU. To train the neural network, we adopt Adam optimizer [85] with an initial learning rate 0.001 which decays every 20 epochs with a decaying rate 0.5. We empirically found Adam optimizer [85] gave better performance than other optimizers like SGD. Overall, we train 60 epochs. We train each method 10 times independently and report the mean value and standard deviation. We do not report the standard deviation explicitly in

Table 3.12: Ablation study on dyadic decomposition efficiency discussion: we compare existing methods with and without dyadic decomposition frontend.

Method	MSE↓	MAE↓
SELDNet [2]	1.35	1.79
SELDNet_Dydec	**1.05**	**1.43**
CRNNNet [18]	1.33	1.77
CRNNNet_Dydec	**1.20**	**1.51**
DND-SED [43]	1.19	1.64
DND-SED_Dydec	**0.89**	**1.40**

Table 3.13: Ablation study on traditional T-F feature for counting task: DyDecNet's dyadic decompostion frontend is replaced by various classic T-F features extractors, such STFT, LogMel, MFCC and Gabor.

Method	MSE↓	MAE↓
DyDecNet_STFT	1.35	1.51
DyDecNet_LogMel	1.33	1.50
DyDecNet_MFCC	1.32	1.49
DyDecNet_Gabor	1.33	1.48
DyDecNet	**0.85**	**1.19**

Table 3.14: Various DyDecNet ablation study variants test result.

Method	MSE	MAE
DyDecNet_SingScale	1.22	1.43
DyDecNet_BN	1.07	1.25
DyDecNet_noNorm	1.15	1.37
DyDecNet	**0.85**	**1.19**

Table 3.15: Various counting methods test result. RegCount means directly regressing the count.

Method	MSE	MAE
DyDecNet_RegCount	1.03	1.39
DyDecNet_SED	2.09	3.06
DyDecNet	**0.85**	**1.19**

the table because we find them very small (about 0.03). We first train the comparing SED methods with both their suggested training strategy and our training strategy, then choose the one with the better performance as the final result. For the energy gain normalization we initialize them as $\alpha = 0.96$, $\delta = 2.$, $\gamma = 0.5$, $\sigma = 0.5$ (they are empirically chosen). The batchsize is 128.

The quantitative result on MSE/MAE is given in Table 3.11. From this table we can learn that DyDecNet outperforms both classic signal processing deterministic methods, comparing SED methods and sound source separation based method by a large margin, under all acoustic scenarios. DyDecNet outperforms all comparing methods in both real-world and synthesized sound datasets. It is capable of learning power-

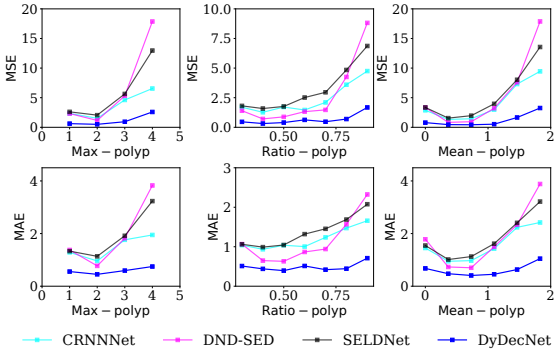


Figure 3.16: MSE and MAE variation against *max*-polyp, *ratio*-polyp and *mean*-polyp on NorthEastUS dataset.

ful representation from both weak sound
signals (NorthEastUS), highly polyphonic (synthesized datasets) and heavy spectrum-overlapping, loudness-varying sound events. Moreover, we find DPTasNet [120] performs worse than the three SED-based methods on the two synthesized bioacoustic datasets where high-polyphony exists, which shows source separation method is not a good counting alternative in highly polyphonic situations.

At the same time, we also observe that the two signal processing deterministic methods (Librosa-onset and Aubio-onset) generate the worst result over both SED based, source separation based methods and DyDecNet. The higher the polyphony level of the dataset, the worse performance the two deterministic methods lead to. For example, in NorthEastUS dataset with a relatively smaller polyphony level, Librosa-onset and Aubio-onset generate relatively good performance with accuracy rate ($p = 1$) reaching 0.58. In our synthesized two datasets with much higher polyphony levels, however, their accuracy drops significantly to near zeros. It thus shows traditional signal processing methods do not fit for sound counting from crowded acoustic scenes.

Moreover, SED-based methods and DyDecNet produce decreasing performance from Polyphony4Birds to Polyphony1Bird and then NorthEastUS. The largest performance drop is observed on real NorthEastUS dataset, which shows counting from real-world dataset is a tough task that desires more future attention. Spectrum-overlap led by intra-class sound events is another potential challenge (better performance on Polyp4Birds than Polyp1Bird).

The MSE/MAE variation against *max*-polyp, *ratio*-polyp and *mean*-polyp difficulty level on NorthEastUS are shown in Fig. 3.16. We can observe that our proposed three metrics *max*-polyp, *ratio*-polyp and *mean*-polyp are effective ways to accurately quantify sound counting tasks difficulty level. The three metrics have observed dramatic performance drop as their difficulty level increases. Nevertheless, DyDecNet remains as the best one across all the three difficulty levels, showing DyDecNet outperforms the comparing methods under difficult levels discussed in this chapter.

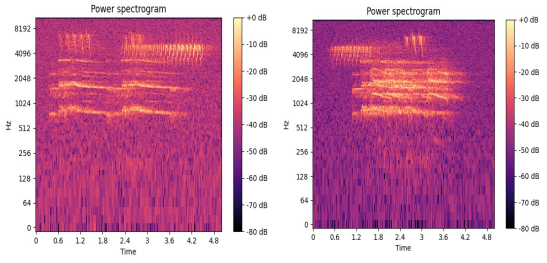


Figure 3.17: Count Example Visualization: we visualize two highly-polyphonic count examples spectrum (magnitude) and their corresponding count number.

We visualize the comparison between dyadic decomposition front-end learned TF feature and classic MFCC [37] TF feature in Fig. 3.18. We can see DydecNet is capable of learning more discriminative feature for two temporally-overlapping and the same class sound events. We further visualize two examples in Fig. 3.17.

We do ablation study on NorthEastUS data. **First**, disentangling our proposed framework's dyadic decomposition frontend and backbone network so as to figure out their individual contribution. To this end, on the one hand, we concatenate dyadic decomposition frontend to the three SED methods backbone networks so that they can learn TF representation from raw waveform. We call them SELDNet_dydec, CRNNNet_dydec and DND-SED_dydec respectively. On the other hand, we feed our backbone neural network with fixed pre-extracted TF features, including short time Fourier transform (STFT), LogMel, MFCC and Gabor Wavelet filter. We call them DyDecNet_STFT, DyDecNet_LogMel and DyDecNet_MFCC, DyDecNet_Gabor, respectively. The results are in Table 3.12 and 3.13. We can observe that: 1) replacing traditional fixed TF feature with dyadic decomposition frontend significantly improves the performance (Table 3.12). The gain stems from two-fold: our dyadic decomposition frontend enables the network to directly learn from the raw waveform so that all frequency-selective filters are adjustable during training process. Second, the dyadic progressive decomposition enables the neural network to learn robust representation for sound counting. Similarly, a huge performance drop is observed if we let our proposed backbone neural network to learn from traditional fixed TF features (Table 3.13). Therefore, it shows that both the dyadic decomposition frontend and backbone neural networks are important for sound counting.

We further want to figure out if the dyadic decomposition is essential for sound counting, and the importance of energy normalization block. We test three variants: our network with simply single scale decomposition which means applying all filters on the raw waveform (DyDecNet_SingScale) which helps validate necessity of hierarchical dyadically decomposition framework; replacing Energy-normalization module with traditional batch normalization [77] (DyDecNet_BN); without any normaliza-

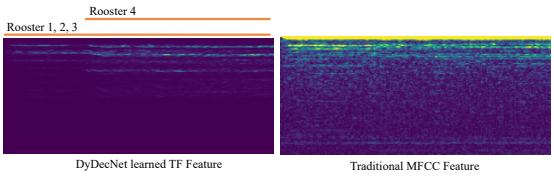


Figure 3.18: We visualize the learned time-frequency comparison between our proposed dyadic decomposition method and traditional MFCC method. We can clearly see that under intra-class polyphonic situation (four rooster audio co-happen at the same time), our proposed method can better discriminate different audio sources while MFCC has mixed them together.

Table 3.16: Dyadic Frontend on SELD Task.

Method	ER (↓)	F (↑)	LE (↓)	LR (↑)
SELDNet [2]	0.63	0.46	23.1	0.69
SELDNet_DyDec	**0.60**	**0.49**	**22.7**	**0.73**
EIN [20]	0.25	0.82	8.0	0.86
EIN_DyDec	**0.21**	**0.86**	**7.4**	**0.88**
SoundDet [68]	0.25	0.81	8.3	0.82
SoundDet_DyDec	**0.21**	**0.88**	**7.2**	**0.86**
SoundDoA [65]	0.23	0.85	7.9	0.87
SoundDoA_DyDec	**0.20**	**0.89**	**7.4**	**0.89**

tion (DyDecNet_noNorm). The result is
in Table 3.14, from which we can clearly observe that either removing energy normalization or replacing it with batch normalization significantly reduces the performance. It thus shows the importance of energy normalization.

Moreover, we run two ablation studies to directly regress the count number and follow SED pipeline, respectively. From the result in Table 3.15, we can conclude that directly regressing sound event count number leads to inferior performance than estimating density map. Treating it as a SED problem leads to the worst performance, because SED needs to first precisely localize each sound event in time domain so as to be able to count it. Localizing in highly polyphonic situation is a challenging task.

To show the dyadic decomposition front-end is a general TF feature extractor, we test it on sound event detection and localisation task (SELD). The dataset we use is TAU-NIGENS [3], and we compare with four main methods: SELDNet [2], EIN [20] that use classic TF feature, SoundDet [68] and SoundDoA [65] use learnable TF-feature. We replace their time frequency (TF) extraction front-end with dyadic decomposition network front-end to see the performance change. The result is given in Table 3.16, we can see that dyadic decomposition front-end exhibits generalization strength to help tackle other acoustic tasks.

3.4.5 Conclusion and Discussion

We can learn from this work that learning from highly polyphonic audio inevitably introduces new research challenges when designing learnable filter bank: loudness variability and frequency overlap. Extracting representative time-frequency representation from such highly polyphonic spatial audio requires specially designed time-frequency extraction strategy. Our proposed dyadic decomposition learnable filter bank is one such solution. We provide such one comparison in Fig. 3.18.

3.5 Conclusion

In this chapter, we present three kinds of learnable filter bank design strategies for SELD task and sound event count task, and we can conclude that carefully designed learnable filter bank performs better than traditional hand-crafted filter bank on SELD task and sound event task. The SoundDet [68] work shows it is feasible to design novel learnable filter bank to jointly encode per-channel time-frequency map and inter-channel time-delay information. The SoundSynp [66] work shows constructing time-frequency map in multi-scale manner in both time and frequency domain is essential for SELD task, especially for audio source's physical position localization task. The SoundCount [62] shows an alternative way of learning time-frequency map in multi-step dyadic decomposition way when it comes to handling highly-polyphonic spatial audio where challenges such as loudness variability and frequency overlap exist. It is worth noting that the time and frequency aware multi-scale extraction strategy introduces extra computation cost due to the pre-constructed multi-scale filter banks. The SoundCount [62] introduced dyadic decomposition time-frequency extraction way is much slower in terms of time efficiency than the one-step extraction way (*e.g.*, traditional time-frequency extraction methods, SoundDet [68] and SoundSynp [66]).

Chapter 4

Spatial Audio Neural Rendering

4.1 Motivation Discussion

The previous Chapter 3 shows incorporating deep neural network into designing learnable filter bank to extract time-frequency representation helps various spatial audio based tasks such as sound source detection and localization (SELD) and sound event counting. It thus shows the potential of deep neural network in renovating classic sound waveform processing methods, especially in the time-frequency extraction side. Followed by these achievements, we stepped further to propose to learn a spatial audio neural rendering field which implicitly models the underlying spatial audio propagation primitives in an enclosed 3D space. We show that deep neural network can also be applied to learn a continuous spatial audio propagation field so that we can access the behaviour change from arbitrary source position to another arbitrary receiver position with high fidelity, avoiding the need to physically measure the data which is tedious and difficult.

Spatial audio neural rendering closely relates to sound synthesis, and can be potentially applied to various inverse problems including room structure/layout estimation, acoustic AR/VR problem.

4.2 Deep Neural Acoustic Primitive

4.2.1 Motivation Discussion and Introduction

Acoustically characterizing an enclosed room scene [95] demands estimating all the acoustic primitives related to the physical space and is key in enabling a wide range of applications, including architectural acoustics [103], audio-based virtual and augmented reality [180, 14] and geometric room structure estimation from audio [94].

One prominent primitive is to identify the sound propagation dynamics underpinning the room that models the interaction of sound waves with entities in the room during its propagation from a source position to a receiver position. Due to the nature of sound waves, this interaction between sound and the room is highly complex and is sensitive to: (i) the sound source and receiver positions, (ii) room architecture, (iii) geometric layout, including furniture placement, and (iv) material properties. Further, sound waves undergo reflection, scattering, and absorption, complicating their dynamics. All of these challenges make quantifying and measuring sound propagation primitives laborious and difficult, especially using classical methods [165].

The acoustic effects of a room scene can be well modeled as a linear time-invariant system [51] and thus the sound propagation primitive can be expressed as a one dimensional room impulse response (RIR) function. The received signal can then be obtained by convolving the source sound with the RIR. Due to the complex nature of sound propagation as described above, the RIR is extremely nonsmooth and arbitrarily long in the time domain. Existing RIR measurement methods either require one to physically collect a discrete RIR in the room scene by sending an impulse sound (*e.g.*, starter pistol, balloon exploding, etc.) or chirp sound at one position and recording the response at another, or to approximate the RIR using geometry or wave-based approaches [153, 9]. While, the former approach based on physical measurements is inefficient and unscalable (*e.g.*, an RIR is available only at the measured locations), the latter approximation methods are computationally expensive and require detailed knowledge of the room scene acoustics. Recent works [142, 108, 159] propose to learn RIRs with deep neural networks in a fully supervised manner, but assume access to massive RIR datasets and evaluate in small room scene [161] settings.

In this work, we take a fresh look at the sound propagation primitive estimation problem. Given the challenge in directly deriving the sound propagation primitive, we propose an indirect framework for which data is much more readily accessible and is also agnostic to the room scene acoustic characteristics. Our underlying insight is that, although directly measuring the primitive response is difficult, the *effects* of the primitive are more easily accessible because it simply requires moving to a room scene to receive sound after propagation. Thus, by analyzing the source and receiver sounds appropriately, we can inversely estimate the propagation primitive. Motivated by this idea, we propose to use two cooperative agents that are temporally synchronized, one serving as a sound source agent and the other as sound receiver agent, to probe a room scene by moving around independently. At arbitrary agents' locations, the source sends a sound signal, which the receiver is assumed to receive. This active

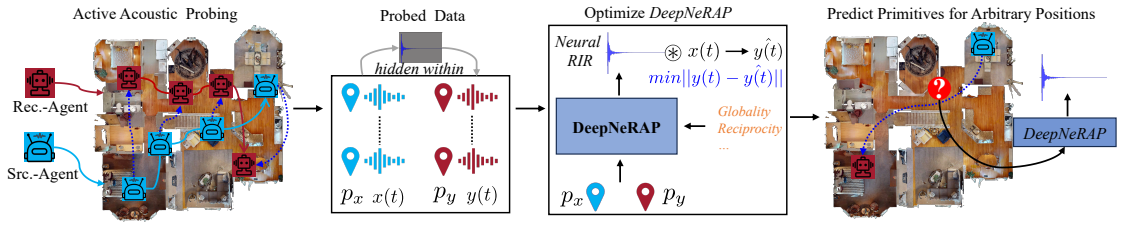


Figure 4.1: **DeepNeRAP Pipeline**: To learn neural sound propagation primitive, we make two agents to actively probe the room scene acoustically by emitting and receiving sounds at varied locations. Such a data collection strategy requires no prior knowledge of room acoustic properties, instead needs only the agents' position. The sound propagation primitive is implicit in the collected dataset. Our DeepNeRAP takes as input two positions and outputs a neural RIR that encodes the primitive. Convolving this neural RIR with the source sound gives the predicted receiver sound. The entire DeepNeRAP model is optimized by minimizing the discrepancy between the receiver recorded sound and predicted neural RIR effected sound. During inference, the learned DeepNeRAP can predict the RIR for any source and receiver positions.

room scene acoustic probing strategy can be easily executed in real scenarios because it requires no prior knowledge of room scene acoustic properties and the agents' position can be easily obtained using existing and mature localization frameworks such as SLAM [84]. Assuming the two agents reach most of the traversable area in the scene, we can easily obtain a probing dataset to inversely learn the sound propagation primitive. We illustrate our approach in Fig. 4.1.

With the probed dataset, we propose a novel framework, called **Deep Ne**ural **R**oom **A**coustics **P**rimitive (DeepNeRAP), to learn sound propagation implicitly in a self-supervised manner. It takes as input two positions, viz., the source and the receiver agents' positions, and predicts the corresponding neural RIR that essentially captures the sound propagation dynamics between the two positions. During training, our *neural RIR* is convolved with the source sound (using 1D convolutions) to predict the receiver sound. DeepNeRAP is then optimized by minimizing the discrepancy between the ground truth received sound and the predicted receiver sound produced using the neural RIR. The DeepNeRAP network design incorporates fundamental room acoustics principles, including *Globality*, *Reciprocity* and *Superposition*; adhering to these principles makes DeepNeRAP primitive closer towards capturing the physics of sound propagation, while being acoustically explainable.

To empirically validate the superiority of DeepNeRAP, we conduct experiments on both synthetic and real-world datasets. For the former, we use the large scale SoundSpaces 2.0 dataset [27, 24] consisting of indoor scenes with an average room

area $> 100m^2$ and enriched with room acoustics. For the latter, we use the real-world MeshRIR dataset [90]. Our experiments on these datasets demonstrate state-of-the-art RIR estimation performances over closely related approaches. Below, we summarize the main contributions of this chapter:

- We present DeepNeRAP to implicitly encode sound propagation primitives in a room scene. It is spatially continuous, capable of predicting a neural RIR for arbitrary source and receiver positions.
- DeepNeRAP is trained in a self-supervised and data efficient manner, requiring neither massive RIRs nor detailed prior knowledge of room scene acoustic properties.
- DeepNeRAP design is guided by fundamental room acoustics physical principles, resulting in the learned primitive being acoustically explainable and meaningful. We show DeepNeRAP's superiority on both synthesized large-scale room scenes and a real-world dataset.

4.2.2 Deep Neural Room Acoustics Primitive

Given an enclosed 3D room scene in $\mathbb{R}^{3D}$, that is assumed to be a linear time invariant (LTI) system, our task is to learn an implicit deep neural room acoustics field $\boldsymbol{\mathcal{F}}_\theta$ that encodes the sound propagation primitive underlying the scene. In our case, the primitive is expressed as a one dimensional neural RIR. $\boldsymbol{\mathcal{F}}_\theta$ is spatially continuous so that it can predict the neural RIR $h(t)_{p_s \to p_r}$ for any arbitrary source position p_s and receiver position p_r. The received sound at p_r then can be derived by convolving $h(t)_{p_s \to p_r}$ with the sound at position p_s and is entirely agnostic to the sound class,

$$h(t)_{p_s \to p_r} = \boldsymbol{\mathcal{F}}_\theta(p_s, p_r); \quad p_s, p_r \in P, \tag{4.1}$$

where P indicates all source or receiver (reachable) positions in the room scene, and θ represents the learnable parameters of $\boldsymbol{\mathcal{F}}$. Due to the high complexity of sound propagation dynamics, the measured RIR $h(t)$ is a highly nonsmooth and long signal (usually more than 20k points), making it difficult to collect RIR data directly to optimize $\boldsymbol{\mathcal{F}}_\theta$. Alternatively, we propose to learn $\boldsymbol{\mathcal{F}}_\theta$ in a more readily accessible way: we use two cooperative agents, one a source agent carrying an omnidirectional loudspeaker and the other a receiver agent carrying an omnidirectional microphone receiver, to actively probe the room scene independently (see Fig. 4.1). At each step, the source agent emits a sound $x(t)$ at one position p_x and the receiver agent receives the response sound $y(t)$ at another position p_y accordingly. This active probing

strategy is practical and easy to execute in real scenarios because it requires neither detailed prior knowledge of room scene's acoustic properties nor direct collection of RIR data (directly collecting RIR data is difficult and inefficient in practice). Since the received sound $y(t)$ implicitly carries the room scene's sound propagation primitive conditioned on the two agents' position, we can utilize it to inversely estimate the sound propagation primitive (*e.g.,* in our case $h(t)$). That is, $\boldsymbol{\mathcal{F}}_\theta$ is learned using N samples of active probing data $\mathcal{D} = \{p_x, p_y, x(t), y(t)\}_{i=1}^{N}$,

$$\boldsymbol{\mathcal{F}}_\theta \leftarrow \boldsymbol{\mathcal{F}}_\theta(\{p_x, p_y, x(t), y(t)\}_{i=1}^{N}); \quad \{p_x\}, \{p_y\} \subseteq P. \tag{4.2}$$

In our setting, the source agent emits a sine sweep [45] sound so as to cover the whole frequency range. The agents' spatial position can be easily retrieved by either SLAM systems [84] or an inertial measurement unit (IMU) with high accuracy. Our framework is "self-supervised" in the sense that $\boldsymbol{\mathcal{F}}_\theta$ is learned using only the data D we collected without involving any sort of data annotation (especially the RIR data). Our self-supervision cue lies in the difference between the emitted sound and the received sound. By enforcing the predicted neural RIR effected sound to be close to the receiver agent recorded sound, we naturally encourage $\boldsymbol{\mathcal{F}}_\theta$ to essentially approximate the room acoustics primitive of the room scene, θ is estimated as:

$$\begin{aligned} \arg\min_{\theta} \sum_{(p_x, p_y, x(t), y(t)) \in \mathcal{D}} & d(\big(h(p_x, p_y) \circledast x\big)(t), y(t)) \\ & \text{where } h(p_x, p_y) = \boldsymbol{\mathcal{F}}_\theta(p_x, p_y), \end{aligned} \tag{4.3}$$

where h is the $\boldsymbol{\mathcal{F}}_\theta$ predicted neural RIR, $\circledast$ is the 1D convolution[1] indicating the RIR effect applied on the sound (in our case, the source agent sound), and $d(\cdot)$ measures the discrepancy between the $\boldsymbol{\mathcal{F}}_\theta$ predicted neural RIR effect and the ground truth observed effect (i.e., the receiver agent recorded sound); in our case, we measure the discrepancy in both frequency and time domain using the ℓ_2 loss.

4.2.3 LTI Room Acoustics Physical Principles

Before introducing DeepNeRAP, we present four fundamental room acoustics physical principles [95, 144] that will guide the DeepNeRAP network design. In an LTI room scene, the measured RIR has to satisfy the following principles.

[1] In room acoustics, the received sound is obtained by convolving the source sound with the corresponding RIR in the time domain. For example, in our case, $y(t) = h(t) \circledast x(t)$.

Principle 1, **Globality:** Unlike in computer vision where a camera captures a localized neighborhood, sound propagation relates to an entire room scene. Once emitted at the source position, sound waveform traverses isotropically[2] to interact with the whole scene before reaching the receiver position. The recorded sound thus acoustically characterizes the whole scene [95].

Principle 2, **Reciprocity**: The reciprocity principle [144, 151] states that in an LTI system, the RIR corresponding to the source and receiver positions is exactly the same in terms of both magnitude and phase should the source and receiver positions be swapped. Therefore, the DeepNeRAP needs to be source-receiver position permutation invariant. For example, in Eqn. 4.1, $h(t)_{p_s \to p_r} = h(t)_{p_r \to p_s}$.

Principle 3, **Superposition**: The superposition principle relating to room acoustics states that the RIR responses caused by more than one sound source is simply the linear combination of the response caused by each single sound source individually. Under this principle, we only need to model the neural RIR for one source and one receiver setting. The polyphonic situation where multiple sound sources are co-emitting sound can be easily derived by linearly adding individual sounds convolved with their associated propagation dynamics (neural RIR) together.

Principle 4, **Sound Independence**: This principle encompasses two key aspects. Firstly, the room acoustics primitive remains intrinsic to a room scene, irrespective of the specific sound used to probe the scene. Secondly, the neural RIR is completely sound-class agnostic. It can be universally applied to any sound to accurately capture propagation effects.

4.2.4 DeepNeRAP Neural Network Architecture

Following the aforementioned physical principles and the recent advances in neural implicit representations, we present the DeepNeRAP network architecture (also illustrated in Fig. 4.2). DeepNeRAP takes as input two spatial positions and outputs a neural RIR that implicitly encodes the sound propagation primitive for this specific position pair. Specifically, DeepNeRAP comprises five modules, which are detailed below in the sequence of data processing flow.

1. A learnable room acoustic representation $\boldsymbol{\mathcal{M}}$, which is represented by a 2D spatial grid representation covering the whole room scene area. Each entry in the spatial

[2]The isotropic assumption is made under the fact that the emitter is not highly directional. If the orientation of a directional emitter was also known, it would be possible to condition the network on this orientation.

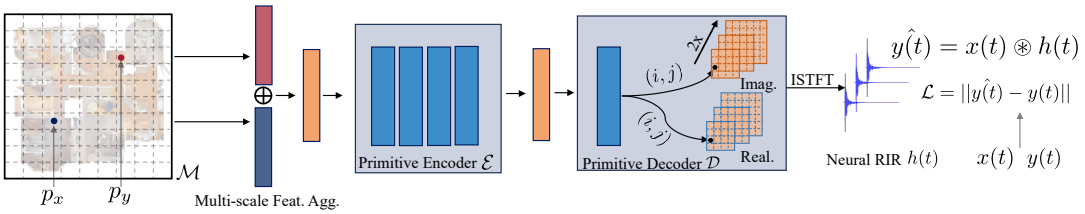


Figure 4.2: **DeepNeRAP Pipeline**: We construct a learnable spatial grid feature $\boldsymbol{\mathcal{M}}$ so that the two agents' positions can be registered to the grid map. By querying features in a multi-scale manner (see Fig. 4.3), we relate the agent's position to the whole room scene satisfying the *globality* principle. The source and receiver position features are merged by a permutation-invariant operator `add`, satisfying the *reciprocity* principle, and then fed to the primitive encoder $\boldsymbol{\mathcal{E}}$ for further refinement. The primitive decoder $\boldsymbol{\mathcal{D}}$ (transposed convolution in our case) then decodes the refined primitive features into a multi-scale neural RIR expressed in the time-frequency domain. The inverse short-time Fourier transform (ISTFT) is used to convert the neural RIR to the time domain. DeepNeRAP is optimized by minimizing the discrepancy between the neural RIR effected receiver sound and the actual recorded sound without accessing RIR data.

grid is registered to a physical position in the room scene and associated with a learnable feature representation.

2. A source and receiver position pair feature extractor $\boldsymbol{\mathcal{P}}$ that emphasizes both each single position's individuality and the global interaction with the whole room scene conditioned on single position.

3. An encoder $\boldsymbol{\mathcal{E}}$ that learns room acoustic primitives. In our case, it is multi-layer perceptrons (MLP).

4. A neural primitive prediction decoder $\boldsymbol{\mathcal{D}}$ that predicts a multi-scale neural RIR in frequency domain, where the neural RIR is represented by a complex 2D map at multiple resolutions.

5. A loss $\boldsymbol{\mathcal{L}}$ that optimizes the whole neural network by minimizing the discrepancy between the neural RIR effected sound and the receiver agent recorded sound at the same height.

In room acoustic representation, $\boldsymbol{\mathcal{M}}$ is expressed as a 2D $N \times N \times k$ spatial grid representation (rather than a 3D voxel grid, $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{N \times N \times k}$) because the two horizontal axes (topdown map) extent of most large room scenes is much larger than the vertical extent. Each entry in the grid $\boldsymbol{\mathcal{M}}$ corresponds to a physical 2D position in the room

scene, and is associated with a learnable small feature of size k. Such constructed grid representation is responsible for learning the sound propagation related representation underpinned by the room scene. Its grid-wise feature organization helps to learn the position-aware representation that is vital for modeling sound propagation. It is worth noting that, although $\boldsymbol{\mathcal{M}}$ is geo-registered to the room scene, we do not explicitly require knowledge of the precise room geometry. We simply need to ensure the grid map covers the whole room. Alternatively, we can construct a grid map simply covering the two agents' traversed area, obviating knowing the prior room scene size information.

The feature extraction procedure for either source position p_x or receiver position p_y should be: 1) position-aware so that the extracted features intrinsically reflect the position's individuality, and 2) related to the global room scene for the sake of the globality principle. To this end, we propose a multi-scale position-aware feature extraction strategy $\boldsymbol{\mathcal{P}}$. Specifically, for the input position p $(= \{p_x, p_y\})$, we retrieve L-scale bounding boxes on the grid map that center at p but are of different sizes. Given a scale resolution r, the l-th bounding box's size is $l \cdot r$. By adjusting the scale resolution r and scale number L, we can correspond p to the whole grid map features. For the l-th scale bounding box, we take the four farthest grid features within the bounding box to p and further adopt bilinear interpolation to get the corresponding feature for p at scale l. By concatenating the interpolated features arising from L scales, we obtain the room acoustic representation $f(p)$ for position p,

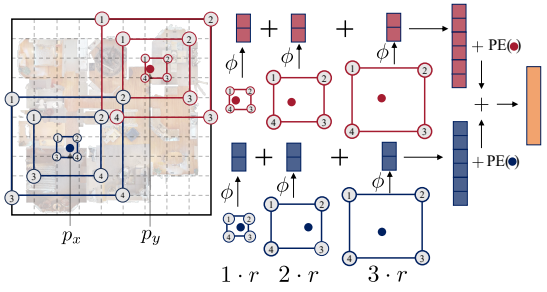


Figure 4.3: Mulit-scale position-aware global grid feature extraction. We just show three scales for simplicity.

$$
\begin{aligned}
f(p) &= f_{l \cdot r}(p) \oplus f_{(l+1) \cdot r}(p), \\
f_{l \cdot r}(p) &= \phi(f^1_{l \cdot r}, f^2_{l \cdot r}, f^3_{l \cdot r}, f^4_{l \cdot r}); l = 1, \cdots, L-1,
\end{aligned}
\tag{4.4}
$$

where $\phi(\cdot)$ indicates the bilinear interpolation operation. $\oplus$ indicates the concatenation operation along feature dimension. $f^i_{l \cdot r}$ indicates the i-th farthest grid feature in the l-th scale bounding box $(i = 1, 2, 3, 4)$. We visualize this multi-scale feature extraction strategy in Fig. 4.3. To reduce the computation overhead and storage cost without sacrificing the expressiveness of $\boldsymbol{\mathcal{M}}$, we instantiate d (the feature dimension

at grid cell) with a small value (in our case $d = 2$) but obtain a much larger and complex representation for p by concatenating small features arising from multiple scales. Such multi-scale aggregation strategy relates p to the whole room scene so it satisfies the globality principle. In addition to just querying room acoustic features (Eqn. 4.4), we explicitly add position encoding [179] to both source and the receiver positions to capture each position's uniqueness.

$$f(p_x) \leftarrow f(p_x) + \text{PE}(p_x);\ f(p_y) \leftarrow f(p_y) + \text{PE}(p_y), \tag{4.5}$$

where $\text{PE}(\cdot)$ is the sine/cosine position encoding. It is worth noting that the position's individuality is encoded by both the position encoding $\text{PE}(\cdot)$ and the concatenation operation $\oplus$ because the concatenation operation concatenates the interpolated features from different scales in an order decided by the position (from closest to farthest). Given the two extracted features, we adopt a permutation-invariant operation (element-wise add) to merge them to get the fused feature representation so that the reciprocity principle is satisfied, i.e., $f(p_{xy}) = f(p_x) + f(p_y)$.

Primitive Encoder $\boldsymbol{\mathcal{E}}$. After obtaining $f(p_{xy})$, we further adopt a multiple layer perceptron (MLP) to further encode the sound propagation essentials. In our case, the MLP consists of 6 fully-connected layers with hidden unit size of 512, batch normalization, and ReLU activation are used.

Primitive Decoder $\boldsymbol{\mathcal{D}}$ takes as input the features learned by $\boldsymbol{\mathcal{E}}$, and directly outputs a neural RIR. Due to the non-smoothness and large dimensionality of neural RIRs in the time domain, we propose to predict it in the time-frequency domain. Unlike prior works that predicts magnitude and phase maps [108] where the phase map is often chaotic, we predict real and imaginary 2D maps because the two maps are comparatively more smooth and easy to predict. By applying inverse short time Fourier transform (ISTFT), we can convert the neural RIR from frequency domain to time domain and the conversion operation is differentiable. Specifically, $\boldsymbol{\mathcal{D}}$ combines $f(p_{xy})$ and position encoded row and column index $[\text{PE}(i), \text{PE}(j)]$ to predict the real and imaginary values indexed at $[i, j]$,

$$\begin{aligned} h(t) &= \text{ISTFT}(H(\omega)), \\ H(\omega)[i,j] &= \boldsymbol{\mathcal{D}}(f(p_{xy}) + \text{PE}(i) + \text{PE}(j)), \end{aligned} \tag{4.6}$$

where $H(\omega)$ is the representation of $h(t)$ in the time-frequency domain, $H(\omega) = [\text{Real}(\omega), \text{Imag}(\omega)]$, which contains a real part map and an imaginary part map of shape $[w, h]$ (with initial $w = h = 128$, then doubled twice with transposed convolution). Instead of just predicting $H(\omega)$ at one resolution, we propose to predict multiple

Table 4.1: Quantitative Result on Matterport3D Dataset. t-MSE: 10^{-7}, f-MSE: 10^{-2}.

Method	Neural RIR						Speech
	t-MSE	SDR	T_{60} Error	f-MSE	PSNR	SSIM	PSEQ
NAF [108]	1.01	5.16	7.84	4.09	15.17	0.996	1.40
IR-MLP [147]	1.02	4.09	8.68	5.68	13.67	0.994	1.40
S2IR-GAN [141]	1.09	3.81	9.19	6.55	12.98	0.994	1.38
DeepNeRAP	**0.93**	**6.62**	**6.04**	**1.68**	**18.95**	**0.998**	**1.53**

$H(\omega)$ to express the same $h(t)$ at multiple time-frequency resolutions. We consecutively predict three $h(t)$ representations in frequency domain: $[H(\omega), H(\omega)_{2x}], H(\omega)_{4x}]$ by doubling the time and frequency dimension. We adopt 2D transposed convolution `TransConv` to 2× scale up (double) $H(\omega)$ resolution. $H(\omega)_{4x} = \texttt{TransConv}(H(\omega)_{2x})$, $H(\omega)_{2x} = \texttt{TransConv}(H(\omega))$. By adjusting the ISTFT parameters such as hop length and window size, we get three $[h(t), h(t)_{2x}, h(t)_{4x}]$ from the three learned time-frequency maps, respectively. During training, we deeply supervise the three neural RIRs [64]. In test, we merge the three neural RIRs to obtain the final neural RIR.

Loss calculator $\boldsymbol{\mathcal{L}}$ computes the discrepancy between $\hat{y}(t)$ and $y(t)$ in both time domain using the ℓ_2 loss) and frequency domain (with multi-resolution time-frequency ℓ_2 loss) [38].

4.2.5 Experiments and Result

Synthetic Dataset. We depend on SoundSpaces 2.0 [27] supported Matterport3D [24] dataset to collect the synthetic data. Matterport3D is a large-scale 3D indoor environment dataset with multiple rooms and complex furniture layout (with averaging size $> 100m^2$), so it contains sophisticated acoustic characteristics. We collect the dataset from all the 54 indoor scenes in the Matterport3D training set designed audio-visual navigation task [26]. For each scene, we randomly sample 100 navigable probing positions covering the whole scene, these positions serve as the positions the two agents can traverse to. By randomly pairing two positions (assume the two agents stand on), we call SoundSpaces 2.0 to simulate the corresponding RIR. Convolving the RIR with sine sweep sound gets the received reberverant sound. Finally, we have obtained 4000 probing data which is further split into 3000/1000 for train/test separately by guaranteeing no position pair in the test set is close enough to any position pair in the training set.

Real-world Dataset. We adopt MeshRIR S32-M441 dataset [90], which contains 32 source positions and 441 receiver positions (14 k data points, with 10k/4k split for train/test). The data collecting room dimension is 7.0 m × 6.4 m × 2.7 m. The

sampling rate is 48kHz and RIR length is 32768 points. For this dataset, we create the sine sweep signal to match the 48 kHz sampling frequency. We predict the same RIR map size but adjusting ISTFT parameters to get longer RIR length in time domain.

There are two main evaluation aspects for measuring the quality of the estimated neural RIR, namely: (i) directly comparing the predicted neural RIR with ground truth RIR, which helps to understand how well the learned primitive approximates the room acoustics primitive, and (ii) comparing the neural RIR effected sound, which helps to test how the learned primitive performs in a real acoustic environment. We adopt VCTK [191] anechoic speech dataset uttered by 110 English speakers with various accents. By convolving RIR (either ground truth RIR or learned neural RIR) with the anechoic speech, we get the corresponding RIR effected (reverberant) speech sound.

For evaluating RIR, we incorporate six metrics, three of which are evaluations on time domain and the other three in the frequency domain. In time domain, we use: (i) **t-MSE**, measuring the difference of the predicted neural RIR and ground truth RIR in time domain with mean square error, (ii) **SDR** (signal-to-distortion ratio) in which, in accordance with the metric outlined in [147], we also report SDR to appraise the fidelity of the predicted neural RIR in comparison to the ground truth RIR, and (iii) $\mathbf{T_{60}}$ **Error**, where T_{60} indicates the time to decay by 60 dB; we measure the T_{60} difference between ground truth RIR and predicted neural RIR. In frequency domain, we convert RIR to time-frequency 2D magnitude map of size 256×256, and evaluate on the magnitude map using: (i) **f-MSE**, we compute mean square error between ground truth magnitude map and predicted neural magnitude map, (ii) **PSNR** (Peak Signal-to-Noise Ratio) quantifying the quality of the magnitude map within the frequency domain (and is widely recognized for its applicability), and (iii) **SSIM** (structural similarity index measure), a perception-based metric which offers insight into the perceptual similarity between the magnitude map originating from the ground truth RIR and that derived from the predicted neural RIR. For the evaluation of sound influenced by RIR, we augment our assessment framework with the **PESQ** (perceptual evaluation of speech quality [148]) metric (using the speech convolved with the true RIR as reference) to have a human-centric perspective on perceptual similarity.

Currently there are no existing methods sharing exactly the same problem setting with our framework. Although sharing the same focus on neural RIR prediction, they largely differ in three aspects: 1. if they require ground truth RIR to train their model, 2. if they require extra prior knowledge of the room scene acoustic

Table 4.2: Quantitative Result on MeshRIR dataset. t-MSE: 10^{-8}, f-MSE: 10^{-2}.

Method	Neural RIR						Speech
	t-MSE	SDR	T_{60} Error	f-MSE	PSNR	SSIM	PSEQ
NAF [108]	2.21	5.36	7.44	4.01	16.19	0.996	1.54
IR-MLP [147]	2.32	4.22	7.87	5.43	14.39	0.995	1.51
S2IR-GAN [141]	2.55	4.27	8.21	6.32	13.88	0.994	1.41
DeepNeRAP	**1.13**	**7.31**	**5.01**	**1.27**	**20.15**	**0.999**	**1.77**

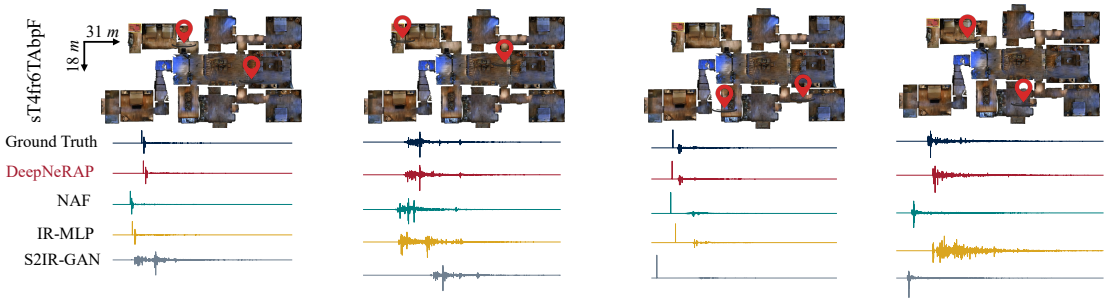


Figure 4.4: Visualization of learned neural RIRs in time domain in one room scene. The source/receiver position is denoted by the red position logo. The positions are different from the positions in the training data.

properties to train their model, and 3. the way to predict RIR (either in time domain or frequency domain, RIR length). For meaningful comparison, we compare with three spatial-position input based methods with appropriate modifications so as to be suitable for our setting. 1. **NAF** [108], a work that is most similar to ours. The main difference is that NAF requires access to massive RIR to train its model and it predicts binaural RIR for relatively small room scenes (Replica Dataset [161]). We modify it to accept two positions and predict monoaural RIR. 2. **IR-MLP** [147], we modify it to accept two positions as input and output the neural RIR. 3. **S2IR-GAN** [141] which is an encoder-decoder architecture, we also modify it to accept two positions and output the neural RIR.

We implement DeepNeRAP in Pytorch. We train DeepNeRAP on A40 GPU with Adam optimizer [85] with an initial learning rate 0.0005 but decays at every 50 epochs with decaying rate 0.5. We train all models for 300 epochs. We experimentally tried other optimizers such as stochastic gradient descent (SGD) and found it led to inferior performance. For the comparing prior methods, we adopt their proposed training strategy. We train each model for each dataset three times independently, and report the mean and variance.

The quantitative results on the synthetic Matterport3D dataset is given in Table 4.1 and on the real-world MeshRIR dataset is given in Table 4.2. From the two

Table 4.3: Ablation study on Matterport3D `17DRP5sb8fy` (t-MSE: 10^{-7}, f-MSE: 10^{-2}) room scene and real-world MeshRIR data (t-MSE: 10^{-8}, f-MSE: 10^{-2}). t-MSE (↓), SDR (↑), T_{60} Error (↓), PESQ (↑).

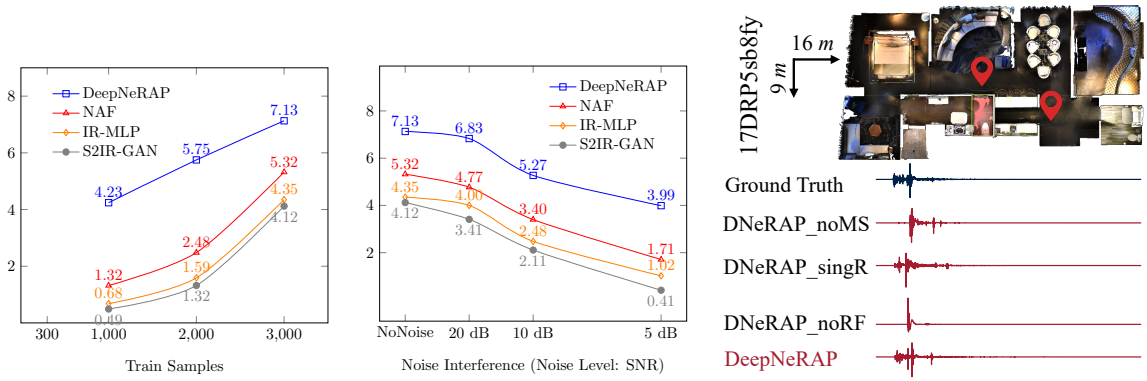
Variants	Matterport3D `17DRP5sb8fy`				MeshRIR Dataset			
	Neural RIR			Speech	Neural RIR			Speech
	t-MSE	SDR	T_{60} Er	PESQ	t-MSE	SDR	T_{60} Er	PESQ
DNeRAP_noRF	1.02	3.10	7.89	1.32	2.54	4.01	8.12	1.43
DNeRAP_noMS	1.01	3.14	7.90	1.36	2.50	4.00	8.13	1.40
DNeRAP_singR	0.94	4.78	7.23	1.51	2.10	5.21	6.33	1.52
DNeRAP_noPE	0.96	4.98	7.45	1.48	2.08	5.24	6.30	1.49
DeepNeRAP	**0.90**	**7.13**	**6.41**	**1.63**	**1.13**	**7.31**	**5.01**	**1.77**

Table 4.4: Ablation study on Matterport3D `17DRP5sb8fy` room scene. t-MSE: 10^{-7}, f-MSE: 10^{-2}

Method	Neural RIR						Speech
	t-MSE	SDR	T_{60} Error	f-MSE	PSNR	SSIM	PSEQ
NDeRAP_noRF	1.02	3.10	7.89	6.85	14.33	0.993	1.31
NDeRAP_noMS	1.01	3.14	7.90	6.50	15.01	0.994	1.32
NDeRAP_singR	0.94	4.78	7.23	3.01	17.90	0.995	1.49
NDeRAP_noPE	0.96	4.98	7.45	3.29	18.18	0.995	1.51
DeepNeRAP	**0.90**	**7.13**	**6.41**	**1.70**	**20.33**	**0.998**	**1.67**

tables, we can clearly see that our proposed DeepNeRAP outperforms all the comparison methods across all evaluation metrics significantly and consistently. In direct neural RIR evaluation, DeepNeRAP receives much higher scores in SDR, PSNR and SSIM, and much lower score in t-MSE, f-MSE and T_{60} error. In terms of neural RIR effect test on speech, DeepNeRAP has achieved higher PESQ score than the three comparing methods, showing the superiority of our learned neural RIR when taking effect on real-world speech dataset. Moreover, we have noticed all methods have achieved slightly better performance on MeshRIR dataset than on Matterport3D dataset. We hypothesize that this is due to the fact the meshRIR dataset is collected in a much smaller and simpler indoor environment and the data in MeshRIR dataset is much more densely collected than the way we used to collect on the Matterport3D dataset. Moreover, owing to the dense sampling data collection strategy in MeshRIR, we have more training data (10k) for MeshRIR than the data (3k) in Matterport3D room scenes.

We further provide qualitative visualizations of the neural predicted RIR across five room scenes in Fig. 4.4. From these figures, we can see that DeepNeRAP is capable of predicting neural RIRs that best match the ground truth RIRs, even under complex room scenes and arbitrary source and receiver positions. Comparing with



(a) SDR variation against training sample (left) and noise interference test (right).

(b) Three DeepNeRAP variants learned neural RIRs Vis..

Figure 4.5: Ablation Study Quantitative Result Report.

the other three methods, we observe that DeepNeRAP better encodes important perceptual properties of reverberation, such as the correct time delay between the source and receiver (alignment with the ground truth RIR), and the diffuse reverberation tail, which contributes most to human perception of reverberation [172]. The ability of DeepNeRAP to more accurately model late reflections, that contribute to diffuse tails, is confirmed by the lower T_{60} errors in Tables 4.1 and 4.2.

We first want to figure out the performance under different training samples or noise interference, which is important to show the robustness of DeepNeRAP. To this end, we run experiment on the scene `17DRP5sb8fy` by either varying the training samples (1,000/2,000/3,000) or adding white noise (the noise level is measured by signal-to-noise ratio, SNR, in dB). We report the SDR variation in Fig. 4.5a, from which we can see that 1) while all methods have observed performance drop with fewer training samples, DeepNeRAP still far outperforms all other methods. It thus shows DeepNeRAP is capable of learning better neural room acoustics primitive with less data. 2) adding noise leads to performance drop, and DeepNeRAP suffers the least from the noise by significantly outperforming all other methods.

We then do six ablations on the room scene with id:`17DRP5sb8fy` and real-world MeshRIR data to assess the necessity of each component in DeepNeRAP.

1. No Room Acoustic Feature $\boldsymbol{\mathcal{M}}$ Learning. We validate if involving a learnable grid feature is necessary; this variant is denoted $\boldsymbol{\mathcal{M}}$ (DNeRAP_noRF).

2. No Multi-Scale Feature Aggregation. In Eqn. 4.4, we adopt a position-aware multi-scale feature aggregation to relate a position to the global room scene. We test one variant without multi-scale aggregation (DNeRAP_noMS).

3. Single resolution RIR Prediction. In Eqn. 4.6, we jointly predict three neural RIR maps in frequency domain. To understand the implications of this choice, we test a variant by just predicting a single resolution (of size 128×128) neural RIR map (DeepNeRAP_singR).

4. No Position Encoding. In Eqn. 4.5, we introduce position encoding to emphasize each position's individuality. We test one variant without position encoding (DNeRAP_noPE).

The quantitative results are given in Table 4.3 for t-MSE, SDR, and T_{60}, Tables 4.4 and 4.5. The tables collectively reveal that all four ablations exhibit a decline in performance. Specifically, DNeRAP_noRF shows the a significant drop on both datasets, underscoring the need to incorporate a learnable grid. DNeRAP_noMS leads to a notable decrease highlighting the benefits of position-aware multi-scale feature aggregation. Reduced performance is also observed for DeRAP_singR and DNeRAP_noPE emphasizing the necessity of multi-resolution neural RIR learning and position encoding. Visualizations of the neural RIRs for these variants in Fig. 4.5b, clearly demonstrating their inferior quality.

Table 4.5: Ablation study on MeshRIR dataset. t-MSE: 10^{-8}, f-MSE: 10^{-2}.

Method	Neural RIR						Speech
	t-MSE	SDR	T_{60} Error	f-MSE	PSNR	SSIM	PSEQ
NDeRAP_noRF	2.54	4.01	8.13	6.65	14.01	0.993	1.43
NDeRAP_noMS	2.50	4.00	8.13	6.33	14.12	0.994	1.40
NDeRAP_singR	2.10	5.21	6.33	4.34	18.01	0.996	1.52
NDeRAP_noPE	2.08	5.24	6.30	4.12	17.11	0.996	1.49
DeepNeRAP	**1.13**	**7.31**	**5.01**	**1.27**	**20.15**	**0.999**	**1.77**

4.2.6 Conclusions and Limitations

In this work, we propose a novel framework DeepNeRAP, to learn sound propagation primitive in a self-supervised way using a data collection approach that is easy to execute. Our approach circumvents the difficulty in modeling room impulse response and we show its superiority on both synthetic data and real-world data. The main limitation is that we assume the two probing agents can actively explore to all areas within a limited step budget, which in real scenario may require implementing efficient exploration algorithms.

4.3 Conclusion

In this chapter, we try to learn a continuous spatial audio neural rendering field that can implicitly model the spatial audio propagation process from arbitrary audio source position to audio receiver position. Such implicit neural field is learned from training much more readily accessible way and the whole learning process is supervised by common physical principles in room acoustics. As a result, the learned spatial audio neural rendering field is relatively interpretable and physically meaningful. Given the fast development of vision-based neural rendering field [119] in recent years, there is huge potential in the spatial audio-visual neural rendering research direction, especially in audio-visual dynamic environment.

Chapter 5

Spatial Audio-Visual Learning

5.1 Motivation Discussion

Throughout the research experience presented in Chapter 3 and Chapter 4, we have gained deep understanding in how to incorporate deep neural network into spatial audio processing to boost the performance of common spatial audio based tasks, as well as the mathematical and physical fundamentals underneath spatial audio. Equipped by these research experience, we further reached out to the spatial audio-visual learning. In this chapter, we focus on audio-visual based 3D sound source detection task (localize each sound source's $[x, y, z]$ coordinate and classify its semantic label). In our setting, we assume the sound source has no visual entity but lies on object's physical surface. It reflects some real-scenarios such as gas-leak and machinery malfunction. We show how the crossmodal visual signal can be exploited to assist one-the-surface sound source detection.

5.2 Multiview based Audio Source Detection

5.2.1 Introduction

In this work, we propose to accurately detect 3D sound sources by jointly exploiting multiview audio-visual cross-modal information. We assume sound sources lie on object's physical surface, constantly and repetitively emitting sounds independently, our goal is to pinpoint its 3D position and class label by "looking at and listening to" the joint visual-acoustic scene. Unlike previous works that assume that the sound is strongly correlated with a visual cue/object (*e.g.,* the sound comes from particular objects like a church bell, a train, or a clock) [170, 164, 74], we assume that the sound source is only weakly associated with vision. For example, the sound source is either

too small to be visually observable or the sound is coming from a novel object. There are a number of real and challenging application scenarios that meet this setting. For example, industrial gas leakage detection requires a robot to pinpoint a leak that shows no visual difference compared with a normal gas pipe - the only clue is the acoustic emission from the defect. Although we may have a rough estimation of 3D sound source position (*e.g.,* we may know the sound comes from a specific area based on some prior knowledge), how to precisely localize this within a local area remains a challenging task.

In this work, we propose to use an acoustic-camera to record the local area from multiple views. The acoustic-camera is a device equipped with a centered pinhole camera and four microphones in a uniform array. The camera and microphones are coplanar and synchronized so that they record the scene from different viewpoints with known camera poses. Similar acoustic-camera has been developed in commercial market [1] where the device is usually equipped with much more microphones (can have as many as more than 100 microphones) distributed coplanarly with the camera in more sophisticated pattern. Those extra introduced microphones naturally cause much more computational burden. At each viewpoint, the RGB image and the multi-channel microphone array signal are recorded simultaneously. The motivations for using multiview audio-visual data are two-fold: first, observing the scene both acoustically and visually from multiple viewpoints enables us to gain a diverse understanding of the sound source; second, multiview RGB images provide useful cues for localizing 3D sound sources. The fundamental idea is to use multiview RGB images to set an "on-the-surface" constraint. A 3D sound source's location when projected onto different

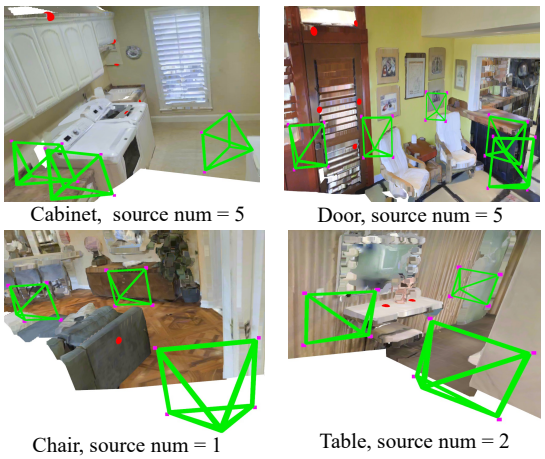


Figure 5.1: **Sound3DVDet Task Illustration**: Multiple 3D sound sources (red ball) are emitted by visually uninformative objects, we use an acoustic-camera device to record the multi-view, multi-modal visual-acoustic scene. Each recording consists of an RGB image at a known pose (green) and a four-channel microphone array (magenta).

[1] see https://soundcam.com/.

RGB image planes are "matching points" when this location lies on the object's surface. Any position shift off the surface (either below or above the surface) leads to the corresponding projections to be "non-matching" points (see Fig. 5.3).

Based on the multiview acoustic and visual recordings (see Fig. 5.1 for sample visualization), we propose Sound3DVDet, a novel 3D sound source localization framework that can efficiently handle arbitrary sources. Drawing inspiration from the Transformer architecture design [179] and the current popular set-based object detection methods [23, 185, 102], *Sound3DVDet* treats 3D sound source detection as a set-prediction problem. It directly predicts a set of 3D sound source queries from multiview acoustic-camera recordings, each query corresponds to a potential 3D sound source. To learn discriminative query representations, *Sound3DVDet* first initializes the 3D sound source queries from an individual microphone array sound signal by explicitly using the inter-channel phase difference. Then it refines these queries using a sequence of Transformer layers by improving the cross-modal consistency between acoustic cues and image matching. The final query representations are decoded into 3D sound source positions and class labels through a detection head neural network. During training, the predicted queries are matched with ground truth via bipartite matching [93] and the whole neural network is optimised by minimizing the discrepancy between prediction and ground truth. To further refine 3D sound sources' locations, we deeply supervise [97] the learning of queries arising from all intermediate layers of Transformer, including the initial queries from the microphone array recording (see Fig. 5.2).

Since there is no publicly available dataset suitable for our task, we use the SoundSpaces 2.0 [27] simulator to create a dataset with 6.2k samples. Experimental results show the our framework outperforms the comparing methods by 20%, 30% and 0.25 in mAP, mAR and mALE metrics, respectively. In summary, we make the three main contributions: **1.** We propose a novel task: 3D sound source detection from a moving acoustic-camera with known camera poses. The acoustic-camera jointly records microphone-array signals and RGB images. The sound source is assumed to lie on an object's physical surface, but may not be visually distinguishable. **2.** We propose **Sound3DVDet**, a novel framework to jointly harness a microphone array and RGB images to accurately detect 3D sound sources. **3.** We introduce a new dataset: *Sound3DVDet dataset*, using which we provide experiments using our model, demonstrating state-of-the-art results on sound source localization and classification.

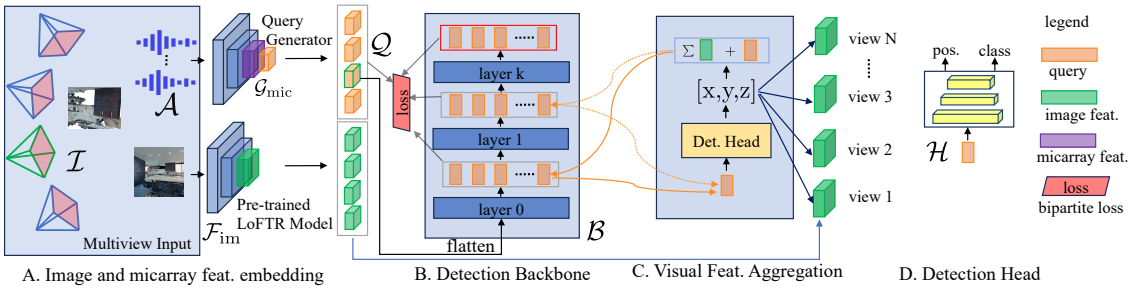


Figure 5.2: **Sound3DVDet Pipeline Illustration**: A. For each single view, we use a learnable sound source query generator to jointly obtain the microphone array signal feature embedding and initial sound source queries, pre-trained image model to get RGB image feature embedding (Sec. 5.2.2), respectively. Then, we randomly choose a reference view (left-most camera in green color) and flatten its initial learned sound source queries. The flattened sound source queries serve as Transformer's tokens and are fed to the detection backbone for further refinement (Sec. 5.2.2). In each intermediate layer in the backbone, we aggregate multiview visual sound source cues for each query by involving "on-the-surface" constraint. This is achieved by first using detection head $\boldsymbol{\mathcal{H}}$ to decode each query into world positions and then projecting this across the multiple views with the known camera poses (sub-figure B). We deeply supervise all sound source queries learning during training. During inference, we use the final queries to predict 3D sound sources.

5.2.2 Multiview based 3D Sound Source Detection

Problem Formulation In this chapter, we assume a 3D enclosed room environment with an arbitrary number of point sound sources lying on indoor objects' physical surface. These sound sources are constantly and repetitively emitting anechoic sound waveforms. The objects we use are commonly seen indoor objects, including furniture (chair, cabinet, table, *etc.*) and architectural structure (wall, door, ceiling, *etc.*). We also assume we have a rough estimation of the sound sources locations either from prior knowledge or other sound source detection techniques. For example, we may know the sound of gas leak comes from a particular wall in a specific room, because the gas pipes traverse along that wall. Moreover, we assume these sources have no apparent visually distinguishable characteristic, which means that we cannot directly detect them from images alone.

In this chapter, we introduce an **acoustic-camera** device to record the local acoustic-visual scene from different viewpoints with known camera poses, each single view recording consists of an RGB image and a microphone array acoustic signal. An acoustic-camera is a device consisting of a pinhole camera and a microphone array that records raw waveforms from each microphone. A microphone array consists of

a spatial arrangement of microphones. As sound propagates at roughly 330 m/s at room temperature, the received sound waveforms by any pair of microphones have a time-delay (or phase difference) due to their different distances to the sound source. Using the recorded multi-channel sound waveforms, the sound sources' spatial location and semantic class can be estimated. We use a small array (four microphones with a 10 cm spacing) in this work which is inexpensive and easy to use. This gives an azimuthal far-field angular uncertainty of approximately $10-15^{\circ}$ for frequencies in the range of 500 Hz to 2000 Hz with a sampling frequency of 22050 Hz - see e.g., [28, 8] for more details. Our aim is to use the movement of the acoustic-camera to precisely locate the positions of multiple sound-sources in 3D and their class labels.

Formally, multiview acoustic-camera recording is denoted $\boldsymbol{\mathcal{R}}_{\text{av}} = \{(\mathcal{A}_i, I_i, T_i)\}_{i=1}^{n}$, where $\mathcal{A}_i \in \mathbb{R}^{4\times w}$ is the i-th view of four-channel microphone array sound waveforms $\mathcal{A}_i = [a_{i1}, a_{i2}, a_{i3}, a_{i4}]$, $I_i \in \mathbb{R}^{C\times H\times W}$ is the i-th RGB image (of size $3 \times 512 \times 512$), $T_i \in \mathbb{R}^{3\times 4}$ is the i-th view camera pose (including both the intrinsic and extrinsic parameters), and n is the number of views. Further, let M be the number of static sound sources, expressed as $\mathcal{S} = \{(p_k, c_k)\}_{k=1}^{M}$, where $p_k \in \mathbb{R}^3$ indicates the 3D position: $p_k = [x_k, y_k, z_k]$ and $c_k \in \mathbb{Z}$ indicates the class label. Our goal is to design a model $\boldsymbol{\Theta}$ to detect 3D sound sources from multiview acoustic-camera recordings, that is:

$$\boldsymbol{\Theta}(\{(\mathcal{A}_i, I_i)|T_i\}_{i=1}^{N}) \rightarrow \mathcal{S}. \tag{5.1}$$

Motivated by [23, 185, 102], we treat 3D sound source detection/localization as a set prediction problem. Given multiview acoustic-recordings $\{(\mathcal{A}_i, I_i)\}_{i=1}^{n}$, our *Sound3DVDet* model $\boldsymbol{\Theta}$ learns a set of sound source *queries*[2] for a reference view (*e.g.*, the i-th view) $\mathcal{Q}_i = \{Q_{i1}, Q_{i2}, \cdots, Q_{iK}\}$. Each query $Q_{ik} \in \mathbb{R}^d$ (d is the feature dimension) is a potential 3D sound source embedding that can be fed to a detection head network $\boldsymbol{\mathcal{H}}$ to be decoded into its corresponding 3D position and class label. During training, we adopt bipartite matching (a.k.a Hungarian algorithm) [93] to find the best assignment between queries and ground truth sound sources, and optimize the whole neural network Θ with the loss incurred by this bipartite matching. During inference, the predicted sound source queries are directly used to output 3D sound sources; we do not assume to use any post-processing (*e.g.* non-maximum suppression (NMS) for detection redundancy removal [146, 101, 145]).

[2]Here and in the subsequent sections, we go by the nomenclature in [23] and call the target variables as *queries*, which correspond to neural representations of the sound sources.

Our *Sound3DVDet* fully embraces the sound source cues arising from a single view microphone array signal and multiview RGB images to detect 3D sound sources. From our empirical observations, we find that usually the microphone array signals from a single view can provide coarse estimations to the sound source locations. Leveraging this observation, we propose to learn initial sound source queries from such a single view of the microphone array signals by a query generator network $\boldsymbol{\mathcal{G}}_{\text{mic}}$, and subsequently optimize these initial queries through a backbone network $\boldsymbol{\mathcal{B}}$. The backbone neural network is a stack of L Transformer encoder layers into which the sound source queries are input as tokens, which then sequentially pass through these L layers. The sound source queries output by a preceding encoder layer is refined by the subsequent encoder layer via: 1) inter-query interaction through Transformer multihead self-attention (MHSA) and feed-forward networks (FFN) and 2) the visual source position cues aggregated from the multiview recordings. Since the same sound source queries are passed through the entire neural network to be refined gradually, we propose to deeply supervise [63, 97] the queries arising from the different intermediate layers. The deep supervision means we supervise all intermediate queries by feeding each of them to detection head network to get their decoded spatial position and semantic label. The loss is then computed for all spatial positions and semantic labels decoded from all intermediate queries. We experimentally find such deep supervision enables the neural network to learn better sound source query representations.

In summary, *Sound3DVDet* $\boldsymbol{\Theta}$ consists of a source query generator $\boldsymbol{\mathcal{G}}_{\text{mic}}$, a detection head $\boldsymbol{\mathcal{H}}$, a backbone $\boldsymbol{\mathcal{B}}$ and an RGB image feature extractor $\boldsymbol{\mathcal{F}}_{\text{im}}$, $\boldsymbol{\Theta} = (\boldsymbol{\mathcal{G}}_{\text{mic}}, \boldsymbol{\mathcal{H}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{F}}_{\text{im}})$. While $\boldsymbol{\mathcal{G}}_{\text{mic}}$, $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mathcal{B}}$ are learnable neural networks, $\boldsymbol{\mathcal{F}}_{\text{im}}$ is pretrained RGB image feature extraction model. Figure 5.2 shows the pipeline, which works as:

1. At each iteration, *Sound3DVDet* takes as input a multiview acoustic-camera recording $\{(\mathcal{A}_i, I_i)\}_{i=1}^{n}$. The multiview images I are fed to $\boldsymbol{\mathcal{F}}_{\text{im}}$ to get the image feature maps. The multiview microphone array signals $\mathcal{A}$ are fed to $\boldsymbol{\mathcal{G}}_{\text{mic}}$ to obtain initial sound source queries $\mathcal{Q}_{\text{init}}$.

2. Go through all initial queries, each time select one reference view $\mathcal{Q}_{\text{init},r}$ (*e.g.* the r-th view, $r = 1, \cdots, N$) and pass it to $\boldsymbol{\mathcal{B}}$ for refinement. For each intermediate output in $\boldsymbol{\mathcal{B}}$, we aggregate source cues from multiview RGB images.

3. During training, we deeply supervise all source queries: 1) from query generator $\boldsymbol{\mathcal{G}}_{\text{mic}}$. 2) from intermediate queries in $\boldsymbol{\mathcal{B}}$. 3) from the final output queries in $\boldsymbol{\mathcal{B}}$.

During inference, we use the final output queries in $\boldsymbol{\mathcal{B}}$ to predict 3D sound source locations and their labels.

A single-view microphone array signal (four-channel sound waveforms) contains enough information for estimating a 3D sound source's spatial position and class label. Specifically, the class label is encoded in each sound-channel waveform's time-frequency (TF) representation and the spatial position is encoded in the inter-channel phase difference (a.k.a time-delay). Following the common practice [2, 20, 57], for each single-channel one dimensional sound waveform, we first apply the short time Fourier transform (STFT) to transform it into a 2D TF representation and then convert it to log-mel scale. To extract the inter-channel phase difference, we compute the generalized cross-correlation phase transform (GCC-Phat [12], represented as a 2D map) feature between any microphone pair. GCC-Phat has been widely used for microphone array signals [20, 2, 178, 21]. In our case, we create 6 GCC-Phat maps as we compute it for all potential microphone pairs from the four microphones ($\binom{4}{2} = 6$). By concatenating the 6 GCC-Phat maps with the four TF representation maps, we obtain a 10-channel 2D feature map, $F_{\text{mic}} \in \mathbb{R}^{10 \times H_1 \times W_1}$ (in our case, $H_1 = W_1 = 256$).

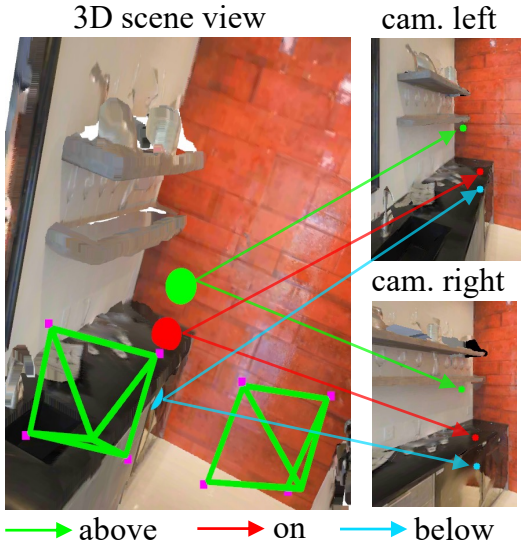


Figure 5.3: **Visual On-the-Surface Constraint:** While a 3D sound source's projections onto images are visually close matching points if the source lies on the surface (red ball), the projections become non-matching points once the source is shifted to either above (green) or below (blue) the surface.

The source query generator $\boldsymbol{\mathcal{G}}_{\text{mic}}$ takes as input the 10-channel feature map F_{mic}, and applies a sequence of 2D convolutions to consecutively reduce the feature spatial resolution while increasing their channel size. The resolution reduction is achieved by setting the 2D convolution *stride=2*. In our case, we treat the last layer output as the initial source queries $\mathcal{Q}_{\text{init}}$. At the same time, we take the penultimate layer output as the microphone array signal feature embedding $f_{\mathcal{A}_i}$. We will use such microphone array signal embedding in one of our ablation studies to test if further aggregating multiview acoustic signal improves the performance.

$$\mathcal{Q}_{\text{init},i}, f_{\mathcal{A}_i} = \boldsymbol{\mathcal{G}}_{\text{mic}}(F_{\text{mic},i}),\ F_{\text{mic},i} \leftarrow \mathcal{A}_i,\ i = 1, \cdots, N \tag{5.2}$$

where $\mathcal{Q}_{\text{init},i}$ is the i-th frame sound source queries. During training, we iterate over all views, each time treat the investigating view initial queries as the reference queries $\mathcal{Q}_{\text{init},r}$ and flatten into tokens before feeding to backbone $\boldsymbol{\mathcal{B}}$ for further refinement.

Visual On-the-Surface Constraint

Since we do not assume the 3D sound source has any obvious visual entity in each single view image, we cannot directly detect the sound source in each image. Thus, to make audio-visual learning feasible and use the multiview visual information, we propose to impose an "on-the-surface" constraint on the sound sources – that is, the sound sources are assumed to lie on the physical surface of some object seen in RGB images from all views. Such an assumption allows for an elegant formulation of an audio-visual location consistency. Specifically, if a 3D sound source lies on an object's surface, its projections onto the multiview RGB images are "matching points" [185, 30, 102]. Any position shift away from the surface (either below or above the surface) makes the projections less likely to be "matching points" (see Fig. 5.3 for illustration). The task then becomes finding a way that is capable of accurately measuring the "matchness" for projections from multiview RGB images.

Unlike traditional image matching methods [201, 189, 174, 152, 192] that focus on finding corresponding points in discriminative image regions, we proceed in the opposite way to decide the "matchness" for multiview 2D pixels from the projections of the predicted locations of the sources. Furthermore, these 2D pixels can lie in regions that may be textured, discriminative, or homogeneous in the 2D images. Therefore, the resulting RGB image embedding needs to be representative enough in providing matching information across multiple views, regardless of the positions of the matching points. To this end, we depend on the pre-trained feature matching model LoFTR [163] to obtain feature embedding for each RGB image. LoFTR [163] is trained for feature matching in a coarse to fine manner, it is capable of finding matching points even in texture homogeneous regions. Benefiting from this advantage, we are able to reasonably measure the matchness for projections on texture homogeneous area (like walls). We extract its coarse-level representation as the initial embedding (of size $256 \times 64 \times 64$), and further introduce an extra Fully-connected layer (FC) to further adjust the embedding to fit our scene dataset (also increase the feature size from 256 to 512),

$$f_{\mathcal{I}} = \text{FC}(\text{LoFTR}(\mathcal{I})) \tag{5.3}$$

where $f_I \in \mathbb{R}^{512\times64\times64}$ ($I \in \mathcal{I}$). $\boldsymbol{\mathcal{F}}_{\text{im}} = \text{FC}(\text{LoFTR}(\cdot))$. We find that adopting the pretrained model for feature matching gives better performances than using the ImageNet [40] pretrained model (*e.g.*, ResNet50 [61], see Experiment).

Transformer-based Detection Backbone

The initial source queries from the r-th reference view are fed to the detection backbone $\mathcal{B}$ for further learning. The backbone network $\mathcal{B}$ consists of L standard Transformer encoder layers, each of which contains a multi-head self-attention (MHSA) and feed forward network (FFN). The queries, working as Transformer tokens, are optimized in two ways: (i) for a single view source, the multihead attention allows the queries to interact among each other allowing implicitly modeling of the dependency and audio dynamics of sound sources within one view, and (ii) the cross-view consistency, allowing all queries arising from Transformer intermediate layers to actively aggregate source cues from crossmodal multiview RGB images,

$$\mathcal{Q}_{l+1,r} = \boldsymbol{\mathcal{B}}_l(\mathcal{Q}_{l,r}|f_{\mathcal{I}}, \boldsymbol{\mathcal{H}}, T), l = 1, \cdots, L-1 \tag{5.4}$$

The source detection head $\boldsymbol{\mathcal{H}}$ decodes any query feature (*e.g.*, $q_l \in \mathcal{Q}_l$) into its designated sound source 3D position p and class label c,

$$[p_{i,k}, c_{i,k}] = \boldsymbol{\mathcal{H}}(\mathcal{Q}_{i,k}), \quad k = 1, \cdots, m \tag{5.5}$$

where $p_{i,k}$ and $c_{i,k}$ indicates the k-th predicted sound source 3D position expressed in the i-th camera coordinate system, $c_{i,k}$ is the class label. In our implementation, $\boldsymbol{\mathcal{H}}$ consists of two parallel fully-connected layers to regress 3D position and predict class label separately.

We aggregate multiview RGB images informed sound source cues to improve the sound queries learning. Such aggregation encourages the queries to predict accurate sound source 3D positions because it directly uses the decoded 3D position (via the detection head $\boldsymbol{\mathcal{H}}$ in Eqn. 5.5) to aggregate source cues. Specifically, given one query $\mathcal{Q}_{l,k}$ arising from the k-th query feature in the l-th detection backbone layer in Eqn. 5.4, we first apply the detection head $\boldsymbol{\mathcal{H}}$ to decode $\mathcal{Q}_{l,k}$ into its corresponding 3D position $p_{l,k}$ expressed in the reference camera coordinate system (the r-th camera system), and then project it to j-th novel view RGB image plane to get its 2D position $[u_{x,j}, u_{y,j}]$ through the camera poses. Afterwards, we adopt bilinear interpolation ϕ to index the cross-view sound source visual clue $f_{I,r\leftarrow j}$ based on $[u_{x,j}, u_{y,j}]$.

$$f_{I,r\leftarrow j} = \phi_{\text{bilinear}}(f_{I_j})_{[u_{x,j},u_{y,j}]}, \quad j = 1, \cdots, N. \tag{5.6}$$

If $[u_{x,j}, u_{y,j}]$ is within the j-th RGB plane, we adopt bilinear interpolation in Eqn. 5.6 to get the feature, otherwise the feature is set 0. Moreover, since the spatial resolution of RGB feature embedding map is much smaller than the original RGB image (RGB image size is 512×512), we follow DETR3D [185] to normalize the valid $[u_{x,j}, u_{y,j}]$ (those lie within the RGB image plane) into $[-1, 1]$ before performing bilinear interpolation. Given all the aggregated multiview RGB image informed source clue features, we merge them into the query through elementwise-add before feeding to next Transformer layer,

$$\mathcal{Q}_{l,k} \leftarrow \mathcal{Q}_{l,k} + \sum_{j=1}^{N} f_{I,r \leftarrow j} \tag{5.7}$$

Specifically, given one query $\mathcal{Q}_{l,k}$ arising from the k-th query feature in the l-th detection backbone layer in Eqn. 5.4, we first apply the detection head $\boldsymbol{\mathcal{H}}$ to decode $\mathcal{Q}_{l,k}$ into its corresponding 3D position $p_{l,k,i}$ in the i-th reference camera coordinate system, which is then projected into j-th ($j \neq i$) novel view camera coordinate system $p_{l,k,j} = T_i p_{l,k,i}$. We finally acquire 2D position in image plane $[u_x, v_y]$ by performing perspective projection on $p_{l,k,j}$ with known intrinsic parameters of i-th camera.

5.2.3 Deeply Supervise All Intermediate Queries

In *Sound3DVDet*, the source queries repetitively appear at different intermediate layers (see Fig. 5.2). We propose to deeply supervise all intermediate sound source queries learning by directly feeding all of them to detection head $\boldsymbol{\mathcal{H}}$ to predict 3D sound source's position and class label, respectively. We then use bipartite matching [93] loss to supervise all predictions learning. Specifically, we deeply supervise three main sound source queries: the initial queries given by query generator $\mathcal{G}_{\text{mic}}$; intermediate queries from each of the L layers in the backbone network $\boldsymbol{\mathcal{B}}$ and the final queries from the last layer of $\boldsymbol{\mathcal{B}}$.

For bipartite matching, since the number of sound source queries is usually larger than the ground truth sound source number ($M < K$), we explicitly pad `no-source` category $\varnothing$ to the ground truth sound sources to reach the number K. Bipartite matching is then applied to find a one-one correspondence σ^* between prediction and ground truth by taking sound source position closeness and label classification score into account, $\sigma^* = \arg\min_{\sigma \in \mathcal{P}} \sum_{k=1}^{K} -1_{\{C_k \neq \varnothing\}} \hat{p}_{\sigma(k)}(C_k) + 1_{\{C_k = \varnothing\}} \mathcal{L}_{\text{pos}}(P_k, \hat{P}_{\sigma(k)})$, where $\hat{p}_{\sigma(k)}$ and $\hat{P}_{\sigma(k)}$ indicate the predicted label classification probability and 3D position, respectively. $\mathcal{P}$ denotes the permutation set. $\mathcal{L}_{\text{pos}}$ is the L_1 loss for position regression. After finding the best correspondence σ^*, we can then compute the final

set prediction loss by combining the classification cross-entropy loss and L_1 position regression loss $\mathcal{L} = \sum_{k=1}^{K} -\log \hat{p}_{\sigma^*(k)}(C_k) + 1_{\{C_k=\varnothing\}}\mathcal{L}_{\text{pos}}(P_k, \hat{P}_{\sigma^*(k)})$.

$$\mathcal{L} = \overbrace{\mathcal{L}(\mathcal{Q}_{\text{init}})}^{\text{initial queries}} + \overbrace{\sum_{l=1}^{L-1}\mathcal{L}(\mathcal{Q}_{\boldsymbol{\mathcal{B}}_l})}^{\text{interm. queries}} + \overbrace{\mathcal{L}(\mathcal{Q}_{\boldsymbol{\mathcal{B}}_L})}^{\text{final queries}} . \tag{5.8}$$

5.2.4 Experiments

Dataset Creation: Given the novelty of our problem setup, currently we do not have any publicly available datasets that fit our experimental setup. To this end, we use the SoundSpaces 2.0 [26] simulator to synthesize a new dataset. We load Matterport3D dataset [24] in SoundSpace 2.0. Matterport3D contains large-scale (with average room area >100 m^2) and complex indoor room scenes, with which we are able to synthesize data with large visual and acoustic diversity. Specifically, we place multiple point sound sources (source emits sound waveform isotropically) on the surface of 6 commonly seen objects: *wall*, *chair*, *table*, *door*, *ceiling*, *cabinet*. Each sound source emits sound independently. Around the object, we let an agent holding an acoustic-camera to record the object from multiple viewpoints. In our implementation, the multiview acoustic-cameras are recorded roughly at the same height because the agent holds the acoustic-camera at a fixed height position (in our case, at a height of 1.5 m).

Specifically, given an object, we randomly place n ($1 \leq n \leq 10$) sound sources on its surface and ensure any two sources are at least 0.3 m apart (no overlap). Each sound source randomly emits one sound class out of five sound class corpus: *telephone-ring*, *siren*, *alarm*, *fireplace* and *horn-beeps*. The sampling frequency is 21k Hz. By varying the number of sound sources, views and sound classes, we can flexibly test their individual impact on sound source detection performance. To further test the impact of visual discriminativeness of the RGB image on detection performance, we divide the sound sources into two main categories according to their position in the images: texture-homogeneous area in which the sound source lies around a textured homogeneous area like wall and table surface, texture-discriminative regions in which the sound source lies around regions like corners. In summary, we have created 5,000/1,250 for train/test, respectively.

Evaluation Metrics: Motivated by existing works on sound event detection [57, 112, 68, 135] and 2D/3D object detection [23, 185, 100], we propose three main evaluation metrics: mean average precision (mAP) and mean average recall (mAR) and mean localization error (ER), to evaluate the performance from various perspectives.

Table 5.1: Quantitative result 1. across all object categories and sound classes (left) comparison between texture homogeneous and texture discrinativative projections of sound sources (right).

Methods	Overall Result			Texture Homogeneous			Texture Discriminative		
	mAP	mAR	mALE	mAP	mAR	mALE	mAP	mAR	mALE
SELDNet [2]	0.101	0.531	0.912	0.107	0.532	0.910	0.100	0.528	0.934
EIN-v2 [20]	0.111	0.612	0.877	0.115	0.620	0.882	0.117	0.600	0.862
SoundDoA [65]	0.123	0.701	0.820	0.125	0.703	0.821	0.122	0.698	0.820
Sound3DVDet	**0.308**	**0.998**	**0.588**	**0.308**	**0.996**	**0.585**	**0.293**	**0.993**	**0.591**

It is worth noting that our *Sound3DVDet* directly outputs all sound sources without any post-processing involved.

We first evaluate within each class separately. Given the detected sound source set and ground truth set for a particular class, we first apply bipartite matching algorithm [93] to assign each detected sound source to one ground truth sound source (in some cases, some detections remain unassigned if the detections outnumber the ground truth, and vice versa). After assignment, a detection is a true positive iff it is within a distance threshold with its assigned ground truth, otherwise a false positive. Given a particular threshold, we can accordingly compute the *precision* and *recall*. In our case, rather than fixing one distance threshold, we instead compute across a set of discrete thresholds and further get the average precision (AP) and average recall (AR) by averaging across all distance thresholds. Finally, we average across all classes to get the mean average precision (mAP) and mean average recall (mAR). mAP and mAR are two widely adopted evaluation metrics in object detection [23, 100, 185, 68]. In our case, we find that mAP and mAR are relatively dependent on the distance threshold we choose, they do not directly give an understanding how close the predicted sound sources are to the ground truth. To this end, we further embrace the localization error (LE) metric that are initially used in sound event detection [112, 135, 57]. LE builds on true positive detections, but it goes further to consider the exact the distance between prediction and ground truth. Following mAP and mAR, we first compute average LE across all distance thresholds and finally compute mean average LE (mALE) across all classes. In this work, we adopt three distance thresholds: $[0.5\ m, 0.8\ m, 1.2\ m]$.

Comparison Methods: There are no existing methods that directly work on our proposed problem. We thus propose to compare with three typical microphone array signals based sound source detection baselines: SELDNet [2], EIN-v2 [20] and SoundDoA [65]. SELDNet serves as the baseline for various microphone array based sound

Table 5.2: Ablation Study on quantitative result on overall quantitative result 1. across all object categories and sound classes (left), 2. on the comparison between texture homogeneous and texture discriminative projections of sound sources (right). SD here means S3DVDet.

Methods	Overall Result			Texture Homogeneous			Texture Discrim.		
	mAP	mAR	mALE	mAP	mAR	mALE	mAP	mAR	mALE
SD_ResNet50	0.236	0.977	0.580	0.235	0.953	0.583	0.240	0.943	0.579
SD_noDeepS	0.167	0.994	0.616	0.171	0.988	0.617	0.164	0.977	0.613
SD_noMVSup	0.253	0.981	0.603	0.254	0.952	0.608	0.168	0.980	0.607
SD_mvSound	0.264	0.994	0.592	0.274	0.993	0.590	0.253	0.984	0.593
SD_wMVIS	0.289	0.997	0.595	0.297	0.994	0.593	0.280	0.989	0.597
Sound3DVDet	0.308	0.998	0.588	0.308	0.996	0.585	0.293	0.993	0.591

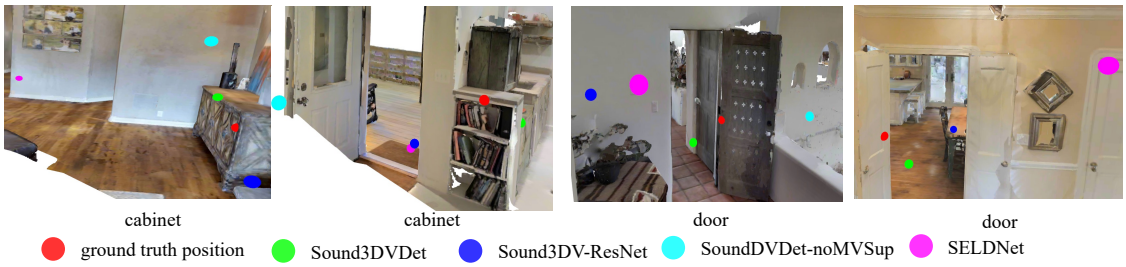


Figure 5.4: **Qualitative Detection Result Visualization**: We visualize the position of one detected sound source position by different methods as well as its ground truth position. We recommend to zoom in for better visualization.

source detection, it combines CNN and GRU [33] to detect sound sources; EIN-v2 [20] and SoundDoA [65] are two more recent work, they further adopt Transformer [179] and permutation invariant training [193] to detect sound source.

Implementation Details We implement *Sound3DVDet* with PyTorch [129] and train it on NVIDIA A40. The model parameter size is 19.9 M. We adopt the AdamW optimizer [104], with an initial learning rate 0.0001 and decays every 100 epochs with a decaying rate 0.5. We train each model variants three times independently, and report the mean and variance for each metric separately. We train all models 100 epochs. We compare with them to test the necessity of involving RGB image and further multiple view recording for 3D sound source detection.

Experiment Results Our quantitative results are given in Table 5.1, from which we can clearly observe that *Sound3DVDet* outperforms all the three comparing methods by a large margin. On average, *Sound3DVDet* outperforms the three comparing methods by 20% on mAP, 30% on mAR and 0.25 on mALE. It thus shows that our proposed framework works well on 3D sound source detection. We also note that all methods have achieved a much higher mAR than mAP, which means set-based

prediction strategy is capable of predicting enough sound sources in each camera view.

The performance in terms of texture difference is shown in Table 5.1. We can observe from this table that 1) the three comparing methods show inconsistency w.r.t. the texture difference, which is reasonable because they do not explicitly depend on vision information to detect 3D sound sources; 2) *Sound3DVDet* can still achieve reasonably good performance in texture homogeneous area with small performance drop.

Ablation Studies. We present five ablation studies. The quantitative results are provided in Table 5.2.

1. Pre-trained Image Matching Feature VS. Classification Feature. As an alternative, we adopt ImageNet [40] pre-trained ResNet50 [61] (S3DVDet_ResNet50) to replace LoFTR [163]. This replacement helps to test what RGB image feature is better for providing "on-the-surface" constraint. From Table 5.2 and 5.2, we can see that such replacement leads to obvious performance drop in mAP (≈ 0.6) and mAR (≈ 0.2). It thus shows pre-trained image matching model is better at setting "on-the-surface" constraint, especially in texture homogeneous area. This is also echoed in Table 5.2, in which we have observed performance drop in detection in texture homogeneous area.

2. Without Deep Supervision When removing the deep supervision from the initial sound source queries and detection backbone intermediate layers (S3DVDet_noDeepS), we have observed significant performance drop (mAP ≈ 1.4, mAR ≈ 0.02, mALE ≈ 0.3), which shows deep supervision strategy is vital to enforce the whole framework to learn more representative sound source queries representation.

3. No Multiview Supervision in which we just rely on single view (microphone array and RGB image) to predict 3D sound sources with cross-view visual feature aggregation (3DVDet_noMVSup). However, the deep supervision module is still kept. We have observed significant overall performance drop. The performance drop becomes significant when the sound sources lie around texture discriminative area. It thus shows multiview supervision is an essential component of *Sound3DVDet*.

4. With Multiview Sound, in which we replace the image feature embedding by the learned microphone array signal embedding (Eqn. 5.2). It helps to test if it is a better choice to use cross-view image supervision than microphone-array signal. We call this variant S3DVDet_mvSound. From the table 5.2, we can clearly see that replacing image with microphone array signal supervision leads to significant performance drop.

5. With Multiview both Image and Sound. In the above test, we show aggregating cross-view acoustic feature leads to inferior performance, but what if we combine image and sound? To this end, we propose a *Sound3DVDet* variant (S3DVDet_wMVIS) that jointly aggregates cross-view image feature and acoustic feature. We have observed performance drop, but the performance drop is not that obvious than other *Sound3DVDet* variants, which in turns shows the importance of involving multiview image feature for 3D sound source prediction.

The ablation studies show the necessity of each component of *Sound3DVDet*. We further qualitative visualization in Fig. 5.4, from which we can see the two *Sound3DVDet* variants and the comparing SELDNet [2] predict sound source incorrectly that either lies in the air or on different object surface. Our proposed framework *Sound3DVDet* can predict the 3D sound source that is closest to the ground truth.

5.2.5 Conclusions and Limitations

In this work, we show how to use multiview acoustic-camera recordings to assist us in localising invisible 3D sound sources. A limitation is that we assume the space between the sources and acoustic-camera is unoccluded, which may not reflect the real-world settings. Another limitation is that we do not consider the situation where the sound sources are moving and dynamic. Using real robotic acoustic-camera setting is also planned for the future.

Chapter 6

Conclusion and Future Work

This chapter concludes my DPhil research and further discusses several potential future research directions. In the conclusion section, we echo the questions presented in the *Abstract* section, and try to answer these questions in the capacity of my DPhil research journey. In the future research direction, I list several potential research topics that deserve attention.

6.1 DPhil Research Conclusion

During my DPhil research, I have systematically explored 1. spatial audio learning by integrating modern deep neural networks and classic acoustic signal processing techniques; 2. adopting deep neural network to implicitly model spatial audio propagation dynamics. 3. spatial audio-visual multimodal learning. After four years' exploration and investigation, I can partially answer the questions raised in the "Abstract".

Q1: What makes spatial sound different from vision? What are the physical principles that govern sound/vision generation and propagation in space. How are the spatial cues revealed from the two modalities, respectively?

A1: Spatial audio differs from vision (or specifically normal RGB image) in multiple ways, including the physical principle behind their generation/propagation and the way they reflect spatial cues. As presented in Sec. 2.3.1 and Sec. 2.1, spatial audio is mechanical waveform travelling through various medium (like air) by causing the particles of the medium to vibrate, while vision mostly derives from light that is defined as electrically charged particles vibration (electronmagnetic radiation). Spatial audio mainly holds wave nature while light exhibits both wave and particle nature. For the vision, the spatial cue lies in the visual object's position, the context jointly revealed by the foreground object and background information. The visual spatial cues can be inferred from either an individual image or multiple images together. For

the sound, the spatial cue jointly lies in single channel amplitude and inter-channel phase difference. The acoustic spatial cue mostly incurs from multi-channel spatial signal (*e.g.*,microphone-array based signal).

Q2: In the modern deep neural network era, how can classic signal processing based sound waveform feature extraction methods benefit from the powerful expressiveness of deep neural network?

A2: The prerequisite for processing a single channel sound waveform is to extract frequency components from one dimensional time domain sound waveform. In other words, the expressiveness of the obtained time-frequency map plays an important role for not only for estimating acoustic spatial task also other relevant acoustic tasks. In the deep neural network era, we can either exploit deep neural to learn to extract more expressive time-frequency representation from single channel sound waveform or directly learn to process multi-channel sound waveform to estimate spatial cue. For single channel sound waveform learn, the deep neural network can be exploited to re-parameterize classic method fixed filter bank to be learnable so as to learn more expressive time-frequency representation in a data-driven way. For multi-channel sound waveform learn, in addition to incorporating single channel learning strategy, deep neural network can be further adopted to encode inter-channel phase difference in a learnable way so that the whole task can be formulated to be end-to-end trainable. My work SoundDet [68], SoundDoA [65] and SoundSynp [66] have preliminarily attested the significant potential in combining classic signal processing methods and modern deep neural networks. The experimental results show that integrating classic signal processing techniques and modern deep neural networks can benefit the acoustic spatial cue estimation task.

Q3: How can we design audio-visual multimodal learning framework to better perceive the 3D environment in embodied AI setting?

A3: Audio and vision jointly provide spatial cues of the surrounding environment. An embodied agent depends on such multimodal signal to better perceive the 3D environment. In real scenarios, there are three main correlation states between sound and vision: strong-correlation where sound and vision indicate the same object of the interest, and the object of interest can be interchangeably detected from either modality; weak-correlation where the object of interest can only be detected from their own modality, but the other modality can provide cross-modal cues to detect the object of interest more precisely. no-correlation where sound and vision are mutually independent, the object of interest can only be detected in its modality. Each correlation state requires specific multimodal learning framework design.

While the first correlation state has been widely explored already, the weak-correlation state has been rarely discussed. During my DPhil research, I have tried to fill in the gap of weak-correlation state based multimodal learning. As presented in Sec. 5.2, we propose to detect visually nonobservable 3D sound sources which lie on object's physical surface from multiview RGB images and microphone-array signals. Our experiment shows, under weak-correlation situation, cross-modal signal can be appropriately utilized to benefit object of interest detection from other modal data. The third no-correlation state mainly happens in dynamic 3D environment where sound sources emit sound waveform intermittently and freely w.r.t. the visual environment. It remains as a future research direction.

6.2 Potential Future Research Directions

There are multiple potential research directions deserve future attention. I list two of them here.

6.2.1 Multimodal Learning Infrastructure Construction

1. *Seek Unified Multimodal Learning Framework*

The research community has experienced a long period of time of separated per-modality research, wherein researchers focusing on different modalities have worked relatively independently to develop their framework. This has led to distinct and relatively isolated processing frameworks tailored to specific modal data types. For example, in natural language processing (NLP), tokenizer is used to embed each word [41]; in vision community, 2D convolutional neural networks have been extensively utilized to extract features from 2D images; and in the realm of acoustics, both learnable and fixed short-time Fourier transform (STFT) are employed to process various acoustic signals. Such significant diversity in multimodal learning frameworks pose challenges to the development of unified multimodal learning framework. Although there have been initial efforts to integrate NLP and image learning [87], a more systematic and theoretically grounded exploration is warranted to design a more unified and cohesive learning framework capable of accommodating multimodal data learning while preserving the distinct characteristics inherent to each individual modality. Such a unified multimodal learning framework has the potential to train multitask agent serving for various purposes.

2. *Novel Audio-Visual Data Acquisition Platform*

Audio and vision are two ubiquitous signals existing in indoor environment. The sound signal comes from either audio sources emitting sound (*e.g.* telephone ringing) or inter-object impact audio (*e.g.* a glass cup falling onto ground). We humans exhaustively exploit such audio-visual multimodal signals to spatially perceive an indoor environment. For example, our paired eyes form stereo vision system to visually perceive objects in the indoor environment, while our paired ears receive binaural sound to inversely infer the sound source information (*e.g.* spatial location and class label), even though the audio source may go beyond sight. We humans synergistically and unconsciously exploit this audio-visual multimodal sensing framework to holistically perceive the indoor environment to perform various tasks.

Building upon the success of our human audio-visual multimodal perception system, a technical question naturally arises: How can we design a unified audio-visual data acquisition framework that is able to achieve comparable spatial perception performance to humans or even surpass human capabilities? Specifically, there are two potential research avenues:

1) **Sensor Spatial Configuration**: Determining the optimal number of microphones and vision sensors (e.g., RGB/RGB-D sensors) to be incorporated and configuring them spatially within a unified framework is essential. This configuration must strike a balance between the perception accuracy and overall computational and energy costs. In the realm of acoustic spatial perception, the use of microphone arrays (comprising two or more microphones grouped together) can enhance spatial perception capabilities. Investigating the ideal microphone-array configuration within an embodied setting demands deeper practical exploration.

2) **Theoretical Foundation on Audio-Vision Synergy**. Both audio sensor (like the microphone-array discussed above) and vision sensors are spatially perceiving the indoor environment. The perceived multimodal signals synergistically encode the same dynamic scene so the signals from different modalities are mutually dependent. For example, a telephone's spatial position and orientation w.r.t. the agent's ego-location reflected in the image is its relative size and imaging angle, in the received microphone-array sound is the magnitude and inter-channel phase difference. Unlike existing methods ([26]) that unanimously process each unimodal signal independently, there is a need for a more comprehensive theoretical analysis elucidating the mathematical and physical correlations between perceived visual and acoustic signals.

6.2.2 Multimodal Learning in Dynamic Setting

A substantial portion of our daily lives are dedicated to indoor environments, whether within personal living spaces or professional workplaces. These indoor settings create an intricate and ever-evolving ecosystem and serve as the backdrop for our multi-faceted daily activities, encompassing work, leisure, and social interactions. We interact with the indoor environments in diverse and multi-sensory ways, involving vision, sound, touch and even language communication, to perform a wide range of complex tasks with ease, such as going to living room to answer the ringing phone (navigation/localization) and brewing coffee (planning/manipulation).

While it may appear straightforward and effortless for humans to execute such tasks within indoor environments, the development of an embodied agent capable of achieving a comparable level of intelligence poses an exceptionally challenging endeavor. Successfully undertaking one task demands the agent to possess a comprehensive repertoire of skills, encompassing, but not limited to, perception, strategic planning, dexterous manipulation and timely adjusting policy w.r.t. dynamics in the environments, by analyzing and integrating data gleaned from an array of sensors. Realizing this ambitious objective requires the collaboration of interdisciplinary experts hailing from a wide spectrum of fields, spanning robotics, electrical and mechanical engineering, cognitive and neural sciences and numerous other pertinent domains. There are two main exploration directions: 1. Embodied multimodal neural implicit representation, with both in-depth theoretical analysis and wide practical applicability exploration. 2. On the new challenges in dynamic indoor environment, including multimodal signal distraction and intermittence.

1. Embodied Multimodal Neural Implicit Representation

Dynamic environments contain multifaceted data. These data come from multimodal sources and are always spatiotemporally evolving, resulting in the difficulty in the dynamic indoor environment data representation, management, maintenance and storage. In this research, I will be dedicated to design and develop neural implicit field to represent dynamic scenes.

The concept of neural implicit field, as introduced in the work ([114]), has emerged as a groundbreaking paradigm for scene representation. Unlike conventional methods that discretely encode and store real-world signals, like images and sound, neural implicit field offers a continuous approach, enabling the reconstruction of signals at arbitrary resolutions. Since deep neural networks (like multi-layer perceptron, MLP) show strong capability to approximate complex functions, using it to implicitly encode a scene can successfully learn the continuous mapping from an arbitrary

resolution query to its corresponding representation. Neural implicit field has demonstrated remarkable success in various domains, including point cloud ([190]) and spatial sound ([69]).

Nonetheless, current methods on neural implicit field remain unimodal in nature and focus on non-embodied context, thereby limiting their real-world indoor environment applicability. In my research, I will be committed to search for exploring and developing novel embodied multimodal neural implicit field. Such field possesses the capability to collectively encode various multimodal signals, such as RGB images and spatial audio, in a manner that aligns with the agent's egocentric perspective within the environment. It entails that the neural implicit field encapsulates the underlying physical principles associated with each distinct modality while remaining sensitive to the embodied agent's precise position and orientation within the room scene. Moreover, the implicit neural field is dynamic, reflecting the spatiotemporal evolving characteristic of dynamic indoor environment.

Learning such a multimodal neural implicit field is intrinsically difficult. The main challenges are two-fold: First, the underlying inherent disparities in the physical mechanisms for each distinct modality requires the multimodal neural implicit field to be able to inherently accommodate all the underlying physical principles from all modalities. Second, the spatiotemporal evolving dynamic characteristic introduces extra challenge for the multimodal neural field to discriminate and unify the dynamics and statics. There are two key challenges in the context of combing vision and audio for dynamic scene understanding:

1. **Vision Locality** versus **Sound Globality**. Light follows a single, linear optical path and ceases to propagate upon encountering an obstacle, allowing it to be amenable to modeling as a neural radiance field ([114]). Conversely, sound waveforms propagate isotropically ([92, 136]), traversing novel pathways and persisting in their journey, even upon collision with physical barriers. It thus requires to search for novel implicit neural field to accommodate the two physically distinctive modalities.

2. **Complex Vision and Audio Association**. In a dynamic environment, vision and audio can be either strongly associated, which means vision and audio indicate each other (*e.g.* telephone ringing associates with telephone), or weakly associated where the sound source may have no or non-obvious visual embodiment ([67]). It requires the mplicit neural field to jointly model the entanglement and disentanglement between vision and audio.

Challenges in Dynamic Indoor Scene

In contrast with most public embodied environments [24, 187] that are predominantly static, real-world indoor environments are dynamic and replete with a multitude of multimodal signals. For instance, individuals move around indoor spaces, engaging in various activities at different times (*e.g.* cooking in the kitchen, working or conversing in the living room). Furthermore, factors like infants' sporadic cries or play, the intermittent sounds generated by appliances like microwaves and ovens contribute to the dynamic nature of indoor environments. These dynamic indoor settings pose numerous new challenges, Among the myriad of potential challenges, I am particularly drawn to investigate two primary areas:

1. **Dynamic Embodied Navigation**. Navigation is one of the most prevalent tasks within the embodied context. In static scene, the navigation goal is to reach to a pre-defined static destination while minimizing the number of steps required (including ImageNav, ObjNav and AudioNav ([24, 187])). However, in dynamic embodied navigation, the agent must contend with intermittently arising environmental dynamics. Examples include scenarios where the destination itself is in motion within the environment, or where the agent must continually receive and interpret signals from the destination to adapt and update its navigation strategy. Furthermore, navigation success is influenced by various factors, such as the need to avoid collisions with other moving agents or pedestrian ([19]). There is a wealth of unexplored territory in the dynamic setting.

2. **Multimodal Dynamic Signal Distraction and Intermittence**. In real-world dynamic environments, multiple multimodal signals coexist, both temporally and spatially, creating a dynamic and evolving scene. Agents operating in such environments must learn to extract useful signals amidst a cacophony of distractions. Moreover, certain signals, such as audio cues, can be intermittent and subject to change over time, necessitating the development of policies that can adapt to such circumstances.

Bibliography

[1] Sharath Adavanne, Giambattista Parascandolo, Pasi Pertilä, Toni Heittola, and Tuomas Virtanen. Sound event detection in multichannel audio using spatial and harmonic features. In *Scenes and Events 2016 Workshop (DCASE2016)*, 2016. 23, 24

[2] Sharath Adavanne, Pasi Pertilä, and Tuomas Virtanen. Sound Event Detection Using Spatial Features and Convolutional Recurrent Neural Network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 5, 23, 32, 33, 41, 42, 43, 44, 54, 55, 56, 57, 58, 59, 60, 62, 66, 69, 70, 73, 97, 102, 105

[3] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. In *IEEE Journal of Selected Topics in Signal Processing*, 2018. 34, 41, 73

[4] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The Conversation: Deep Audio-Visual Speech Enhancement. *arXiv preprint arXiv:1804.04121*, 2018. 28

[5] Rafael Aguiar, Gianluca Maguolo, Loris Nanni, Yandre Costa, and Carlos Silla. On the importance of passive acoustic monitoring filters. *Journal of Marine Science and Engineering (JMSE)*, 2021. 60

[6] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. In *The Journal of the Acoustical Society of America*, 1979. 18, 19

[7] Juan Pablo Bello, Charlie Mydlarz, and Justin Salamon. *Sound Analysis in Smart Cities*. Springer International Publishing, 2018. 60

[8] Oded Bialer, Noa Garnett, and Tom Tirer. Performance advantages of deep neural networks for angle of arrival estimation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3907–3911. IEEE, 2019. 95

[9] Stefan Bilbao and Brian Hamilton. Wave-based room acoustics simulation: Explicit/implicit finite volume modeling of viscothermal losses and frequency-dependent boundaries. *Journal of the Audio Engineering Society*, 2017. 18, 76

[10] D. Botteldoore. Finite-difference time-domain simulation of low-frequency room acoustic problems. *Journal of the Acoustical Society of America*, 1995. 18

[11] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier Transform and its Applications*, volume 31999. McGraw-Hill New York, 1986. 17

[12] M. S. Brandstein and H. F. Silverman. A Robust Method for Speech Signal Time-Delay Estimation in Reverberant Rooms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997. 23, 33, 47, 97

[13] Breitkreutz, Christiane and Brade, Jennifer and Winkler, Sven and Bendixen, Alexandra and Klimant, Philipp and Jahn, Georg. Spatial Updating in Virtual Reality – Auditory and Visual Cues in a Cave Automatic Virtual Environment. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2022. 4

[14] Willem-Paul Brinkman, Allart Hoekstra, and Rene Vanegmond. The Effect Of 3D Audio And Other Audio Techniques On Virtual Reality Experience. In *Studies in Health Technology and Informatics*, 2015. 4, 75

[15] James Broderick, Jim Duggan, and Sam Redfern. The Importance of Spatial Audio in Modern Games and Virtual Environments. In *IEEE Games, Entertainment, Media Conference (GEM)*, 2018. 4

[16] Paul Brossier. *Automatic annotation of musical audio for interactive system.* PhD thesis, Queen Mary University of London, 2006. 69

[17] Emre Cakir, Ezgi Can Ozan, and Tuomas Virtanen. Filterbank learning for deep neural network based polyphonic sound event detection. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3399–3406. IEEE, 2016. 23, 24

[18] Emre Cakir, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. Convolutional recurrent neural networks for polyphonic sound event detection. In *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 2017. 60, 66, 69, 70

[19] Tommaso Cancelli, Enrico abd Campari, Luciano Serafini, Angel X. Chang, and Lamberto Ballan. Exploiting Proximity-Aware Tasks for Embodied Social Navigation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 112

[20] Yin Cao, Turab Iqbal, Qiuqiang Kong, Fengyan An, Wenwu Wang, and Mark D Plumbley. An Improved Event-Independent Network for Polyphonic Sound Event Localization and Detection. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. 6, 10, 23, 24, 30, 41, 42, 43, 44, 47, 53, 54, 55, 56, 57, 58, 59, 61, 73, 97, 102, 103

[21] Yin Cao, Turab Iqbal, Qiuqiang Kong, Yue Zhong, Wenwu Wang, and Mark D Plumbley. Event-Independent Network for Polyphonic Sound Event Localization and Detection. In *DCASE Workshop*, 2020. 6, 23, 30, 54, 97

[22] Yin Cao, Qiuqiang Kong, Turab Iqbal, Fengyan An, Wenwu Wang, and Mark Plumbley. Polyphonic Sound Event Detection and Localization using a Two-Stage Strategy. In *DCASE Workshop*, 2019. 42

[23] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with Transformers. In *European Conference on Computer Vision (ECCV)*, 2020. 93, 95, 101, 102

[24] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)*, 2017. 77, 84, 101, 112

[25] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic Audio-Visual Navigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 10, 28

[26] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman.

SoundSpaces: Audio-Visual Navigation in 3D Environments. In *European Conference on Computer Vision (ECCV)*, 2020. 6, 9, 10, 28, 30, 84, 101, 109

[27] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning. In *NeurIPS 2022 Datasets and Benchmarks Track*, 2022. 6, 10, 28, 30, 77, 84, 93

[28] Joe C Chen, Kung Yao, and Ralph E Hudson. Acoustic Source Localization and Beamforming: Theory and Practice. *EURASIP journal on advances in signal processing*, 2003:1–12, 2003. 95

[29] Yang Chen, Wenwu Wang, Zhe Wang, and Bingyin Xia. A Source Counting Method Using Acoustic Vector Sensor Based on Sparse Modeling of DOA Histogram. *IEEE Signal Processing Letters*, 2019. 60

[30] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Graph-DETR3D: Rethinking Overlapping Regions for Multi-view 3D Object Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 98

[31] François Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 54

[32] Lauren M. Chronister, Tessa A. Rhinehart, Aidan Place, and Justin Kitzes. An Annotated Set of Audio Recordings of Eastern North American Birds Containing Frequency, Time, and Species Information, 2021. 60, 68

[33] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Ntworks on Sequence Modelling. In *Advances Neural Information Processing System (NeurIPS)*, 2014. 23, 33, 54, 69, 103

[34] Giuseppe Ciaburro and Gino Iannace. Improving smart cities safety using sound events detection based on deep neural network algorithms. *Informatics*, 7(3), 2020. 3

[35] Samuel Clarke, Negin Heravi, Mark Rau, Ruohan Gao, Jiajun Wu, Doug James, and Jeannette Bohg. DiffImpact: Differentiable Rendering and Identification of Impact Sounds. In *Annual Conference on Robot Learning (CoRL)*, 2021. 3

[36] L. Cremer. Die wissenschaftlichen Grundlagen der Raumakustik: Geometrische Raumakustik. In *Stuttgart, Germany*, 1948. 19

[37] Steven Davis and Paul Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In *IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)*, 1980. 47, 49, 57, 61, 62, 72

[38] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real Time Speech Enhancement in the Waveform Domain. In *Interspeech*, 2020. 84

[39] Jia Deng, Wei Dong, Richard Socher, Jia Li, Kai Li, and Feifei Li. ImageNet: A large-scale hierarchical image database. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 25

[40] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A Large-Scale Hierarchical Image Database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 99, 104

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 2019. 108

[42] Kota Dohi, Takashi Endo, Harsh Purohit, Ryo Tanabe, and Yohei Kawaguchi. Flow-Based Self-Supervised Density Estimation for Anomalous Sound Detection. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. 60

[43] Konstantinos Drossos, Stylianos I. Mimilakis, Shayan Gharib, Yanxiong Li, and Tuomas Virtanen. Sound Event Detection with Depthwise Separable and Dilated Convolutions. In *International Joint Conference on Neural Networks (IJCNN)*, 2020. 53, 69, 70

[44] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-visual Model for Speech Separation. *arXiv preprint arXiv:1804.03619*, 2018. 28

[45] Angelo Farina. Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique. In *Audio Engineering Society Convention*, 2020. 79

[46] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 2022. 58

[47] J. Fu, H. Zheng, and T. Mei. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 36

[48] Thomas Funkhouser, Nicolas Tsingos, Ingrid Carlbom, Gary Elko, Mohan Sondhi, James West, Gopal Pingali, Patrick Min, and Addy Ngan. A Beam Tracing Method for Interactive Architectural Acoustics. *Journal of the Acoustical Society of America*, 2003. 18

[49] Aviv Gabbay, Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Seeing through Noise: Visually driven Speaker Separation and Enhancement. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. 28

[50] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to Separate Object Sounds by Watching Unlabeled Video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 28

[51] William G Gardner. Reverberation Algorithms. In *Applications of digital signal processing to audio and acoustics*, pages 85–131. Springer, 1998. 76

[52] M. S. Gazzaniga. *The Cognitive Neuroscience of Mind: A Tribute to Michael S. Gazzaniga*. the MIT press, 2010. 16

[53] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 68

[54] Francois Grondin and James Glass. A Study of the Complexity and Accuracy of Direction of Arrival Estimation Methods Based on GCC-PHAT for a Pair of Close Microphones. In *arxiv:1811.11787*, 2018. 32

[55] Francois Grondin, James Glass, Iwona Sobieraj, and Plumbley Mark D. A study of the complexity and accuracy of direction of arrival estimation methods based on gcc-phat for a pair of close microphones. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, 2019. 10, 23, 24

[56] Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Lianwu Chen, Yuexian Zou, and Dong Yu. Multi-modal Multi-channel Target Speech Separation. In *IEEE Journal of Selected Topics in Signal Processing*, 2020. 28

[57] Eric Guizzo, Christian Marinoni, Marco Pennese, Xinlei Ren, Xiguang Zheng, Chen Zhang, Bruno Masiero, Aurelio Uncini, and Danilo Comminiello. L3DAS22 Challenge: Learning 3D Audio Sources in a Real Office Environment. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 23, 97, 101, 102

[58] Eric Guizzo, Christian Marinoni, Marco Pennese Pennese, Xinlei Ren, Xiguang Zheng, Chen Zheng, Bruno Masiero, Aurelio Uncini, and Danilo Comminiello. L3DAS22 Challenge: Learning 3D Audio Sources in a Real Office Environment. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 31, 53, 58

[59] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, and Masahiro Yasuda. First-shot anomaly detection for machine condition monitoring: A domain generalization baseline. *In arXiv e-prints: 2303.00455*, 2023. 6

[60] Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Takaaki Hori, Jonathan Le Roux, and Kazuya Takeda. Duration-controlled lstm for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 25(11):2059–2070, 2017. 23, 24

[61] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, 2016. 8, 33, 54, 99, 104

[62] Yuhang He, Zhuangzhuang Dai, Long Chen, Niki Trigoni, and Andrew Markham. SoundCount: Sound Counting from Raw Audio with Dyadic Decomposition Neural Network. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2024. 13, 31, 74

[63] Yuhang He, Irving Fang, Yiming Li, Rushi Bhavesh Shah, and Chen Feng. Metric-Free Exploration for Topological Mapping by Task and Motion Imitation in Feature Space. In *Robotics: Science and Systems (RSS)*, 2023. 96

[64] Yuhang He, Irving Fang, Yiming Li, Rushi Bhavesh Shah, and Chen Feng. Metric-Free Exploration for Topological Mapping by Task and Motion Imitation in Feature Space. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. 84

[65] Yuhang He and Andrew Markham. SoundDoA: Learn Sound Source Direction of Arrival and Semantics from Sound Raw Waveforms. In *Interspeech*, 2022. 47, 54, 55, 56, 57, 58, 59, 61, 73, 102, 103, 107

[66] Yuhang He and Andrew Markham. SoundSynp: Sound Source Detection from Raw Waveforms with Multi-Scale Synperiodic Filterbanks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023. 13, 31, 74, 107

[67] Yuhang He, Sangyun Shin, Anoop Cherian, Niki Trigoni, and Andrew Markham. Sound3DVDet: 3D Sound Source Detection Using Multiview Microphone Array and RGB Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5496–5507, January 2024. 111

[68] Yuhang He, Niki Trigoni, and Andrew Markham. SoundDet: Polyphonic Moving Sound Event Detection and Localization from Raw Waveform. In *International Conference on Machine Learning (ICML)*, 2021. 13, 31, 47, 49, 54, 55, 56, 57, 58, 59, 60, 61, 62, 64, 73, 74, 101, 102, 107

[69] Yuhang He, Niki Trigoni, and Andrew Markham. Deep Neural Room Acoustic Primitive. In *International Conference on Machine Learning (ICML)*, 2024. 14, 111

[70] Eugene Hecht. *Optics*. Addison Wesley, 4th intern edition, 2002. 26

[71] Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen. Audio context recognition using audio event histograms. In *European Signal Processing Conference (EUSIPCO)*, 2010. 60

[72] Sepp Hochreiter and Schmidhuber Jugen. Long short-term memory. In *Neural Computation*, 1997. 23, 33, 69

[73] Murray Hodgson and Eva-Marie Nosal. Experimental evaluation of radiosity for room sound-field prediction. *Journal of the Acoustical Society of America*, 2006. 18

[74] Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. Watch or Listen: Robust Audio-Visual Speech Recognition with Visual Corruption Modeling and Reliability Scoring. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 9, 91

[75] Di Hu, Lichao Mou, Qingzhong Wang, Junyu Gao, Yuansheng Hua, Dejing Dou, and Xiao Xiang Zhu. Ambient sound helps: Audiovisual crowd counting in extreme conditions. *arXiv preprint*, 2020. 60

[76] Jinbo Hu, Yin Cao, Ming Wu, Qiuqiang Kong, Feiran Yang, Mark D. Plumbley, and Jun Yang. A track-wise ensemble event independent network for polyphonic sound event localization and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 54, 58, 59

[77] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015. 72

[78] Eric J. Humphrey, Simon Durand, and Brian McFee. Openmic-2018: An open dataset for multiple instrument recognition. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018. 60, 68

[79] Navdeep Jaitly and Geoffrey Hinton. Learning a Better Representation of Speech Soundwaves Using Restricted Boltzmann Machines. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011. 24

[80] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torraba. ConceptFusion: Open-set Multimodal 3D Mapping. *Robotics: Science and Systems (RSS)*, 2023. 28

[81] Jeffrey Jeffrey Borish. Extension of the Image Model to Arbitrary Polyhedra. In *The Journal of Acoustical Socienty of America*, 1984. 19

[82] Jean-Marc Jot and Keun Sup Lee. Augmented Reality Headphone Environment Rendering. In *International Conference on Audio for Virtual and Augmented Reality*, 2016. 4

[83] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2018. 53

[84] Alif Ridzuan Khairuddin, Mohamad Shukor Talib, and Habibollah Haron. Review on simultaneous localization and mapping (slam). In *IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 2015. 77, 79

[85] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representation (ICLR)*, 2015. 55, 69, 86

[86] Mendel Kleiner, Bengtinge Dalenbäck, and Peter Svensson. Auralization-an overview. *Journal of the Audio Engineering Society*, 1993. 18

[87] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. 8, 108

[88] Haru Kondoh and Asako Kanezaki. Multi-goal Audio-visual Navigation using Sound Direction Map. *ArXiv*, abs/2308.00219, 2016. 28

[89] Sandeep Kothinti, Keisuke Imoto, Debmalya Chakrabarty, Gregory Sell, Shinji Watanabe, and Mounya Elhilali. Joint acoustic and class inference for weakly supervised sound event detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 36–40. IEEE, 2019. 23, 24

[90] Shoichi Koyama, Tomoya Nishida, Keisuke Kimura, Takumi Abe, Natsuki Ueno, and Jesper Brunnström. MESHRIR: A Dataset of Room Impulse Responses on Meshed Grid Points for Evaluating Sound Field Analysis and Synthesis Methods. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021. 78, 84

[91] Daniel Krause and Konrad Kowalczyk. Arborescent neural network architectures for sound event detection and localization. Technical report, DCASE2019 Challenge, Tech. Rep, 2019. 23, 24

[92] Asbjørn Krokstad, S. Strøm, and Svein Sorsdal. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration*, 1968. 18, 111

[93] Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955. 93, 95, 100, 102

[94] Martin Kuster. Reliability of Estimating the Room Volume from a Single Room Impulse Response. In *Journal of the Acoustics Society of America*, 2008. 75

[95] Heinrich Kuttruff. Room Acoustics. In *Applied Science Publishers*, 1979. 11, 75, 79, 80

[96] Heinrich Kuttruff. *Room Acoustics, Fourth Edition.* Elsevier Science Publishers, 2016. 18

[97] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-Supervised Nets. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015. 93, 96

[98] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 60

[99] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *European Conference on Computer Vision (ECCV)*, 2018. 38

[100] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne

Tuytelaars, editors, *European Conference on Computer Vision (ECCV)*, 2014. 34, 55, 101, 102

[101] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single Shot Multibox Detector. In *European Conference on Computer Vision (ECCV)*, 2016. 95

[102] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position Embedding Transformation for Multi-View 3D Object Detection. *European Conference on Computer Vision (ECCV)*, 2022. 93, 95, 98

[103] Marshall Long. Architectural Acoustics (Second Edition). In *Architectural Acoustics (Second Edition)*, page xxix, Boston, 2014. Academic Press. 75

[104] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representation (ICLR)*, 2019. 103

[105] Vincent Lostanlen, Justin Salamon, Mark Cartwright, Brian McFee, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello. Per-channel energy normalization: Why and how. *IEEE Signal Processing Letters (SPL)*, 2019. 65

[106] Chen Change Loy, Ke Chen, Shaogang Gong, and Tao Xiang. Crowd counting and profiling: Methodology and evaluation. In *Modeling, Simulation and Visual Analysis of Crowds*, 2013. 66

[107] Rui Lu, Zhiyao Duan, and Changshui Zhang. Listen and Look: Audio-visual Matching assisted Speech Source Separation. In *IEEE Signal Processing Letters*, 2018. 28

[108] Andrew Luo, Yilun Du, Michael J Tarr, Joshua B Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *arXiv preprint arXiv:2204.00628*, 2022. 76, 83, 84, 86

[109] Stéphane Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, Inc., USA, 3rd edition, 2008. 17, 21, 49, 51, 57, 62

[110] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th python in science conference*, volume 8, 2015. 68

[111] Annamaria Mesaros, Sharath Adavanne, Archontis Politis, Toni Heittola, and Tuomas Virtanen. Joint Measurement of Localization and of Detection of Sound Event. In *9 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019. 33, 41, 42, 58

[112] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for Polyphonic Sound Event Detection. *Applied Sciences*, 2016. 101, 102

[113] Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, and Mark D. Plumbley. Sound event detection: A tutorial. *IEEE Signal Processing Magazine,*, 2021. 60

[114] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 110, 111

[115] David Mintzer. Transient Sounds in Rooms. In *The Journal of Acoustical Socienty of America*, 1950. 19

[116] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-Stitch Networks for Multi-task Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 53

[117] Brian Moore. *Psychoacoustics*, pages 459–501. Springer New York, New York, NY, 2007. 2

[118] Giovanni Morrone, Sonia Bergamaschi, Luca Pasa, Luciano Fadiga, Vadim Tikhanoff, and Leonardo Badino. Face Landmark-based Speaker-independent Audio-visual Speech Enhancement in Multi-talker Environments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 28

[119] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H. Mueller, Chakravarty R. Alla Chaitanya, Anton Kaplanyan, and Markus Steinberger. DONeRF: Towards real-time rendering of compact neural radiance fields using depth oracle networks. *Computer Graphics Forum*, 40(4):45–59, 2021. 90

[120] Thil von Neumann, Christoph Boeddeker, Lukas Drude, Keisuke Kinoshita, Marc Delcroix, Tomohiro Nakatani, and Reinhold Haeb-Umbach. Multi-talker

ASR for an Unknown Number of Sources: Joint Training of Source Counting, Separation and ASR. In *Interspeech*, 2020. 60, 69, 71

[121] Thi Ngoc Tho Nguyen, Karn N. Watcharasupat, Ngoc Khanh Nguyen, Douglas L. Jones, and Woon-Seng Gan. SALSA: Spatial Cue-Augmented Log-Spectrogram Features for Polyphonic Sound Event Localization and Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022. 10, 23, 24, 30

[122] Eva-Marie Nosal, Murray Hodgson, and Ian Ashdown. Investigation of the validity of radiosity for sound-field prediction in cubic rooms. *Journal of the Acoustical Society of America*, 2004. 18

[123] Dimitri Palaz, Ronan Collobert, and Mathew Magimai-Doss. Estimating Phoneme Class Conditional Probabilities from Raw Speech Signal Using Convolutional Neural Networks. In *Interspeech*, 2013. 24

[124] Stephen E. Palmer. *Vision Science: Photons to Phenomenology*. the MIT press, 1999. 26

[125] Arjun Pankajakshan, Helen L. Bear, and Emmanouil Benetos. Polyphonic sound event and sound activity detection: A multi-task approach. *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 323–327, 2019. 60

[126] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Acoustic Event Detection in Real Life Recordings. In *18th European Signal Processing Conference (EUSIPCO)*, 2010. 24

[127] Sanjeel Parekh, Alexey Ozerov, Slim Essid, Ngoc QK Duong, Patrick Pérez, and Gaël Richard. Identify, Locate and Separate: Audio-visual Object Extraction in Large Video Collections using Weak Supervision. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019. 28

[128] Daniel Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, 2019. 55

[129] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga,

Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2019. 55, 69, 103

[130] Phuong Pham, Juncheng Li, Joseph Szurley, and Samarjit Das. Eventness: Object detection on spectrograms for temporal localization of audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. 23, 24

[131] H. Phan, T. N. Tho Nguyen, P. Koch, and A. Mertins. Polyphonic audio event detection: Multi-label or multi-class multi-task classification problem? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 60

[132] Allan D. Pierce. *Acoustics: An Introduction to Its Physical Principles and Applications*. McGraw-Hill Book Company, 1989. 20

[133] Andrzej Pietrzyk. Computer modeling of the sound field in small rooms. *Journal of the Audio Engineering Society*, 1998. 18

[134] Archontis Politis, Sharath Adavanne, and Tuomas Virtanen. A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection. In *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020. 25, 31, 53

[135] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas Virtanen. Overview and Evaluation of Sound Event Localization and Detection in DCASE 2019. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020. 101, 102

[136] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. WaveGlow: a Flow-based Generative Network for Speech Synthesis. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 111

[137] Jie Pu, Yannis Panagakis, Stavros Petridis, and Maja Pantic. Audio-visual Object Localization and Separation using Low-rank and Sparsity. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 28

[138] Ville Pulkki, Symeon Delikaris-Manias, and Archontis Politis. *Spatial Sound Scene Synthesis and Manipulation for Virtual Reality and Audio Effects*, pages 347–361. 2018. 6

[139] Senthil Purushwalkam, Sebastian Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. Audio-visual floorplan reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, 2020. 10

[140] Stephen Mark Purves, Dale Williams. *Neuroscience. 2nd edition.* Sinauer Associates 2001, 2001. 26

[141] Anton Ratnarajah, Ishwarya Ananthabhotla, Vamsi Krishna Ithapu, Pablo Hoffmann, Dinesh Manocha, and Paul Calamia. Towards Improved Room Impulse Response Estimation for Speech Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 84, 86

[142] Anton Ratnarajah, Shi-Xiong Zhang Zhang, Meng Yu, Zhenyu Tang, Dinesh Manocha, and Dong Yu. Fast-RIR: Fast neural diffuse room impulse response generator. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 76

[143] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *In IEEE Workshop on Spoken Language Technology (SLT)*, 2018. 24, 35, 36, 47, 49, 57, 64

[144] John William Strutt Rayleigh and Robert Bruce Lindsay. *The Theory of Sound.* Dover Publications, New York, 2nd Edition Revised and Enlarged edition, 1945. 79, 80

[145] Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical Depth DistributionNetwork for Monocular 3D Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 95

[146] Shaoqing Ren, Kaiming He, Ross Girshich, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 8, 95

[147] Alexander Richard, Peter Dodds, and Vamsi Krishna Ithapu. Deep impulse responses: Estimating and parameterizing filters with deep networks. In *IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022. 84, 85, 86

[148] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of telephone Networks and Codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, 2001. 85

[149] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised Audio-visual Co-segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 28

[150] Tara N. Sainath, Brian Kingsbury, Abdel-rahman Mohamed, and Bhuvana Ramabhadran. Learning filter banks within a deep neural network framework. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013. 24

[151] Prasanga Samarasinghea and Thushara D. Abhayapala. Acoustic Reciprocity: An Extension to Spherical Harmonics Domain. In *The Journal of the Acoustical Society of America*, 2017. 80

[152] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning Feature Matching with Graph Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 98

[153] Lauri Savioja and U. Peter Svensson. Overview of Geometrical Room Acoustic Modeling Techniques. *Journal of the Acoustical Society of America*, 2015. 18, 76

[154] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulations and array processing algorithms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. 9, 61

[155] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech*, 2019. 24

[156] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to Localize Sound Source in Visual Scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 28

[157] Rahul Sharma, Krishna Somandepalli, and Shrikanth Narayanan. Cross-modal Learning for Audio-visual Speech Event Localization. *arXiv preprint arXiv:2003.04358*, 2020. 28

[158] Sandhya Sharma, Kazuhiko Sato, and Bishnu Prasad Gautam. A methodological literature review of acoustic wildlife monitoring using artificial intelligence tools and techniques. *Sustainability*, 15(9), 2023. 3

[159] Christian J. Steinmetz, Vamsi Krishna Ithapu, and Paul Calamia. Filtered Noise Shaping for Time Domain Room Impulse Response Estimation from Reverberant Speech. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2021. 76

[160] Chris Stockbridge. Acoustics of Teleconferencing. *The Journal of the Acoustical Society of America*, 74(S1):S5–S5, 08 2005. 5

[161] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797*, 2019. 76, 86

[162] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Frédéric Lepoutre, and François Grondin. Resource-Efficient Separation Transformer, 2022. 60

[163] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-Free Local Feature Matching with Transformers. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 98, 104

[164] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning Audio-Visual Source Localization via False Negative Aware Contrastive Learning.

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 9, 91

[165] Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Černocký. Building and Evaluation of a Real Room Impulse Response Dataset. *IEEE Journal of Selected Topics in Signal Processing*, 2019. 76

[166] Tho Nguyen Thi Ngoc, Ngoc Khanh Nguyen, Huy Phan, Lam Pham, Kenneth Ooi, Douglas Jones, and Woon-Seng Gan. A general network architecture for sound event localization and detection using transfer learning and recurrent neural network. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021. 23, 24

[167] Nguyen Thi Ngoc Tho, Shengkui Zhao, and Douglas L. Jones. Robust doa estimation of multiple speech sources. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 2287–2291, 2014. 32, 41

[168] T. N. Tho Nguyen, D. L. Jones, and W. Gan. A Sequence Matching Network for Polyphonic Sound Event Localization and Detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. 23, 32, 42

[169] Thi Ngoc Tho Nguyen, Douglas L. Jones, Karn N. Watcharasupat, Huy Phan, and Woon-Seng Gan. SALSA-Lite: A Fast and Effective Feature for Polyphonic Sound Event Localization and Detection with Microphone Arrays. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 6, 23, 24, 30

[170] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-Visual Event Localization in Unconstrained Videos. In *European Conference on Computer Vision (ECCV)*, 2018. 9, 91

[171] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual Event Localization in Unconstrained Videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 28

[172] James Traer and Josh H McDermott. Statistics of Natural Reverberation Enable Perceptual Separation of Sound and Space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016. 88

[173] Nicolas Turpault, Romain Serizel, Scott Wisdom, Hakan Erdogan, John R. Hershey, Eduardo Fonseca, Prem Seetharaman, and Justin Salamon. Sound Event Detection and Separation: A Benchmark on Desed Synthetic Soundscapes. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. 60

[174] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning Local Features with Policy Gradient. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 98

[175] Efthymios Tzinis, Shrikant Venkataramani, Zhepei Wang, Cem Subakan, and Paris Smaragdis. Two-Step Sound Source Separation: Training On Learned Latent Targets. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. 60

[176] Efthymios Tzinis, Xilin Wang, Zhepei amd Jiang, and Paris Smaragdis. Compute and Memory Efficient Universal Sound Source Separation. In *Journal of Signal Processing Systems*, 2022. 60

[177] Michele Valenti, Stefano Squartini, Aleksandr Diment, Giambattista Parascandolo, and Tuomas Virtanen. A convolutional neural network approach for acoustic scene classification. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1547–1554, 2017. 6

[178] Bert Van Den Broeck, Alexander Bertrand, Peter Karsmakers, Bart Vanrumste, Hugo Van hamme, and Marc Moonen. Time-domain Generalized Cross Correlation Phase Transform Sound Source Localization for Small Microphone Arrays. In *The 5th European DSP in Education and Research Conference*, 2012. 23, 36, 97

[179] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jacob Uszkoreit, Llion Jones Jone, Aidan N. Gomez, and Lukasz Kaiser. Attention is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 43, 54, 83, 93, 103

[180] Charles Verron, Mitsuko Aramaki, Richard Kronland-Martinet, and Grégory Pallone. A 3-d immersive synthesizer for environmental sounds. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2010. 75

[181] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 60

[182] Qing Wang, Huaxin Wu, Zijun Jing, Feng Ma, Yi Fang, Yuxuan Wang, Tairan Chen, Jia Pan, Jun Du, and Chin-hui Lee. The ustc-iflytek system for sound event localization and detection of dcase2020 challenge. In *DCASE workshop*, 2020. 23

[183] Qing Wang, Huaxin Wu, Zijun Jing, Feng Ma, Yi Fang, Yuxuan Wang, Tairan Chen, Jia Pan, Jun Du, and Chin-Hui Lee. The ustc-iflytek system for sound event localization and detection of dcase2020 challenge. In *Tech. Report of SELD Decase2020 Challenge*, 2020. 54, 56

[184] Weimin Wang, Weiran Wang, Ming Sun, and Chao Wang. Acoustic scene analysis with multi-head attention networks. In *Interspeech*, 2020. 5

[185] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, , and Justin M. Solomon. DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries. In *The Conference on Robot Learning*, 2021. 93, 95, 98, 100, 101, 102

[186] Yuxuan Wang, Pascal Getreuer, Thad Hughes, Richard Lyon, and Rif Saurous. Trainable Frontend for Robust and Far-Field Keyword Spotting. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 65

[187] Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: Real-World Perception for Embodied Agents. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 112

[188] Xianjun Xia, Roberto Togneri, Ferdous Sohel, Yuanjun Zhao, and Defeng Huang. Multi-task learning for acoustic event detection using event and frame position information. *IEEE Transactions on Multimedia*, 22(3):569–578, 2019. 23, 24

[189] Tao Xie, Kun Dai, Ke Wang, Ruifeng Li, and Lijun Zhao. DeepMatcher: A Deep Transformer-based Network for Robust and Accurate Local Feature Matching. *arXiv preprint arXiv:2301.02993*, 2023. 98

[190] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-NeRF: Point-based neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 111

[191] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92), 2019. 85

[192] Pei Yan, Yihua Tan, Shengzhou Xiong, Yuan Tai, and Yansheng Li. Learning Soft Estimator of Keypoint Scale and Orientation with Probabilistic Covariant Loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 98

[193] Dong Yu, Morten Kolbæ k, Zheng-Hua Tan, and Jesper Jensen. Permutation Invariant Training of Deep Models for Speaker-Independent Multi-Talker Speech Separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 103

[194] Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. Leaf: A learnable frontend for audio classification. *International Conference on Learning Representations (ICLR)*, 2021. 24, 47, 49, 57, 61, 62

[195] Neil Zeghidour, Nicolas Usunier, Iasonas Kokkinos, Thomas Schatz, Gabriel Synnaeve, and Emmanuel Dupoux. Learning Filterbanks from Raw Speech for Phone Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. 24

[196] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 60, 66

[197] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The Sound of Motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 28

[198] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The Sound of Pixels. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 28

[199] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep Audio-visual Learning: A Survey. *International Journal of Automation and Computing*, 2021. 28

[200] Lingyu Zhu and Esa Rahtu. Separating Sounds from a Single Image. In *ArXiv, Arxiv 2007.07984*, 2020. 28

[201] Shengjie Zhu and Xiaoming Liu. Pmatch: Paired masked image modeling for dense geometric matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 98