LEVERAGING REVERBERATION AND VISUAL DEPTH CUES FOR SOUND EVENT LOCALIZATION AND DETECTION WITH DISTANCE ESTIMATION

Technical Report

Davide Berghi

CVSSP, University of Surrey,
Guildford, UK
davide.berghi@surrey.ac.uk

Philip J. B. Jackson

CVSSP, University of Surrey,
Guildford, UK
p.jackson@surrey.ac.uk

ABSTRACT

This report describes our systems submitted for the DCASE2024 Task 3 challenge: Audio and Audiovisual Sound Event Localization and Detection with Source Distance Estimation (Track B). Our main model is based on the audio-visual (AV) Conformer, which processes video and audio embeddings extracted with ResNet50 and with an audio encoder pre-trained on SELD, respectively. This model outperformed the audio-visual baseline of the development set of the STARSS23 dataset by a wide margin, halving its DOAE and improving the F1 by more than 3x. Our second system performs a temporal ensemble from the outputs of the AV-Conformer. We then extended the model with features for distance estimation, such as direct and reverberant signal components extracted from the omnidirectional audio channel, and depth maps extracted from the video frames. While the new system improved the RDE of our previous model by about 3 percentage points, it achieved a lower F1 score. This may be caused by sound classes that rarely appear in the training set and that the more complex system does not detect, as analysis can determine. To overcome this problem, our fourth and final system consists of an ensemble strategy combining the predictions of the other three. Many opportunities to refine the system and training strategy can be tested in future ablation experiments, and likely achieve incremental performance gains for this audio-visual task.

Index Terms— Sound Event Localization and Detection, Audio-Visual Machine Learning, Multimodal, Distance Estimation, Sound Event Detection

1. INTRODUCTION

Sound Event Localization and Detection (SELD) consists of simultaneously detecting and classifying the active sound sources over time (sound event detection (SED)) while predicting their position or direction of arrival DOA [1]. This task is essential for a variety of applications, such as human-robot interaction, augmented reality, navigation, smart home, and security, to name a few. Over time, researchers have been addressing more and more challenges related to the task, including the detection of polyphonic sounds [2], simultaneous same-class events and moving sources [3], and external interfering sounds that must not be detected [4]. This year, the challenge also involves predicting the distance of the active sources [5]. While SELD has been typically formulated as an audio-only task, a parallel audio-visual track has been included in the past two editions of the DCASE challenge, leveraging the Sony-TAu Realistic

Spatial Soundscapes 2023 (STARSS23) dataset [6]. STARSS23 includes 360° video recordings spatially and temporally aligned to the acoustic sound-field captured by the microphone array. This allows the exploration of SELD as a multimodal audio-visual problem (AV-SELD). The two modalities are complementary and can be beneficial to the task: vision provides high spatial accuracy whereas audio can detect occluded objects.

This report describes our systems submitted to the audio-visual track of the challenge. Specifically, we adopted two main models and methods. The first extends our previous work on SELD where an audio-visual (AV) conformer takes as input the concatenation of audio and video embeddings extracted with pre-trained encoders [7]. The extension consists of adapting the existing model to support distance estimation, as described in Section 2. In Section 3, the second method investigates additional input features and extends the model architecture to improve distance estimation and take more advantage of the visual modality. We report our results on the STARSS23 development set [6] in Section 4, including a temporal ensemble and an ensemble of all three variants, improving substantially over the challenge baselines. Section 5 concludes.

2. AUDIO-VISUAL CONFORMER

As depicted in the left part of Fig. 1, this method extracts audio and visual feature embeddings with an audio and a visual encoder, respectively. The embeddings are then concatenated and processed by a Conformer module with 4 layers. Finally, the output features are fed to two fully connected layers to predict multi-ACCDDOA vectors [5]. Such representation is the extension of the multi-ACCDOA vectors proposed by Shimada *et al.* [8], adopted to include distance predictions. Specifically, at each time frame, the output presents $N = 3$ tracks, each predicting 4 positional values (x, y, z, and distance), for each of the $C = 13$ classes, resulting in a 156-dimensional vector. We employed the hyperbolic tangent as the activation function for the x, y, z predictions, while ReLU for the distance. In this report, we will refer to this system as per "*AV-Conformer*". The same audio-visual architecture and methods were presented in our recent work [7]. The main difference with the system presented for this challenge consists in the integration of the distance estimation and the use of the new metrics. Below, we describe the audio and visual encoders, and the data augmentation and pre-training strategies.

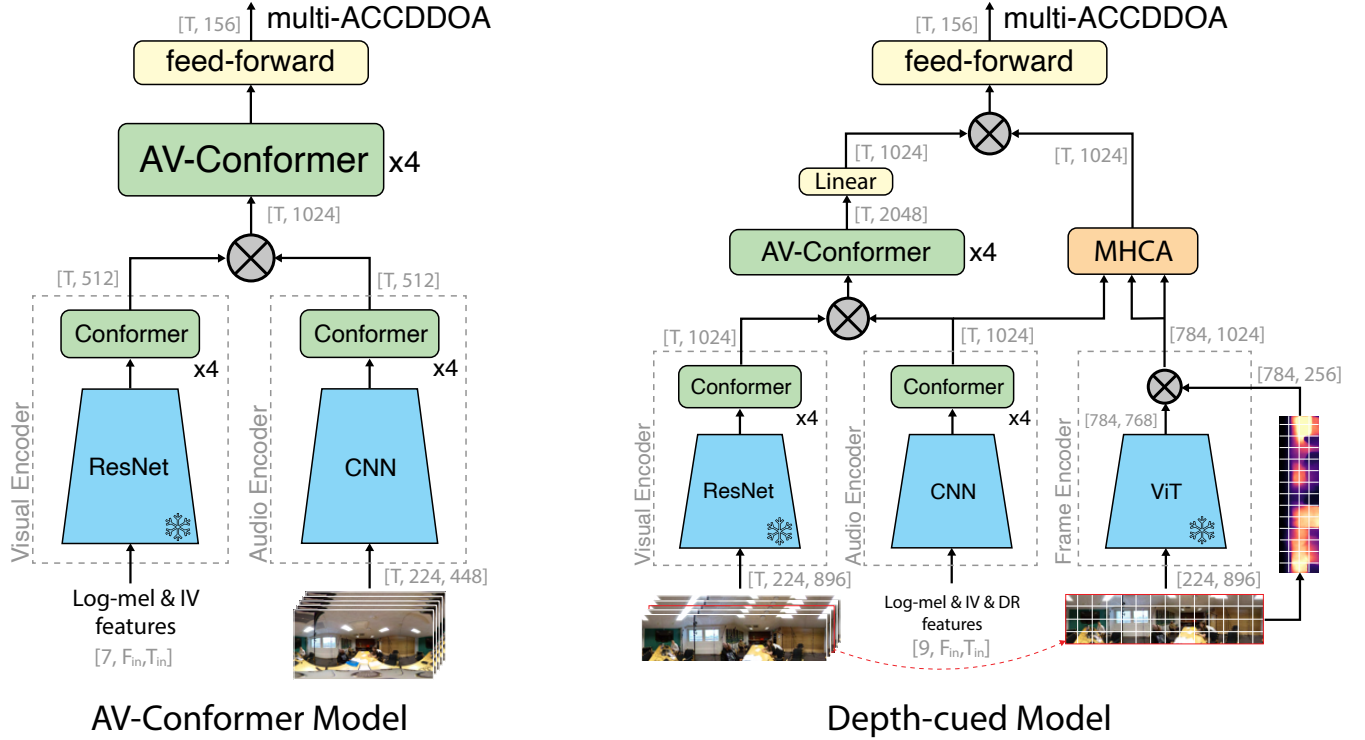AV-Conformer Model

Depth-cued Model

Figure 1: Diagram of the two main architectures submitted to the challenge. On the left, the AV-Conformer with the respective audio and visual encoders. On the right, the Depth-cued model that includes a frame encoder employed to extract visual features from the central frame. The Depth-cued model leverages cubemap views whereas the AV-Conformer equirectangular views. The snowflake symbol indicates that the weights of ResNet50 and ViT are fixed during training.

2.1. Audio Encoder

The audio encoder takes as input acoustic features extracted from the FOA spatial sound. We employed intensity vectors (IV) in the log-mel space concatenated with the log-mel spectrograms extracted from the FOA channels, yielding 7-dimensional input features with shape $7 \times T_{in} \times F_{in}$, where T_{in} corresponds to the number of temporal bins and F_{in} the frequency bins. The audio encoder includes a CNN architecture followed by a Conformer [9]. The CNN presents 4 convolutional blocks with residual connections. Each block consists of two 3×3 convolutional layers followed by average pooling, batch normalization, and ReLu activation. The average pooling layer is applied with a stride of 2, halving the temporal and frequency dimension at each block. The resulting tensor of shape $512 \times T_{in}/16 \times F_{in}/16$ is then reshaped and frequency average pooling is applied to achieve a $T_{in}/16 \times 512$ dimensional feature embedding. T_{in} is chosen so that $T_{in}/16$ matches the label frame rate (10 labels per second). The feature embedding is further processed by a Conformer module with 4 layers and 8 heads each. The size of the kernel for the depthwise convolutions is set to 51.

2.2. Visual Encoder

As visual encoder, we employed a ResNet-Conformer. Specifically, we fed each video frame to ResNet50 [10] at a frame rate of 10 fps. In such a way, we extract a number of frame embeddings that match the label frame rate as well as the audio embedding temporal resolution. We then process the frame embeddings with a Conformer module identical to the one employed in the audio en-

coder. The video segments used as inputs to the visual encoders were resized to 448×224p. The original ResNet50 model is pretrained on squared 224×224p frames and its last layer is a global average pooling applied on a 7×7 feature map. Since, in our case, the input horizontal dimension is 448, the resulting feature map has shape 14×7. Therefore, we replaced the last global average pooling with a 7×7 average pooling kernel with stride equal to 7, obtaining two output vectors that are then concatenated. Before being fed to the Conformer module, the output of ResNet50 is reduced to 512 dimensions with a linear layer.

2.3. Data Augmentation and Pre-Training

The audio CNN-Conformer was pre-trained on SELD employing the simulated data generator script provided for the DCASE 2022 challenge [11]. This allowed the generation of ~30h of spatial recordings, including noiseless and noisy versions. The ResNet50 model employed to process 2D frames is available with the Torchvision library and it is pre-trained on ImageNet [12]. The synthetic SELD dataset used to pre-train the audio encoder was augmented by a factor of 8 with the audio channels swap (ACS) technique [13]. For the AV-SELD dataset, we also augmented the visual modality consistently with the ACS transformation, i.e., the audio-visual channel swap (AVCS) [7, 14]. This generates new video frames by flipping and rotating the original ones, creating an effective audio-visual spatial transformation. An example of a transformed frame is shown in Fig. 2
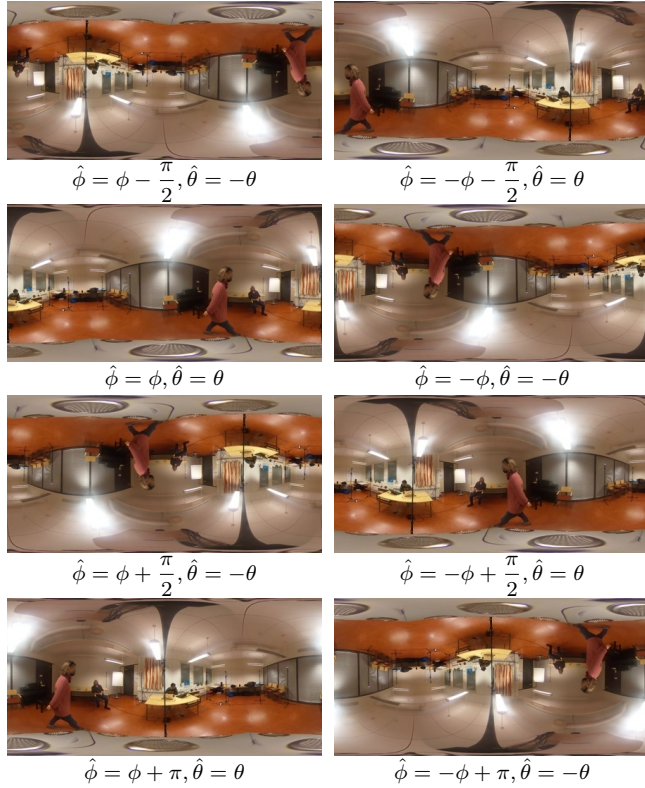
$\hat{\phi} = \phi - \frac{\pi}{2}, \hat{\theta} = -\theta$ $\hat{\phi} = -\phi - \frac{\pi}{2}, \hat{\theta} = \theta$

$\hat{\phi} = \phi, \hat{\theta} = \theta$ $\hat{\phi} = -\phi, \hat{\theta} = -\theta$

$\hat{\phi} = \phi + \frac{\pi}{2}, \hat{\theta} = -\theta$ $\hat{\phi} = -\phi + \frac{\pi}{2}, \hat{\theta} = \theta$

$\hat{\phi} = \phi + \pi, \hat{\theta} = \theta$ $\hat{\phi} = -\phi + \pi, \hat{\theta} = -\theta$

Figure 2: Examples of visual transformation in relation to the respective DOA augmentation for "fold3_room6_mix006.mp4".

3. DEPTH-CUED MODEL

The other main system submitted to the challenge differs significantly from the previous one in many aspects, including model architecture and size, audio and visual data pre-processing, and audio pre-training. Consequently, their overall performances cannot be directly compared. A step-by-step ablation study would be necessary to understand the impact of each change. We refer to this method as per "*Depth-cued model*", since particular attention was given to incorporating audio and visual depth features.

3.1. Direct and Reverberant Audio Features

While log-mel spectrograms and intensity vectors (IV) are effective input features for the SED and DOAE subtasks, they are not ideal for tackling the distance estimation task. Distance cues can be extracted from the relationship between the direct and reverberant components of the captured audio signals. Specifically, the later tail of the reverberant signal (late reverberation) carries information about the sound's apparent distance. We decided to include both direct and reverberant components ("DR", for compactness) extracted from the Omnidirectional audio channel to the set of input features, in the form of two additional log-mel spectrograms. In order to estimate the direct sound, we applied the Weighted Prediction Error (WPE) dereverberation algorithm [15]. Specifically, we adopted the python implementation of the WPE algorithm released by Drude *et al.* [16] (taps=60; delay=5; iterations=5). The reverberant component is then estimated by taking the difference between the original and the direct signal in the temporal domain. RD fea-

tures are then concatenated to the 4 log-mel spectrograms and the 3 intensity vectors used before, producing audio input features with shape $9 \times T_{in} \times F_{in}$.

3.2. Cubemap Conversion and Depth Map Extraction

The 360° videos of STARSS23 are in the form of 2:1 equirectangular views. Such representations produce heavy distortions that become more severe closer to the borders of the frame. We argue that this might penalize the ability of the visual encoder to understand the scene since ResNet50 is pre-trained on frontal views. To this end, we converted the video frames to cubemap representations. Such image representation consists of mapping the 360° image data onto the six faces of a cube. Since the sound events are primarily located within the four horizontal faces (left-right-front-back) and little information can be obtained from the top (ceiling) and the bottom (EigenMike) faces, we kept only the horizontal faces. The resulting frames present an aspect ratio of 4:1, with resolution 896×224p. Note, the conversion to cubemap is applied after the AVCS technique is performed to the equirectangular views.

To leverage the visual modality in support of the distance estimation task, we extracted depth features from video frames too. Specifically, we applied the recent "Depth Anything" model [17] to generate depth views of the scene. Fig. 3 shows (a) a frame with the original equirectangular representation, (b) its cubemap representation, and (c) the depth map extracted from the horizontal faces of the cubemap representation.

3.3. Depth-cued Model's Architecture

Part of the Depth-cued model architecture presents the same audio and visual encoders and Conformer unit employed in the AV-Conformer model. However, the input frames used with the Depth-cued model are cubemap views, and the audio encoder is pre-trained on the synthetic mixtures provided by the challenge's organizers and takes as input also DR features. Additionally, both audio and visual encoders produce 1024-dimensional embeddings.

An additional frame encoder is included too. It consists of a vision transformer (ViT) [18] pre-trained on ImageNet [12]. The input tokens fed to the ViT are obtained by dividing the central frame into patches of 16×16p. Therefore, a total of $14 \times 56 = 784$ tokens are extracted from the 896×224p frame, and each token presents $16 \times 16 \times 3 = 768$ dimensions, where 3 corresponds to the number of color channels. The same patch extraction technique is applied to the depth map frame feature too, which only presents a single channel. The depth patches are then concatenated to the output of the ViT to produce 782×1024 dimensional features. These are employed to generate a set of Key-Value pairs for a multi-head cross-attention (MHCA) unit applied against the audio features (Queries).

The main motivation for employing the ViT and the MHCA unit is to try to leverage the spatial accuracy provided by the visual modality. Unlike ResNet50 which downsamples the spatial resolution of the input frames with the risk of losing spatial information, we argue that the ViT should preserve such information encoded within the processed patches. The outputs generated by the AV-Conformer and by the MHCA are then concatenated and fed to a feed-forward network that predicts the output multi-ACCDDOA vectors [5].
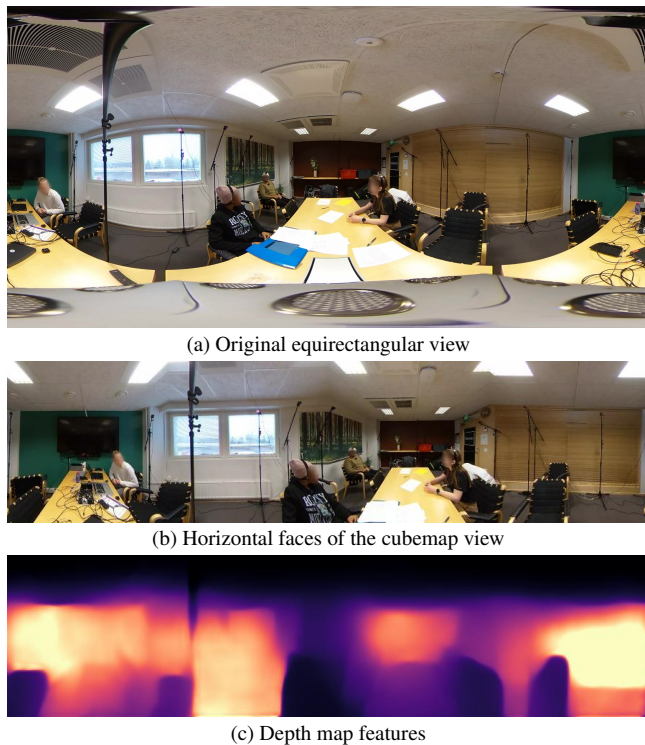
(a) Original equirectangular view

(b) Horizontal faces of the cubemap view

(c) Depth map features

Figure 3: Examples of cubemap transformation and depth map features for "fold3_room13_mix003.mp4". Note how by removing the top and bottom faces from the cubemap representation the Eigen-Mike and a good portion of the ceiling are no longer in the frame.

4. EXPERIMENTS

4.1. Implementation Details

To train our models, we divided the dataset into chunks of 3 seconds, extracted at steps of 1s for training and with no overlap for testing. An STFT with 512-point Hann window and hop size of 150 samples generates spectrograms discretizing the 3-second audio chunks (24kHz) into 480 temporal bins (T_{in}). We used 128 frequency bins to generate log-mel spectrograms and intensity vectors (IV) in log-mel space. We trained our models with batches of 32 inputs and Adam optimizer for 40 epochs, then we selected the best epoch. The learning rate is set to 0.00005 for the first 30 epochs, then and it is decreased by 5% every epoch. The metrics adopted for the evaluation are the ones proposed in the DCASE 2024 Task3 Challenge [6].

4.2. Ensemble strategies

Inference is conducted on 3-second video segments without any overlap, meaning each segment starts exactly 3 seconds after the previous one. To enhance spatial accuracy, we tested a temporal ensemble (TE) strategy. For this, inference was performed with a 1-second hop size, generating 3 predictions for each second of the sequence (except for the first 2 seconds). A sound event is considered active only if at least 2 out of the 3 predictions detect it. To determine if the sounds detected by different predictions are from the same event, we applied a spatial threshold of 15°. Predictions are considered related if their positions are within this threshold.

Table 1: Results on the development set of STARSS23 [6].

Method	$F_{\leq 20°/1}\uparrow$	DOAE $\downarrow$	RDE $\downarrow$
Baseline AO	13.1%	36.9°	33.0%
Baseline AV	11.3%	38.4°	36.0%
AV-Conformer	**40.8%**	17.7°	30.5%
AV-Conformer (TE)	38.7%	**16.8°**	30.4%
Depth-cued	30.7%	18.9°	**27.0%**
Ensemble	40.3%	18.0°	29.7%

The average of their x, y, z, and distance predictions is employed as the final temporal ensemble. This approach improved spatial accuracy but reduced the F1 score. As a result, we submitted a single system using TE, specifically applied to the AV-Conformer model.

As a result, our submitted systems are (1) the AV-Conformer method, (2) the AV-Conformer with TE, and (3) the Depth-cued model. We found that the Depth-cued model rarely detects "Water tap" and "Bell" sounds, and it never detects "Knock" sounds. Therefore, we submitted a fourth system, which is an ensemble of the other three. This ensemble strategy follows the same approach as the temporal ensemble. However, for the "Water tap", "Bell", and "Knock" classes, if any one of the three models detects the sound, we consider it active, even if the condition of having at least 2 out of 3 predictions is not met.

4.3. Results

The results achieved on the development set of the STARSS23 [6] dataset are reported in Table 1. The AV-Conformer method achieves the highest F1 score, while the Depth-cued model has the lowest RDE. However, the Depth-cued model's F1 score is approximately 10 percentage points lower than that of the AV-Conformer, mainly due to its failure to detect certain rare classes, like "Knock". The ensemble system recovers these 10 points in the F1 score. However, it does not provide the improvement in localization accuracy that is achieved with the temporal ensemble. All submitted systems achieve superior performance than the audio-only (AO) and audio-visual (AV) baselines that employ the FOA audio format.

5. CONCLUSION

This technical report describes the 4 systems submitted to the Task 3 of DCASE 2024 Challenge (Track B). Specifically, two main models are explored: an AV-Conformer and a Depth-cued model, which correspond respectively to system one and system three of the four submitted. The second system applies a temporal ensemble strategy to the output of the AV-Conformer, and the fourth system consists of an ensemble of the other three. All our systems outperform the audio-only and audio-visual challenge baselines on the development set of the STARSS23 dataset.

6. ACKNOWLEDGMENT

7. 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, pp. 34–48, 2019.

[2] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events Workshop*, 2019.

[3] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events Workshop*, 2020.

[4] A. Politis, S. Adavanne, D. A. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events Workshop*, 2021.

[5] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," *ArXiv*, vol. 2403.11827, 2024.

[6] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *International Conference on Neural Information Processing Systems*, 2023.

[7] D. Berghi, P. Wu, J. Zhao, W. Wang, and P. J. B. Jackson, "Fusion of audio and visual embeddings for sound event localization and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[8] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 316–320.

[9] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020, pp. 5036–5040.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[11] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. A. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *Detection and Classification of Acoustic Scenes and Events Workshop*, 2022.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale image database," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[13] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.

[14] A. S. Roman, B. Balamurugan, and R. Pothuganti, "Enhanced sound event localization and detection in real 360-degree audio-visual soundscapes," *ArXiv*, vol. 2401.17129, 2024.

[15] T. Yoshioka and T. Nakatani, "Generalization of multichannel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[16] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[17] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth Anything: Unleashing the power of large-scale unlabeled data," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.