A Lightweight Temporal Attention Module for Frequency Dynamic Sound Event Detection

Yuliang Zhang, Defeng (David) Huang, Roberto Togneri,

School of Electrical, Electronic and Computer Engineering, The University of Western Australia
yuliang.zhang@research.uwa.edu.au, {david.huang, roberto.togneri}@uwa.edu.au

***Abstract*—Recent advances in Sound Event Detection (SED) have leveraged frequency-dynamic convolution to address the shift-variant nature of audio representations in the frequency domain. However, most existing methods overlook the temporal importance of individual frames during early feature extraction, which is critical for accurate event boundary detection. In this paper, we propose a lightweight temporal attention module integrated into convolutional SED architectures. The module computes temporal weights by compressing the frequency axis and applying per-frame attention using one of three strategies: MLP (frame-wise), Conv1D (local context), and MultiHead Attention (global context). These weights are injected either before or after the convolutional operation to enhance time-sensitive representations. Through comprehensive ablation experiments on the DCASE2021 Task4 dataset, we show that introducing temporal attention, with only about a 1% increase in parameters, consistently improves model performance. Specifically, averaged over 10 independent runs, the proposed temporal attention module increases PSDS1 from 0.4241 to 0.4383 on FDY-CRNN, from 0.4327 to 0.4395 on DFD-CRNN, and from 0.4376 to 0.4452 on MDFD-CRNN. These improvements demonstrate that even lightweight attention mechanisms targeting temporal saliency can significantly enhance the event boundary modeling capabilities of frequency-dynamic SED systems.**

***Index Terms*—Sound event detection, frequency dynamic convolution, temporal dynamic attention.**

1. INTRODUCTION

Sound Event Detection (SED) involves identifying and temporally localizing acoustic events within audio recordings, serving as a critical component for various applications such as audio surveillance, urban sound monitoring, and multimedia indexing [1]–[6]. Deep neural networks, particularly convolutional recurrent neural networks (CRNNs), have emerged as the predominant approach in the SED domain due to their powerful capability to simultaneously capture spatial (frequency) patterns and temporal (time) dependencies from audio spectrogram representations [7] [8]. Despite these advances, accurately identifying the temporal boundaries of audio events continues to be challenging, posing significant limitations to event localization precision and overall detection performance.

Recent research in this area has largely focused on enhancing the frequency-adaptive properties of convolutional neural networks, specifically addressing the shift-variant nature of audio signals along the frequency domain. For example, Frequency Dynamic Convolution (FDY-CRNN) dynamically generates convolutional kernels customized for different frequency bands, substantially improving feature extraction and event classification accuracy [9]. Extensions of this approach, such as Dilated Frequency Dynamic Convolution (DFD-CRNN) [10] and Multi-Dilated Frequency Dynamic Convolution (MDFD-CRNN) [11], have introduced convolution kernels with varying dilation rates. These methods effectively expand receptive fields and enhance spectral coverage, thereby further increasing robustness. However, a significant limitation remains: these frequency-focused techniques implicitly assume uniform temporal importance across all frames, neglecting the distinct temporal significance of individual frames, especially those crucial to precise event boundary detection.

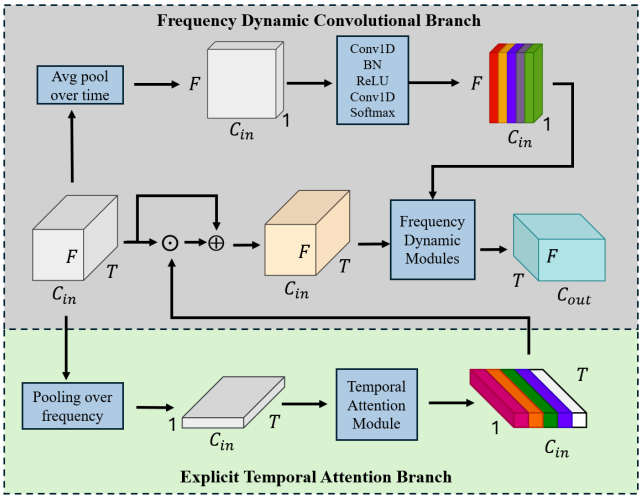

Fig. 1: The proposed architecture of frequency dynamic convolution with temporal attention branch integrated before convolution.

While [9] explored a *Temporal Dynamic Convolution* (TDY) variant, they found it offered little standalone benefit, attributing this to the bi-directional GRU layer in CRNN already modeling sequential dependencies. Their focus was limited to either frequency or temporal attention. We believe that simultaneously addressing both in early convolutional feature extraction will significantly boost Sound Event Detection (SED) performance.

To overcome this limitation, this paper introduces an explicit, lightweight temporal attention module designed to be seamlessly integrated into existing frequency dynamic convolution frameworks, as illustrated in Fig. 1. The proposed module explicitly emphasizes informative temporal segments by first compressing the frequency dimension through pooling operations, followed by applying a temporal attention mechanism that assigns distinct attention weights to individual time frames. To comprehensively explore the temporal context, we evaluate three attention strategies: a Multi-Layer Perceptron (MLP) focusing on frame-level weighting without context, a one-dimensional convolutional network (Conv1D) capturing local temporal relationships, and Multi-Head Attention mechanisms that model global temporal dependencies. Importantly, our architectural design clearly separates the temporal attention branch from the frequency dynamic convolution branch, effectively mitigating potential interference between frequency-adaptive and temporal weighting operations.

Comprehensive experiments conducted on the DCASE2021 Task4 dataset validate the effectiveness and generalizability of our proposed approach. Averaged over ten independent experimental runs, the integration of the proposed temporal attention module consistently results in substantial performance improvements across multiple frequency dynamic convolution architectures with a minimum in-

crease in parameters. Specifically, we observe absolute PSDS1 score improvements from 0.4241 to 0.4383 for FDY-CRNN, from 0.4327 to 0.4395 for DFD-CRNN, and from 0.4376 to 0.4452 for MDFD-CRNN. These significant enhancements confirm that explicitly incorporating temporal saliency into frequency-dynamic convolutional architectures markedly improves the precision of sound event boundary localization without introducing substantial computational complexity.

2. METHODS

Our proposed framework integrates an explicit temporal attention mechanism into frequency dynamic convolution (FDY-Conv) architectures, enhancing temporal feature extraction while retaining frequency-adaptive capabilities. The overall approach consists of two main components: (1) Integration with frequency dynamic convolution, and (2) computation of temporal attention weights.

2.1. Integration with Frequency Dynamic Convolution

We explore two different integration strategies for incorporating temporal attention: before and after the frequency dynamic convolution layer. These are shown in Fig. 2 and can be described as follows:

Before Frequency-Dynamic Convolution: In this strategy, temporal attention weights are applied directly to the input features prior to frequency-dynamic convolution. Given the input feature tensor $X \in \mathbb{R}^{B\times C_{in}\times T\times F}$, the attention weight tensor $A \in \mathbb{R}^{B\times C_{in}\times T\times 1}$ is computed and then element-wise multiplied with the input in a residual manner:

$$X' = X \odot (1 + A), \tag{1}$$

where B, C_{in}, T, and F represent the batch size, input channels, time dimension, and frequency dimension, respectively. The symbol $\odot$ signifies element-wise multiplication, with broadcasting applied along the frequency dimension. This approach aims to enhance the temporal saliency of input features before subsequent frequency-specific convolution operations.

After Frequency Dynamic Convolution. In contrast, this strategy applies temporal attention weights after frequency dynamic convolution, emphasizing frames based on convolutional output features. Specifically, given convolutional output features $Y \in \mathbb{R}^{B\times C_{out}\times T\times F}$, attention weights $A \in \mathbb{R}^{B\times C_{out}\times T\times 1}$ are computed and applied:

$$Y' = Y \odot (1 + A), \tag{2}$$

highlighting important temporal frames after frequency adaptation. Here, C_{out} denotes the output channels of convolution.

2.2. Computation of Temporal Attention Weights

Our temporal attention module, illustrated in the lower (temporal) branch of Fig. 1, processes information in two key stages: first, frequency-axis pooling, and then temporal attention calculation.

Pooling over Frequency: To compress the frequency dimension, we experiment with three simple pooling methods: mean pooling, max pooling, and the combined (mean+max) pooling strategy. Formally, given input features $X \in \mathbb{R}^{B\times C\times T\times F}$, pooling across the frequency dimension yields $Z \in \mathbb{R}^{B\times C\times T}$ as:

$$Z_{mean} = \frac{1}{F}\sum_{f=1}^{F} X_{:,:,:,f}, \tag{3}$$

$$Z_{max} = \max_{f\in[1,F]} X_{:,:,:,f}, \tag{4}$$

$$Z_{mean+max} = Z_{mean} + Z_{max}. \tag{5}$$

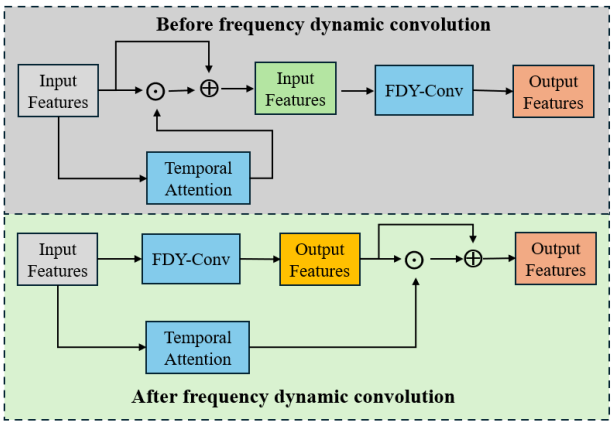

Fig. 2: Two integration strategies for the temporal attention module: before and after frequency dynamic convolution.

These pooling methods provide compact temporal representations for subsequent attention calculations.

Temporal Attention Module: We evaluate three different attention strategies: MLP, Conv1D, and Multi-Head Attention.

(1) MLP-based Attention: The simplest method applies a frame-wise multilayer perceptron (MLP) independently on each temporal frame:

$$A = \sigma(\text{MLP}(Z)), \quad A \in \mathbb{R}^{B\times C\times T}, \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid activation, ensuring attention values within (0,1). The MLP itself consists of two feed-forward layers with a ReLU activation function positioned between them.

(2) Conv1D-based Attention: This strategy leverages 1D convolutions to capture local temporal contexts by convolving across neighboring frames. Our current configuration for the Conv1D-based attention module consists of a sequence of operations: an initial 1D convolution layer $Conv1D_{\text{in}}$, followed by Batch Normalization (BN), a ReLU activation function, and finally another 1D convolution layer $Conv1D_{\text{out}}$.

The overall attention mechanism can be expressed as:

$$A = \sigma(\text{Conv1D}_{\text{out}}(\text{ReLU}(\text{BN}(\text{Conv1D}_{\text{in}}(Z))))), \quad A \in \mathbb{R}^{B\times C\times T} \tag{7}$$

This method effectively exploits local temporal continuity, thereby emphasizing temporally coherent segments within the feature sequence.

(3) Multi-Head Attention: This global strategy employs self-attention across all temporal frames, explicitly modeling global temporal relationships. With positional encoding P, the attention computation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{8}$$

where Q, K, V are query, key, and value matrices obtained from linear transformations of $(Z + P)$, and d_k is the dimensionality of the keys. The final attention weight A is derived by aggregating outputs across multiple attention heads:

$$A = \sigma\left(W_o\left[\text{head}_1; \text{head}_2; ...; \text{head}_h\right]\right), \tag{9}$$

where W_o is the final linear projection, and each head independently computes self-attention as defined above. This method effectively

captures global temporal dependencies and enhances model sensitivity to event boundaries.

3. EXPERIMENTAL SETUPS

3.1. Implementation Details

The experiments are conducted on the *Domestic Environment Sound Event Detection* (DESED) dataset [12]. This dataset comprises synthetic strongly labeled data, real weakly labeled data, and real unlabeled data. All audio recordings are 10 seconds long and sampled at 16 kHz.

For audio feature extraction, log-mel spectrograms are generated. This process involves a Short-Time Fourier Transform (STFT) with an FFT size of 2048, a hop length of 256, and a Hamming window. Subsequently, a mel filterbank with 128 mel bins is applied to the STFT magnitude to produce the final spectrogram representation.

Data augmentation methods employed in this work include frame shifting [12], mixup [13], time masking [14], and FilterAugment [15]. We apply a 7-frame median filter as a post-processing step to all classes. While class-specific filtering could optimize results further, we use a fixed-length filter across all classes to ensure a fair, post-processing-minimized comparison among models.

Our baseline model is the FDY-CRNN, which consists of seven convolutional layers. The first layer utilizes conventional 2D convolution, while the remaining six layers employ Frequency Dynamic Convolution (FDY conv). Other training parameters align with those of the original FDY-conv or its variants, with the key distinction being the addition of our proposed explicit temporal attention branch. Specifically, for the MLP-based and Conv1D-based temporal attention variants, the hidden dimensionality of the first layer is set to $\frac{1}{4}C_{\text{in}}$. In the Multi-Head Attention configuration, we use two heads with an attention dimension of 32 and a dropout rate of 0.1.

Models are trained for up to 200 epochs with a batch size of 48. Experiments for FDY-CRNN and DFD-CRNN are conducted on a single NVIDIA P100 GPU, whereas MDFD-CRNN experiments are trained on an NVIDIA V100 GPU.

3.2. Evaluation Metrics

To evaluate SED performance, the polyphonic sound detection score (PSDS) was used [16]. For DCASE Challenges 2021-2023 Task 4, two types of PSDS were utilized [17]. PSDS1 emphasizes the accuracy of event boundaries, rewarding systems that produce precise onset and offset timestamps. In contrast, PSDS2 is more tolerant to minor temporal deviations and focuses on reducing cross-triggering errors.

In our experiments, the **sum of PSDS1 and PSDS2** serves as the primary optimization objective. To ensure statistical robustness, each model configuration is trained independently ten times. We report the average PSDS1 and PSDS2 scores, alongside the intersection-based F1 score (IN-F1). Adhering to the mean teacher framework established in the DCASE challenge baseline [18] [19], we employ a teacher-student training strategy. All reported results are derived from the teacher model's predictions, providing a consistent and fair basis for comparing different model architectures.

4. RESULTS AND DISCUSSION

4.1. Overall Results

To evaluate the effectiveness of the proposed temporal attention module, comprehensive experiments were conducted on the DESED dataset. The results obtained by integrating different temporal attention strategies (MLP, Conv1D, Multi-Head Attention) and pooling methods (mean, max, mean+max) are summarized in Table 1, using the FDY-CRNN baseline model for comparison.

Table 1: Comparison of temporal attention integration strategies, attention types, and pooling methods on FDY-CRNN.

Integration	AttnType	Pooling	PSDS1	PSDS2	IN-F1
Baseline	–	–	0.4241	0.6516	0.7263
Before	Conv1D	max	0.4213	0.6413	0.7346
	Conv1D	mean	**0.4360**	**0.6635**	**0.7475**
	Conv1D	mean+max	0.4286	0.6551	0.7397
	MLP	max	0.4303	0.6459	0.7356
	MLP	mean	<u>0.4383</u>	<u>0.6604</u>	<u>0.7427</u>
	MLP	mean+max	0.4321	0.6525	0.7433
	MultiHead	max	0.4276	0.6452	0.7376
	MultiHead	mean	0.4244	0.6408	0.7399
	MultiHead	mean+max	0.4258	0.6508	0.7372
After	Conv1D	max	0.4129	0.6249	0.7197
	Conv1D	mean	0.4352	0.6493	0.7436
	Conv1D	mean+max	0.4189	0.6314	0.7252
	MLP	max	0.4245	0.6540	0.7388
	MLP	mean	0.4304	0.6569	0.7378
	MLP	mean+max	0.4267	0.6514	0.7346
	MultiHead	max	0.4285	0.6444	0.7310
	MultiHead	mean	0.4349	0.6533	0.7424
	MultiHead	mean+max	0.4366	0.6498	0.7376

Overall, the introduction of explicit temporal attention consistently improves model performance relative to the baseline FDY-CRNN model across various metrics when utilizing mean pooling strategies. The best performing configuration integrates temporal attention before frequency dynamic convolution, employing mean pooling combined with either Conv1D or MLP-based attention. Specifically, the *before*-Conv1D-mean and *before*-MLP-mean configurations achieve the highest PSDS1 and PSDS2 scores. We attribute this superiority to their balanced ability to capture local temporal contexts (Conv1D) and individual frame-level temporal emphasis (MLP). Additionally, mean pooling appears most effective, likely due to its stable and representative summarization of spectral information across frequency bins.

4.2. Ablation Study

To further investigate the contribution of each design choice, we perform detailed ablation analyses:

Integration Strategies: As evidenced in Table 1, the placement of the temporal attention module significantly impacts performance. Specifically, integrating temporal attention *before* FDY-Conv consistently outperforms the *after* placement in both the **Conv1D-based** and **MLP-based attention** variants. This observation suggests that applying frame-level or local temporal attention directly to the input features facilitates a more effective pre-selection of temporally salient frames. This, in turn, guides the subsequent frequency-dynamic convolution towards more relevant temporal segments. Conversely, applying attention after FDY-Conv might inadvertently overemphasize features already shaped by the convolution, potentially limiting the model's adaptive capacity.

A contrasting trend is observed for the **Multi-Head Attention** variant, where the *after* configuration achieves higher PSDS scores than its *before* counterpart. We hypothesize that performing global self-attention directly on raw input features tends to homogenize the time dimension, potentially smoothing out critical local energy peaks that FDY-Conv relies on for adaptive frequency filtering. Deferring this global re-weighting step until after convolution allows the model to first distill discriminative local time-frequency structures. Only then is their overall temporal prominence adjusted, effectively balancing local selectivity with broader global context.

Pooling Methods: Among pooling methods, mean pooling consistently achieves the best overall performance, surpassing both max pooling and combined mean+max pooling. While max pooling can

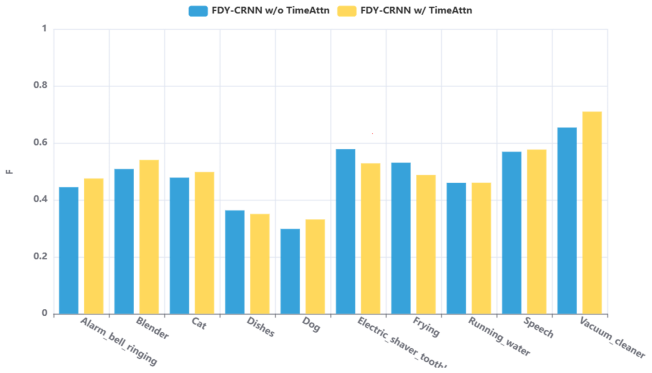

Fig. 3: Quantitative comparison of event-based F1 scores for each sound event class w/ and w/o the temporal attention branch.

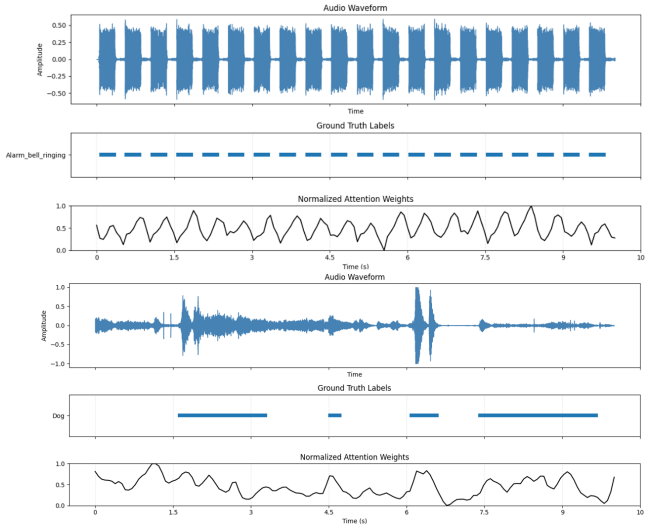

Fig. 4: Visualization of temporal attention variations for audio sample of "alarm" and "dog" event.

overly focus on peak activations, potentially introducing instability, mean pooling provides a stable frequency representation beneficial for robust temporal weighting. Interestingly, combining mean and max pooling does not yield additional performance benefits, indicating that the simplicity and robustness of mean pooling alone are sufficient.

Attention Types. Comparing attention mechanisms, Conv1D and MLP consistently outperform Multi-Head Attention. This could be explained by the simplicity and locality advantages of Conv1D and MLP, which sufficiently capture essential temporal patterns without overly complex global relationships. The Multi-Head Attention, despite theoretically stronger global context modeling, might introduce unnecessary complexity and potential overfitting given the dataset size and task complexity.

In conclusion, our detailed experiments validate that a simple, explicit temporal attention module effectively improves temporal boundary localization in sound event detection, with the optimal configuration involving integration before frequency convolution, mean pooling, and either Conv1D or MLP-based attention modules.

4.3. Analysis and Discussion

To better understand the contribution of the proposed temporal attention module, we conduct both quantitative and qualitative analyses. Class-wise event-based F1 scores, shown in Fig. 3, reveal that the temporal attention mechanism yields notable improvements for transient and short-duration events such as *alarm_bell_ringing*, *blender*, *cat*, and *dog*. These types of events are typically characterized by sharp onsets and brief durations, making precise temporal boundary modeling essential. By dynamically adjusting the importance of individual frames, our module appears to enhance the network's sensitivity to critical onset and offset regions, leading to more accurate event localization.

Table 2: Impact of temporal attention module with different attention types on various frequency dynamic convolution architectures.

Model	AttnType	Params	PSDS1	PSDS2	IN-F1
FDY-CRNN	–	11.061M	0.4241	0.6516	0.7263
	MLP	11.172M	**0.4383**	**0.6604**	**0.74277**
DFD-CRNN	–	11.061M	0.4327	**0.6624**	0.7314
	Conv1D	11.281M	**0.4395**	0.6620	**0.7432**
MDFD-CRNN	–	18.157M	0.4376	0.6504	0.7416
	MLP	18.365M	**0.4452**	**0.6613**	**0.7435**

In contrast, obvious performance degradation is observed for long-duration and relatively stationary events such as *electric_shaver_toothbrush* and *frying*. These events tend to exhibit stable and continuous spectral characteristics over time. In such cases, uniform temporal weighting may already be sufficient, and the introduction of temporal modulation might introduce unnecessary variability, potentially disrupting stable temporal features.

We further examine the effectiveness of the proposed attention mechanism. Visual inspection of the learned temporal attention weights shows that our module consistently assigns higher importance to frames near event onsets and/or offsets, especially for short-duration events such as "alarm" and "dog", as illustrated in Fig. 4. This indicates that the model successfully learns boundary-aware temporal cues, which helps enhance event localization accuracy.

4.4. Performance on FDY-Conv and Its Variants

Table 2 presents the optimal configurations for integrating our proposed temporal attention module ("before" integration with mean pooling) across different FDY-Conv variants. Clearly, the inclusion of explicit temporal attention consistently enhances performance metrics (PSDS1, PSDS2, and IN-F1) across all architectures. Specifically, FDY-CRNN achieves an absolute PSDS1 improvement of 1.42%, DFD-CRNN of 0.68%, and MDFD-CRNN of 0.76%. These notable improvements underscore the general applicability and effectiveness of temporal attention in diverse frequency-dynamic convolution frameworks.

Furthermore, the additional parameter overhead introduced by our attention module is minimal, ranging approximately from 1.00% (FDY-CRNN) to 1.15% (DFD-CRNN) and 1.15% (MDFD-CRNN). This modest increment demonstrates the computational efficiency of our attention mechanism, validating its practical utility in enhancing model performance without significantly increasing computational costs.

5. CONCLUSION

In this paper, we proposed a lightweight temporal attention module integrated with frequency dynamic convolution (FDY-Conv) architectures to improve sound event detection, especially for accurate event boundary localization. Our module explicitly emphasizes temporally salient frames through frequency-axis compression and frame-level attention weighting using MLP, Conv1D, or Multi-Head Attention mechanisms. Comprehensive evaluations on the DESED dataset demonstrate consistent and significant performance improvements across various FDY-Conv architectures (FDY-CRNN, DFD-CRNN, and MDFD-CRNN), demonstrating both effectiveness and computational efficiency.

REFERENCES

[1] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2007, pp. 21–26.

[2] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, pp. 1–46, 2016.

[3] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 1306–1309.

[4] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.

[5] J. Salamon, J. P. Bello, A. Farnsworth, M. Robbins, S. Keen, H. Klinck, and S. Kelling, "Towards the automatic classification of avian flight calls for bioacoustic monitoring," *PloS one*, vol. 11, no. 11, p. e0166866, 2016.

[6] S. Krstulović, "Audio event recognition in the smart home," *Computational Analysis of Sound Scenes and Events*, pp. 335–371, 2017.

[7] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[8] 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.

[9] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," *arXiv preprint arXiv:2203.15296*, 2022.

[10] H. Nam, S.-H. Kim, D. Min, J. Lee, and Y.-H. Park, "Diversifying and expanding frequency-adaptive convolution kernels for sound event detection," *arXiv preprint arXiv:2406.05341*, 2024.

[11] H. Nam and Y.-H. Park, "Pushing the limit of sound event detection with multi-dilated frequency dynamic convolution," *arXiv preprint arXiv:2406.13312*, 2024.

[12] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

[13] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[15] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4308–4312.

[16] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 61–65.

[17] DCASE Community, "DCASE 2021 Challenge Task 4: Sound Event Detection in Domestic Environments," https://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments, 2022, accessed: [Current Date, e.g., 2025-07-06].

[18] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

[19] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.