Audio-Visual Sound Event Localization and Detection Based on CRNN Using Depth-Wise Separable Convolution

Yi Wang[1(✉)], Hongqing Liu[2], Yu Zhao[1], and Yi Zhou[2]

[1] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China
s210101136@stu.cqupt.edu.cn

[2] Intelligent Speech and Audio Research Lab., Chongqing University of Posts and Telecommunications, Chongqing, China

Abstract. Sound event localization and detection (SELD) focuses on the simultaneous detection of various sound events along with their sptial and temporal localization. Recent work shows that audio-visual fusion methods, rarely involved in SELD research, show promising results than the single modality. For this, we proposed an audio and visual signals fusion mechanism for SELD based on convolutional recurrent neural network (CRNN). Object detection and pre-trained model processing on the corresponding image at the start frame of the audio feature sequence is utilized to acquire visual cues passed into models. Compared to traditional convolution, we devise a depth-wise separable convolution block to better learn the relevant information of different sound event categories in audio features. Experimental results on STARSS23 of DCASE (2023) indicate that the introducing of visual cues do improve the SELD performance compared to the audio-only system. The convolution block devised in our proposed work further enhances the model's performance as it achieves higher SELD score.

Keywords: Sound event localization and detection · Deep learning · Audio-visual fusion · Convolutional recurrent neural network

1 Introduction

Sound event localization and detection (SELD) involves the integration of sound event detection (SED) and direction-of-arrival (DoA) estimation, forming a collaborative task. The primary objective of sound event detection (SED) is to identify the start and end times of sound events associated with specific target categories. In contrast, direction-of-arrival (DoA) estimation aims to predict the spatial locations of distinct active sound events. Polyphonic SELD refers to scenarios where multiple sound events occur simultaneously and overlap in time. Due to its ability to characterize sound sources spatially and temporally, SELD

F. Gao et al. (Eds.): Chinacom 2023, LNICST 590, pp. 364–378, 2024.
https://doi.org/10.1007/978-3-031-67162-3_23

can be used to automatically describe social and human activities and presents a wide range of applications, such as audio surveillance [1], automatic speech recognition (ASR) [2,3] and environmental sound classification [4].

In early years, Gaussian mixture models (GMM) [5], hidden Markov models (HMM) [6,7] and array processing techniques, such as multiple signal classification [8] and steered response power [9,10] were used for SED and DoA estimation task, respectively. However, these methods are usually not robust in noisy environments. Recent research has shown notable advancements in SELD performance through the utilization of deep neural network (DNN) based approaches. Adavanne et al. introduced a pioneering approach in the field of SELD by proposing SELDnet [11]. SELDnet is a convolutional recurrent neural network (CRNN) model specifically designed for polyphonic SELD, where it addresses the joint tasks of sound event detection (SED) and direction-of-arrival (DoA) estimation. Shimada et al. [12] employed a consolidated output representation known as the activity-coupled Cartesian DOA (ACCDOA) representation, which combines the predictions of sound event detection (SED) and direction-of-arrival (DOA). To address the issue of overlapping events from the same class, this method was further developed into a multi-ACCDOA target format [13]. The Detection and Classification of Acoustic Scenes and Events (DCASE) challenge, which is a very popular sound detection challenge today, was founded in 2019, and SELD was first as task 3 in the challenge. With the annual continuation of the DCASE Challenge, state-of-the-art methods in the sound event localization and detection (SELD) task often leverage intricate network architectures and extensive data augmentations [14–16].

As is known to all, human brain usually utilizes the combined senses of hearing and vision to perceive its surroundings and extract valuable complementary information. Recently, audio-visual systems arise in applications such as speaker diarization [17], speaker detection [18] with the release of these competition related datasets. Qian et al. [19] collects a realistic dataset on a robotic platform and adapts audio-visual fusion method for tracking speaker spatially by estimating their Direction of Arrival (DoA). For audio-visual sound source localization, most current researches localize sound sources in a video frame assuming they are visible [20,21], instead of their spatial positions. These approaches about different tasks with different network structures but employing audio-visual fusion methods all presented good performance. In contrast to relying solely on audio data, video data holds the potential to alleviate challenges and uncertainties in the spatio-temporal characterization of acoustic scenes. However, due to the absence of a formal, carefully labeled audio-visual dataset and its more complex sound event categories and scenarios, there are few literature on audio-visual SELD. DCASE 2023 released Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) extended form STARSS22 [22], a manually annotated dataset made of recordings of real sound scenes. It adds an additional 4 h of material captured in Tampere University distributed between the training and evaluation sets. It further includes simultaneous 360° video recordings for all the audio recordings and it augments the respective labels with source distance informa-

tion, apart from the direction-of-arrival to stimulate further developments on SELD research.

In this work, we propose an audio-visual network based on CRNN for sound event localization and detection. We improve CRNN networks by introducing visual modules and use the corresponding image at the start frame of the acoustic feature to conduct objection detection and process by pre-trained models for obtaining visual feature. We devise a convolution block based on depth-wise convolution with multi-scale kernel size instead of conventional 2D convolution. In the following sections, we will detail the proposed network. Then, comparisons were conducted to investigate the performance of our proposed audio-visual method between DCASE 2023 official baseline and audio-only from experimental results on STARSS23.

2 The Proposed Method

In this part, we introduce our proposed approach for audio-visual SELD based on CRNN. We employ a late fusion strategy and Fig. 1 shows the overall process of our framework which consists of feature extractor, audio encoder, video encoder, and decoder. We extract features from audio and visual input separately and send them into unimodal encoder. Then, we fusion these embeddings and utilize decoder to perform temporal modeling along with output mapping.

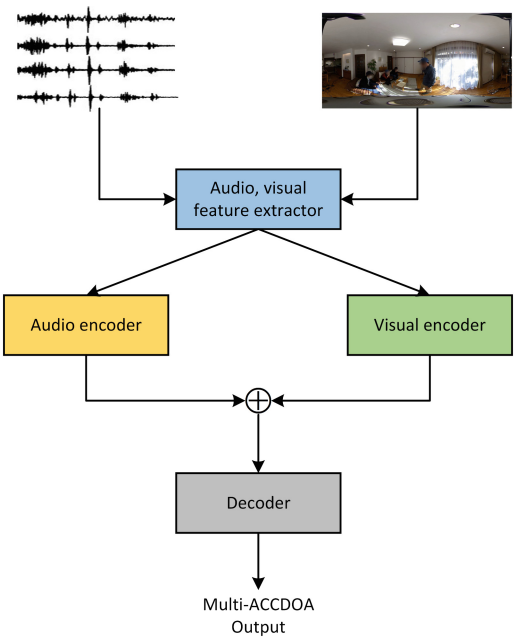

Fig. 1. The diagram of audio-visual fusion SELD network.

2.1 Feature Extraction

Audio Feature. SELDnet was trained using multichannel magnitude and phase spectrogram as the input data during its initial introduction [11]. Afterward, alternative feature representations were explored and demonstrated to be more effective for SELD. These included utilizing multichannel log-spectrograms and intensity vectors (IVs) for the FOA format, as well as employing the generalized cross-correlation with phase transform (GCC-PHAT) on the mel-scale for the MIC format. Compared with IVs, inter-channel phase differences (IPDs) also contain the directional information associated with different sound source. It had been used as spatial feature in [12,23], which showed a promising performance. In our proposed work, we also use IPDs together with the amplitude of complex spectrograms as audio feature. For audio feature extraction as showed in Fig. 2, short-time Fourier transform (STFT) is first utilized to transform the raw audio signal into complex spectrograms. Then, the amplitude of complex spectrograms and IPDs are calculated respectively. The IPD can be computed as

$$IPD_{t,f,p,q} = \angle x_{t,f,p} - \angle x_{t,f,q}, \tag{1}$$

where $x_{t,f,p}$ and $x_{t,f,q}$ are STFT coefficients, t, f, p, and q denote the time frame, the frequency bin, the Ambisonic channel p, and the channel q, respectively. We fix $p = 0$ to compute relative IPDs between all the other channels, i.e., $q \neq 0$. The overall output dimension of audio feature extraction is $C \times T \times M$, where C is sum of channels from the amplitude of complex spectrograms and IPDs, M is half of FFT point size for taking positive frequencies only, T represents all frames included in one audio clip.

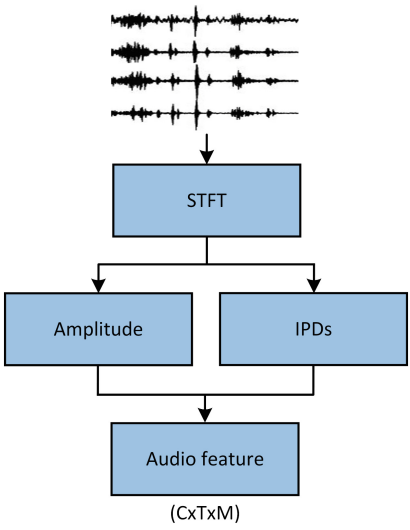

Fig. 2. The overall process diagram of audio feature extraction.

Visual Feature. Object detection, being a fundamental concern in the realm of computer vision, has facilitated the development of numerous practical applications, including but not limited to autonomous driving, robot vision, and surveillance systems [24]. Its result represented by bounding box not only show the object category, but also contains spatial information of the object. MMDetection [25] is a comprehensive object detection toolbox that offers a diverse array of methods for object detection and instance segmentation, along with associated components and modules. It encompasses functionalities for both training and inference, while also supplying pre-trained weights for over 200 network models. Based on mmdetection, we select YOLOX-Tiny, a light model of YOLOX [26] pre-trained on COCO dataset, as object detector to detect person in the corresponding image at the start frame of the audio feature sequence. The detection results are visualized in Fig. 3.

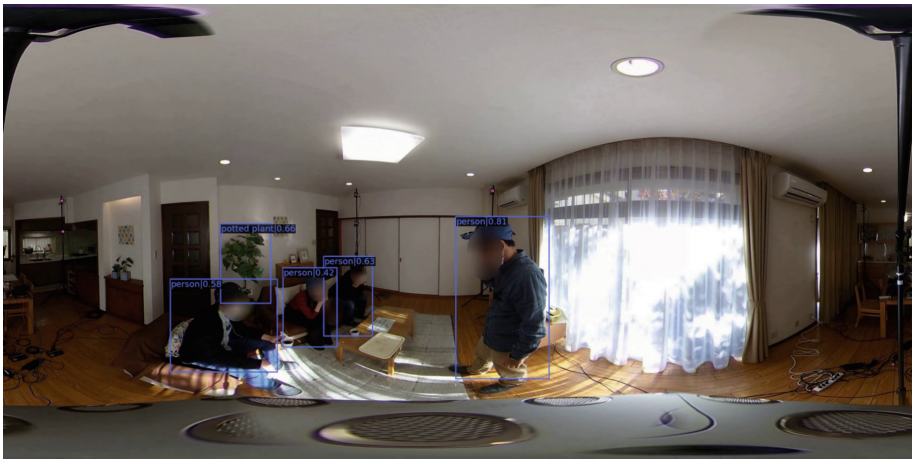

Fig. 3. The visualization of object detection results

Assume that the detected bounding box of a person at time t is represented by b_t, given by

$$b_t = (u_t, v_t, w_t, h_t), \tag{2}$$

where (u_t, v_t) represents the position of the top-left corner, and (w_t, h_t) indicates the width and height of the bounding box, respectively. The bounding boxes of these people are transformed to a concatenation of two Gaussian-like vectors $\rho_t(u), \rho_t(v)$ as visual feature, which represents likelihoods of objects present along the image's horizontal axis u and vertical axis v, respectively [19]. For example, the horizontal Gaussian-like vector is formulated as

$$\rho_t(u) = \begin{cases} exp(-\frac{|u-\mu_{u,t}|^2}{\sigma_u^2}), & b_t \neq \emptyset \\ 0, & otherwise \end{cases} \tag{3}$$

where $\mu_{u,t} = u_t + \frac{1}{2}w_t$ is the horizontal center of b_t, and σ_u is the standard deviation. The formulation for vertical vectors is the same by replacing u with v, $\mu_{u,t}$ with $\mu_{v,t} = v_t + \frac{1}{2}h_t$, and σ_u with σ_v in (3). Then, We can get Gaussian-like vectors with a shape of $2 \times B \times L$, where 2 stands for horizontal and vertical axis, B, L represent the preset number of bounding boxes and length of Gaussian distribution, respectively. Consider the fact that there are some sound event classes in STARSS23 that are not related to people, we pass raw video frame which is also the start frame of the audio feature sequence into a pre-trained ResNet-18 [27] model to produce a visual embedding with one dimension of 1000 as additional visual feature. In Fig. 4, we show our overall process for extracting visual features.

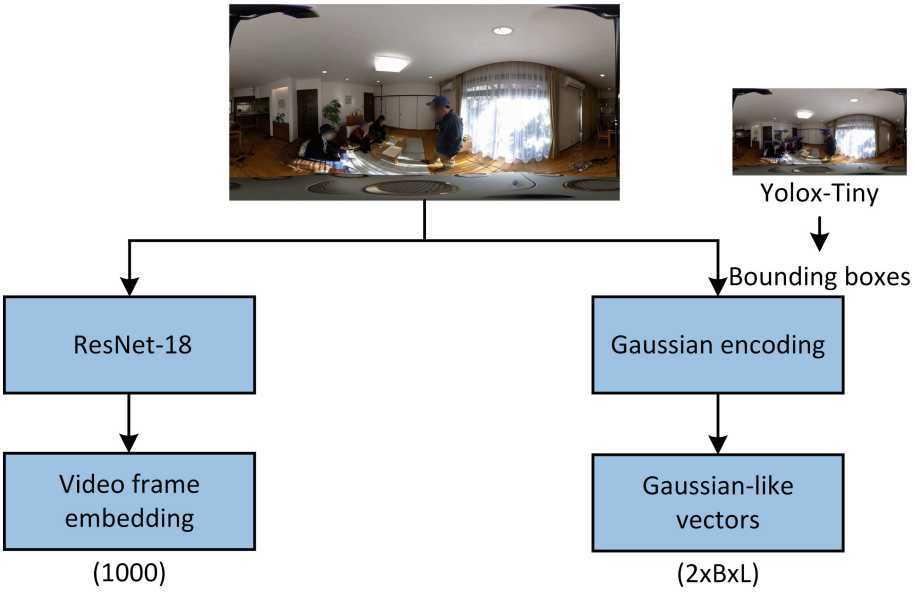

Fig. 4. The overall process diagram of visual feature extraction.

2.2 Network Architecture

Audio Encoder. The audio encoder architecture in our audio-visual fusion SELD network is shown in Fig. 5. To better catch the relevant information of different sound event categories in audio features, we introduce a depth-wise separable convolution (DSC) block composed of two depth-wise 2D convolution with kernel sizes of $3 * 3$ and $5 * 5$, respectively, and a point-wise 2D convolution with kernel size of $1 * 1$ to perform channel fusion, shown in Fig. 5(right). In audio encoder, a DSC block with F filters is used to up-sample the input audio channel. After that, time and frequency down-sampling blocks also employing

F filters are used to learn intra-channel features and reduce dimension size for feature fusion. While the use of filters spanning all the channels make the encoder to learn inter-channel features. For time and frequency down-sampling block in Fig. 5(left), a DSC block is first used to further encode the audio feature. Then, we use both 2D convolution with different strides and kernel sizes, and max-pooling to down-sample audio feature in time and frequency dimensions. The feature after 2D convolution and max-pooling are added as a residual connection. The kernel size used in 2D convolution is $5*5$, $3*5$, $3*3$ with stride $4*4$, $2*4$, $2*2$ respectively. Padding is adaptively used for getting a proper dimension. The kernel size for Max-pool is same as the stride in 2D convolution. Among these blocks, rectified linear unit activation and batch normalization are inserted to introduce nonlinear characteristics. After audio encoder, we average the output along the time dimension to generate audio embedding F_a with $T_1 \times F$ dimension.

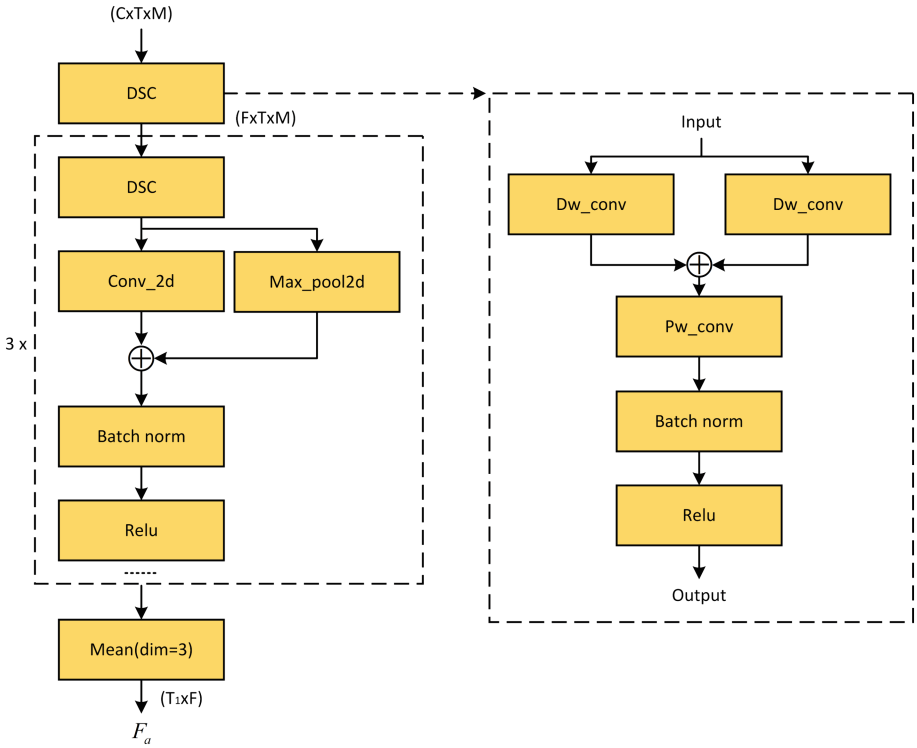

Fig. 5. The modules of audio encoder (left), the structure of DSC block (right).

Visual Encoder. As the video object Gaussian-like vectors and raw video frame embedding have been both pre-processed, we use two fully connected layers to encode them into visual embedding. It should be noted that the dimension of different modalities embedding must be consistent in order to be fused. For Gaussian-like vectors, we first adjust them to one-dimensional and map them to the same dimension as audio embedding using one of fully connected layers. In order to further encode the video frame embedding of the pre-trained ResNet-18 model, an intermediate dimension is set. Then, We expand and repeat their time dimensions to get F_{vo}, F_{vf} with $T_1 \times F$ dimension same as F_a. We display the visual encoder structure in Fig. 6.

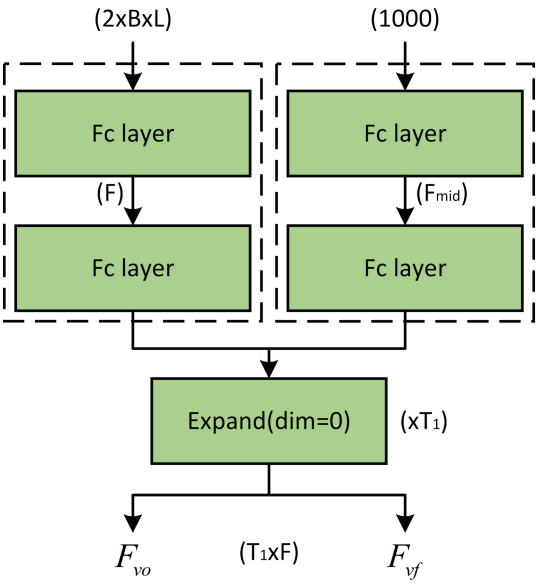

Fig. 6. The structure of visual encoder.

Decoder. In Fig. 7 we display the composition of the decoder. For data fusion, we adopts a late fusion paradigm as mentioned before. We straightly concatenate F_a with F_{vo} and F_{vf} respectively considering the complexity of the model. Then, we separately pass them into two bidirectional RNN layers for learning temporal context information. Each of RNN layers use 256 nodes of gated recurrent units (GRU) in our work specifically. After that, the two parts are added to obtain the fusion embeddings F_{av}. A fully connected layer is used subsequently to map the F_{av} into output dimension. Due to time down-sampling used in audio encoder, an up-sampling operation called interpolate in the temporal dimension is conducted to ensure the output size is consistent with label temporal dimension T. At last, the output result is reshaped into Multi-ACCDOA format, $3 \times 3 \times N \times T$, where

3 represent the maximum number of simultaneous sound events in a frame and the coordinates of x, y, and z, respectively, N is the number of sound event categories.

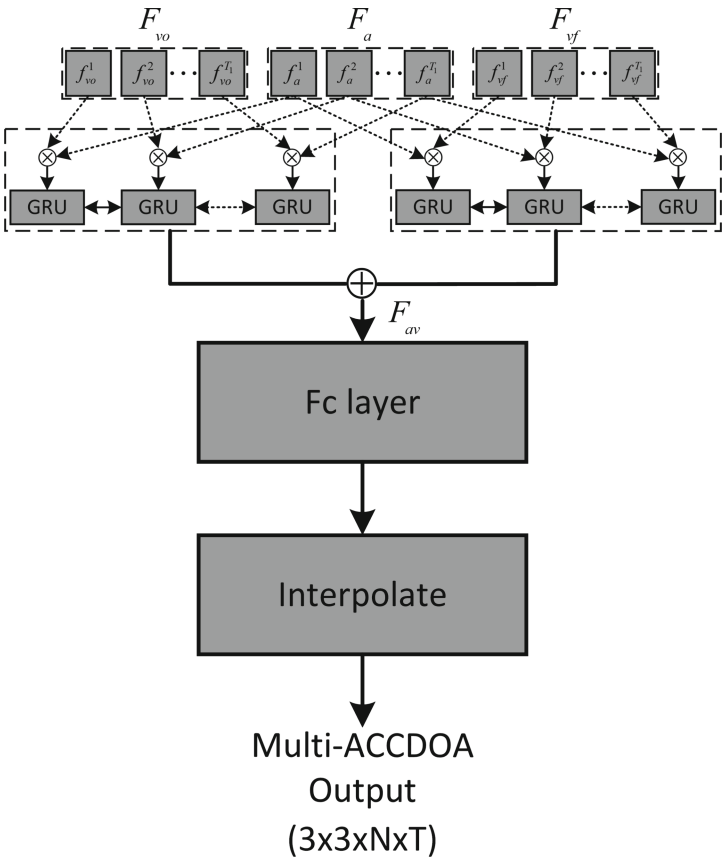

Fig. 7. The structure of decoder.

Loss Function. Auxiliary duplicating permutation invariant training (ADPIT) [13] is used as loss function in our work. ADPIT extends class-wise PIT by duplicating the original target as the auxiliary targets of the other tracks, rather than zero vectors might interfere the optimization of multi-ACCDOA training. The formula for class-wise ADPIT loss is:

$$\mathcal{L}^{PIT} = \frac{1}{CT} \sum_{c}^{C} \sum_{t}^{T} \min_{\alpha \in Perm(ct)} l_{\alpha,ct}^{ACCDOA}, \tag{4}$$

$$l_{\alpha,ct}^{ACCDOA} = \frac{1}{N}\sum_{n}^{N} MSE(\mathbf{P}^{*}_{\alpha,nct}, \hat{\mathbf{P}}_{nct}), \tag{5}$$

where $\alpha \in Perm(ct)$ is one possible class-wise frame-level permutation at class c and frame t. $\mathbf{P}^{*}_{\alpha,nct}$ and $\hat{\mathbf{P}}_{nct}$ represent an ACCDOA target of a permutation α, an ACCDOA prediction at track n, class c, frame t, respectively.

3 Experiments

3.1 Dataset

The Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset comprises multichannel recordings of sound scenes captured in diverse rooms and environments. It includes temporal and spatial annotations of significant events associated with a predefined set of target classes. The dataset provides two formats of audio data: 1) First-Order of Ambisonics; 2) tetrahedral microphone array. Furthermore, STARSS23 provides simultaneous 360° video recordings that are precisely aligned in both space and time with the microphone array recordings. This additional feature allows for a comprehensive understanding of the audio-visual scene. The STARSS23 dataset consists of real-world scene records, displaying notable variations in sound event sample density and the occurrence of each class. There are total 13 annotated target sound event classes and in which up to 3 simultaneous events can occur. During the development stage, we train our proposed model on the training set of STARSS23, and evaluate those systems using test set of STARSS23. Finally, the best models of our proposed method are evaluated on the blind test set.

3.2 Hyper-parameters

The First-Order of Ambisonics (FOA) format audio data used in all experiments is sampled at a frequency of 24 kHz and have four channels. A random audio clip of 1.27 s is chosen and subjected to the Short-time Fourier Transform using a hop size of 240 and an FFT size of 512. During testing, the audio clips are divided into non-overlapping segments of fixed length, specifically 1.2 s each. The video frames per second is 29.97 and resolution is 1920*960. We reshape the video resolution to 360*180 to reduce the calculation cost. The first video frame of every audio clip is used for object detection and raw frame input. We set the threshold of detected bounding box confidence as 0.1 and up to six boxes corresponding to person are selected to obtain Gaussian-like vectors. When there are fewer results than six in current frame, we assign zero vectors as auxiliary targets. The threshold of each sound class for mapping the predicted results to true or false is all set to 0.3. Adam optimizer is used with a 1e-6 weight decay. The learning rate is set to 0.001, reducing by 0.5 times every 10000 epochs. The max epochs for training the model is 40000.

3.3 Evaluation Metrics

For evaluation, we use official evaluation metrics [28] to evaluate the SELD performance. The SELD score is computed by

$$SELD = \frac{1}{4}(ER_{20^\circ} + (1 - F_{20^\circ}) + \frac{LE_{CD}}{180^\circ} + (1 - LR_{CD})), \tag{6}$$

where location-dependent error rate (ER_{20°), location-dependent F-score (F_{20°), class-dependent localization error (LE_{CD}), localization recall metric (LR_{CD}) are the official SELD metrics. Here, LE_{CD} represents the average angular distance between predictions and references belonging to the same sound class, LR_{CD} measures the true positive rate, indicating the proportion of predicted positions correctly identified within a specific class, relative to the total number of occurrences of that class. The metrics ER_{20°, F_{20° are utilized to evaluate the detection performance, where 20° represents spatial threshold. To be considered a true positive, the predicted angular error from the ground truth must be less than 20° for both ER_{20° and F_{20°. The evaluation metrics employ a class-wise average known as macro-averaging, which assigns equal weight to each class and places greater emphasis on the system's performance in smaller classes.

3.4 Experimental Results

Due to the recent introduction of SELD's audio-visual track in DCASE 2023 and the limited literature on it, we have chosen the official baseline to test the performance of our proposed method. Table 1 shows the performance of our proposed method compared with the baseline model in development stage including audio-visual and audio-only. The first column indicates the principal changes differentiating from baseline model. From the table, all the modified models outperform the baseline model. From the third row, we can see that residual connection in audio encoder brings improvement compared to audio-visual baseline model. The addition of DSC blocks further reduces the ER_{20° and raises the F_{20°, which indicates the DSC block is able to help distinguish the different sound classes. After that, the use of additional video frame embedding improves localization performance, which means video frame embedding can provide additional location cues. Finally, the model ensemble strategies are utilized to improve the generalization ability and better results are achieved by combining the predictions of different models. In our work, we average the predicted results of two models with lowest SELD score to obtain the best performance showed in the last row of Table 1 audio-visual part.

Since the limited quantity of the dataset and the lack of effective data augmentation methods, the performance of audio-visual track is constrained compared to previous audio-only methods usually benefitting from extensive data augmentations. For comparison, we also show the evaluation results on development dataset with audio-only input below in Table 1. The audio-only systems are implemented with the same network architecture on the same training data, i.e., STARSS23 development set, similar to audio-visual system but lack of visual

Table 1. Development Evaluation SELD Results.

Audio-Visual Model	SELD score	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
MIC_Baseline	0.755	1.08	10.8%	59.8°	28.5%
FOA_Baseline	0.710	1.07	14.3%	48.4°	35.5%
With Res	0.689	1.03	15.4%	46.1°	37.8%
With DSC	0.672	**0.96**	16.6%	46.6°	36.3%
With F_{vf} (Proposed)	0.660	0.97	16.0%	44.9°	41.8%
With ensemble (Proposed)	**0.654**	0.97	**17.1**%	**44.1°**	**42.7%**
Audio-only Model	SELD score	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
MIC_Baseline	0.768	1.06	9.7%	70.3°	28.1%
FOA_Baseline	0.716	1.00	14.4%	60.0°	32.7%
Proposed	**0.698**	**0.98**	14.4%	61.6°	**38.6%**

module. Compared with audio-visual models in Table 1, the results indicate that visual cues are able to make predictions more accurate by providing complementary information. In addition, the introducing of DSC block and other modification in our proposed method indeed improve SELD performance, even better than audio-visual baseline model.

For DCASE challenge, the most important criterion for evaluating the performance of a model is to compare the results of the model on a blind test set. We used the predictions of the blind test set based on our proposed method and its ensemble way mentioned earlier as our submitted systems. The evaluation results on the blind test set are summarized in Table 2. It is observed that the performance of our proposed method outperforms the baseline methods on the blind test set.

Table 2. Evaluation SELD Results on Blind Test Set.

Model	SELD score	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
MIC_Baseline	0.776	1.20	9.9%	58.5°	32.3%
FOA_Baseline	0.725	1.10	11.1%	47.2°	35.2%
Proposed	0.718	**1.04**	11.9%	49.5°	32.4%
With ensemble (Proposed)	**0.715**	1.05	**12.7%**	47.8°	33.0%

4 Conclusion

This work have introduced our proposed audio-visual fusion method based on CRNN for SELD task. For introducing visual cues, we perform object detection on the video frame corresponding to the start frame of the audio feature sequence and pass it into a pre-trained ResNet-18 model to produce visual feature. Then, fully connected layers are used to align the visual and audio embedding dimensions for subsequent feature fusion. Additionally, we devise a more powerful convolution block by exploring a depth-wise separable convolution with multi kernel sizes, aimed at distinguishing different sound event categories in SELD. We have evaluated our method during development stage and the experimental results show the introduction of visual cues indeed outperform audio-only approach. Compared to the task3 of DCASE 2023 baseline, the modification in our method further improve the SELD performance. On the official blind test set of the task3 of DCASE 2023 challenge, our proposed method also outperforms the baseline method and ranks fourth place.

References

1. Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., Vento, M.: Audio surveillance of roads: a system for detecting anomalous sounds. IEEE Trans. Intell. Transp. Syst. **17**(1), 279–288 (2015)
2. Wang, H., Chu, P.: Voice source localization for automatic camera pointing system in videoconferencing. In: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 187–190. IEEE (1997)
3. Lee, H.Y., Cho, J.W., Kim, M., Park, H.M.: DNN-based feature enhancement using DOA-constrained ICA for robust speech recognition. IEEE Signal Process. Lett. **23**(8), 1091–1095 (2016)
4. Barchiesi, D., Giannoulis, D., Stowell, D., Plumbley, M.D.: Acoustic scene classification: classifying environments from the sounds they produce. IEEE Signal Process. Mag. **32**(3), 16–34 (2015)
5. Heittola, T., Mesaros, A., Eronen, A., Virtanen, T.: Context-dependent sound event detection. EURASIP J. Audio Speech Music Process. **2013**(1), 1–13 (2013)
6. Mesaros, A., Heittola, T., Eronen, A., Virtanen, T.: Acoustic event detection in real life recordings. In: 2010 18th European Signal Processing Conference, pp. 1267–1271. IEEE (2010)
7. Butko, T., Pla, F.G., Segura, C., Nadeu, C., Hernando, J.: Two-source acoustic event detection and localization: Online implementation in a smart-room. In: 2011 19th European Signal Processing Conference, pp. 1317–1321. IEEE (2011)
8. Schmidt, R.: Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
9. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. **24**(4), 320–327 (1976)
10. Do, H., Silverman, H.F., Yu, Y.: A real-time SRP-phat source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP 2007, vol. 1, pp. I–121. IEEE (2007)

11. Adavanne, S., Politis, A., Nikunen, J., Virtanen, T.: Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE J. Sel. Topics Signal Process. **13**(1), 34–48 (2018)
12. Shimada, K., Koyama, Y., Takahashi, N., Takahashi, S., Mitsufuji, Y.: Accdoa: activity-coupled cartesian direction of arrival representation for sound event localization and detection. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 915–919. IEEE (2021)
13. Shimada, K., Koyama, Y., Takahashi, S., Takahashi, N., Tsunoo, E., Mitsufuji, Y.: Multi-accdoa: localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 316–320. IEEE (2022)
14. Wang, Q., et al.: The ustc-iflytek system for sound event localization and detection of DCASE2020 challenge. IEEE AASP Chall. Detect. Classif. Acoust. Scenes Events (2020)
15. Shimada, K., Takahashi, N., Koyama, Y., Takahashi, S., Tsunoo, E., Takahashi, M., Mitsufuji, Y.: Ensemble of accdoa-and einv2-based systems with d3nets and impulse response simulation for sound event localization and detection. arXiv preprint arXiv:2106.10806 (2021)
16. Wang, Q., et al.: The nerc-slip system for sound event localization and detection of dcase2022 challenge. DCASE2022 Challenge, Technical Report (2022)
17. Wang, Z., et al.: The multimodal information based speech processing (misp) 2022 challenge: audio-visual diarization and recognition. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)
18. Roth, J., et al.: Ava active speaker: an audio-visual dataset for active speaker detection. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4492–4496. IEEE (2020)
19. Qian, X., Wang, Z., Wang, J., Guan, G., Li, H.: Audio-visual cross-attention network for robotic speaker tracking. IEEE/ACM Trans. Audio Speech Lang. Process. **31**, 550–562 (2022)
20. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 247–263 (2018)
21. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16867–16876 (2021)
22. Politis, A., et al.: Starss22: a dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. arXiv preprint arXiv:2206.01948 (2022)
23. Nguyen, T.N.T., Jones, D.L., Watcharasupat, K.N., Phan, H., Gan, W.S.: Salsalite: a fast and effective feature for polyphonic sound event localization and detection with microphone arrays. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 716–720. IEEE (2022)
24. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: a survey. In: Proceedings of the IEEE (2023)
25. Chen, K., et al.: Mmdetection: open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
26. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)

27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
28. Politis, A., Mesaros, A., Adavanne, S., Heittola, T., Virtanen, T.: Overview and evaluation of sound event localization and detection in DCASE2019. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 684–698 (2020)