FUSION OF AUDIO AND VISUAL EMBEDDINGS FOR SOUND EVENT LOCALIZATION AND DETECTION

Davide Berghi, Peipei Wu, Jinzheng Zhao, Wenwu Wang, Philip J. B. Jackson

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, U.K.

ABSTRACT

Sound event localization and detection (SELD) combines two subtasks: sound event detection (SED) and direction of arrival (DOA) estimation. SELD is usually tackled as an audio-only problem, but visual information has been recently included. Few audio-visual (AV)-SELD works have been published and most employ vision via face/object bounding boxes, or human pose keypoints. In contrast, we explore the integration of audio and visual feature embeddings extracted with pre-trained deep networks. For the visual modality, we tested ResNet50 and Inflated 3D ConvNet (I3D). Our comparison of AV fusion methods includes the AV-Conformer and Cross-Modal Attentive Fusion (CMAF) model. Our best models outperform the DCASE 2023 Task3 audio-only and AV baselines by a wide margin on the development set of the STARSS23 dataset, making them competitive amongst state-of-the-art results of the AV challenge, without model ensembling, heavy data augmentation, or prediction post-processing. Such techniques and further pre-training could be applied as next steps to improve performance.

Index Terms— microphone array, 360 video, sound event localization and detection, audio-visual fusion, cross-modal attention

1. INTRODUCTION

SELD simultaneously predicts the position or DOA of the active sound sources over time, while recognizing their event class [1]. This task is vital 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 and simultaneous same-class events, dealing with moving targets, and ignoring external interfering sounds. In 2019, SELD was proposed for the first time for the DCASE challenge Task 3 [2]. Recently, the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset [3] was released. In contrast to previous datasets for SELD, STARSS23 also includes 360° video recordings spatially and temporally aligned to the acoustic sound-field captured by the microphone array. This allows exploration of SELD as a multimodal audio-

visual problem, we refer to it as audio-visual (AV)-SELD. The two modalities are complementary: vision provides high spatial accuracy whereas audio can detect occluded objects [4]. A different scenario for AV-SELD was proposed for the L3DAS23 challenge [5], with a single frontal color image for each microphone array position. Thus, the visual component offered a static narrow view.

In this paper, we tackle AV-SELD by combining audio and visual feature embeddings extracted with pre-trained architectures. We compare two attention-based approaches to fuse the two modalities and demonstrate a substantial improvement over the audio-only and audio-visual DCASE baselines on the STARSS23 dataset. Code to replicate our models is available on https://github.com/dberghi/AV-SELD.

2. RELATED WORKS

Sound scenes for SELD are usually provided in two spatial acoustic formats: First-Order Ambisonics (FOA), or tetrahedral microphone array (MIC). Initially, convolutional recurrent neural networks (CRNN) were the most adopted choice to tackle the task [1, 6, 7, 8]. However, attention-based architectures, such as Multi-Head Self-Attention (MHSA) [9], Transformers [10], or Conformers [11, 12], are often preferred lately. As audio input features, Cao *et al.* [6] proposed to concatenate multichannel log-mel spectrograms with the GCC-PHAT computed at each pair of microphones. For the FOA format, intensity vectors (IV) in log-mel space are adopted [13]. Nguyen *et al.* [7] proposed the Spatial Cue-Augmented Log-Spectrogram (SALSA) features, and SALSA-Lite [14] for the MIC format. To solve the problem of concurrent same-class events, an Event Independent Network [9] with permutation invariant training was proposed. Inspired by this, Shimada *et al.* [15] proposed multiple activity-coupled Cartesian DOA (m-ACCDOA) vectors as SELD output representation. With m-ACCDOA the network is trained with class-wise auxiliary duplicating permutation invariant training (ADPIT).

To the best of our knowledge, there is no prior work on AV-SELD before the DCASE 2023 Task 3 challenge, or at least not with a similar setting. Therefore, we will describe the systems proposed by the teams that participated in the challenge, even though the technical reports are too short to
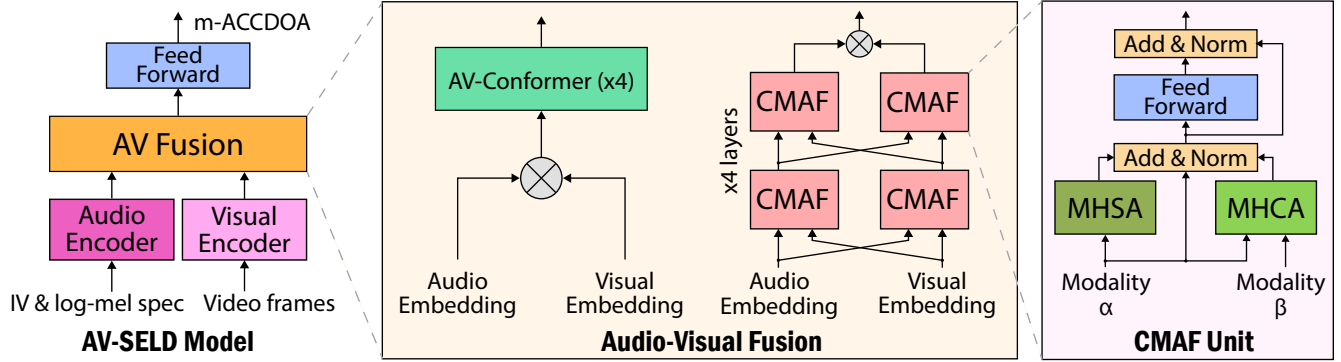
Fig. 1. (left) AV-SELD pipeline diagram with CNN-Conformer audio encoder [11], and I3D [16] or ResNet-Conformer [17][11] visual encoder; (center) four layers of attention-based fusion of audio and visual embeddings: AV-Conformer [11] or Cross-Modal Attentive Fusion, where $\otimes$ is concatenation; (right) CMAF unit diagram [18].

provide a thorough description. The AV baseline system consists of a CRNN that leverages visual object detection. Objects bounding boxes are represented as Gaussian-like vectors, as in [18], and concatenated to the features extracted by the audio CNN, before being fed to the recurrent unit. Wang *et al.* [19] adopted a similar approach. However, alongside the Gaussian-like vectors, they also included a frame embedding extracted with ResNet18 from the first video frame. Wang *et al.* [20] performed human keypoint detection and object detection with a ResNet-Conformer model for the AV-SELD predictions. The predictions are then post-processed by taking full advantage of the spatial accuracy of the visual modality. For instance, if the model predicts the class "Footsteps", the predicted location is replaced with the location of the closest foot keypoint. Kang *et al.* [21] adopted a student-teacher distillation to train a ResNet-Conformer from a ResNet-GRU network. For the visual modality, they employed object detection. Kim *et al.* [22] trained a network similar to the audio-only baseline, i.e., a CRNN followed by MHSA. In contrast to the other systems, they employed a 3D CNN to extract a visual embedding from raw video, optical flow frames, and object detection features. The so-extracted visual embedding is then concatenated with the audio embedding after the MHSA block, before the prediction of m-ACCDOA vectors, preventing attention-based modality fusion.

In this work, we extract an audio feature embedding using a CNN-Conformer network and a visual embedding employing either a ResNet-Conformer or the Inflated 3D ConvNet (I3D) [16]. The two feature embeddings are then fused by an attention-based approach: we compare an AV-Conformer and the Cross-Modal Attentive Fusion (CMAF) network [18].

3. METHOD

As depicted in Fig. 1, our method extracts audio and visual feature embeddings with an audio and a visual encoder, respectively. The embeddings are then combined in an attention-based mechanism. Finally, the output features are

fed to two fully connected layers to predict m-ACCDOA vectors for SELD [15]. Among our experiments, we studied two different visual encoders and two fusion mechanisms. Below we describe each component and the AV dataset.

Dataset: Results supporting this study are generated on the development set of the STARSS23 dataset [3]. It consists of about 7.5h of real AV spatial recordings of acoustic scenes temporally and spatially annotated with the acoustic class of the active events. It presents a total of 13 event classes. Our experiments employ the FOA audio format. Video recordings consist of 360° views of the environment captured from the microphone array perspective at a resolution of 1920x960p, 30fps. Audio is sampled at 24kHz, 16 bits. Labels are provided at a resolution of 100ms. The dataset includes directional interferes that do not belong to the target classes and must not be detected. It is common to encounter up to 3 simultaneous events, but more simultaneous occurrences, up to 5, are possible. We employed the provided training-testing split of the development set to train and test our models.

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 [11]. 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. 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.

Visual Encoder: As visual encoders, we compared the ResNet-Conformer and I3D. In the former, we fed each video frame to ResNet50 [17] 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 encoder. Additionally, we tested a 3D CNN to take full advantage of the video temporal dimension. We opted for I3D [16] as it is a widely employed yet effective model for action recognition and 3D feature extraction. The original model comprises two branches: one for the RGB video and one for the optical flow frames. For simplicity, we only employed the RGB branch. For computational reasons, we kept the input frame rate adopted with the ResNet-Conformer (10fps). 3D CNNs have the effect of downsampling the temporal dimension. Therefore, the output of I3D is then temporally interpolated to reconstruct the label frame rate. The video segments used as inputs to the visual encoders were resized to 448x224p. Both ResNet50 and I3D were pre-trained on squared 224x224p frames. Therefore, to avoid losing spatial information, we divided the input frames into a 'left' and a 'right' side of 224x224p each, and concatenated the two resulting vectors. In practice, the last layer of ResNet50 is a global average pooling applied on a 7x7 feature map. Since the input horizontal dimension is 448, the resulting feature map has shape 14x7. Therefore we replaced the last global average pooling with a 7x7 average pooling kernel with stride equal to 7, obtaining two output vectors that are then concatenated. In both cases, the outputs of the encoders are fed to a linear layer to achieve a 512-dimensional video embedding.

Audio-Visual Fusion: We considered two attention-based fusion approaches. The first is an AV-Conformer that takes as input the concatenation of the audio and visual embeddings, i.e., a 1024-dimensional vector at each time frame. The AV-Conformer is identical to the one employed in the audio CNN-Conformer and in the ResNet-Conformer. The second approach leverages the CMAF network proposed by Qian *et al.* [18]. The audio and visual embeddings are the inputs of two parallel CMAF blocks. The structure of each block for two generic modalities α and β is depicted in Fig. 1 (right). In contrast to a traditional Transformer [23], in parallel to the MHSA module, they present Multi-Head Cross-Attention (MHCA) where α is used to generate the query $\mathbf{Q}_\alpha$ to be mapped to the set of key $\mathbf{K}_\beta$ and value $\mathbf{V}_\beta$ pairs generated from β. The input α is then added to the outputs of the MHSA and the MHCA modules. Two parallel CMAF blocks represent one layer. Their outputs are the inputs for the following layer. We employed a total of 4 layers. Both MHSA and MHCA contain 8 heads. The two outputs of the last layer are concatenated to generate a 1024-dimensional vector at each time frame. We generate $N = 3$ tracks as m-ACCDOA vectors via two fully-connected layers.
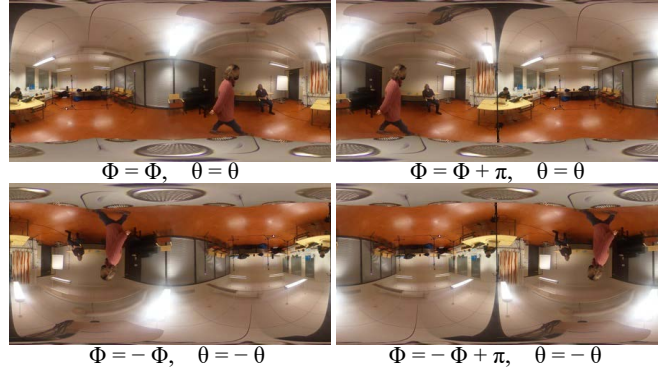

Fig. 2. Visual transformations for DOA data augmentation.

The network is trained with class-wise ADPIT loss [15].

4. EXPERIMENTS

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 [24]. 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. I3D was pre-trained on action detection on the Kinetics dataset. 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 [12]. For the AV-SELD dataset, we also augmented the visual modality consistently with the ACS transformation as indicated by Wang *et al.* [20]. However, while they only augment the positions of the detected human keypoints or objects bounding boxes, we generate complete new video frames by flipping and rotating the original ones. This creates an effective audio-visual spatial transformation: the audio-visual channel swap (AVCS). Examples of a transformed frame are shown in Fig. 2.

Implementation Details and Evaluation Metrics: To train our models, we divided the dataset into chunks of 3 seconds, extracted at steps of 0.5s 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 50 epochs, then we selected the best epoch. The learning rate is set at a fixed value for the first 30 epochs, then and it is decreased by 5% every epoch. The initial value is in the range of 1E-4 and 1E-3, depending on the one that generates the best results for each model. The metrics adopted for the evaluation are the ones proposed in the DCASE23 Task3 Challenge [3]. They include location-dependent detection metrics, such as the Error Rate ($ER_{\leq 20°}$) and F1-

Table 1. Results on STARSS23 development set. $F1$ and LR are in percent (%), LE in deg (°). CNN-Conf is employed as audio encoder. I3D: Inflated 3D ConvNet; RC: ResNet-Conformer; Conf: AV-Conformer; CMAF: Cross-Modal Attentive Fusion; CA: Cross-Attention; GRU: Gated Recurrent Units; AO: Audio-only; AV: audio-visual.

Visual	Fusion	ER↓	F1↑	LE↓	LR↑	SELD↓
I3D	CMAF	0.57	40.5	33.2	55.3	0.45
I3D	Conf	0.52	46.4	16.9	60.2	0.39
RC	CMAF	0.54	41.3	31.6	53.4	0.44
RC	Conf	0.51	49.5	15.8	60.2	**0.38**
RC	CA	0.55	34.7	30.4	47.8	0.47
RC	GRU	**0.50**	49.4	16.2	56.8	**0.38**
Both	Conf	0.52	48.0	16.2	**60.8**	**0.38**
Visual-only		1.03	0.9	103	11.4	0.87
Audio-only		0.51	**50.2**	**15.4**	56.4	**0.38**
Baseline AO		0.57	29.9	22.0	47.7	0.48
Baseline AV		1.07	14.3	48.0	35.5	0.71

score ($F1_{\leq 20°}$), computed with respect to a spatial threshold $T = 20°$, and the Localization Error (LE_{CD}) and Localization Recall (LR_{CD}). $F1_{\leq 20°}$, LE_{CD}, and LR_{CD} are macro-averaged across classes. For compactness, we will omit the subscripts $\leq 20°$ and CD from now on. Results also include the overall $SELD$ score that encodes all other four metrics: $SELD = mean(ER, (1 - F1), LE/180, (1 - LR))$.

Experiments: In addition to the combinations of ResNet-Conformer/I3D and CMAF/AV-Conformer introduced previously, we tested a simplified cross-attention (CA) mechanism that combines the audio and visual embeddings like in CMAF. However, each CA block only presents MHCA without parallel self-attention. We also compared our attention-based fusion methods with a recurrent network, i.e., two bidirectional gated recurrent units that take as input the concatenation of the embeddings. Furthermore, we conducted an experiment where both I3D and ResNet-Conformer visual embeddings are concatenated to the audio embedding before being fed to the AV-Conformer. Results are reported in Tab. 1. The table also includes visual-only (VO) and audio-only (AO) methods to understand the contribution of each modality. For fairness, we matched the networks' depth of the AV methods by adding an additional 4-layer Conformer to the CNN-Conformer and the ResNet-Conformer, respectively. The bottom rows report the results achieved by the AO and AV challenge baselines.

Results: Comparing the single-modal methods to the AV methods, it is clear that most of the task is accomplished by the audio modality, which alone got the best LE and $F1$. Nevertheless, results from VO suggest that the visual modality can provide a partial contribution to that task, especially in terms of LR which is above 11%. In fact, the AV methods that employ the AV-Conformer fusion achieved up to a 4 percentage points gain in LR over the AO system ($LR > 60\%$).

However, vision did not contribute to improving the ER and spatial accuracy. Comparing I3D and ResNet-Conformer, it is possible to appreciate a slight benefit with the latter, even though it is marginal. Nevertheless, the method that leverages both visual encoders is the one with the highest gain in LR.

While CMAF proved to be preferable over direct cross-attention, the methods trained with this type of AV fusion produced weak results compared to the AV-Conformer or GRU, with $LE > 30°$. From a thorough analysis, we noticed that the problem is caused by the 'Knock' class, which CMAF and CA are not detecting. This particularly penalizes the $F1$, LE, and LR as they are macro-averaged across classes. For CMAF with ResNet-Conformer as visual encoder, we re-computed the micro-averaged metrics and got $F1 = 54.3\%$, $LE = 19.0°$, and $LR = 71.4\%$. For comparison, the system that employs the AV-Conformer and both visual encoders achieves micro-averaged $F1 = 55.4\%$, $LE = 18.6°$, and $LR = 72.6\%$. This suggests that CMAF and CA are more competitive than what appears in the table. While it is not a justification for not detecting the 'Knock' class, it should be noted that 'Knock' appears in only 90 frames out of the entire training set (9 seconds). We argue that results might be different with a more balanced dataset.

AV-SELD is a new and non-trivial task. This can be appreciated from the weak results achieved by the AV baseline. Yet, all our methods outperformed the baselines by a wide margin. To our knowledge, only Wang *et al.* [20] and Kang *et al.* [21], the challenge top-2 teams, achieved higher performance on the AV-SELD task. We believe that integrating visually-guided prediction post-processing, as in [20], could help improve DOA estimation of our systems alongside ER and $F1$. Further improvement can be achieved from additional pre-training, data augmentation, and model ensembling.

5. CONCLUSION

In this paper, we proposed and compared some of the first architectures for AV-SELD. We tackled the task by combining audio and visual feature embeddings extracted with pre-trained deep networks, comparing two well-established visual encoders. In terms of AV fusion, tests include but are not limited to, the AV-Conformer and CMAF. Our methods substantially outperformed the audio-visual and audio-only baselines of the DCASE 2023 Task 3 challenge.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. of Selected Topics in Signal Processing*, vol. 13, pp. 34–48, 2019.

[2] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *DCASE Workshop*, 2019.

[3] Kazuki Shimada *et al.*, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *ArXiv*, vol. abs/2306.09126, 2023.

[4] Jinzheng Zhao *et al.*, "Audio-visual speaker tracking: Progress, challenges, and future directions," *ArXiv*, vol. abs/2310.14778, 2023.

[5] Eric Guizzo *et al.*, "L3DAS22 challenge: Learning 3D audio sources in a real office environment," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2022.

[6] Yin Cao *et al.*, "Polyphonic sound event detection and localization using a two-stage strategy," in *DCASE Workshop*, 2019.

[7] Thi Ngoc Tho Nguyen *et al.*, "SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022.

[8] Qing Wang *et al.*, "The USTC-iFlytek system for sound event localization and detection of DCASE2020 challenge technical report," in *Tech. Rep. of DCASE Challenge*, 2020.

[9] Yin Cao *et al.*, "An improved event-independent network for polyphonic sound event localization and detection," in *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*, 2021.

[10] Peipei Wu *et al.*, "PLDISET: Probabilistic localization and detection of independent sound events with transformers," *DCASE Workshop*, 2023.

[11] Anmol Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *ArXiv*, vol. abs/2005.08100, 2020.

[12] Qing Wang *et al.*, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.

[13] Yin Cao *et al.*, "Event-independent network for polyphonic sound event localization and detection," in *DCASE Workshop*, 2020.

[14] Thi Ngoc Tho Nguyen *et al.*, "SALSA-lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*, 2022.

[15] Kazuki Shimada *et al.*, "Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*, 2022.

[16] João Carreira and Andrew Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.

[18] Xinyuan Qian *et al.*, "Audio-visual cross-attention network for robotic speaker tracking," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 550–562, 2023.

[19] Yi Wang, Hongqing Liu, and Yi Zhou, "Audio-visual sound event localization and detection based on CRNN using depth-wise separable convolution," in *Tech. Report of DCASE Challenge*, 2023.

[20] Qing Wang *et al.*, "The NERC-SLIP system for sound event localization and detection of DCASE2023 challenge," in *Tech. Report of DCASE Challenge*, 2023.

[21] Sang-Ick Kang, Kyongil Cho, Myungchul Keum, and Yeonseok Park, "The distillation system for sound event localization and detection of DCASE2023 challenge," in *Tech. Report of DCASE Challenge*, 2023.

[22] Gwantae Kim and Hanseok Ko, "Data augmentation, neural networks, and ensemble methods for sound event localization and detection," in *Tech. Report of DCASE Challenge*, 2023.

[23] Ashish Vaswani *et al.*, "Attention is all you need," in *Int. Conf. on Neural Information Processing Systems*, 2017.

[24] Archontis Politis *et al.*, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *DCASE Workshop*, 2022.