Exploring Audio-Visual Information Fusion for Sound Event Localization and Detection In Low-Resource Realistic Scenarios

1st Ya Jiang
University of Science and Technology of China
Hefei, China

2nd Qing Wang*
University of Science and Technology of China
Hefei, China

3rd Jun Du
University of Science and Technology of China
Hefei, China

4th Maocheng Hu
National Intelligent Voice Innovation Center
Hefei, China

5th Pengfei Hu
University of Science and Technology of China
Hefei, China

6th Zeyan Liu
University of Science and Technology of China
Hefei, China

7th Shi Cheng
University of Science and Technology of China
Hefei, China

8th Zhaoxu Nian
University of Science and Technology of China
Hefei, China

9th Yuxuan Dong
University of Science and Technology of China
Hefei, China

10th Mingqi Cai
iFlytek Research
Hefei, China

11th Xin Fang
iFlytek Research
Hefei, China

12th Chin-Hui Lee
Georgia Institute of Technology
Atlanta, USA

***Abstract*—This study presents an audio-visual information fusion approach to sound event localization and detection (SELD) in low-resource scenarios. We aim at utilizing audio and video modality information through cross-modal learning and multi-modal fusion. First, we propose a cross-modal teacher-student learning (TSL) framework to transfer information from an audio-only teacher model, trained on a rich collection of audio data with multiple data augmentation techniques, to an audio-visual student model trained with only a limited set of multi-modal data. Next, we propose a two-stage audio-visual fusion strategy, consisting of an early feature fusion and a late video-guided decision fusion to exploit synergies between audio and video modalities. Finally, we introduce an innovative video pixel swapping (VPS) technique to extend an audio channel swapping (ACS) method to an audio-visual joint augmentation. Evaluation results on the Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 Challenge data set demonstrate significant improvements in SELD performances. Furthermore, our submission to the SELD task of the DCASE 2023 Challenge ranks first place by effectively integrating the proposed techniques into a model ensemble.**

***Index Terms*—DCASE, sound event localization and detection, cross-modal teacher-student learning, multi-modal fusion, audio channel swapping, video pixel swapping**

I. Introduction

The goal of sound event localization and detection (SELD) is to detect the onsets and offsets of occurrences of specific sound events over time, while simultaneously tracking their spatial positions during the active period. The spatiotemporal information acquired from SELD systems holds broad applications in different audio-visual applications, such as self-localization, smart home, and audio surveillance [1].

SELD usually consists of two subtasks: sound event detection (SED) and sound source localization (SSL) which primarily focuses on direction-of-arrival (DOA) estimation in this paper. The SED task typically employs various classification methods such as Gaussian mixture models (GMMs) [2], hidden Markov models (HMMs) [3] and recurrent neural networks (RNNs) [4] to identify the temporal occurrences of each sound event of interest. DOA estimation methods generally include parametric-based approaches, such as subspace-based methods [5], time-difference-of-arrival (TDOA) techniques [6], signal synchronization-based approaches [7] and data-driven neural-network(NN)-based approaches [8] [9]. Recently, there has been a considerable amount of attention on jointly performing SED and DOA estimation, as introduced in the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Challenge [10]. SELDnet [11] was then proposed to utilize two parallel branches for SED and DOA estimation, respectively. Afterward, various NN-based solutions for SELD have emerged. In [12], Shimada *et al.* proposed an activity-coupled Cartesian DOA (ACCDOA) and expanded it to multi-ACCDOA [13] to distinguish multiple overlapping sound events of the same category. Currently, the majority of SELD networks process multi-channel audio

*corresponding author

inputs.

Considering that humans perceive multi-modal cues to explore real-world events, some studies have suggested integrating audio and video modalities to gain better insight into the SELD problem. However, most audio-visual detection and localization studies primarily focus on localization within video frames. This has led to the development of multiple techniques [14], [15], represented by the publicly available audio-visual event localization (AVE) task. These researches concentrate on categorizing events occurring in each video segment, as well as locating event boundaries based on a given visual or auditory query, where an audio-visual event is both audible and visible in the video. Nonetheless, it is challenging to localize a visual object emitting sound in spatial positions, because audio operates in a 3D space while video operates on a 2D image plane. Some studies proposed transformations between a spatial DOA and a target image location using calibrated sensors [16], [17]. Specifically, [18] attempted to localize and track speaker DOAs using an audio-visual cross-modal attentive fusion framework. [19] proposed to localize active speakers in a full 360° field of view.

Recently, the DCASE 2023 Challenge introduced an audio-visual track utilizing synchronized audio and video recordings. Although video data can provide valuable cues to mitigate the challenges in characterizing the spatiotemporal features of acoustic scenes, the availability of real audio-visual data is extremely limited, in this case with merely 3.83 hours, further adding to the challenges of effectively utilizing the information from video modality for solving the SELD problem. Human faces are blurred due to privacy concerns, making it impossible to extract speech activities from the video. Additionally, the information embedded in video frames is too redundant, leading to the potential duplication of sound targets. Therefore, we propose the use of cross-modal transfer learning followed by multi-modal fusions for audio-visual SELD (AV-SELD) in low-resource realistic scenarios. We highlight the three key contributions of this work as follows:

(1) We design a cross-modal teacher-student learning (TSL) framework to perform transfer learning from the teacher model trained on abundant external audio data to the student model with limited audio-visual data.
(2) We introduce a two-stage process employing feature fusion and video-guided decision fusion to further improve the localization precision in SELD models.
(3) We propose an efficient video pixel swapping (VPS) method to jointly augment multi-modal data eightfold and enhance the robustness of the student model.

II. Proposed Method for AV-SELD

The overall flowchart of the proposed AV-SELD framework using cross-modal teacher-student learning is illustrated in Fig. 1, consisting of the teacher model trained with extensive audio data and the student model built with limited audio-visual data. Besides, a two-stage audio-visual deep fusion strategy and joint audio-visual data augmentation are incorporated to further improve the SELD performance. The details are elaborated in the following subsections.

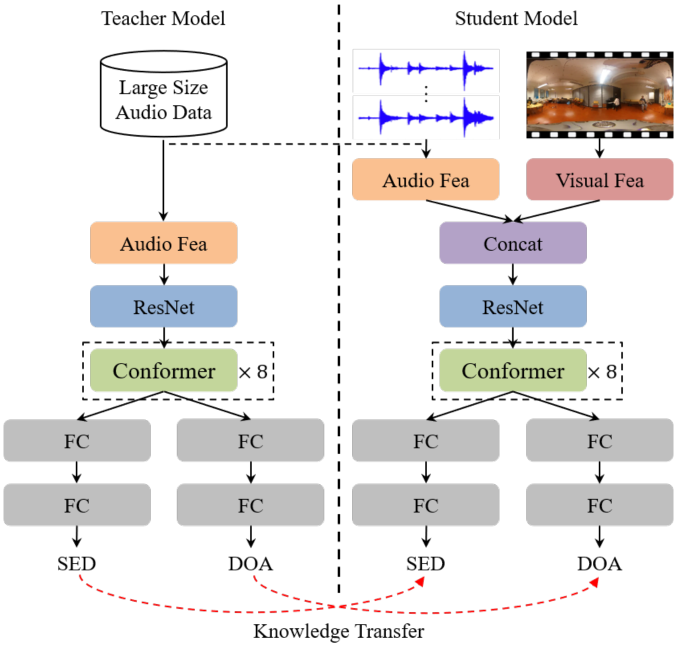


Fig. 1. The proposed AV-SELD model is based on cross-modal teacher-student learning and multi-modal fusion, as shown in the overall architecture.

A. Audio-only teacher model training

The audio-only teacher model is constructed as the ResNet-Conformer (RC) architecture proposed in our previous work [20] in the DCASE 2022 Challenge, utilizing an 18-layer ResNet network followed by an 8-layer Conformer module as illustrated in the left branch of Fig. 1. We slice the 4-channel audio data into 10-second segments and extract 4-channel log Mel-spectrograms and 3-channel intensity vector (IV) features by applying short-term Fourier transform (STFT) with a frame length of 40 ms and a frame hop of 20 ms. Consequently, the concatenated 500×7×64 audio features are fed into the model. The ResNet-Conformer follows two parallel branches to perform SED and DOA estimation respectively, each containing two fully connected (FC) layers. The audio-only teacher network is trained using a multi-objective learning framework:

$$\mathcal{L}_{\mathrm{SELD}}^{\mathrm{M}} = \beta_1 \mathcal{L}_{\mathrm{SED}}^{\mathrm{M}} + \beta_2 \mathcal{L}_{\mathrm{DOA}}^{\mathrm{M}} \quad (1)$$

where $\mathrm{M} \in \{\mathrm{T}, \mathrm{S}, \mathrm{TS}\}$. When $\mathrm{M} = \mathrm{T}$, $\mathcal{L}_{\mathrm{SELD}}^{\mathrm{T}}$ represents the loss function used in teacher model training. When $\mathrm{M} = \mathrm{S}$, $\mathcal{L}_{\mathrm{SELD}}^{\mathrm{S}}$ denotes the original loss function used in student model training in Section II-B. When $\mathrm{M} = \mathrm{TS}$, $\mathcal{L}_{\mathrm{SELD}}^{\mathrm{TS}}$ indicates the teacher-instructed loss function used in cross-modal teacher-student learning in Section II-C. $\beta_1 = 0.1$ and $\beta_2 = 1$ are the weighting factors set for SED loss $\mathcal{L}_{\mathrm{SED}}$ and DOA loss $\mathcal{L}_{\mathrm{DOA}}$ respectively. The SED and DOA subtasks are optimized by minimizing the binary cross entropy (BCE) and mean squared error (MSE) criteria respectively, as defined below:

$$\mathcal{L}_{\mathrm{SED}}^{\mathrm{T}} = -\frac{1}{KN}\sum_{k,n}\left[y_{k,n}^{\mathrm{T}}\log\hat{y}_{k,n}^{\mathrm{T}}\right. \tag{2}$$
$$\left. + \left(1-y_{k,n}^{\mathrm{T}}\right)\log\left(1-\hat{y}_{k,n}^{\mathrm{T}}\right)\right]$$
$$\mathcal{L}_{\mathrm{DOA}}^{\mathrm{T}} = \frac{1}{KN}\sum_{k,n}\left\|\left(\hat{\mathbf{o}}_{k,n}^{\mathrm{T}}-\mathbf{o}_{k,n}^{\mathrm{T}}\right)y_{k,n}^{\mathrm{T}}\right\|^{2} \tag{3}$$

where $\{y_{k,n}^{\mathrm{T}}, \hat{y}_{k,n}^{\mathrm{T}}\}$ and $\{\mathbf{o}_{k,n}^{\mathrm{T}}, \hat{\mathbf{o}}_{k,n}^{\mathrm{T}}\}$ represent the ground truth and model output for SED and DOA of the n-th sound event at the k-th frame, respectively. $\mathbf{o}_{k,n}^{\mathrm{T}}$ and $\hat{\mathbf{o}}_{k,n}^{\mathrm{T}}$ represent Cartesian position vectors. K denotes the total number of frames in a batch and N denotes the number of classes.

During the training of the teacher model, we incorporate external audio data with multiple data augmentation techniques to enhance the model's stability. Firstly, we leverage the additional 20 hours of simulated data provided by the DCASE 2023 Challenge and apply the ACS augmentation to generate extensive audio data, following the approach described in our previous work [21]. The ACS augmentation involves performing transformations directly to four audio channels, resulting in eight different DOA representations and effectively increasing the amount of audio data eightfold. Additionally, we employ mixup [22] augmentation to enhance the diversity of the model input features.

B. Audio-visual student model training

The student model is an audio-visual SELD network sharing the same backbone structure as the teacher model and incorporating the audio-visual feature fusion to use spatiotemporal information from multiple modalities. The audio features are extracted in the same way as the teacher model. For the visual modality, we select 10 video frames per second evenly to perform object detection using Faster-RCNN [23], which can predict various classes of objects. Based on the detected boxes of the human class, a keypoint detection model, e.g., HRNet [24] pre-trained on COCO-WholeBody dataset [25], then predicts five keypoints of visible speakers, namely mouths, left and right hands, and left and right feet for each corresponding image. We use the pixel coordinate of the mouth of each person to generate two Gaussian-like vectors, indicating likelihoods of speakers appearing along horizontal and vertical axes [18]. The center is the same as the mouth position and the standard deviation is proportional to a predefined width and height. We extract visual features for each person and pad with zeros in cases where fewer than six people are present simultaneously, resulting in a 100×6×2×64 Gaussian feature vector for a 10-second clip. The visual features are replicated five times along the temporal dimension and reshaped to 500×12×64 to be concatenated with audio features, leading to 19-channel spatiotemporal representations. The fused multi-modal features are fed into ResNet-Conformer as shown in the right branch of Fig. 1.

To address the problem of data sparsity, we propose a novel VPS method to extend ACS approach to audio-visual joint data augmentation, which matches multi-modal data in pairs. Given that the ACS method is inherently realized by swapping the four microphones arranged on a spherical baffle spatially, we can regard a 360° panoramic video frame image as a cylindrical surface with the camera located at the center of the cylinder. Therefore, the pixel resolution of the image, e.g., 1920 × 960, corresponds to an azimuth angle range of $\phi \in [180^{\circ}, -180^{\circ}]$ and an elevation angle range of $\theta \in [-90^{\circ}, 90^{\circ}]$. Based on the above analysis, we can design eight corresponding translations and inversions of pixel coordinates to jointly expand audio-visual data efficiently. Further details can be found in our technical report [26]. Taking one transformation, $\phi = \phi + \pi, \theta = \theta$ for example, which means the original DOA azimuth angle ϕ is rotated by 180° while the elevation angle remains the same. Accordingly in the process of VPS, the horizontal pixel points are translated by 960 pixel points along the negative direction, while the vertical pixel points remain unchanged in the corresponding video image.

C. Cross-modal teacher-student learning

We construct a cross-modal teacher-student learning (TSL) framework based on teacher weights transferring and cross-modal loss function. In the first step, pre-trained weights of the teacher model are utilized to initialize the first 7 channels of the student model's input feature maps with the remaining 12 channels randomly initialized. Secondly, we design a teacher-instructed loss function $\mathcal{L}_{\mathrm{SELD}}^{\mathrm{TS}}$ as a regularization term of the original student loss function $\mathcal{L}_{\mathrm{SELD}}^{\mathrm{S}}$ calculated in the same way as Eq. 1. The final TSL loss function is formulated as Eq. 4. Specifically, we employ Kullback-Leibler (KL) divergence to regularize SED loss inspired by [27], computed with the SED output of the teacher and student network as Eq. 5. Meanwhile, the DOA loss of the student model $\mathcal{L}_{\mathrm{DOA}}^{\mathrm{S}}$ is regularized using the outputs in the teacher model as Eq. 6. TSL framework enables the model to assimilate information from the video modality alongside the guidance and regularization imparted by the audio teacher model, which ensures a balanced contribution of multi-modal features.

$$\mathcal{L}_{\mathrm{SELD}} = \gamma_1 \times \mathcal{L}_{\mathrm{SELD}}^{\mathrm{S}} + \gamma_2 \times \mathcal{L}_{\mathrm{SELD}}^{\mathrm{TS}} \tag{4}$$
$$\mathcal{L}_{\mathrm{SED}}^{\mathrm{TS}} = \frac{1}{KN}\sum_{k,n}\left[\hat{y}_{k,n}^{\mathrm{T}}\log\frac{\hat{y}_{k,n}^{\mathrm{T}}}{\hat{y}_{k,n}^{\mathrm{S}}}\right. \tag{5}$$
$$\left. + \left(1-\hat{y}_{k,n}^{\mathrm{T}}\right)\log\frac{\left(1-\hat{y}_{k,n}^{\mathrm{T}}\right)}{\left(1-\hat{y}_{k,n}^{\mathrm{S}}\right)}\right]$$
$$\mathcal{L}_{\mathrm{DOA}}^{\mathrm{TS}} = \frac{1}{KN}\sum_{k,n}\left\|\left(\hat{\mathbf{o}}_{k,n}^{\mathrm{S}}-\hat{\mathbf{o}}_{k,n}^{\mathrm{T}}\right)\hat{y}_{k,n}^{\mathrm{T}}\right\|^{2} \tag{6}$$

where $\gamma_1 = 1$ and $\gamma_2 = 0.5$ are weighting factors for the student loss and teacher-regularized loss, separately. $\hat{y}_{k,n}^{\mathrm{S}}$ and $\hat{\mathbf{o}}_{k,n}^{\mathrm{S}}$ are the estimated active probabilities and DOA for the n-th sound event at the k-th frame in the student network, respectively. The parameters of the teacher model are fixed during the cross-modal learning.

D. Video-guided decision fusion

In the late stage, we employ a decision fusion approach to further uncover crucial information from video detection in order to enhance the model's localization precision. Video detection-based localization tends to yield more accurate results compared to network predictions, especially for cases involving rapid movements of sound sources, due to potential drift and fluctuation. Therefore, we propose a video-guided decision fusion rule to rectify inaccurate DOA estimations using visual cues in the following three steps:

Firstly, the human keypoints detected in Section II-B can be associated with specific sound classes. Specifically, mouths are associated with male speech, female speech, clapping and laughter classes, left and right hands are associated with water tap class, and left and right feet are associated with walk class. Other sound event classes are excluded due to the lack of corresponding visual objects (e.g., Knock) or poor detection performance (e.g., Door). Secondly, given a Cartesian vector DOA estimation $\hat{\mathbf{o}} = \left(\hat{a}, \hat{b}, \hat{c}\right)$ of a sound event at t-th frame predicted by model and all coordinates of human keypoints $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p\}$ detected from the current video frame, we can compute the angular distances [10] between them, as defined below:

$$d\left(\hat{\mathbf{o}}, \mathbf{v}_i\right) = \arccos\left(\frac{\langle \hat{\mathbf{o}}, \mathbf{v}_i \rangle}{\|\hat{\mathbf{o}}\| \, \|\mathbf{v}_i\|}\right) \tag{7}$$

where $\mathbf{v}_i = (a_i, b_i, c_i), i \in \{1, \dots, p\}$, and p is the number of detected keypoints. Thirdly, we select the keypoint candidate $\hat{\mathbf{v}}$ with the smallest angular distance $\hat{d}$. If $\hat{d}$ is less than a predefined threshold $\sigma = 30°$, we consider the video detection outcomes to be more accurate and replace $\hat{\mathbf{o}}$ with $\hat{\mathbf{v}}$ as the final DOA estimation. Utilizing video-guided decision fusion, the localization accuracy and precision of the network can be further improved.

III. EXPERIMENTS

A. Experimental setup

We evaluate the SELD task on the official development set of the Sony-Tau Realistic Spatial Soundscapes 2023 (STARSS23) dataset [28] recorded in realistic spatial soundscapes in DCASE 2023 Challenge. The development set of STARSS23 is divided into a training part (dev-set-train) about 3.83 hours and a testing part (dev-set-test) about 3.22 hours. The labels of the official evaluation set have not been released, and our results on the evaluation set are available on the DCASE 2023 Task3 Challenge results page[1]. The basic dataset contains a total of 13 sound event classes. The audio recordings are collected in a 4-channel spatial format with a 24 kHz sampling rate. The video data are captured using a 360° camera with a resolution of 1920 × 960 at 29.97 frames per second. The details of audio-visual feature extraction and fusion are discussed in Section II-A and Section II-B. In the DCASE 2023 SELD task, a convolutional recurrent neural network (CRNN) is utilized as the audio-visual baseline structure. Our main model architecture is ResNet-Conformer consisting of 8 attention heads [20], and the dimensions of the input, key and value vectors are set to 256, 32 and 32, respectively. Adam [29] is adopted as the optimizer. The tri-stage learning rate scheduler [30] is used with an upper limit of 0.001 without TSL and 0.0001 with TSL, respectively. We evaluate all methods with SELD$_{score}$ [31], as defined below:

TABLE I
Performances comparison among different data augmentation methods on audio-only (AO) teacher model and audio-visual (AV) student model. CRNN is used as the official baseline. ‘DA1’: AO Basic data, ‘DA2’: AO Basic data + simulated data, ‘DA3’: AO Basic + simulated data + ACS, ‘DA4’: AO Basic + simulated data + ACS + mixup, ‘DA5’: AV Basic data, ‘DA6’: AV Basic data + ACS-VPS, ‘DA7’: AV Basic data + ACS-VPS + mixup.

Model	Setup	ER$_{20°}$	F$_{20°}$	LE$_{\mathrm{CD}}$	LR$_{\mathrm{CD}}$	SELD$_{score}$
Teacher	CRNN+DA1	1.00	0.14	60.00°	0.33	0.72
	RC+DA1	0.75	0.17	39.82°	0.38	0.61 (15.3% ↓)
	RC+DA2	0.56	0.42	18.65°	0.67	0.39 (45.8% ↓)
	RC+DA3	0.45	0.55	14.28°	0.66	0.33 (54.2% ↓)
	RC+DA4	0.42	0.57	14.30°	0.67	0.31 (56.9% ↓)
Student	CRNN+DA5	1.07	0.14	48.00°	0.36	0.71
	RC+DA5	0.76	0.18	34.13°	0.42	0.59 (16.9% ↓)
	RC+DA6	0.57	0.36	18.82°	0.51	0.45 (36.6% ↓)
	RC+DA7	0.53	0.49	15.80°	0.64	0.37 (47.9% ↓)

$$\mathrm{SELD}_{score} = \frac{1}{4}\left[ER_{20^{\circ}} + (1 - F_{20^{\circ}}) + LE_{\mathrm{CD}}^{'} + (1 - LR_{\mathrm{CD}})\right] \tag{8}$$

where $ER_{20^{\circ}}$ and $F_{20^{\circ}}$ represent location-dependent error rate and F-score when the spatial error is within 20°. $LE_{\mathrm{CD}}^{'} = LE_{\mathrm{CD}}/\pi$, where LE_{CD} denotes the localization error between predictions and references of the same class. LR_{CD} is a simple localization recall metric.

B. Experimental results and analysis

1) Experiments on audio-visual data augmentation: We evaluate the effectiveness of various audio-visual data augmentation methods on teacher and student models trained from scratch. As described in Table I, we first apply our self-developed RC network to the basic training data and its good contextual and global modeling capabilities produce 15.3% and 16.9% improvements on the AO teacher and AV student models relative to the official CRNN baselines, respectively. The early audio-visual feature fusion combines the acoustic and visual cues to favor a 0.02 improvement in SELD scores for the student model relative to the teacher model. Subsequently, we step by step employ different data augmentation techniques on the teacher and student models, respectively. In teacher model training, we continually add simulation data, ACS method and mixup augmentation on top of the official basic data, leading to four different audio data configurations, identified as ‘DA1’, ‘DA2’, ‘DA3’ and ‘DA4’. The injection of substantial external data and effective data augmentation techniques progressively enhance the model's robustness, however, it should be noted that the simulated

[1] https://dcase.community/challenge2023/task-sound-event-localization-and-detection-evaluated-in-real-spatial-sound-scenes-results

TABLE II
PERFORMANCES COMPARISON ON TSL USING AO TEACHER MODEL 'T' TO INSTRUCT AV STUDENT MODEL 'S' WITH DIFFERENT CONFIGURATIONS. 'T1': RC + DA2, 'T2': RC + DA3, 'T3': RC + DA4, 'T4': CRNN + DA4, 'S1': RC + DA5, 'S2': RC + DA6, 'S3': RC + DA7.

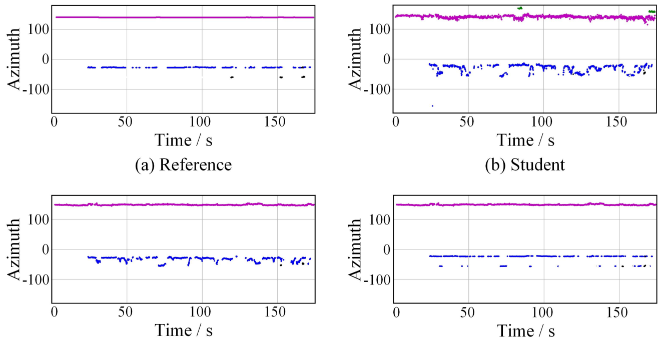
Model	$ER_{20°}$	$F_{20°}$	LE_{CD}	LR_{CD}	$SELD_{score}$
T1-S1	0.64	0.28	33.93°	0.57	0.49 (31.0% ↓)
T2-S1	0.51	0.39	28.37°	0.58	0.42 (40.8% ↓)
T3-S1	0.51	0.41	27.48°	0.64	0.40 (43.7% ↓)
T3-S2	0.43	0.55	14.42°	0.68	0.32 (54.9% ↓)
T3-S3	0.41	0.59	14.10°	0.73	0.29 (59.2% ↓)
T4	0.53	0.42	18.33°	0.60	0.40
T4-S3	0.47	0.49	15.91°	0.63	0.36 (49.3% ↓)

audio cannot be aligned with the video data, making the student model lose considerable external data compared to the teacher model. Therefore, we initially apply the proposed audio-visual joint data augmentation by simultaneously performing VPS with ACS augmentation, expanding the duration of the audio-visual data from 3.83 hours to 30.64 hours, represented as 'DA6'. The ACS-VPS method significantly advances the student model with a 36.6% SELD improvement, since the ACS-VPS method creates more diverse spatial positions through rotation and flipping without disturbing the realism and continuity of the recorded video pixels and reverberation conditions and overlapping segments in multi-channel audio. Then, applying the mixup method, denoted as 'DA7', at the input layer further enhances the model's performance. Not only in our experimental phenomena but also in the official baseline result, AV models outperform AO models slightly due to redundancy and obstacles in extracting video cues, which motivates us to further enhance the student model's capability by fusing audio and visual information.

2) Experiments on teacher-student learning: Based on the teacher and student models trained as described above, we conduct experiments with cross-modal teacher-student learning. Following the TSL framework we proposed in Section II-C, the learning process of student models can be instructed by various teacher models. As illustrated in Table II, given the student model 'S1' with the limited basic audio-visual dataset 'DA1', different teacher models 'T1', 'T2' and 'T3' are adopted first. As the guidance from teacher models grows, the SELD performance of the student models consistently improves from 0.49 to 0.40 (as depicted in the first three rows of Table II). Consequently, we fix the optimal audio model with 'DA4' in Table I as the teacher and conduct ablation experiments to assess the student model's performance (as exhibited in the third to fifth rows of Table II). By incorporating data augmentation techniques into the student model incrementally, TSL demonstrates a notable enhancement of 0.11 in SELD score on the AV student model, culminating in a 59.2% improvement relative to the official baseline. Additionally, in order to explore the generalization of our proposed TSL framework, we train an AO teacher model with the largest dataset 'DA4' based on the official CRNN network, which is denoted as 'T4'. Then the teacher model 'T4' is used to guide the AV student model with the RC network. Analogously, our TSL framework showcases a development of 49.3%. Cross-modal TSL contributes to utilizing rich information from audio data and balancing original student loss $\mathcal{L}^{S}_{SELD}$ with the teacher-instructed regularized loss $\mathcal{L}^{TS}_{SELD}$, which helps alleviate the overfitting problem on low-resource realistic audio-visual dataset. Moreover, TSL framework is robust to different model architectures.

TABLE III
PERFORMANCES COMPARISON FOR VIDEO-GUIDED DECISION FUSION (DENOTED AS 'VGDF') ON TEACHER AND STUDENT MODELS.

Model	VGDF	$ER_{20°}$	$F_{20°}$	LE_{CD}	LR_{CD}	$SELD_{score}$
Teacher	-	0.42	0.57	14.30°	0.67	0.31
	✓	0.40	0.58	12.64 °	0.67	0.30
Student	-	0.41	0.59	14.10°	0.73	0.29
	✓	0.40	0.61	12.25°	0.73	0.28



(a) Reference
(b) Student
(c) TSL
(d) TSL + VGDF

Fig. 2. The visualized comparison of azimuth DOA results of different methods, including student model trained with RC network and DA7 in Table I, TSL model denoted as T3-S3 in Table II and TSL + VGDF method in Table III. Different colors identify different sound event classes.

3) Experiments on video-guided decision fusion: At the late stage, we apply the video-guided decision fusion (VGDF) method to mine DOA information from video data on teacher and student models in Section III-B2, almost all of which demonstrate stable improvements in SELD performance. As shown in Table III, we present the VGDF performance on the teacher model (refer to 'RC+DA4' in Table I) and the student model (refer to 'T3-S3' in Table II). Upon closer inspection of metrics, it is evident that the VGDF method primarily optimizes the LE_{CD} metric, which aligns with our motivation for designing it. The localization error of the student model decreases from 14.10° to 12.25° thanks to the accurate video target detection algorithms and the matching rules we designed. Furthermore, the localization-dependent $ER_{20°}$ and $F_{20°}$ metrics are also optimized accordingly due to the correlation between the localization and detection metrics.

The final single AV student model with VGDF achieves a SELD score of 0.28, which outperforms the second place in the DCASE 2023 Challenge by 15.2%.

We visualize azimuth results of one recording in Fig. 2 to illustrate impacts of the proposed methods. Fig. 2 (b) is the result of the student model with AV joint augmentation, while TSL in Fig. 2 (c) gives precise SELD estimations, such as correcting *Music* prediction (compared to the green line in Fig. 2 (b)). By further performing VGDF, Fig. 2 (d) exhibits the closest SELD performance to the reference in Fig. 2 (a) among all of the predictions.

IV. Conclusion

We propose an audio-visual information fusion framework for SELD in low-resource realistic scenarios. By incorporating cross-modal TSL, multi-modal fusion and novel AV joint augmentation, the proposed framework demonstrates significant improvements and our submission to the SELD task of the DCASE 2023 Challenge won 1st place with model ensembles, standing out as the only one whose AV systems outperformed AO systems of all other teams in the Challenge. In the future, we will explore more cross-modal fusion strategies and intermodal relationships to learn useful cues embedded in multimodality data.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427.

References

[1] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, et al., “Audio surveillance of roads: A system for detecting anomalous sounds,” *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 279–288, 2015.

[2] Toni Heittola, Annamaria Mesaros, Antti Eronen, et al., “Context-dependent sound event detection,” *Eurasip J. Audio Speech Music Process.*, vol. 2013, no. 1, pp. 1–13, 2013.

[3] Taras Butko, Fran González Pla, Carlos Segura, et al., “Two-source acoustic event detection and localization: Online implementation in a smart-room,” in *IEEE EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, 2011, pp. 1317–1321.

[4] Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, et al., “Duration-controlled lstm for polyphonic sound event detection,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 11, pp. 2059–2070, 2017.

[5] Ralph Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas, Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[6] Charles Knapp and Glifford Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[7] Jacek P Dmochowski, Jacob Benesty, and Sofiene Affes, “A generalized steered response power method for computationally viable source localization,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, 2007.

[8] Harshavardhan Sundar, Weiran Wang, Ming Sun, and Chao Wang, “Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources,” in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 4642–4646.

[9] Thi Ngoc Tho Nguyen, Woon-Seng Gan, Rishabh Ranjan, and Douglas L Jones, “Robust source counting and doa estimation using spatial pseudo-spectrum and convolutional neural network,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2626–2637, 2020.

[10] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, et al., “Overview and evaluation of sound event localization and detection in dcase 2019,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 684–698, 2020.

[11] Sharath Adavanne, Archontis Politis, Joonas Nikunen, et al., “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” *IEEE J. of Selected Topics in Signal Process.*, vol. 13, no. 1, pp. 34–48, 2018.

[12] Kazuki Shimada, Yuichiro Koyama, Naoya Takahashi, et al., “Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection,” in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 915–919.

[13] Kazuki Shimada, Yuichiro Koyama, Shusuke Takahashi, et al., “Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training,” in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022, pp. 316–320.

[14] Yapeng Tian, Jing Shi, Bochen Li, et al., “Audio-visual event localization in unconstrained videos,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 247–263.

[15] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, et al., “Cross-modal attention network for temporal inconsistent audio-visual event localization,” in *AAAI - AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 279–286.

[16] Yang Liu, Volkan Kılıç, Jian Guan, et al., “Audio–visual particle flow smc-phd filtering for multi-speaker tracking,” *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 934–948, 2019.

[17] Xinyuan Qian, Alessio Brutti, Oswald Lanz, et al., “Multi-speaker tracking from an audio–visual sensing device,” *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2576–2588, 2019.

[18] Xinyuan Qian, Zhengdong Wang, Jiadong Wang, et al., “Audio-visual cross-attention network for robotic speaker tracking,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 550–562, 2022.

[19] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu, “Egocentric deep multi-channel audio-visual active speaker localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2022, pp. 10544–10552.

[20] Shutong Niu, Jun Du, Qing Wang, et al., “An experimental study on sound event localization and detection under realistic testing conditions,” in *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[21] Qing Wang, Jun Du, Hua-Xin Wu, et al., “A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 1251–1264, 2023.

[22] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, et al., “mixup: Beyond empirical risk minimization,” in *Int. Conf. Learn. Represent.*, 2018.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, et al., “Faster R-CNN: Towards real-time object detection with region proposal networks,” *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[24] Ke Sun, Bin Xiao, Dong Liu, et al., “Deep high-resolution representation learning for human pose estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 5693–5703.

[25] Sheng Jin, Lumin Xu, Jin Xu, et al., “Whole-body human pose estimation in the wild,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[26] Qing Wang, Ya Jiang, Shi Cheng, Maocheng Hu, Zhaoxu Nian, Pengfei Hu, Zeyan Liu, Yuxuan Dong, Mingqi Cai, Jun Du, and Chin-Hui Lee, “The nerc-slip system for sound event localization and detection of dcase2023 challenge,” Tech. Rep., DCASE2023 Challenge, June 2023.

[27] Hengshun Zhou, Jun Du, Hang Chen, et al., “Audio-Visual Information Fusion Using Cross-Modal Teacher-Student Learning for Voice Activity Detection in Realistic Environments,” in *Proc. Interspeech 2021*, pp. 341–345.

[28] Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, et al., “Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events,” *arXiv preprint arXiv:2306.09126*, 2023.

[29] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[30] Daniel S. Park, William Chan, Yu Zhang, et al., “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” in *Proc. Interspeech 2019*, pp. 2613–2617.

[31] Annamaria Mesaros, Sharath Adavanne, Archontis Politis, et al., “Joint measurement of localization and detection of sound events,” in *IEEE Workshop Appl. Signal Process. Audio, Acoust. (WASPAA)*, 2019, pp. 333–337.