MANY-TO-MANY AUDIO SPECTROGRAM TRANSFORMER: TRANSFORMER FOR SOUND EVENT LOCALIZATION AND DETECTION

Sooyoung Park, Youngho Jeong, Taejin Lee

Electronics and Telecommunications Research Institute, Media Coding Research Section, Daejeon, Republic of Korea, {sooyoung, yhcheong, tjlee}@etri.re.kr

ABSTRACT

Over the past few years, convolutional neural networks (CNNs) have been established as the core architecture for audio classification and detection. In particular, a hybrid model that combines a recurrent neural network or a self-attention mechanism with CNNs to deal with longer-range contexts has been widely used. Recently, Transformers, which are pure attention-based architectures, have achieved excellent performance in various fields, showing that CNNs are not essential. In this paper, we investigate the reliance on CNNs for sound event localization and detection by introducing the *Many-to-Many Audio Spectrogram Transformer* (M2M-AST), a pure attention-based architecture. We adopt multiple classification tokens in the Transformer architecture to easily handle various output resolutions. We empirically show that the proposed M2M-AST outperforms the conventional hybrid model on TAU-NIGENS Spatial Sound Events 2021 dataset.

***Index Terms*—** Sound event localization and detection, self-attention, Transformer

1. INTRODUCTION

Convolutional neural networks (CNNs) have become essential for designing deep neural networks for image understanding tasks. The translation equivariance and locality of CNNs are known to be effective for image understanding. Due to the success of CNNs in image understanding, CNNs have also been used in other pattern recognition fields [1, 2]. Especially in audio understanding, CNNs have been applied to spectrogram images which are extracted from audio recordings by applying short-time Fourier transform to recognize image patterns. However, it is necessary to understand the longer context as well as the local context of the spectrogram in audio understanding fields. To understand this longer context, networks combining recurrent neural networks (RNNs) or self-attention with CNNs have been widely used [3, 4].

Self-attention mechanisms [5], especially Transformers, have become the new standard for natural language processing (NLP). The main approach of NLP is to fine-tune large pre-trained networks on small task-specific datasets. Transformers are well known for their computing efficiency and scalability. Using these Transformers, large models trained on large-scale text corpus datasets have been released [6]. These large models are known to extract generality from large amounts of training data. With the success of Transformers in NLP, Transformers are starting to be utilized in other fields [7, 8, 9, 10]. However, architectures combined with CNN rather than pure transformer architectures are mainly used.

Recently, *Vision Transformer* (ViT) [7, 8] using only pure Transformers for image understanding has been introduced. The outstanding performance of ViT is starting to question whether CNNs are still essential in many applications. Since then, research on Transformers replacing CNNs has become a trend in various fields. The *Keyword Transformer* (KWT) [9] and *Audio Spectrogram Transformer* (AST) [10] have been introduced as the first attempts to replace CNNs with Transformers in audio understanding. These studies demonstrate the potential of a pure Transformer to lower the dependence on CNNs in audio understanding. Inspired by the strength of the simple Transformer model in computer vision and audio classification, we propose an adaptation of this architecture to sound event localization and detection (SELD) [11].

In this paper, we propose a pure transformer architecture, *Many-to-Many Audio Spectrogram Transformer* (M2M-AST), for sound event localization and detection (SELD). M2M-AST enables efficient training of large models through transfer learning from large pre-trained models. AST provides one audio classification output for a single channel audio input (one-to-one). M2M-AST can have different resolution output sequences for multi-channel audio inputs (many-to-many).

2. RELATED WORK

2.1. Sound Event Localization and Detection

SELD [11] is the task of classifying multiple sound events with temporal activity into specific classes and detecting their directions. Therefore, SELD can be separated into two small tasks: sound event detection (SED) and direction of arrival estimation (DOAE). Specifically, SED is the task of classifying sound events into specified target classes to identify their onsets and offsets when sound events occur. DOAE is the task of detecting directions in which sound events occur in every frame. The DCASE challenge has published datasets for SELD since 2018. The TAU-NIGENS Spatial Sound Events 2020 dataset [12] consists of data that allow up to two simultaneous occurrences of sound events with directional activities. Additionally, up to three target sound events can occur simultaneously in the TAU-NIGENS Spatial Sound Events 2021 dataset [13]. Also, TAU-NIGENS Spatial Sound Events 2021 dataset is more difficult than TAU-NIGENS Spatial Sound Events 2020 dataset because there is background noise from unknown spatial acoustic events.

In SELD, the two-stage approach [14] and the joint modeling approach are dominant. The two-stage approach splits SELD into two models, SED and DOAE, and trains each separately. In the joint modeling approach, SED and DOAE are co-trained or integrated into a single system. Both methods commonly use convolutional recurrent neural networks (CRNNs) [4, 15, 16] or hybrid networks [17] that combine CNNs with self-attention layers. For CNNs, the output resolution depends on the pooling size. Therefore, SELD models using CNNs have limited output resolution by pooling size

and cannot freely construct output resolution depending on the application.

2.2. Self-Attention and Transformers

As pure self-attention-based networks, Transformers became the standard for NLP. Then, with the advent of ViT [7], the pure Transformer model expanded to the field of image understanding. ViT outperforms CNNs in image classification with self-attention computations between different image patches. However, ViT requires a significant amount of training data. To improve this, DeiT [8], which uses data augmentations and a knowledge distillation token to improve data efficiency, has been proposed. With the success of understanding images without CNNs, other research fields are also studying the reliance on CNNs.

KWT [9] and AST [10] are the first studies using a pure transformer in the field of audio understanding. These studies show that the pure Transformer models can replace CNNs in audio classification. In particular, KWT is a model that has adjusted the structure of the DeiT model for audio classification. AST is a model for efficiently training large-scale Transformer networks using ImageNet pre-trained models. The above studies are about a one-to-one structure that performs one classification on one audio recording. We propose an M2M-AST architecture that outputs sequences of varying resolutions from multi-channel audio recordings.

3. MANY-TO-MANY AUDIO SPECTROGRAM TRANSFORMER

3.1. Features

We use logmel and intensity vectors as input features [13] for SELD. The proposed SELD system is based on two-stage approach. The proposed SED network and DOAE network take different input features. The SED network uses logmel energy extracted from the microphone array data segmented into a single channel as input features. The DOAE network uses 7-channel inputs by extracting logmel and intensity vectors from Ambisonic data. This is summarized in Table 1. Table 2 shows the pre-processing parameters to extract input features

	Format	Feature	# Channels (C)	Label
SED	Microphone array	Logmel	1	Multi label binarization
DOAE	Ambisonic	Logmel, intensity vector	7	Cartesian coordinate (xyz)

Table 1: Feature and label configuration for SED and DOAE

Pre-processing	
Time window length	20 ms
Time window stride	10 ms
Frame length (T)	300 (3 sec)
# Mel-bins (M)	128

Table 2: Pre-processing parameters

3.2. Model Architecture

As shown in Figure 1, the proposed M2M-AST uses a Transformer encoder in the same way as AST [10]. M2M-AST uses only the encoder layer of the Transformer for classification and regression. Compared to AST, M2M-AST has differences in input feature and classification token configuration. Neural Networks for SELD extract multi-channel feature images from 4-channel audio recordings and use them as input features. Therefore, M2M-AST uses multi-channel feature images extracted from 4-channel audio recordings. Then we segment the extracted multi-channel feature images into the 16x16 patch sequence. At this point, we split the feature images by applying the same stride to the time and frequency dimensions. Afterwards, patch tokens are extracted through a linear projection for each patch. Like ViT [7, 8], the learnable classification token for classification is appended at the beginning of the patch token sequence. However, since SELD performs SED and DOAE every 100 ms, M2M-AST should output a series of outputs rather than a single output like AST. Therefore, patch embedding consists of appending a classification token sequence of equal length to the length of the output sequence at the beginning of the patch token sequence. The length of the classification token sequence determines the output resolution. For example, configuring an output resolution of 100 ms for 3 seconds input data would use a sequence of 30 classification tokens. On the other hand, configuring an output resolution of 20 ms for 3 seconds of input data would use a sequence of 150 classification tokens. The Transformer has no convolution or recurrence, so it cannot leverage the relative spatial information of the patch tokens in the 2D feature images. To take advantage of the position information of the patch tokens, we add a learnable positional embedding $\mathbf{p}_i (\in \mathbb{R}^d)$ to the patch embedding. The Transformer encoder's outputs of the classification token sequence learn the audio spectrogram representation by computing the self-attention between each patch token. Then, we use a dense layer with an activation layer for SED and DOAE from the Transformer encoder's output of classification tokens. M2M-AST's model parameters are the same as AST and are summarized in Table 3.

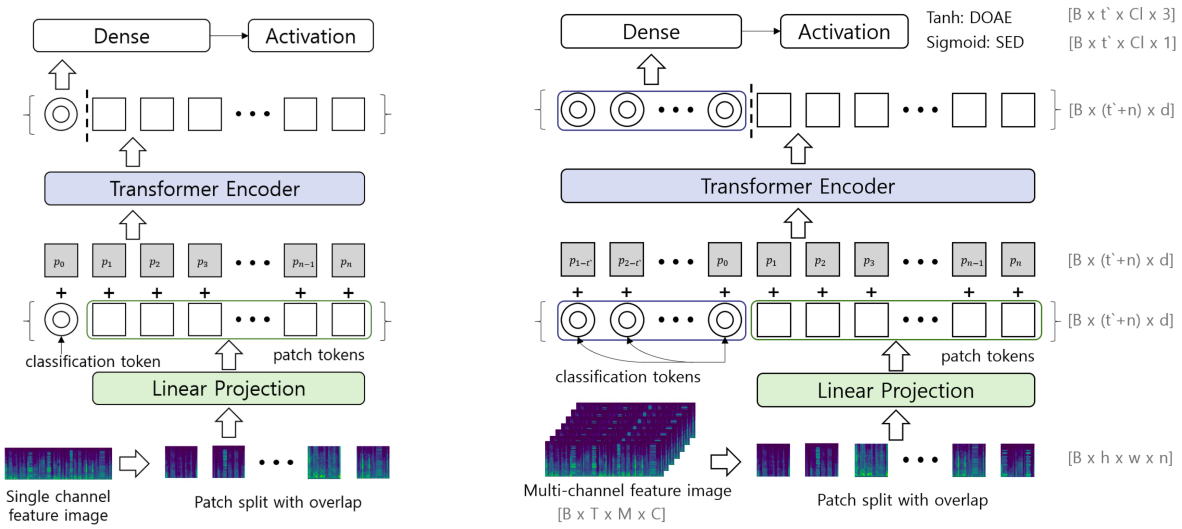
Model parameter	
Patch shape (h x w)	16x16
Patch overlap	6
# Patches (n)	348
Patch dimension (d)	768
# Encoder layer (L)	12
# Attention head	12
Output size (t')	20
Dropout	0.1

Table 3: Model parameters

3.3. Transfer Learning

M2M-AST uses only a Transformer encoder like AST [10]. M2M-AST uses the same Transformer encoders as ViT [7, 8]. ViT requires a large dataset for sufficient performance. To overcome this problem, DeiT [8] uses knowledge distillation. Unlike image datasets, audio datasets contain relatively small amounts of data. Therefore, AST uses transfer learning to distill knowledge from the ImageNet pre-trained model. M2M-AST uses transfer learning in the same way as AST. The weight can be easily transferred because the same Transformer encoder is used. However, the layer learning patch embeddings vary in size and require some adjustments.

DeiT uses 3-channel input images, while M2M-AST uses variable multi-channel input images. In M2M-AST, the weight corresponding to each channel in the linear projection layer uses the



(a) AST (b) Many-to-Many AST

Figure 1: Architecture of AST and Many-to-Many AST; B: batch size, T: time, M: # mel-bins, C: channel, t': output size, (h x w): patch shape, n: # patchs, d: patch dimension, Cl: # class

average weight of the three channels in DeiT. While DeiT has a fixed length of patch embedding sequence, M2M-AST has not fixed length of patch embedding sequence because the classification token sequence that determines the output resolution is variable. Therefore, the positional embeddings for the patch tokens in M2M-AST $[\mathbf{p}_1, \cdots, \mathbf{p}_n]$ are transferred as scaled values through cut and bilinear interpolation to map the relative positions of the positional embeddings in DeiT to the input feature. Individual positional embeddings of classification token sequence $[\mathbf{p}_{1-t'}, \cdots, \mathbf{p}_0]$ are equally initialized by the average value of classification token and distillation token in DeiT. This transfer learning method makes it easy to extract pre-trained network knowledge for ImageNet into the audio domain.

3.4. Post-processing

The input time window for our system is 3 seconds. We slide this window with a small hop size to create many overlapped results and average these results during the inference [15]. Additionally, we apply median filtering and tuning the threshold for each class during SED inference. Finally, we apply a 16-way rotation augmentation to infer the test data and average the values obtained by rotating the results in reverse [15, 16].

4. EXPERIMENTS

We provide experimental results on TAU-NIGENS Spatial Sound Events 2021 development dataset [13]. The development dataset consists of 600 1-minute wave files. We use 400 minutes of data for training, 100 minutes for validation, and the remaining 100 minutes for testing. Our system is trained using the hyper-parameters in Table 4. We use transfer learning with the pre-trained model. The pre-trained model used in our system is shown in Table 5. We fine-tune the SED model with 85M parameters and the DOAE model with 86M parameters for 50 epochs independently. We use the Adam optimizer. For the development dataset, the training time consumed by the M2M-AST is 4 hours for SED and 2 hours for DOAE at 4-TITAN Xp. The model mentioned in Table 5 is used for the experiment. The models mentioned in Table 5 are for each SED and DOAE task. Because of the large model size of M2M-AST, we use a two-stage approach rather than joint training for SELD.

Training	
Epoch	50
# Batch (B)	24
Learning rate	0.0001
Optimizer	Adam

Table 4: Hyper-parameters for proposed system

	Task	Pre-trained model	Loss
M2M-AST1	SED	DeiT	BCE
M2M-AST2	SED	M2M-AST1	soft f-loss [18, 19]
M2M-AST3	DOAE	DeiT	MSE
M2M-AST4	DOAE	M2M-AST3	masked MSE

Table 5: Model configuration

4.1. Results

Table 6 reports the results on the TAU-NIGENS Spatial Sound Events 2021 dataset [13]. All results are based on logmel energy and intensity vectors as input features. Baseline-Large is a model in which the filter size of the baseline is increased to be similar to the model size of M2M-AST. Using M2M-AST with the two-stage approach significantly improves the performance of all metrics over the baseline. In addition, the proposed pure Transformer model, M2M-AST, outperforms the CRNN-based models listed in Table 6, demonstrating that it is a sufficient replacement for CRNNs. Therefore, we show that self-attention computing within the Transformer on SELD, SED, and DOAE can reduce the reliance on CNNs.

	# Params	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
CRNN (Baseline FOA)	0.5M	0.69	33.9 %	24.1°	43.9 %
CRNN (Baseline-Large)	184M	0.65	45.6 %	22.6°	55.0 %
CRNN [20]	14M	0.65	48.3 %	22.0°	62.6 %
M2M-AST1&3	172M	0.55	62.6 %	17.5°	74.0 %
M2M-AST1&4	172M	0.52	64.4 %	**16.0°**	74.0 %
M2M-AST2&3	172M	0.52	64.0 %	17.7°	**74.7 %**
M2M-AST2&4	172M	**0.50**	**65.7 %**	**16.3°**	**74.7 %**

Table 6: Experimental results for development dataset

4.2. Ablation Study

We perform a series of ablation studies to explain M2M-AST design choices. We conduct ablation studies based on M2M-AST1 and M2M-AST3 initialized with ImageNet pre-training models while using loss functions commonly used in SELD.

4.2.1. Batch size and frame length

We compare the performance of M2M-AST with different batch sizes and frame lengths of input features through grid search. Table 7 shows the results of this comparison. Performance for SED under the ideal DOAE condition is evaluated through the F_1 score and LR_{CD}(F_∞ score). F_1 score represents the balanced score of precision and recall. Besides LR_{CD} represents the recall dominant score. On the other hand, longer input frames improve both precision and calls. This is because longer input frames make M2M-AST use more patches for training. For DOAE, smaller batch sizes and longer input frame lengths improve performance.

	SED (F_1, LR_{CD})			DOAE (LE_{CD})		
# Batch	1 sec	2 sec	3 sec (Used)	1 sec	2 sec	3 sec (Used)
24 (Used)	(68.3, 66.3)	(75.0, 73.2)	(74.0, **74.0**)	26.3°	22.2°	**21.8°**
48	(69.5, 70.9)	(75.7, 72.1)	(75.2, 73.6)	27.9°	23.1°	23.0°
96	(70.7, 70.3)	(**75.8**, 68.7)	-	27.0°	24.4°	-

Table 7: Experimental results with different batch sizes and input frame lengths

4.2.2. Patch split with overlap

Table 8 shows a performance comparison with patch splits of 16x16 sizes using various sizes of strides. Configuring dense patch segmentation with large overlap helps both SED and DOAE improve performance. However, for SED, performance improvements converge on overlap size 6. Thus, exploiting patch splits with a larger overlap size than 6 leads to the burden of memory and computation cost.

	# Patches	SED (F_1, LR_{CD})	DOAE (LE_{CD})
No Overlap	144	(71.6, 60.2)	27.3°
Overap-2	189	(73.8, 68.6)	24.6°
Overap-4	240	(74.1, 70.6)	24.1°
Overap-6 (Used)	348	(74.0, **74.0**)	**21.8°**
Overap-8	540	(**74.9**, 72.5)	**21.0°**

Table 8: Experimental results with different lengths of patch overlap

4.2.3. Output resolution

M2M-AST can adjust the number of classification tokens to have a variety of output resolutions. Table 9 shows a performance comparison of M2M-AST with output resolution from 100 ms to 25 ms. Since the resolution of the ground truth data is 100 ms, we use the nearest-neighbor interpolation to construct labels with high resolution and use them for training. Then we apply a median filter to construct an output of 100 ms. For SED, smaller resolution results in slight performance gains due to median filtering. On the other hand, for DOAE, the results do not vary significantly with changes in output resolution.

Output resolution	Output size (t')	SED (F_1, LR_{CD})	DOAE (LE_{CD})
25 ms	120	(75.3, 73.8)	22.2°
33.3 ms	90	(76.5, 75.1)	22.1°
50 ms	60	(74.4, 72.8)	22.7°
100 ms (Used)	30	(74.0, 74.0)	21.8°

Table 9: Experimental results with different output resolutions

4.2.4. Pre-training and loss function

We compare the performance of randomly initialized M2M-AST and M2M-AST transferred from pre-trained models. As shown in Table 10, the weight transferred model from ImageNet pre-trained model outperforms the randomly initialized model in SED. On the other hand, transfer learning from ImageNet pre-trained model improves DOAE performance slightly. In addition, we compare M2M-AST using different loss functions while using a pre-trained model. In SED, soft f-loss [18, 19] is slightly better than binary cross-entropy (BCE), but there is no significant difference. On the other hand, with the DOAE pre-trained model, masked MSE improves performance by 2.7 degrees over BCE.

	Pre-trained model	Loss	SED (F_1, LR_{CD})	DOAE (LE_{CD})
No pre-train (SED)	-	BCE	(60.4, 54.5)	-
ImageNet pre-train (M2M-AST1)	DeiT	BCE	(74.0, 74.0)	-
SELD pre-train (M2M-AST2)	M2M-AST1	soft f-loss	(**75.8**, **74.7**)	-
No pre-train (DOAE)	-	MSE	-	22.5
ImageNet pre-train (M2M-AST3)	DeiT	MSE	-	21.8
SELD pre-train (M2M-AST4)	M2M-AST3	masked MSE	-	**19.1**

Table 10: Experimental results with different loss functions and pre-trained models

5. CONCLUSIONS

In this paper, we describe how to apply the standard Transformer architecture to SELD. As a consequence, we introduce M2M-AST, a pure Transformer model for SELD. Existing SELD networks have commonly used hybrid architectures that combine CNNs with RNNs or self-attention layers. We empirically show that M2M-AST can replace these hybrid networks in SELD, SED, and DOAE. The experimental results represent the potential of a pure Transformer to lower the reliance on CNNs in SELD. Traditional neural networks use pooling layers to change the output shape. However, due to the pooling size of this pooling layer, the output resolution cannot be configured freely. On the other hand, M2M-AST has the advantage of being able to easily design to have a variety of output resolutions.

6. ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00050, Development of Human Enhancement Technology for auditory and muscle support).

7. REFERENCES

[1] S. Suh, S. Park, Y. Jeong, and T. Lee, "Designing acoustic scene classification models with CNN variants," DCASE2020 Challenge, Tech. Rep., June 2020.

[2] R. Giri, S. V. Tenneti, K. Helwani, F. Cheng, U. Isik, and A. Krishnaswamy, "Unsupervised anomalous sound detection using self-supervised classification and group masked autoencoder for density estimation," DCASE2020 Challenge, Tech. Rep., July 2020.

[3] Q. Wang, H. Wu, Z. Jing, F. Ma, Y. Fang, Y. Wang, T. Chen, J. Pan, J. Du, and C.-H. Lee, "The ustc-iflytek system for sound event localization and detection of dcase2020 challenge," DCASE2020 Challenge, Tech. Rep., July 2020.

[4] Y. Cao, T. Iqbal, Q. Kong, Y. Zhong, W. Wang, and M. D. Plumbley, "Event-independent network for polyphonic sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 11–15.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 6000–6010.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[7] 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.

[8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers amp; distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 18–24 Jul 2021, pp. 10 347–10 357.

[9] A. Berg, M. O'Connor, and M. T. Cruz, "Keyword transformer: A self-attention model for keyword spotting," *arXiv preprint arXiv:2104.00769*, 2021.

[10] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[11] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020.

[12] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2020)*, November 2020.

[13] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," *arXiv preprint arXiv:2106.06999*, 2021.

[14] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, October 2019, pp. 30–34.

[15] K. Shimada, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Sound event localization and detection using activity-coupled cartesian doa vector and rd3net," DCASE2020 Challenge, Tech. Rep., July 2020.

[16] L. Mazzon, M. Yasuda, Y. Koizumi, and N. Harada, "Sound event localization and detection using foa domain spatial augmentation," DCASE2019 Challenge, Tech. Rep., June 2019.

[17] Q. Wang, J. Du, H. 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," *arXiv preprint arXiv:2101.02919*, 2021.

[18] S. Park, Y. Jeong, and T. Lee, "Metric optimization for sound event localization and detection," in *2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, 2020, pp. 1–4.

[19] T. Tanaka and T. Shinozaki, "F-measure based end-to-end optimization of neural network keyword detectors," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1456–1461.

[20] T. N. T. Nguyen, K. Watcharasupat, N. K. Nguyen, D. L. Jones, and W. S. Gan, "Dcase 2021 task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection," DCASE2021 Challenge, Tech. Rep., November 2021.