FEATURE FUSION BASED ON CROSS-FEATURE TRANSFORMER FOR SOUND EVENT LOCALIZATION AND DETECTION WITH SOURCE DISTANCE ESTIMATION

Technical Report

Jishen Tao

East China University of Science and Technology
School of Information Science and Engineering
130 Meilong Road, Shanghai, 200237, China
y80220002@mail.ecust.edu.cn

Ning Chen

East China University of Science and Technology
School of Information Science and Engineering
130 Meilong Road, Shanghai, 200237, China
chenning_750210@163.com

ABSTRACT

Since the audio of many sound events contains rich high-frequency components, the Log-Mel, which compacts the high-frequency components seriously, cannot represent the essential feature of sound event entirely. In this paper, the Log-Mel Spectrogram + Intensity Vector (LMSIV) and Magnitude Spectrogram (MS) are fused to solve this problem. First, the Cross-Feature Transformer (CFT) is performed on each feature to inspire the other feature to reinforce itself through directly attending to latent relevance revealed in the other feature to fuse the features while ensuring awareness of their interaction introduced. Then Self-Attention Transformer (SAT) is performed on the concatenation of the obtained embeddings to further prioritize contextual information in it. The experimental results show that our proposed system outperform the baseline system on the development dataset of DCASE 2024 task3.

Index Terms— sound event localization and detection, feature fusion, cross-feature attention, transformer

1. INTRODUCTION

Sound Event Localization and Detection (SELD) is the task of identifying the different types of sound events that occur in audio and locating where they are active [1]. Specifically, it consists of two subtasks: Sound Event Detection (SED) and Direction-Of-Arrival (DOA) estimation [1, 2]. SELD has a wide range of applications in a variety of fields, such as smart home, security and surveillance, and voice assistants [3, 4]. Through the use of microphone arrays and sophisticated algorithms, SELD can provide a detailed understanding of sound event in the environment, achieving smarter and more accurate responses to smart devices[5]. In the DCASE2023 Challenge task3 added estimating the distance of sound to the task. The task currently consists of three subtasks namely DOA, SED and source distance estimation.

In the early stage, the two subtasks of SELD were treated separately, and different models are designed for each task. The classic methods for SED task were based on Gaussian Mixture Models (GMM) [6], Hidden Markov Models (HMM) [7, 8], Support Vector Machines (SVMs) [9], and Non-negative Matrix Factorization (NMF) [10]. While for DOA task, the parametric methods were based on Time-Difference-Of-Arrival (TDOA) [11], the Steered-Response-Power (SRP) [12], Multiple Signal Classification (MUSIC) [13], and the Estimation of Signal Parameters via Rotational Invariance Technique (ESPRIT) [14].

To enhance the performance of one task with the information of the other task, and meet the requirements of real application scenarios, the SELD models were constructed to jointly predict the SED and DOA. According to the joint strategy, SELD models can be classified into three types: i) The SELD models with two parallel branches, each of which achieves SED or DOA task. In [15], an end-to-end Convolutional Recurrent Neural Network (CRNN) with two parallel branches was constructed for SELD. Each branch achieves one task. ii) The SELD models with two parallel and one parameter sharing branch. For example, in [16], besides the two parallel branches for SED and DOA, another branch, which shares the information between SED branch and DOA branch to achieve event activity detection (EAD), is introduced to enhance both tasks, especially in polyphonic sound event localization and detection scenario. In [17], the hard parameter sharing strategy is replaced with a soft parameter sharing to enhance the SED and DOA performance further. iii) The SELD based on Activity Coupled Cartesian Direction Of Arrival (ACCDOA) and Multi-ACCDOA. Considering that the SELD models based on two or more branches have to balance the two objections during optimization, which increases the system complexity and network size as well, ACCDOA representation, which assigns a sound event activity to the length of a corresponding cartesian DOA vector, is proposed [18, 19]. It makes it possible to solve SELD task with a single target, avoid the necessity of balancing the objections, and reduce the model size, at the same time. So, the proposed model is based on Multi-ACCDOA.

The feature adopted will affect the performance of the SELD greatly. Various feature extraction methods have been proposed for SELD according to the formats of the input audio, First-Order Ambisonics (FOA) or multichannel Microphone Array (MIC) [2]. In this paper, we focus on the feature extraction for FOA audio. At present, Log-Mel Spectrogram and Intensity Vector (LMSIV) is one of the most successful features designed for SELD task. However, the Log-Mel Spectrogram will compact the high-frequency component to a large extent, while many sound events, such as piano, scream, and bell, have rich high-frequency components. So LMSIV feature is not good enough to describe the properties of the sound event entirely and precisely. Considering that Magnitude Spectrogram (MS) describes the content in all frequency bands of the sound equally, it is fused with LMSIV in this paper for SELD task. Specifically, first, the Cross-Feature Transformer (CFT) is performed on each feature (LMSIV or MS) to inspire the other feature to reinforce itself through directly attending to latent relevance revealed in the other feature to fuse the feature while ensuring awareness

of their interaction introduced. In addition, Self-Attention Transformer (SAT) is applied on the concatenation of the reinforced features to further prioritize contextual information in it to enhance the performance of SELD task further. Experimental results on the official dataset of DCASE 2024 task3 Challenge demonstrate that the proposed model outperforms the baseline system, and the introduction of both spectrogram and the fusion strategy contributes to the performance enhancement of the proposed model.

2. METHOD

2.1. Feature Extraction

Both LMSIV and MS are adopted to represent the input audio signal.

• **MS**. Considering that many sound events contain rich high-frequency components, MS is adopted to represent the input audio. Specifically, Short-Time Fourier Transform (STFT) is performed on the audio signal received by a microphone array of arbitrary geometry in a real sound scene to obtain $\mathbf{X}(t,f)$ (see Eq. (1)); where t and f are time and frequency indices, respectively:

$$\mathbf{X}(t,f) = \sum_{l=1}^{L} \mathbf{I}_l(t,f) \times \mathbf{H}(t,f) + \mathbf{V}(t,f) \tag{1}$$

where L is the number of sound sources; $\mathbf{H}(t,f)$ is the frequency domain steering vector; $\mathbf{I}_l(t,f)$ and $\mathbf{V}(t,f)$ are the spectrogram of l-th sound source signal and that of the noise component in the environment where the source is located. The absolute value of $\mathbf{X}(t,f)$, $|\mathbf{X}(t,f)|$, is then calculated to obtain MS feature.

• **LMSIV**. The Log-Mel spectrogram, denoted as $\mathbf{Y}_{mel}(t,k)$, is calculated from $\mathbf{X}(t,f)$ with Eq. (2), where k is the mel index:

$$\mathbf{Y}_{mel}(t,k) = \log\left(\left|\mathbf{X}(t,f)\right|^2 \times \mathbf{W}_{mel}(f,k)\right) \tag{2}$$

where $\mathbf{W}_{mel}$ is the mel filter. The audio in FOA format has four channels, which contain the omnidirectional components of the audio, the distribution of the audio in the left and right directions, the distribution of the audio in the forward and backward directions, and the distribution of the audio in the vertical direction, respectively. The STFT of these four channels are denoted as $\mathbf{X}_W(t,f)$, $\mathbf{X}_X(t,f)$, $\mathbf{X}_Y(t,f)$ and $\mathbf{X}_Z(t,f)$, respectively. Then, the Intensity Vector (IV) denoted as $\mathbf{Y}_{IV}(t,k)$, which is calculated based on the differences in the intensities among different channels (see Eqs. (3–4)), provides information of the relative intensity distribution of the sound in each direction [2, 20, 21].

$$\mathbf{Y}_{IV}(t,k) = \frac{\mathbf{Y}(t,f)}{\|\mathbf{Y}(t,f)\|_2} \times \mathbf{W}_{mel}(f,k) \tag{3}$$

$$\mathbf{Y}(t,f) = -\frac{1}{\rho \cdot c} \mathrm{Re}\left(\mathbf{X}_W^*(t,f) \times \begin{bmatrix}\mathbf{X}_X(t,f)\\ \mathbf{X}_Y(t,f)\\ \mathbf{X}_Z(t,f)\end{bmatrix}\right)^\intercal \tag{4}$$

where $\|\cdot\|_2$is L_2 norm; ρ and c are the sound density and velocity, respectively; $^\intercal$ represents the transpose; $\mathrm{Re}(\cdot)$ indicates the real part calculator; and * denotes conjugate. Then, the LMSIV is obtained by concatenating of $\mathbf{Y}_{mel}(t,k)$ and $\mathbf{Y}_{IV}(t,k)$.

2.2. Network Architecture

The architecture of the proposed SELD model is shown in Figure 1. It is composed of three parts, feature compactness, feature fusion, and prediction.

• **Feature Compactness**. The input of this model are the MS and LMSIV of the input audio. To extract the local shift-invariant embeddings in each input feature and reduce their dimensions efficiently, three convolution blocks, each of which is composed of 3×3 convolutional layer, 2D BatchNorm, ReLU, 2D Max-Pooling and 2D dropout, and Tanh are performed on the input feature sequently to obtain the compacted feature of LMSIV and that of MS, which are denoted as $\mathbf{F}^1$ and $\mathbf{F}^2$, respectively.

• **Feature Fusion**. To take full advantage of the common as well as complementary properties of $\mathbf{F}^1$ and $\mathbf{F}^2$ in representing sound event related feature, the CFT and SAT are combined in the proposed model to fuse them as follows.

First, $\mathbf{F}^1$ and $\mathbf{F}^2$ are concatenated to obtain the fused representation $\mathbf{F}$.

Next, $\mathbf{F}^i$ $(i = 1, 2)$ and $\mathbf{F}$ are fed into the CFT, which includes Multi-Head Cross-Feature Attention (MHCFA), to facilitate sufficient complementor and interactions between $\mathbf{F}^1$ and $\mathbf{F}^2$. Thus, the cross-feature attention module can learn the attention score between a target compacted feature $\mathbf{F}^i$ and the fused feature $\mathbf{F}$, which will then serves to control the adaptation and reinforcement of one feature to the other in the fusion representation [22]. For $\mathbf{F}_u$ Query $\mathbf{Q}_u$, fusion Key $\mathbf{K}_u$, and Value $\mathbf{V}_u$ can be obtained with Eq. (5):

$$\begin{aligned}\mathbf{Q}_u &= \mathbf{F}_u \times \mathbf{W}_Q\\ \mathbf{K}_u &= \mathbf{F} \times \mathbf{W}_K\\ \mathbf{V}_u &= \mathbf{F} \times \mathbf{W}_V\end{aligned} \tag{5}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable weights; $u = 1, \ldots, U$; $\mathbf{F}_u$ represents the input of the u-th layer MHCFA in CFT; U is the number of the layer of MHCFA; $\mathbf{F}^i$ is adopted as $\mathbf{F}_u$ in the first layer of MHCFA. And, the attention weights of the j-th cross-feature attention head for $\mathbf{F}_u$, denoted as $\mathbf{H}_u^j$, can be obtained with Eq. (6):

$$\begin{aligned}\mathbf{H}_u^j &= \text{Attention}(\mathbf{Q}_u^j, \mathbf{K}_f^j, \mathbf{V}_f^j)\\ &= \text{Softmax}\left(\frac{\mathbf{Q}_u^j \times (\mathbf{K}_f^j)^\intercal}{\sqrt{d_k}}\right) \times \mathbf{V}_f^j\end{aligned} \tag{6}$$

where d_k is the dimension of $\mathbf{K}_f^j$; $j = 1, \ldots, m$; m is the number of attention heads. Then, the MHCFA value between $\mathbf{F}^i$ and $\mathbf{F}$ is obtained by Eq. (7):

$$\mathbf{H}_u = Concat(\mathbf{H}_u^1, ..., \mathbf{H}_u^m). \tag{7}$$

As shown in Figure 1, the CFT is constructed based on the above the MHCFT. The CFT is composed of multiple identical cross-feature attention encoder layers, each encoder layer consists of a MHCFT block and normalization layer and a position-wise feed-forward network with residual connection. The output of each cross-feature attention encoder layer serves as one of the inputs to the next encoder layer. Specifically, part of $\mathbf{F}$ related to $\mathbf{F}^i$ is transformed into $\mathbf{K}_f$ and $\mathbf{V}_f$ pairs in the CFT to compute $\mathbf{H}_u$. Then each $\mathbf{F}^i$ is merged with other features by the position-wise feed-forward layers of each CFT.

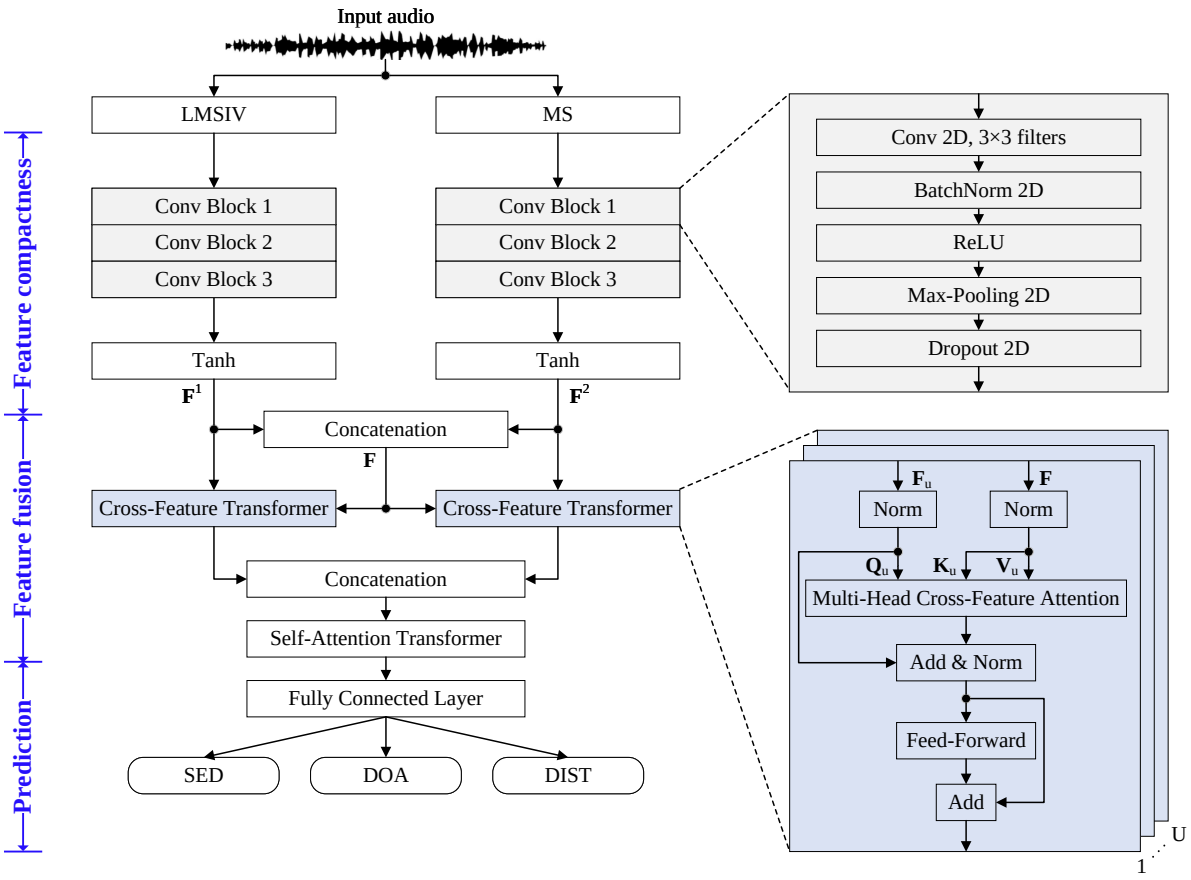

Figure 1: The architecture of the proposed model.

Finally, the output of each CFT are concatenated to obtain the fused feature. And, SAT is performed on the fused feature to integrate the global contextual information contained in it to obtain the enhanced fused feature, which is a more comprehensive global representation.

• **Prediction**. A fully connected layer, which is composed of Conv1d, Linear and Tanh, is performed on the enhanced fused feature to keep the outputs of the network in a similar range. The output of the fully connected layer is a tensor of size ($batch-size$, $label-sequence-length$, $classes*4*3$) where $batch-size$ is the number of samples included in each batch during the training process; $label-sequence-length$ is the length of the label sequence; $classes$ is the number of sound classes in the dataset. $classes*4*3$ indicates that the output can represent a maximum of three simultaneous events of the same class and gives the x, y, z coordinates of each sound event relative to the Cartesian coordinate system with the origin as the center and distance of sound event. This output format is Multi-ACCDOA [19].

The Auxiliary Duplicating Permutation Invariant Training (ADPIT) method [19] is adopted to train the model. The corresponding loss, donted as $\mathbb{L}$ is show in Eqs. (8–9):

$$\mathbb{L} = \frac{1}{C \cdot S} \sum_{c=1}^{C} \sum_{s=1}^{S} \min_{\alpha \in Perm(c,s)} l_{\alpha,c,s}^{ACCDOA} \tag{8}$$

$$l_{\alpha,c,s}^{ACCDOA} = \frac{1}{N} \sum_{n}^{N} MSE(\mathbf{P}_{\alpha,n,c,s}, \hat{\mathbf{P}}_{n,c,s}) \tag{9}$$

where S is the number of frames; C is the number of sound event classes; $Perm(c,s)$ is the set of all possible permutations; $\alpha \in Perm(c,s)$ is one permutation of frame s of class c. N is the number of tracks; $\mathbf{P}_{\alpha,n,c,s}$ is the ground truth of permutation α; $\hat{\mathbf{P}}_{n,c,s}$ is the prediction of the model for track n, class c and frame s, obtained by the proposed model.

3. EXPERIMENTAL SETUP

We conducted experiments for both Track A and Track B of Task 3, where the experimental principle of Track A is as described above, and we did not study the processing of video in Track B. For Track B, we did not study the video processing in Track B. For the model of Track B, the video processing part is provided by the baseline.

3.1. Data Aaugmentation

The size of the official training dataset is too small, so we did some data enhancement operations on the dataset. For audio data we did Audio Channel Swapping (ACS) operation. This operation enlarges the audio dataset to 8 times of the original dataset. For the related video dataset we mimic the principle of ACS and do video pixel swapping (VPS) operation for data enhancement.

3.2. Experiment Results

Table 1 shows the performance of our proposed method on the development dataset. As shown in Table 1, our proposed model outperforms the baseline model on both Track A and Track B. The difference between Model 1 and Model 2 is whether the same fully connected layer is used for the prediction of SED, DOA and DIST at the end of the model, where Model 2 separates the prediction of DIST from DOA and SED.

Table 1: The performance of our system for dev-test set.

System	Track	$F_{\leq 20^\circ}$	$DOAE_{CD}$	RDE_{CD}
Baseline	Track A	13.1%	36.9°	0.33
Baseline	Track B	11.3%	38.4°	0.36
Model 1	Track A	19.2%	22.9°	0.32
Model 1	Track B	16.2%	26.2°	0.41
Model 2	Track B	17.9%	24.2°	0.38

4. CONCLUSIONS

In this study, CFT-based fusion strategy is introduced to fuse LMSIV and MS to take advantage of common and complementary properties contained in them to enhance the SELD performance. In addition, SAT is combined to further prioritize the contextual information contained in the fused feature. Experimental results on the official dataset of Task 3 of the DCASE 2024 Challenge demonstrate that the proposed model outperforms the baseline, the CFT-based fusion strategy contributes to the performance enhancement greatly. And LMSIV and MS are complementary in SELD task.

5. REFERENCES

[1] Q. Wang, J. Du, Z. Nian, S. Niu, L. Chai, H. Wu, J. Pan, and C.-H. Lee, "Loss function design for dnn-based sound event localization and detection on low-resource realistic data," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[2] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "Salsa: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022.

[3] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE transactions on intelligent transportation systems*, vol. 17, no. 1, pp. 279–288, 2015.

[4] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," *IEEE International Conference on Robotics and Automation*, vol. 1, pp. 1033–1038, 2004.

[5] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[6] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on audio, speech, and music processing*, vol. 2013, no. 1, pp. 1–13, 2013.

[7] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," *European Signal Processing Conference*, pp. 1317–1321, 2011.

[8] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," *European signal processing conference*, pp. 1267–1271, 2010.

[9] K. Lopatka, J. Kotus, and A. Czyzewski, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimedia Tools and Applications*, vol. 75, pp. 10 407–10 439, 2016.

[10] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, *et al.*, "An exemplar-based nmf approach to audio event detection," *IEEE workshop on applications of signal processing to audio and acoustics*, pp. 1–4, 2013.

[11] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[12] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," *Brown University*, 2000.

[13] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[14] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 37, no. 7, pp. 984–995, 1989.

[15] 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, no. 1, pp. 34–48, 2018.

[16] Y. Cao, T. Iqbal, Q. Kong, Y. Zhong, W. Wang, and M. D. Plumbley, "Event-independent network for polyphonic sound event localization and detection," *arXiv preprint arXiv:2010.00140*, 2020.

[17] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 885–889, 2021.

[18] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 915–919, 2021.

[19] 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," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 316–320, 2022.

[20] S. Zhao, T. Saluev, and D. L. Jones, "Underdetermined direction of arrival estimation using acoustic vector sensor," *Signal Processing*, vol. 100, pp. 160–168, 2014.

[21] Y. Cao, T. Iqbal, Q. Kong, M. Galindo, W. Wang, and M. D. Plumbley, "Two-stage sound event localization and detection using intensity vector and generalized cross-correlation," *Detection Classification Acoustic Scenes Events (DCASE) Challange*, 2019.

[22] R. Wang, W. Jo, D. Zhao, W. Wang, A. Gupte, B. Yang, G. Chen, and B.-C. Min, "Husformer: A multi-modal transformer for multi-modal human state recognition," *IEEE Transactions on Cognitive and Developmental Systems*, 2024.