Universiteit Leiden

Master Computer Science

Lightweight Audio-Visual Sound Event Localization and Detection: An Audio-Dominant Architecture Design

Name:	Suzhen Deng
Student ID:	s4017501
Date:	29/08/2025
Specialisation:	Data Science
1st supervisor:	Dr. Qinyu Chen
2nd supervisor:	Dr. Lu Cao

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

ACKNOWLEDGMENTS

First, I would like to sincerely thank my primary supervisor, Dr. Qinyu Chen, for giving me this research topic and guiding me through every step of the project. Your clear thinking and logical approach helped me focus on the key problems, and your kindness and support made this journey much easier. It has been a privilege to have such a dedicated mentor during my studies abroad.

I am also deeply grateful to my second supervisor, Dr. Lu Cao, for your valuable comments and suggestions on my thesis. Your patient and thoughtful feedback, as well as your careful review, have greatly improved this work.

Many thanks to PhD candidate Jiawen Qi for your patience and technical guidance. Your support, together with Dr. Chen's, has greatly enhanced my research and programming skills.

I owe heartfelt thanks to my parents, who have always been my strongest support. Thank you for your love, trust, and encouragement, which gave me the courage to pursue this path with confidence.

I would also like to thank my close friends Xiaona and Yuxian. Your companionship and encouragement made my life abroad warmer and less lonely.

Finally, I want to thank myself for staying strong and persistent. This journey has taught me valuable lessons and helped me grow, both academically and personally.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ADPIT Auxiliary Duplicating Permutation Invariant Training

AV-SELD Audio-Visual Sound Event Localization and Detection

AVSR Audio-Visual Speech Recognition

CNN Convolutional Neural Network

dBFS decibels relative to full scale

DOA Direction of Arrival

FMA Free Music Archive

FOA First-Order Ambisonics

FSD50K Freesound Dataset 50K

HVAC Heating, Ventilation, and Air Conditioning

IRs impulse responses

MHCA multi-head cross-attention

MHSA multi-head self-attention

MSE Mean Squared Error

RIRs room impulse responses

SED Sound Event Detection

SELD Sound Event Localization and Detection

SSL Sound Source Localization

STARSS23 Sony-TAu Realistic Spatial Soundscapes 2023

ABSTRACT

Audio-Visual Sound Event Localization and Detection (AV-SELD) is a multimodal task recently introduced in the DCASE Challenge. It requires systems to jointly detect, classify, and localize sound events in 3D space by combining audio and visual information. While recent methods have achieved promising performance, they typically rely on excessively large models, limiting their practical deployment. Furthermore, little systematic investigation has been conducted into the optimal balance between audio and visual modalities. To address these challenges, we propose a dynamic dimension adjustment strategy for systematically exploring audio-visual modality ratios. Building upon recent AV-Conformer architectures, we optimize the visual encoder and network architecture, achieving substantial parameter reduction. Experiments show that balanced (1:1) or slightly audio-biased configurations yield the best overall performance. Notably, we find that heavy visual encoders are unnecessary for this task, and that a single fusion Conformer layer is sufficient. Overall, our approach reduces the number of parameters by 88% while improving F1 performance by 9.6%.

In summary, our work systematically investigates the impact of audio-visual modality ratios and provides guidance for future research. In addition, we introduce a lightweight and effective model with clear potential for deployment on edge devices, which sets a direction for future lightweight research in AV-SELD and related multimodal tasks.

CHAPTER 1
INTRODUCTION

1.1 Research Background

Sound Event Localization and Detection (SELD) is a multi-task learning problem that can be divided into two sub-tasks [1]: Sound Event Detection (SED) and Sound Source Localization (SSL). Of these, SED focuses on identifying the temporal activities and textual labels of sound events [2], which are individual sounds that describe what is happening in the environment [3], such as knocking, music, or human speech. In most real-world scenarios, the system needs to detect multiple overlapping sound events. Meanwhile, SSL aims to estimate the direction and position of sound sources with respect to the microphone [1]. The direction is commonly expressed using the Direction of Arrival (DOA), which includes both azimuth and elevation angles. In some cases, the position may also involve estimating the distance between the sound source and the microphones.

Consequently, the SELD task has been applied in many real-world domains. One important application is wearable devices that convey sound location information through vibration. Such devices, including belts, earphones, and hats [4][5], can help users with hearing impairments. Beyond wearables, SELD is applied in robotics for human-robot interaction [6], enhancing both response efficiency and accuracy. It also provides spatial acoustic information to support environmental perception in tasks such as navigation and autonomous driving [7][8]. Furthermore, SELD is crucial for automated monitoring systems, enabling audio surveillance, safety alerts, and anomaly detection.[9][10][11]

This thesis is based on the SELD task of the DCASE 2024 Challenge[1]. The task was first introduced in 2019. A synthesized dataset was generated by convolving random sound

[1]https://dcase.community/challenge2024/

events with randomly chosen impulse responses (IRs) at fixed locations [12]. From 2020 to 2022, the SELD task continuously evolved by introducing real-world acoustic environments, multilingual speech events, and diverse microphone array configurations. In 2023, DCASE incorporated visual modality into the SELD task for the first time, enabling models to learn from both audio and visual sources. Building on this, the 2024 task continued with the multimodal audio-visual setting and introduced distance information to support both localization and distance-aware evaluation [2]. In our work, following the DCASE 2024 requirements, we implement audio-visual 3D spatial localization by estimating both DOA and distance.

1.2 Motivation and Contributions

Since 2023, DCASE has introduced video into the task to enhance the spatial sensitivity of audio-based systems. Previous studies have shown that the visual modality provides spatial cues, which improve stability when fused with audio in cases where the audio signal degrades [13]. Visual information is important as it complements audio. There are many studies investigating modality complementarity, particularly in the tasks of audio classification [14] and automatic speech recognition [15][16]. However, in the field of AV-SELD, studies on modality complementarity remain relatively scarce.

In the DCASE 2024 challenge, many teams have proposed AV-SELD models that achieved good performance. However, their parameter sizes are typically between 60M and 90M, with some models reaching up to 400M. Considering that AV-SELD is often applied in resource-constrained scenarios such as robots, AR/VR headsets and other edge devices, excessive model complexity would hinder energy efficiency and real-time responsiveness. Therefore, it is important to design an efficient and compact model that can still achieve comparable performance for real-world applications.

Based on the above background and challenges, we propose an audio-dominant lightweight AV-SELD model. The main contributions of this work are as follows:

- We conduct a systematic analysis of audio-visual feature ratios and fusion dimensions in AV-SELD. The results reveal the dominant role of the audio modality and show that a balanced ratio with an appropriate total feature dimension achieves optimal performance. These findings provide practical guidance for future multimodal architecture design.

- We propose an audio-dominant lightweight AV-SELD model based on the baseline of Berghi et al. [17], achieving an 88% parameter reduction while improving F1 score from 40.8% to 44.7%. This demonstrates that our design successfully balances efficiency with accuracy.

CHAPTER 2

RELATED WORK

This chapter reviews related work on the AV-SELD task. First, we introduce different feature extraction methods, including commonly used approaches for both audio and visual features. Then, we discuss audio-visual fusion strategies along with their implementation approaches, such as feature concatenation and attention mechanisms. Finally, we review existing research on modality ratio optimization in domains such as Audio-Visual Speech Recognition (AVSR) and identify the research gap in AV-SELD tasks.

2.1 Feature Extraction

Audio and visual modalities complement each other in both temporal and spatial aspects. Audio provides strong temporal cues and remains effective in detecting sound events even when the source is occluded or not visible. Meanwhile, visual information can assist with localization and detection when audio signals are weak or absent[13]. Given this complementarity, many studies in AV-SELD have applied different feature extraction strategies to leverage both modalities.

In the AV-SELD task, audio feature extraction typically combines spectral content with spatial directional information. A common approach is to extract log-Mel spectrograms and intensity vectors to capture these two types of features, respectively. Therefore, some studies[18] [19] have adopted the combination of log-Mel spectrograms and intensity vectors as input features. Additionally, Berg et al.[20] explored both Mel spectrogram and MFCC for spectral feature extraction. For spatial feature extraction, they employed Neural GCC-PHAT (NGCC-PHAT) to extract the TDOA between microphone channels.

Pretrained models for video have been shown to be effective for extracting visual features. For example, ResNet-50 pretrained on ImageNet is widely used as a visual feature extractor

in AV-SELD tasks. In addition, some studies, such as the one by Berg et al.[20]., introduced Panoformer, a depth estimation model pretrained on panoramic images, to assist in distance prediction.

2.2 Audio-Visual Fusion Strategies

Feature fusion has been approached in various ways. Jiang et al. [21] propose a two-stage strategy. First, a ResNet-Conformer backbone fuses audio features with Gaussian-based visual features at the feature level. Subsequently, they employ visual cues to refine the audio predictions, specifically using object detection and human keypoints to correct sound source localization.

Some studies employ mid-level fusion, where high-level features are extracted from audio and visual modalities using separate encoders, and then fused in the feature space. This method preserves modality-specific information and enables interaction at a higher and more abstract level. For example, Kim et al. used 3D convolutional networks to extract multi-source visual features, which were then fused with audio features via element-wise addition after the audio encoder[19]. Another study proposed two fusion strategies: feature-level fusion by concatenating audio and visual embeddings; and CMAF, which employs multi-head self-attention (MHSA) and multi-head cross-attention (MHCA) modules to dynamically model inter-modal relationships[17].

2.3 Audio-Visual Complementarity and Ratio Studies

In some domains, many studies have demonstrated that audio and visual modalities are complementary, and their combination often outperforms using a single modality. Nanni et al. designed an audio classification framework that integrates acoustic and visual features, achieving superior performance compared to unimodal approaches[14]. In audio-visual speech recognition, Petridis et al. proposed an end-to-end model based on BLSTM that extracts features from both raw pixels and spectrograms[15], showing that visual cues can

complement the audio modality in noisy environments, thereby improving overall system robustness. Many studies have also examined how the ratio between modalities affects performance. A study compared the contribution of audio and visual modalities to performance using MSHMM[16]. They found that the best performance was achieved with an 80% audio and 20% visual weighting. However, after normalizing to eliminate the scale differences between the two modalities, the actual contributions were found to be approximately equal. Besides, Gimeno-Gómez and Martínez-Hinarejos proposed a parameter-efficient AVSR model based on Branchformer. They employed an adaptive fusion module to automatically learn the contribution ratio of audio and visual modalities, which converged to approximately 70% for audio and 30% for video in their experiments[22]. The ratio increases for the visual modality in noisy environments, while audio remains dominant in clean conditions.

Although the importance of using visual information to complement audio has been widely acknowledged, studies on optimal modality ratios have led to notable improvements in certain fields. However, in AV-SELD, determining the optimal audio-visual ratio remains largely unexplored. Prior AV-SELD work has largely relied on fixed 1:1 audio-visual ratios without investigating alternative configurations. This limitation restricts our understanding of multimodal fusion mechanisms and may lead to suboptimal performance. Therefore, we explore various audio-visual feature configurations for AV-SELD, aiming to address this gap and provide insights for future lightweight model design.

CHAPTER 3

DATASET CONSTRUCTION

This chapter introduces the datasets used for the AV-SELD task. The primary dataset is Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23), which provides synchronized First-Order Ambisonics (FOA) audio and 360° panoramic video, serving as a real-world benchmark for AV-SELD. In addition, we used the SpatialScaper library [23] to generate approximately 30 hours of synthetic data to address dataset imbalance and scarcity issues. The following sections describe the characteristics and generation procedures of each dataset.

3.1 Dataset Overview

This study employs the STARSS23[24] dataset for training and evaluation. STARSS23 consists of real indoor acoustic scenes recorded in Tampere, Finland, and Tokyo, Japan, covering 13 classes of targeted sound events: *female speech, male speech, clapping, telephone, laughter, domestic sounds, footsteps, door, music, musical instruments, water tap, bell*, and *knock*. The audio is recorded in 4-channel FOA format with a sampling rate of 24 kHz, and the video consists of 360° panoramic recordings at 1920×960 resolution.

The STARSS23 dataset provides precisely synchronized multimodal audio-visual data, along with annotations for each sound event, including class labels, temporal activity, spatial direction (azimuth and elevation angles), and distance from the microphone array. The development set contains 7 hours and 22 minutes of audio-visual recordings, consisting of 168 clips, which are split into 90 for training and 78 for testing. The dataset features a high occurrence frequency of common daily sound events such as female and male speech, music, and domestic sounds. While certain categories, like *knock*, appear in only 9 seconds of the entire training set. It also includes complex overlapping scenarios and multiple sources

Figure 3.1: A 360° video frame from the STARSS23 dataset illustrating a domestic acoustic scene containing various potential sound sources.

of the same class, with some scenes containing more than five simultaneous sound sources. Therefore, STARSS23 is both highly suitable for and challenging to the AV-SELD task.

3.2 Synthetic Dataset

To address the class imbalance in the STARSS23 and improve the model's generalization capability, we used an external synthetic dataset generated with the SpatialScaper library[23]. The synthetic dataset is based on existing sound events from datasets such as Freesound Dataset 50K (FSD50K) and Free Music Archive (FMA), combined with room impulse responses (RIRs) and random room configurations. Parameters for DOA and distance are also included to simulate realistic spatial sound scenes. In addition, we augment 70% of the synthesized data with background noise to simulate real-world conditions. The background noise is sampled from various environmental sources within the SpatialScaper framework, such as ambient sounds, crowd babble, and Heating, Ventilation, and Air Conditioning (HVAC) systems. Based on the reference level in decibels relative to full scale (dBFS), we divide the noise intensity into three levels: low noise (-75 to -65 dBFS, 30%), medium noise (-65 to -50 dBFS, 50%), and high noise (-50 to -40 dBFS, 20%). In total, we generated approximately 30 hours of FOA-format audio, with each clip lasting 60 seconds, and generated labels aligned with the STARSS23 format.

The synthetic dataset consists of 70% noisy and 30% clean audio recordings. To match the polyphonic nature of STARSS23, we created synthetic data with three scene complexity levels: simple (one source), medium (1–2 overlapping sources), and complex (up to 3 overlapping sources), distributed at ratios of 55%, 35%, and 10%, respectively, as shown in Figure 3.2. This distribution is intended to support robust learning of individual sound events while providing sufficient training samples for multi-source overlapping scenarios.

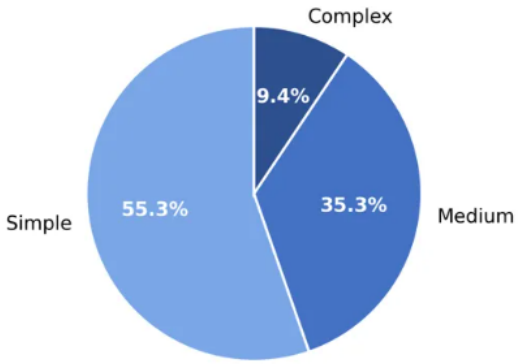


Figure 3.2: Scene complexity distribution of the synthetic dataset. The pie chart illustrates the proportions of the three complexity levels: simple (55%), medium (35%), and complex (10%).

The number of sound events in each scene follows a normal distribution, and dynamic class weighting is applied to ensure the inclusion of rare classes such as water tap, bell, and footsteps. Figure 3.3 shows the class distribution of the synthetic dataset. The 13 sound event classes are relatively balanced, with an average of 1,257 events per class. The *footsteps* class has a higher count (1,807 events) to compensate for its scarcity in the original STARSS23 dataset. Some infrequent sound events, such as *bell* and *water tap*, were also supplemented in the synthetic dataset.

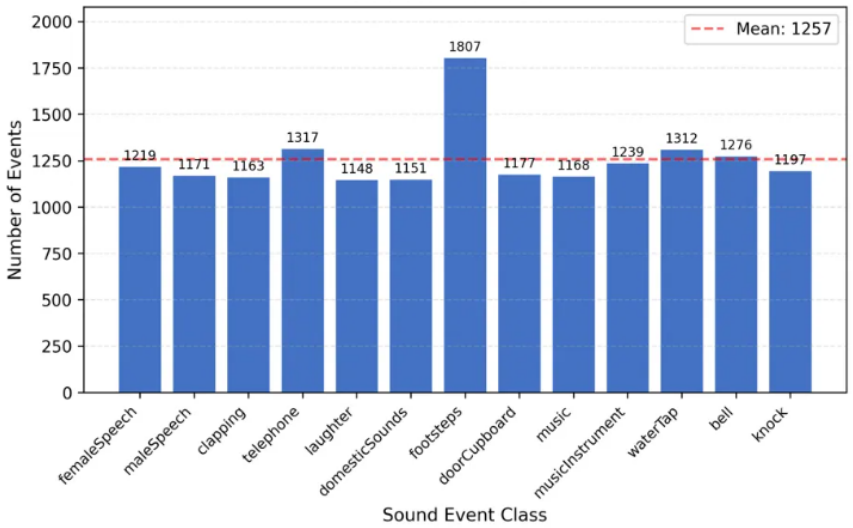


Figure 3.3: Class distribution of the synthetic dataset, with balanced representation across 13 classes of sound events.

CHAPTER 4

METHOD

In this chapter, we introduce the lightweight AV-SELD model proposed in this work, which builds upon the work of Berghi et al[17]. By optimizing the architecture, reducing feature dimensions, and adopting a lightweight visual encoder, we reduce the number of parameters from 87M to 10M, achieving an 8.7× compression rate while maintaining competitive performance.

We selected this baseline for three main reasons: (i) it is open-source and has been published in a peer-reviewed ICASSP paper, (ii) in the DCASE 2023 challenge it ranked 5th overall but was effectively the second-best among distinct teams. Together, these reasons make it a competitive yet tractable baseline.

This chapter is structured as follows: Section 4.1 introduces the overall architecture. Section 4.2 describes the audio and visual encoders, where the audio encoder employs a pretrained Convolutional Neural Network (CNN)-Conformer architecture (Section 4.2.1) and the visual encoder adopts an efficient ShuffleNet-Conformer architecture (Section 4.2.2). Section 4.3 introduces our adaptive dimension adjustment strategy and the fusion layer. Section 4.4 presents the ADPIT loss function, which is specifically designed to handle polyphonic sound events and overlapping sources.

4.1 Overall Architecture

As shown in Figure 4.1 , the proposed AV-SELD model takes as input 4 channel FOA audio signals and 360° videos, employs audio and video encoders, and uses an AV-Conformer for multimodal feature fusion. The audio encoder is based on a CNN-Conformer architecture, where four convolutional layers are used to extract spectral features and a single Conformer layer models temporal dependencies. The visual encoder adopts ShuffleNet with a single

Conformer layer to extract visual features. After concatenating the audio and visual features, we employ a single Conformer layer as the fusion module to model cross-modal dependencies. Finally, a fully connected layer produces a 156-dimensional multi-ACCDOA output. The output includes three detection tracks, each providing four-dimensional information for 13 sound event classes, to achieve parallel detection and localization of multiple sound sources.

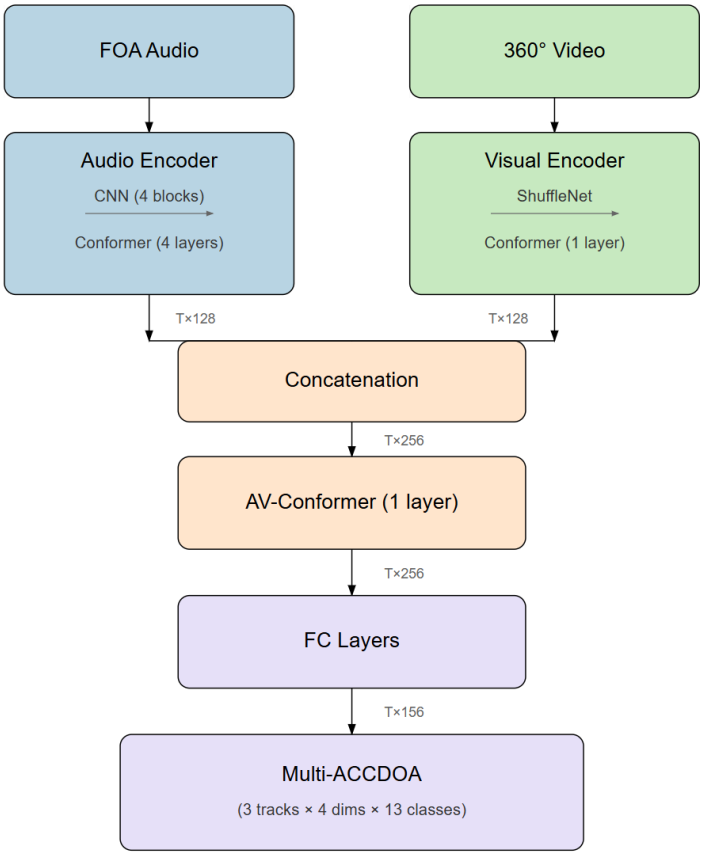


Figure 4.1: System architecture of the lightweight AV-SELD framework

4.2 Audio and Visual Encoder

Audio Encoder The audio encoder is based on a CNN-Conformer architecture. It first converts the 4-channel FOA signals into log-Mel spectrograms and computes 3 intensity vectors, forming a 7-channel audio feature representation $\mathbf{X} \in \mathbb{R}^{7\times T_{\text{in}}\times F_{\text{in}}}$, which serves as

the input to the audio encoder. The network employs a CNN with four convolutional blocks to extract multiscale spectral features. Each block consists of two 3×3 convolutional layers with residual connections, followed by batch normalization and ReLU activation, and uses average pooling with a stride of 2 for temporal–spectral downsampling. The channel dimensions increase from $64 \rightarrow 128 \rightarrow 256 \rightarrow 512$, resulting in feature maps of size $\mathbb{R}^{512 \times T_{\text{in}}/16 \times F_{\text{in}}/16}$. Subsequently, the feature maps are averaged along the frequency dimension to obtain a representation of size $\mathbb{R}^{512 \times T_{\text{in}}/16}$. This representation is then transposed to $\mathbb{R}^{T_{\text{in}}/16 \times 512}$ to form the temporal feature representation. To obtain a compact feature representation, a linear layer reduces the feature dimension to $\mathbb{R}^{T_{\text{in}}/16 \times 128}$. Finally, the features are processed by a 4-layer Conformer with 8 attention heads, a kernel size of 51, and a feed-forward dimension of 1024. The output audio features have dimensions of $\mathbb{R}^{T \times 128}$ with temporal alignment for subsequent fusion.

Visual Encoder To reduce computational complexity for practical deployment, the visual encoder employs a lightweight ShuffleNet v2 architecture in place of the ResNet50 used in baseline approaches, combined with a single Conformer layer. We chose ShuffleNet v2 because it is an efficient CNN designed for mobile deployment. It balances speed, accuracy, and parameter size, which matches the goal of our lightweight AV-SELD model. ShuffleNet introduces pointwise group convolution to reduce computational cost and employs channel shuffle to ensure effective cross-channel information flow within 1×1 convolutions[25]. As illustrated in Figure 4.2, each ShuffleNet unit consists of a 1×1 pointwise group convolution, a 3×3 depthwise convolution, and a residual connection. This design significantly reduces computation while preserving accuracy. In our design, the visual encoder consists of ShuffleNet v2 pretrained on ImageNet, with the classification layer removed for visual feature extraction.

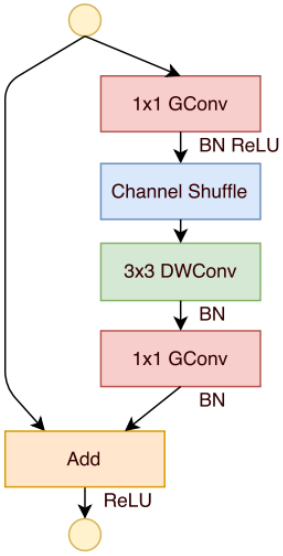


Figure 4.2: ShuffleNet unit architecture with pointwise group convolution and channel shuffle[25]

The ShuffleNet is combined with a single Conformer. The visual encoder processes video frames with a shape of $\mathbb{R}^{T\times224\times448\times3}$, where T is the number of frames, 224×448 is the frame resolution, and 3 represents the RGB channels. Since ShuffleNet v2 $\times0.5$ is designed for 224×224 input, while the preprocessed video frames have a resolution of 224×448, we split each frame into two 224×224 sub-images. After feature extraction, each sub-image produces a 1024-dimensional feature vector, which is concatenated to form $\mathbb{R}^{T\times2048}$. Finally, the concatenated features are fed into a single Conformer layer with 8 attention heads, a kernel size of 51, and a feed-forward dimension of 1024, producing an output of size $\mathbb{R}^{T\times128}$.

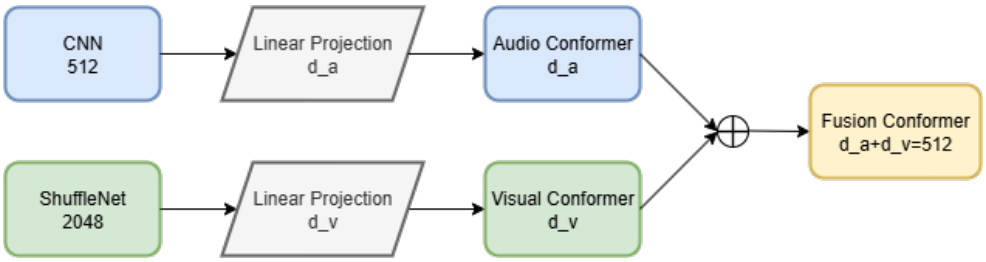
4.3 Audio-Visual Feature Fusion Strategy

4.3.1 Adaptive Dimension Adjustment

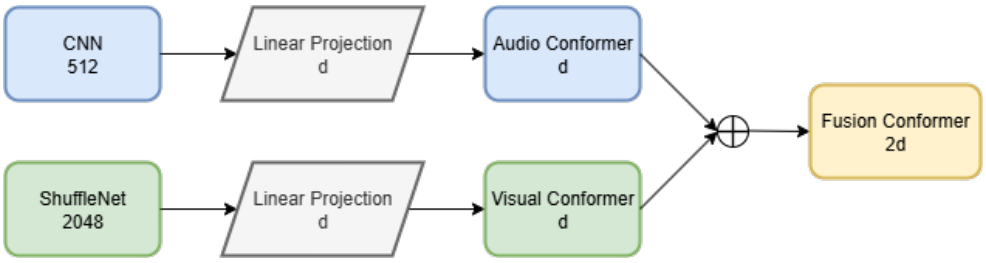
To explore the optimal allocation ratio of audio and visual features, we introduce an adaptive dimension adjustment strategy under the constraint of a fixed total fusion dimension. As shown in Figure 4.3(a), we add a learnable linear projection layer after the CNN en-

coder and before the audio Conformer. A similar linear projection layer is also added in the visual branch, between the output of ShuffleNet and the visual Conformer.

The projection layers adjust the 512-dimensional CNN output and 2048-dimensional ShuffleNet output to the target dimensions d_a and d_v for audio and visual features, respectively, under the constraint $d_a + d_v = 512$. With this design, the Conformer parameters are computed on the reduced feature dimensions, enabling flexible feature allocation while maintaining computational efficiency.



(a) Fixed total dimension with varying ratios



(b) Variable total dimensions with fixed ratio (1:1)

Figure 4.3: Two dimension adjustment strategies: (a) explores audio-visual ratios under fixed total dimension; (b) varies total fusion dimension with balanced allocation.

Beyond allocation ratios, we also study the effect of the total fusion dimensionality. As shown in Figure 4.3(b), we map both audio and visual features to a common dimension $d \in \{64, 128, 256, 512, 1024\}$. The fused representation is their concatenation with size $2d$. This setup tests whether increasing the fusion space improves performance or instead induces overfitting, given the limited training data. This strategy enables systematic exploration of both audio-visual dimension ratios and total fusion capacity within a unified

framework. The projection layers balance efficiency and representation capacity by flexibly adjusting features while preserving critical information.

4.3.2 Fusion and Output Representation

After adjustment, the aligned audio and visual features, with each having a size $\mathbb{R}^{T\times 128}$, are concatenated along the feature dimension to form a fused representation of size $\mathbb{R}^{T\times 256}$. These features are then fed into a single Conformer layer, with 8 attention heads, a kernel size of 51, and a feed-forward dimension of 1024. The Conformer block captures cross-modal relationships between audio and visual features through its attention mechanism and convolutional components.

The resulting bimodal feature representation is mapped to a 156-dimensional output via two fully connected layers. The output applies the Multi-track Activity-Coupled Cartesian Direction of Arrival with Distance (Multi-ACCDOA) format. In this representation, the 156 dimensions correspond to 3 tracks, 4 parameters, and 13 sound event classes. For direction prediction, we use the $\tanh$ activation function to constrain the output to the range $(-1, 1)$, whereas for distance prediction, a ReLU activation is applied to enforce non-negativity.

4.4 ADPIT Loss

In our task, up to three sources of the same class may occur simultaneously. To handle this multi-source scenario, we adopt the Auxiliary Duplicating Permutation Invariant Training (ADPIT) loss function. In the ADPIT strategy, the training labels are organized into a five-dimensional tensor, referred to as target, whose shape is defined as:

$$\texttt{target} \in \mathbb{R}^{B\times T\times N\times A\times C}$$

where B is the batch size, T is the number of time frames, N is the number of dummy tracks, A is the feature dimension, and C is the number of different sound event classes.

Here, we set $B = 32$, $T = 30$, and $N = 6$ in our implementation. The six dummy tracks ($N = 6$) correspond to A0, B0, B1, C0, C1, and C2, where A, B, and C represent cases with one, two, and three overlapping sound sources of the same class per frame, respectively. For each time frame $t \in \{1, \ldots, T\}$, dummy track $n \in \{1, \ldots, N\}$, and sound event class $c \in \{1, \ldots, C\}$, the corresponding label consists of five elements: a flag $a_{nct} \in \{0, 1\}$ indicating the detection activity, a DOA vector $R_{nct} \in [-1, 1]^3$ representing the 3D position (x, y, z) of the sound source with $|R_{nct}| = 1$, and a distance value $D_{nct} \in [0, \infty)$.

Permutation Construction Since each sound event class allows at most three simultaneously active sources, the total number of distinguishable permutations is:

Table 4.1: All 13 distinguishable permutations used in ADPIT training

Type	**Permutations**					
A-type (1 source)	A0A0A0					
B-type (2 sources)	B0B0B1	B0B1B0	B0B1B1	B1B0B0	B1B0B1	B1B1B0
C-type (3 sources)	C0C1C2	C0C2C1	C1C0C2	C1C2C0	C2C0C1	C2C1C0

We extract `target_A0` to `target_C2` from the original tensor $\texttt{target} \in \mathbb{R}^{B \times T \times N \times A \times C}$ by multiplying the activity flag with the corresponding DOA and distance values. Each resulting tensor has the shape $\mathbb{R}^{B \times T \times 4 \times C}$ where the 4 dimensions represent (x, y, z, dis). Then, we construct permutations by concatenating three selected tracks along the third dimension. The 13 permutations, used to align the target labels with the model output in the presence of overlapping sources, are summarized in Table Table 4.1. Each combined target tensor (from `target_A0A0A0` to `target_C2C1C0`) has a shape of $\mathbb{R}^{B \times T \times 12 \times C}$.

MSE Loss Computation The ADPIT loss function is based on the Mean Squared Error

(MSE), which is defined as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Each permutation (from `target_A0A0A0` to `target_C2C1C0`) is compared with the output tensor by computing the MSE along the third dimension. This involves averaging the feature-wise differences (x,y,z and distance) across all tracks for each frame and class. As a result, we obtain 13 loss tensors, each with a shape of $\mathbb{R}^{B\times T\times C}$.

Padding To avoid misleading zero values from other tracks when only one overlap case is active, each permutation is padded with content from the other dummy tracks. For example, `target_A0A0A0` is augmented with information from `target_B0B0B1` and `target_C0C1C2`. The dummy padding strategy may introduce overlapping content across different permutations, but it ensures non-zero targets and enables a branch-free implementation [5].

Minimum Loss Selection The 13 loss tensors are stacked along a new first dimension into a tensor of shape $\mathbb{R}^{13\times B\times T\times C}$. By selecting the minimum across the permutation dimension, we then find the best-matching permutation index for each frame and class and obtain the corresponding minimum loss index tensor with shape $\mathbb{R}^{B\times T\times C}$. Based on these indices, the corresponding loss values are selected, and the overall batch loss is computed by averaging the selected values.

CHAPTER 5

EXPERIMENTAL RESULTS

In this chapter, we investigate the impact of audio–visual feature ratios on performance and validate the effectiveness of our lightweight model. Sections 5.1 and 5.2 describe the training settings and evaluation metrics, which establish the foundation for the subsequent analysis. Then, the experiments are approached from three perspectives. Section 5.3 examines how the balance between audio and visual features influences model performance. Section 5.4 investigates the impact of total fusion dimensions. Section 5.5 integrates these findings and systematically explores different architectural configurations to determine the optimal lightweight design.

5.1 Training Settings

For training setup, we use an NVIDIA RTX 4090 GPU with 24 GB memory and employ a two-stage training strategy: the audio encoder is first trained on the synthetic dataset for 60 epochs; the full audio-visual model is then trained on the STARSS23 dataset for 20 epochs, initialized with the pretrained audio encoder. For hyperparameters, we set the batch size to 32 and the learning rate to 5×10^{-5}, with a decay factor of 0.95 applied after the 30^{th} epoch. To increase sample diversity during training, the audio is split into 3-second segments with 0.5-second overlaps. During testing, 3-second segments without overlap are used to ensure consistent evaluation. The audio sampling rate is set to 24 kHz, and the video frame rate is 10 fps. For feature extraction, audio features are obtained from 128-bin log-Mel spectrograms and 7-dimensional intensity vectors, while visual features are extracted using ShuffleNet. The audio features are initially computed at 160 Hz and then downsampled to 10 Hz to match the label resolution and video frame rate.

5.2 Evaluation Metrics

In our experiment, we use the macro-averaged F1 score, DOA error, and relative distance error as the main evaluation metrics. The F1 score evaluates detection accuracy under a location constraint, DOA error measures the angular localization accuracy, and relative distance error measures the distance component of localization accuracy. These three metrics combine to form the SELD score, with lower values indicating better overall performance. The final model is selected based on the lowest SELD score to optimize both detection and localization.

$$F_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_{\text{spatial},c} + 0.5 \times (FP_c + FN_c)} \tag{5.1}$$

Here, $\sum_{c=1}^{C} \text{TP}_c$ is the number of correctly detected sound events across all frames and classes. A detection is considered correct if it exists in both prediction and ground truth labels, and satisfies the spatial localization constraints: $\theta_{ij} \leq 20^\circ$ AND $\rho_{ij} \leq 1.0$, where θ_{ij} is angular error and ρ_{ij} is relative distance error.

$\text{FP}_{\text{spatial}}$ is the number of false positives caused by incorrect spatial localization. Specifically, if a predicted event matches the correct class label in the ground truth but fails to meet the spatial localization constraints, it is counted as $\text{FP}_{\text{spatial}}$. FP_c counts false positives, including cases where a class is over-predicted compared to the ground truth and cases where the predicted class doesn't exist in the ground truth at all. Conversely, FN_c is the number of undetected events, which are present in the ground truth labels but missed by the prediction.

The DOA error measures the angular accuracy of predicted sound source directions and is computed only for sound events that are present in both the prediction and ground truth labels. It is defined as:

$$\text{DOA}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \frac{\sum_{i=1}^{N_{TP,c}} \theta_i}{N_{TP,c}} \tag{5.2}$$

Here, $N_{TP,c}$ denotes the number of correctly matched event pairs for class c. θ_i is the angular error of the i-th matched pair. It is defined as the angular distance between the ground-truth and predicted direction vectors in 3D space:

$$\theta = \arccos(\hat{\mathbf{v}}_{gt} \cdot \hat{\mathbf{v}}_{pred}) \times \frac{180}{\pi} \tag{5.3}$$

In this equation, $\hat{\mathbf{v}}_{gt}$ and $\hat{\mathbf{v}}_{pred}$ denote the ground-truth and predicted direction vectors after normalization, respectively.

$$\text{RelDistE}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \frac{\sum_{i=1}^{N_{TP,c}} \delta_i}{N_{TP,c}} \tag{5.4}$$

The relative distance error calculation is also computed for successfully matched event pairs, where δ_i denotes the relative error for the i-th matched pair, defined as:

$$\delta_i = \frac{|d_{gt,i} - d_{pred,i}|}{d_{gt,i}} \tag{5.5}$$

where $d_{gt,i}$ and $d_{pred,i}$ are the ground-truth and predicted distances of the i-th matched pair, respectively. The error is divided by the true distance to normalize it and make different distance scales comparable.

The SELD (Sound Event Localization and Detection) score is employed as an overall performance metric. It is computed as the arithmetic average of three normalized metrics, providing a balance between detection and localization:

$$\text{SELD} = \frac{1}{3} \left[(1 - F1) + \frac{\text{DOA Error}}{180} + \text{RelDistE} \right] \tag{5.6}$$

5.3 Audio-Visual Comparison

We designed an audio-visual comparison experiment to quantify the contributions of different audio and visual feature ratios and to explore the optimal dimensional configuration for

feature fusion. With the total fusion dimension fixed at 512, the allocation between audio and visual representations was varied to find an optimal trade-off between performance and efficiency.

In this setup, both visual and audio features are compressed to target dimensions through linear layers, from ShuffleNet outputs and 512-dimensional CNN encoder outputs respectively. Both features are then concatenated to form a 512-dimensional input to the fusion Conformer module. All experiments are conducted using the same 4-layer Conformer architecture, with only feature dimensions varied to ensure a fair comparison. The specific allocation is summarized in the table below, covering configurations from audio-dominant to visual-dominant.

Table 5.1: Performance of audio–visual feature dimension allocations with the total fusion dimension fixed at 512. Bold values indicate the best in each metric. Overall, balanced or mildly skewed allocations provide more stable performance than extreme audio- or visual-dominant settings.

Model	F1 Score (%)↑	DOA Error (°)↓	Relative Distance Error (%)↓
A64_V448	34.0	20.6	**28.37**
A128_V384	43.5	17.9	30.91
A192_V320	42.8	17.7	30.91
A256_V256	43.9	**17.5**	30.26
A288_V224	**44.9**	17.9	30.34
A320_V192	42.7	18.5	30.84
A384_V128	43.9	18.3	30.51
A448_V64	44.5	18.1	30.27

Table 5.1 summarizes the performance of different audio and visual feature dimension configurations. The results show a complex, non-linear relationship between the feature dimensions and the overall performance. Specifically, increasing the audio dimension from 64 to 288 improves the F1 score from 34% to 44.9%, representing a gain of approximately 10 percentage points. However, when further increasing it to 448, the F1 score first decreases and then gradually recovers towards 44.5%, showing a non-monotonic

trend. Meanwhile, as the audio dimension increases, the relative distance error remains around 30%. Although the trend is non-linear, the audio-dominant configurations achieve overall superior F1 scores. For example, A288_V224 achieves the best performance with an F1 score of 44.9%, followed by A448_V64 with 44.5%. Conversely, reducing the audio feature dimension to a minimal size leads to a sharp performance degradation, with A64_V448 achieving only 34.0% F1—10.9 percentage points lower than the optimal configuration. This shows that compressing the audio dimension to 64 severely impairs model performance and suggests a practical lower bound for effective audio representation.

Meanwhile, different evaluation metrics exhibit varying sensitivities to the dimension allocation. The DOA error achieves its lowest value when the audio and visual feature dimensions are balanced, as in the A256_V256 configuration. The relative distance error performs best when the visual feature dimension is high. Excessively increasing the visual dimension, such as to 448, presents a trade-off: while it improves distance estimation, the F1 score drops significantly. This highlights the need to balance multiple evaluation metrics. Such a non-linear relationship may result from a multi-objective training strategy, where multiple evaluation metrics are optimized jointly. These results suggest that the model implicitly balances detection precision, localization accuracy, and distance estimation.

Experimental observations indicate that a balanced audio-visual configuration achieves better trade-offs among detection, localization, and distance estimation. Therefore, we choose A256_V256, a 1:1 ratio setup, as the baseline to explore the impact of overall fusion dimensionality. This ratio is kept constant in all subsequent experiments.

5.4 Fusion Dimension Analysis

This experiment is designed to analyze the impact of the total feature dimension on model performance and to explore its potential for lightweight model design. In contrast to the experiment in Section 5.3, which varies the dimensional allocation ratio between audio and

visual features, here the ratio is fixed while the total dimension is adjusted. We designed six total dimension configurations ranging from 128 to 2048, as shown in Table 5.2.All configurations share the same network architecture.

Table 5.2: Performance of AV-SELD models with different total fusion dimensions under a fixed audio-visual ratio.

Model	**F1 score (%)↑**	**DOA Error (°)↓**	**Relative Distance Error (%)↓**	**Parameters**
AV_64_64	34.7	21.0	30.53	10,054,452
AV_96_96	41.4	18.8	36.49	13,140,468
AV_128_128	42.2	17.2	30.44	16,570,548
AV_256_256	43.9	17.5	30.46	33,731,508
AV_512_512	43.4	17.1	30.54	84,305,844
AV_1024_1024	43.0	16.8	30.78	252,302,772

From the Table 5.2, the model performance first increases and then declines as the total fusion dimension increases, especially in terms of the F1 score, which shows signs of saturation beyond a certain point. The F1 score reaches its peak at AV_256_256 and then declines, while the DOA error continues to improve as the fusion dimension increases.The relative distance error fluctuates around 30%, suggesting that this metric is not particularly sensitive to changes in fusion dimension.

Besides, some outliers such as AV_64_64 with a low F1 score and AV_96_96 with a high relative distance error may be due to the limited model capacity caused by the small feature dimensions. Therefore, small feature dimensions are insufficient for effective modeling, and increasing the total fusion dimension enhances performance until it reaches a saturation point.

The model's parameter size grows exponentially as the fusion dimension increases. When the performance reaches saturation, further increasing the dimension leads to an unnecessary computational burden. To balance performance and model complexity, we choose AV_128_128, with a total fusion dimension of 256, as the lightweight configuration. This model achieves 96% of the highest F1 score (AV_256_256) while reducing parameters by

approximately 51%, and even slightly improving DOA error and relative distance error, making it a strong candidate for deployment in resource-constrained scenarios.

5.5 Model Efficiency Analysis

To develop a lightweight model for the AV-SELD task, we conducted a series of progressive lightweight experiments that systematically reduced model parameter complexity while evaluating performance changes.

First, we compared different visual encoders, including ResNet18 and ShuffleNetV2, with the baseline ResNet50. With significantly fewer parameters, these encoders allow us to evaluate the performance impact of reduced visual complexity. From Table 5.3, it can be observed that the F1 scores remain consistent across different visual encoders, all around 42%. For DOA error, ResNet18 achieves the best performance at 17.1°, followed closely by ShuffleNet (17.2°), while ResNet50 shows slightly higher error at 17.6°. Moreover, ShuffleNet achieves the lowest relative distance error, 30.44%, outperforming the other models. While maintaining competitive performance, ShuffleNet requires only 341,792 parameters, representing a 98.5% reduction compared to ResNet50. These findings suggest that high-capacity visual encoders do not necessarily lead to performance improvements in AV-SELD. In contrast, lightweight and efficient designs can substantially reduce computational and storage costs while maintaining competitive performance.

Table 5.3: Comparison of visual encoder architectures in terms of AV-SELD performance and parameter efficiency.

Visual Encoder	**F1 score (%)↑**	**DOA Error (°)↓**	**Relative Distance Error (%)↓**	**Parameters**
ResNet50	42.1	17.6	30.91	23,508,032
ResNet18	42.0	**17.1**	31.43	11,176,512
ShuffleNet_V2	**42.2**	17.2	**30.44**	341,792

Then, we conducted an ablation study to investigate the effect of varying the number of

Conformer layers on the overall performance. Based on the optimal dimension configuration (A128_V128), we first varied the number of visual Conformer layers, evaluating four structural variants: 4–4–4 (baseline), 4–3–4, 4–2–4, and 4–1–4. The three numbers represent the number of audio Conformer layers, visual Conformer layers, and fusion Conformer layers, respectively. Next, we varied the audio Conformer layers, including experiments on audio-only models that excluded the visual encoder. Finally, we tested different numbers of fusion Conformer layers, ultimately reducing them to a single layer in the 4–1–1 configuration. The results are summarized in Table 5.4. By comparing different visual Conformer

Table 5.4: Performance comparison under different architecture configurations. Each configuration is denoted as A–V–F, where A, V, and F indicate the number of Conformer layers in the audio encoder, visual encoder, and fusion module, respectively.

Configuration	**F1 score (%)↑**	**DOA Error (°)↓**	**Relative Distance Error (%)↓**	**Parameters**
Baseline	40.8	17.7	30.50	*85,354,420*
4_4_4	42.2	17.2	30.44	16,570,548
4_2_4	43.0	18.4	29.28	15,269,298
4_1_4	44.3	18.5	29.76	14,618,673
4_1_1	**44.7**	17.7	31.15	10,032,942
2_1_1	30.3	19.6	30.38	–

layer configurations from 4-4-4 to 4-1-4, we observe that F-score improves from 42.2% to 44.3% as visual layers decrease, despite a slight increase in DOA error. This suggests that, for AV-SELD, stacking multiple visual layers offers little benefit. A single-layer visual Conformer is sufficient to capture the essential spatial cues from 360° panoramic videos. Deeper visual layers may introduce noise or lead to overfitting, which could explain the observed performance degradation.

We also observe that reducing the fusion layer further improves performance, with the F-score increasing from 44.3% to 44.7% while the DOA error decreases from 18.5° to 17.7°. Although the relative distance error increases slightly from 29.76% to 31.15%, the parameter count is reduced by approximately 38%, from 16.57M to 10.03M. This finding suggests that efficient cross-modal interaction does not require extensive fusion mechanisms, achiev-

ing parameter reduction while preserving performance.

Notably, reducing the audio encoder depth causes substantial performance decline. In the 2_1_1 configuration, F1 score drops from 44.7% to 30.3% and DOA error increases from 17.7° to 19.6°. This result confirms that audio dominates the AV-SELD task, and the 4-layer CNN-Conformer structure is essential for extracting spatial audio features and preserving performance.

Overall, the 4_1_1 configuration achieves the optimal balance between performance and efficiency. It significantly reduces parameters from 85.35M to 10.03M, improves the F1 score from 40.8% to 44.7%, and maintains the DOA error at 17.7°, with only a slight increase in relative distance error. It substantially reduces parameters, not only maintaining baseline performance but also improving detection accuracy. These results validate the effectiveness of the audio-dominated lightweight design for AV-SELD and highlight its practicality for deployment on resource-constrained devices.

CHAPTER 6

CONCLUSION

In this work, we propose an audio-dominant lightweight AV-SELD model that reduces parameters by 88% while significantly improving detection performance. This design successfully achieves the dual goals of efficiency and accuracy.

Our study revealed several key findings. First, a balanced 1:1 ratio between audio and visual features achieves the best overall performance, particularly in DOA error and relative distance error. With this ratio fixed, performance improves as the total feature dimension increases, but reaches saturation at 256 dimensions. Beyond this point, further increases only add computational overhead without yielding meaningful gains. Second, the results show the dominance of the audio modality in AV-SELD tasks. This allows us to employ a lightweight visual encoder using a simple feature extractor with a single-layer Conformer, which still outperforms deeper counterparts in our experiments. Third, a single fusion layer is sufficient to model audio-visual complementarity, while excessive stacking of layers and features can introduce noise or lead to overfitting.

We believe that large models tend to overfit on small datasets, often memorizing noise and dataset-specific patterns instead of learning generalizable features. While we compress the model to 10 million parameters, the limited capacity forces it to focus on the essential acoustic features and spatial cues. Therefore, the smaller model enhances generalization and robustness against environmental noise and irrelevant visual distractors. Additionally, excessive parameters lead to optimization difficulties, such as gradient vanishing/exploding and poor local optima. In contrast, lightweight models reduce training complexity, making optimization more stable and convergence faster.

Our work has important practical implications. AV-SELD is widely used in smart surveillance, robotics, and smart home applications, which often operate on resource-constrained

edge devices. The proposed lightweight model, with only 10M parameters, can be directly deployed on embedded systems, mobile devices, and IoT platforms, enabling practical deployment of multimodal perception systems.Meanwhile, our findings offer valuable insights for designing lightweight architectures in other multimodal tasks, highlighting the importance of tailoring architectures according to the relative contributions of different modalities to the target task.

Despite the promising results, several aspects of this work remain to be improved. First, the ADPIT loss function for multi-task joint learning requires balancing multiple objectives, which may lead to trade-offs in individual metrics. Second, the model was only evaluated on STARSS23, so its generalization to other datasets remains to be tested. Therefore, future work may focus on the following directions:

- Exploring knowledge distillation and quantization techniques to achieve further model compression.
- Designing adaptive loss weighting strategies to better balance the objectives in multi-task learning
- Validating the model's generalization performance on additional SELD datasets beyond STARSS23.

REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[2] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, IEEE, 2024, pp. 286–290.

[3] A. Mesaros *et al.*, "Sound event detection in the dcase 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 2019.

[4] S. Michaud, B. Moffett, A. T. Rousiouk, V. Duda, and F. Grondin, "Smartbelt: A wearable microphone array for sound source localization with haptic feedback," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2023, pp. 1950–1955.

[5] M. Nakata, Y. Hirao, M. Perusquía-Hernández, N. Isoyama, H. Uchiyama, and K. Kiyokawa, "A vibrotactile device for enabling sound localization and identification for deaf and hard of hearing individuals," in *Proceedings of the 16th Asia-Pacific Workshop on Mixed and Augmented Reality (APMAR 2024)*, ser. CEUR Workshop Proceedings, vol. 3907, CEUR-WS.org, 2024.

[6] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2018, pp. 74–79.

[7] Z. Shi, L. Zhang, and D. Wang, "Audio–visual sound source localization and tracking based on mobile robot for the cocktail party problem," *Applied Sciences*, vol. 13, no. 10, p. 6056, 2023.

[8] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[9] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, pp. 1–46, 2016.

[10] Z. Dosbayev *et al.*, "Audio surveillance: Detection of audio-based emergency situations," in *International Conference on Computational Collective Intelligence*, Springer, 2021, pp. 413–424.

[11] C. Grobler, C. P. Kruger, B. J. Silva, and G. P. Hancke, "Sound based localization and identification in industrial environments," in *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*, IEEE, 2017, pp. 6119–6124.

[12] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," *arXiv preprint arXiv:1905.08546*, 2019.

[13] J. Zhao *et al.*, "Audio-visual speaker tracking: Progress, challenges, and future directions," *arXiv preprint arXiv:2310.14778*, 2023.

[14] L. Nanni, Y. M. Costa, D. R. Lucio, C. N. Silla Jr, and S. Brahnam, "Combining visual and acoustic features for audio classification tasks," *Pattern Recognition Letters*, vol. 88, pp. 49–56, 2017.

[15] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end audiovisual fusion with lstms," *arXiv preprint arXiv:1709.04343*, 2017.

[16] D. Dean, P. Lucey, S. Sridharan, and T. Wark, "Weighting and normalisation of synchronous hmms for audio-visual speech recognition," in *Proceedings of the Workshop on Audio-Visual Speech Processing Cognitive and Computational Approaches*, Tilburg University, 2007, pp. 110–115.

[17] D. Berghi, P. Wu, J. Zhao, W. Wang, and P. J. Jackson, "Fusion of audio and visual embeddings for sound event localization and detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 8816–8820.

[18] Q. Wang *et al.*, "The nerc-slip system for sound event localization and detection with source distance estimation of dcase 2024 challenge," *DCASE2024 Challenge, Tech. Rep.*, 2024.

[19] G. Kim and H. Ko, "Data augmentation, neural networks, and ensemble methods for sound event localization and detection," *DCASE2023 Challenge, Tech. Rep.*, 2023.

[20] A. Berg, J. Engman, J. Gulin, K. Aström, M. Oskarsson, and B. Sony Europe, "The lu system for dcase 2024 sound event localization and detection challenge," DCASE2024 Challenge, Tech. Rep, Tech. Rep., 2024.

[21] Y. Jiang *et al.*, "Exploring audio-visual information fusion for sound event localization and detection in low-resource realistic scenarios," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2024, pp. 1–6.

[22] D. Gimeno-Gómez and C. D. Martinez-Hinarejos, "Tailored design of audio-visual speech recognition models using branchformers," *Computer Speech & Language*, p. 101 811, 2025.

[23] I. R. Roman, C. Ick, S. Ding, A. S. Roman, B. McFee, and J. P. Bello, "Spatial scaper: A library to simulate and augment soundscapes for sound event localization and detection in realistic rooms," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 1221–1225.

[24] K. Shimada *et al.*, "Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *Advances in neural information processing systems*, vol. 36, pp. 72 931–72 957, 2023.

[25] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.