Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Dynamic prompting class distribution optimization for semi-supervised sound event detection*

Lijian GAO†1, Qing ZHU[1], Yaxin SHEN[1], Qirong MAO†‡1,2, Yongzhao ZHAN[1]

[1]*School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212016, China*

[2]*Jiangsu Engineering Research Center of Big Data Ubiquitous Perception and Intelligent Agricultural Applications, Zhenjiang 212016, China*

†E-mail: ljgao@ujs.edu.cn; mao_qr@ujs.edu.cn

Received Jan. 27, 2024; Revision accepted June 27, 2024; Crosschecked Feb. 21, 2025

Abstract: Semi-supervised sound event detection (SSED) tasks typically leverage a large amount of unlabeled and synthetic data to facilitate model generalization during training, reducing overfitting on a limited set of labeled data. However, the generalization training process often encounters challenges from noisy interference introduced by pseudo-labels or domain knowledge gaps. To alleviate noisy interference in class distribution learning, we propose an efficient semi-supervised class distribution learning method through dynamic prompt tuning, named prompting class distribution optimization (PADO). Specifically, when modeling real labeled data, PADO dynamically incorporates independent learnable prompt tokens to explore prior knowledge about the true distribution. Then, the prior knowledge serves as prompt information, dynamically interacting with the posterior noisy-class distribution information. In this case, PADO achieves class distribution optimization while maintaining model generalization, leading to a significant improvement in the efficiency of class distribution learning. Compared with state-of-the-art methods on the SSED datasets from DCASE 2019, 2020, and 2021 challenges, PADO achieves significant performance improvements. Furthermore, it is readily extendable to other benchmark models.

Key words: Prompt tuning; Class distribution learning; Semi-supervised learning; Sound event detection
https://doi.org/10.1631/FITEE.2400061 **CLC number:** TP391.4

1 Introduction

Sound event detection (SED) has gained significant attention due to its practical relevance in various real-world applications such as audio surveillance (Crocco et al., 2016; Park and Kim, 2020), acoustic scene understanding (Imoto et al., 2020), and human–machine interaction (Fu et al., 2019). SED tasks require recognizing the categories of events and marking the onset and offset time for each event in a mixed audio recording. This generally involves two separate sub-tasks: audio tagging and audio localization (Mesaros et al., 2021). Specifically, when given an input spectrogram, the SED model will output two-level predictions: clip-level probabilities for audio tagging and frame-level probabilities for event localization.

Traditionally, training a well-performing SED model requires an ample amount of manually labeled training data (Gao LJ et al., 2019, 2022, 2024; Li et al., 2020; Serizel et al., 2020). However, the fine-grained manual annotation at the frame level is extremely time-consuming and results in a severe

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 62176106 and U1836220), the Special Scientific Research Project of School of Emergency Management of Jiangsu University (No. KY-A-01), the Project of Faculty of Agricultural Engineering of Jiangsu University (No. NGXB20240101), the Post-graduate Research & Practice Innovation Program of Jiangsu Province (Nos. KYCX22_3668 and KYCX21_3373), and the Jiangsu Key Research and Development Plan (No. BE2020036)

ORCID: Lijian GAO, https://orcid.org/0000-0002-6458-0660; Qirong MAO https://orcid.org/0000-0002-0616-4431

shortage of annotated samples, presenting a major hurdle for SED research in the big data era. To address this issue, researchers have shifted their focus towards semi-supervised sound event detection (SSED) tasks, leveraging large-scale unlabeled and synthetic data for generalization learning to effectively mitigate overfitting on the limited labeled data. Recently, a teacher–student framework (Yan et al., 2020; Koh et al., 2021; Zheng et al., 2021b; Gao LJ et al., 2023) was established by the most popular methods in SSED commonly based on the consistency regularization assumption (Tarvainen and Valpola, 2017). In this framework, the teacher model provides pseudo-labels to guide the student model for generalization training on the unlabeled and synthetic data, while the student model exploits real labeled data concurrently for supervised training.

Despite the success of the teacher–student framework in generalization learning, noisy interference introduced by pseudo-labels or domain knowledge bias of synthetic data in class distribution learning is often under-considered, resulting in suboptimal performance in SED tasks. Therefore, one of the most challenging tasks in SSED is alleviating the noisy interference in class distribution learning. To this end, recent works have achieved performance gains through effective pseudo-labeling (PL) strategies (Chan and Chin, 2021; Koh et al., 2021), or try to transfer the domain knowledge from synthetic data domain to real domain (Zheng et al., 2021a). However, there is a trade-off between the quality and quantity of pseudo-labels, leading to a reduction in the utilization of unlabeled data. Additionally, effective domain adaptation (DA) algorithms typically require careful design to avoid overfitting to a specific domain, which often comes with increased algorithm complexity.

More recently, an advanced class distribution optimization method called prompt tuning (PT) has been proposed and widely adopted to adapt pre-trained models in downstream tasks (Brown et al., 2020). PT freezes the parameters of pre-trained models and incorporates additional learnable prompt tokens into the models when training on downstream task data. By optimizing only a small portion of the parameters (i.e., prompt tokens), PT effectively tailors class distribution information for specific downstream tasks while retaining the generalization capacity of pre-trained models. Notably, this technique has demonstrated significant achievements in natural language processing (NLP) (Gao TY et al., 2021; Gu YX et al., 2022; Singhal et al., 2023; Xu et al., 2023; Gu ZD and He, 2024), computer vision (Jia et al., 2022; Sohn et al., 2023), and other related domains (Wang et al., 2023; Murugesan et al., 2024). However, in semi-supervised learning tasks that typically exclude the involvement of pre-trained models, the application of PT to alleviate noisy-class distribution information emerges as a critical challenge demanding immediate attention.

In this paper, we propose an efficient class distribution learning method by performing dynamic PT in the mean teacher (MT) architecture (an advanced teacher–student-based semi-supervised learning framework) for SSED tasks, referred to as prompting class distribution optimization (PADO), to effectively alleviate noisy interference. Specifically, building upon the Transformer model as the baseline, PADO introduces the MT framework. During generalization training on the unlabeled and synthetic data, PADO employs class tokens in Transformers to model the noisy-class distribution information. Concurrently, when modeling real labeled data (i.e., supervised training), PADO embeds extra prompt tokens to explore the prior information from the real labeled data. The prior information learned by prompt tokens serves as prompt knowledge, dynamically interacting with class tokens to effectively optimize the noisy-class distribution information. The contributions of our work are summarized as follows:

1. We integrate PT into semi-supervised learning and introduce an advanced semi-supervised class distribution learning method, referred to as PADO. PADO leverages real labeled data to explore prior knowledge of class distribution, dynamically interacts with noisy-class distribution information learned from the unlabeled and synthetic data, and effectively alleviates noisy interference for semi-supervised learning.

2. PADO avoids directly optimizing the noisy-class distribution information modeled by class tokens, thus preventing the decline in generalization capability. Instead, it models prior knowledge as prompt information during only supervised training to dynamically guide the learning of class distribution, which greatly enhances the effectiveness of semi-supervised learning.

3. Extensive experiments on DCASE 2019, 2020, and 2021 SSED datasets show that PADO significantly outperforms the current state-of-the-art (SOTA) methods. Moreover, the significant performance improvements across various backbone models underscore the remarkable generality of PADO.

2 Related works

In this section, we will briefly review recent literature related to our work.

2.1 Feature learning and class distribution modeling for SED

The primary task in SED is to design efficient neural networks for learning acoustic features and modeling class distributions. Recently, the convolutional recurrent neural network (CRNN) is often adopted as the backbone model for SSED tasks, which employs a convolutional neural network (CNN) as the front-end network for the recurrent neural network (RNN) model. With its remarkable spatial and temporal feature modeling capabilities, CRNN has become one of the mainstream models in the SED field (Dinkel et al., 2021; Mesaros et al., 2021; Gao LJ et al., 2022).

More recently, Transformer-based models have been gradually introduced and achieved significant performance in SED tasks (Miyazaki et al., 2020b; Guan et al., 2022), leveraging a self-attention mechanism to directly explore global context information from sequential spectrograms. To jointly model local spatial and global context information, some recent works have attempted to combine a CNN and a Transformer (Kong et al., 2020; Miyazaki et al., 2020a; Wakayama and Saito, 2022; Gao LJ et al., 2023). One of the most representative works is the Conformer model (Gulati et al., 2020), which ingeniously embeds convolutional layers into each self-attention block, facilitating the local spatial information modeling. Subsequently, Conformer was introduced into SED tasks (Miyazaki et al., 2020a) and won the championship in the DCASE 2020 challenge (task 4) SSED task, serving as inspiration for subsequent research. For example, Joint-Former (Gao LJ et al., 2023), a recent work we have completed, has combined a masked auto-encoder with the Conformer model, further improving the performance of SSED.

In class distribution learning, Transformer-based models offer an explicit and more efficient approach compared to the aforementioned models (e.g., CRNN), which implicitly learn category information. In contrast, Transformer explicitly models class statistical information injecting independent learnable parameters (usually noted as class tokens) in the input spectrogram. These class tokens then interact with the spectrogram through self-attention for efficient class prediction.

2.2 Class distribution optimization for SSED

In alleviating noisy interference during class distribution learning for SSED tasks, recent efforts have been directed toward two main aspects to optimize the noised class distribution information: PL and DA. For PL, Koh et al. (2021) proposed a shift consistency training (SCT) approach with a two-stage PL strategy, using a pre-trained pseudo-label generation model for unlabeled data. Additionally, Chan and Chin (2021) applied a convolutive non-negative matrix factorization (CNMF) algorithm to annotate pseudo-labels. For DA, Zheng et al. (2021a) developed a mutual mean teacher (MMT) method that uses two different MT models to guide student training, integrating an adversarially trained domain classifier.

3 Methodology

In this section, we first define the SSED task through the basic MT framework as shown in Fig. 1, and then describe our PADO algorithm for SSED in detail (as shown in Fig. 2).

3.1 Definitions of SSED task

SED contains two sub-tasks: audio tagging and audio localization. Typically, training an SSED model involves a large amount of unlabeled data X_{u}, strongly labeled synthetic data X_{s}, and a small amount of real labeled data X_{l}. As illustrated in Fig. 1, the training set can be defined as $X = \{X_{\mathrm{u}}, X_{\mathrm{s}}, X_{\mathrm{l}}\}$, where $\boldsymbol{x} \in X$ represents spectrograms of audio signals. The real labeled data $X_{\mathrm{l}} \in X$ only contains clip-level ground truth, denoted as weakly labeled data. Typically, the basic MT framework for SSED tasks contains three components as shown in Fig. 2, including an encoder $\boldsymbol{E}_{\alpha}$ for raw input

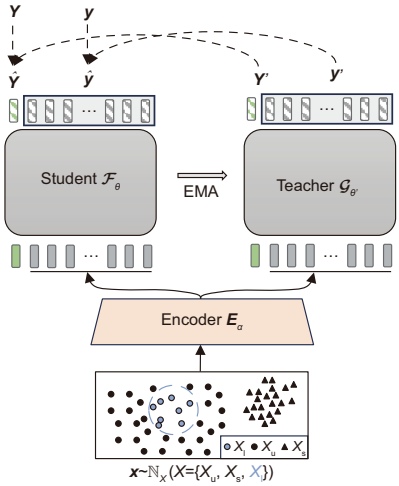


Fig. 1 Framework of the MT-based SSED. MT: mean teacher; SSED: semi-supervised sound event detection; EMA: the exponential moving average algorithm. References to color refer to the online version of this figure

modeling, a student model $\boldsymbol{\mathcal{F}}_\theta$, and a teacher model $\boldsymbol{\mathcal{G}}_{\theta'}$, where α, θ, and θ' denote the parameters corresponding to each module, respectively. The detailed descriptions of the MT-based SSED architecture are as follows:

Assume that $\boldsymbol{x} \in \mathbb{R}^{d\times T}$, where T is the length of each sequential data $\boldsymbol{x}$ and d is the dimension of $\boldsymbol{x}$. For a randomly sampled spectrogram feature $\boldsymbol{x}$ from the training set X (denoted as $\boldsymbol{x} \sim \mathbb{N}_X$), $\boldsymbol{E}_\alpha$ is required to down-sample the long-time sequential input to reduce computation costs and learn the latent feature $\boldsymbol{h}$ for $\boldsymbol{x}$ (the gray block sequence in Fig. 1), that is, $\boldsymbol{h}_{\boldsymbol{x}} = \boldsymbol{E}_\alpha(\boldsymbol{x})$. $\boldsymbol{h}_{\boldsymbol{x}} \in \mathbb{R}^{d'\times T'}$, where d' represents the dimension of $\boldsymbol{h}_{\boldsymbol{x}}$ and T' is the sequence length after down-sampling. Then the latent features will be concatenated with the independent learnable class token $\boldsymbol{h}_{\text{cls}} \in \mathbb{R}^{d'\times 1}$ (the green block in Fig. 1) and fed into both the student model $\boldsymbol{\mathcal{F}}_\theta$ and the teacher model $\boldsymbol{\mathcal{G}}_{\theta'}$ to obtain the SED predictions $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{y}})$ and $(\boldsymbol{Y}', \boldsymbol{y}')$, respectively:

$$(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{y}}) = \boldsymbol{\mathcal{F}}_\theta(\mathcal{C}[\boldsymbol{h}_{\text{cls}}, \boldsymbol{E}_\alpha(\boldsymbol{x})]), \tag{1}$$

$$(\boldsymbol{Y}', \boldsymbol{y}') = \boldsymbol{\mathcal{G}}'_\theta(\mathcal{C}[\boldsymbol{h}_{\text{cls}}, \boldsymbol{E}_\alpha(\boldsymbol{x})]), \tag{2}$$

where $\mathcal{C}[\cdot]$ represents the concatenation operator alongside the time dimension. Specifically, predictions of audio tagging $(\hat{\boldsymbol{Y}}, \boldsymbol{Y}')$ are obtained through the class token, while the results of audio localization $(\hat{\boldsymbol{y}}, \boldsymbol{y}')$ are predicted by the latent features. That is, the class distribution information is modeled in the class token, providing guidance for locating each sound event through the self-attention mechanism.

According to the consistency assumption, $\boldsymbol{\mathcal{F}}_\theta$ is trained under joint constraints $\mathcal{L}_{\text{joint}}$ as shown in Eq. (3), where $\omega(n) = \mathrm{e}^{-5\times(1-n)^2}$ represents the weight of the consistency loss. Because the teacher model cannot provide effective pseudo-labels in the early stages of training, $\omega(n)$ is initialized close to 0 (i.e., $\omega(0) = \mathrm{e}^{-5} \approx 0.007$) and gradually increased with the training epochs, where $n \in [0, 1]$ represents the proportion of training epochs. Specifically, $\mathcal{L}_{\text{joint}}$ contains a supervised loss $\mathcal{L}_{\text{sup}}$ (Eq. (4)) and a consistency loss $\mathcal{L}_{\text{con}}$ (Eq. (5)), both of which are the sum of tagging error ℓ_{tag} and localization error ℓ_{loc}, where $(\boldsymbol{Y}, \boldsymbol{y})$ are the real-labeled ground truth.

$$\begin{aligned}\mathcal{L}_{\text{joint}}(\boldsymbol{x}) =& \mathcal{L}_{\text{sup}}(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{y}}; \boldsymbol{Y}, \boldsymbol{y}) \\ &+ \omega(n)\mathcal{L}_{\text{con}}(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{y}}; \boldsymbol{Y}', \boldsymbol{y}'),\end{aligned} \tag{3}$$

$$\mathcal{L}_{\text{sup}}(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{y}}; \boldsymbol{Y}, \boldsymbol{y}) = \ell_{\text{tag}}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) + \ell_{\text{loc}}(\hat{\boldsymbol{y}}, \boldsymbol{y}), \tag{4}$$

$$\mathcal{L}_{\text{con}}(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{y}}; \boldsymbol{Y}', \boldsymbol{y}') = \ell_{\text{tag}}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}') + \ell_{\text{loc}}(\hat{\boldsymbol{y}}, \boldsymbol{y}'). \tag{5}$$

Following the MT framework setup, the teacher model is an ensemble of historical student models and updates its parameter θ' using the exponential moving average (EMA) algorithm as follows:

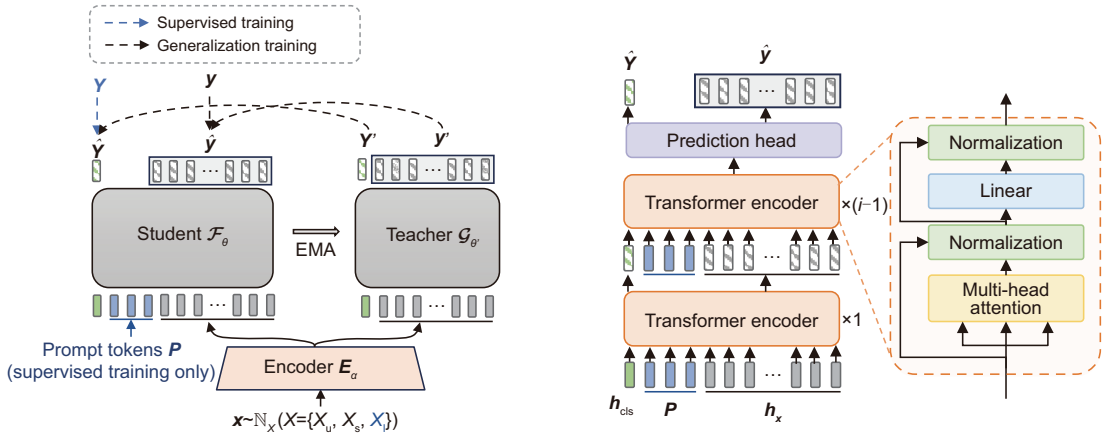
$$\theta'_t = \gamma\theta'_t + (1-\gamma)\theta_{t-1}, \tag{6}$$

where t is the current training epoch index and γ is the smoothing coefficient, typically set to 0.999.

Finally, adopting the binary cross-entropy (BCE) loss function, we calculate the aforementioned tagging and localization errors (i.e., ℓ_{tag} and ℓ_{loc}), and optimize the joint loss $\mathcal{L}_{\text{joint}}$ to complete the training of the MT-based SSED framework. However, the noisy interference from pseudo-labels and that from knowledge gaps of the synthetic data domain in the class distribution learning process are under-considered. To this end, our proposed solution PADO for SSED will be discussed below.

3.2 PADO-based SSED framework

The pipeline of our PADO-based SSED framework is given in Fig. 2a, which divides the semi-supervised learning process into two parallel stages: generalization training and supervised training (In practice, these two processes run synchronously until convergence). The generalization training stage



(a) (b)

Fig. 2 Framework of the PADO-based SSED: (a) pipeline of the PADO; (b) architecture of the student model embedded with prompt tokens. PADO: prompting class distribution optimization; SSED: semi-supervised sound event detection; EMA: the exponential moving average algorithm

is responsible for improving the model's generalization performance when modeling the unlabeled and synthetic data. In contrast, the supervised training stage on the real labeled data introduces extra prompt tokens to model the real distribution information, serving as prior knowledge to optimize the noisy-class distribution introduced during the generalization training stage.

3.2.1 Generalization training process in PADO

First, given the unlabeled and synthetic data (i.e., $\boldsymbol{x} \sim \mathbb{N}_{\{X_\mathrm{u}, X_\mathrm{s}\}}$), consistent with the basic MT framework, PADO optimizes the consistency loss $\mathcal{L}_\mathrm{con}$ as described in Eq. (5) for generalization training. In this stage, the model's generalization performance is improved, reducing the over-fitting on a small amount of real labeled data. However, the noisy interference from pseudo-labels and that from domain bias of synthetic data is still underconsidered, resulting in a noisy-class distribution. Therefore, an efficient class distribution optimization strategy is proposed by PADO through a dynamically prompted supervised training process.

3.2.2 Supervised training process in PADO

Building on the insights gained during generalization training, the supervised training process in PADO aims to alleviate the potential noisy interference in class distribution information. Here, PADO ingeniously addresses this challenge by introducing prompt tokens $\boldsymbol{P}$, a group of independent learnable parameters, during supervised training on real labeled data (i.e., $\boldsymbol{x} \sim \mathbb{N}_{X_\mathrm{l}}$). Specifically, as illustrated in Fig. 2b, prompt tokens $\boldsymbol{P}$ are embedded into the latent feature sequences $\boldsymbol{h}_{\boldsymbol{x}} = \boldsymbol{E}_\alpha(\boldsymbol{x})$ of each layer of the Transformer encoders. Note that the prompt tokens are randomly initialized at each layer. Then, efficient interactions are performed among $\boldsymbol{P}$, $\boldsymbol{h}_\mathrm{cls}$, and $\boldsymbol{h}_{\boldsymbol{x}}$ of the input spectrogram through the self-attention mechanism in Transformer encoders, jointly predicting the SED results (as shown in Eq. (7) in the case of $\boldsymbol{x} \in \{X_\mathrm{l}\}$).

$$(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{y}}) = \begin{cases} \boldsymbol{\mathcal{F}}_\theta\left(\mathcal{C}\left[\boldsymbol{h}_\mathrm{cls}, \boldsymbol{h}_{\boldsymbol{x}}\right]\right), & \text{if } \boldsymbol{x} \in \{X_\mathrm{u}, X_\mathrm{s}\}, \\ \boldsymbol{\mathcal{F}}_\theta\left(\mathcal{C}\left[\boldsymbol{h}_\mathrm{cls}, \boldsymbol{P}, \boldsymbol{h}_{\boldsymbol{x}}\right]\right), & \text{if } \boldsymbol{x} \in \{X_\mathrm{l}\}. \end{cases} \tag{7}$$

More specifically, taking the i^th $(i = 0, 1, \ldots, I)$ layer encoder $\mathrm{Att}^i\{\cdot\}$ as an example, where I is the maximum number of encoder layers, Eqs. (8)–(10) define the interaction process when embedding $\boldsymbol{P}$ through the self-attention mechanism. $\mathrm{PE}_\mathrm{sinusoidal}$ denotes the sinusoidal position encoding, where the positional indices of $\boldsymbol{h}_{\boldsymbol{x}}$ should remain consistent during both generalization training and supervised training. For example, assuming that the length of $\boldsymbol{P}$ is 3 and the sequence length of $\boldsymbol{h}_{\boldsymbol{x}}$ is 10, then during generalization training, the positional indices of $[\boldsymbol{h}_\mathrm{cls}, \boldsymbol{h}_{\boldsymbol{x}}]$ should be $\{0, 4, 5, \ldots, 13\}$, while during supervised training, the positional indices of $[\boldsymbol{h}_\mathrm{cls}, \boldsymbol{P}, \boldsymbol{h}_{\boldsymbol{x}}]$ are set to $\{0, 1, \ldots, 13\}$ after prompt tokens are embedded. Therefore, the positional index

of $\boldsymbol{h}_{\text{cls}}$ is set to 0, the positional indices of $\boldsymbol{P}$ are $\{1,2,3\}$, and the positional indices of $\boldsymbol{h}_{\boldsymbol{x}}$ remain fixed at $\{4,5,\ldots,13\}$. Additionally, $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ represent query, key, and value vectors in the attention mechanism, respectively, while $W_{\boldsymbol{Q}}$, $W_{\boldsymbol{K}}$, and $W_{\boldsymbol{V}}$ are the corresponding trainable parameters, respectively. MultiHead$\{\cdot\}$ denotes multi-head attention and σ is the softmax function. After interaction, the SED prediction results can be obtained as Eq. (11), where Pred denotes the linear prediction head. Finally, after training with $\mathcal{L}_{\text{joint}}$ (Eq. (3)), only the student model embedded with $\boldsymbol{P}$ (as shown in Fig. 2b) is required for inference.

$$\boldsymbol{h}^i = \mathcal{C}[\boldsymbol{h}^i_{\text{cls}}, \boldsymbol{P}^i, \boldsymbol{h}^i_{\boldsymbol{x}}] + \text{PE}_{\text{sinusoidal}}, \tag{8}$$

$$\begin{cases} \boldsymbol{Q}^i = \boldsymbol{h}^i W^i_{\boldsymbol{Q}}, \\ \boldsymbol{K}^i = \boldsymbol{h}^i W^i_{\boldsymbol{K}}, \\ \boldsymbol{V}^i = \boldsymbol{h}^i W^i_{\boldsymbol{V}}, \end{cases} \tag{9}$$

$$\begin{aligned} \boldsymbol{h}^i_{\text{cls}}, \_, \boldsymbol{h}^i_{\boldsymbol{x}} &= \text{Att}^i\{\boldsymbol{Q}^i, \boldsymbol{K}^i, \boldsymbol{V}^i\} \\ &= \text{MultiHead}\left\{\sigma\left(\frac{\boldsymbol{Q}^i(\boldsymbol{K}^i)^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}^i\right\}, \end{aligned} \tag{10}$$

$$\begin{cases} \hat{\boldsymbol{Y}} = \text{Pred}(\boldsymbol{h}^I_{\text{cls}}), \\ \hat{\boldsymbol{y}} = \text{Pred}(\boldsymbol{h}^I_{\boldsymbol{x}}). \end{cases} \tag{11}$$

Based on the simple yet powerful training strategy offered by PADO, we can easily train an SED model in a semi-supervised manner under the constraint of joint loss (Eq. (3)).

So far, the PADO-based SSED framework leverages a set of prompt tokens visible only during supervised training to learn prior information about real labeled data, thus assisting in optimizing the noisy-class distribution dynamically. Recall that the noisy-class distribution of sound events is learned by the class tokens, which are roughly optimized through generation learning on the unlabeled data. In this case, the prior information about real labeled data can be regarded as prompt knowledge, which effectively interacts with the class tokens through self-attention. Finally, the noisy-class distribution can be optimized under the constraint of supervised SED loss (Eq. (4)). Notably, PADO avoids directly optimizing the noisy-class distribution information modeled in class tokens, preventing the decline in generalization capability obtained by generalization training. In contrast, this indirect and dynamic guidance ultimately enhances the efficiency of semi-supervised class distribution learning, showcasing the remarkable solutions performed by PADO to navigate the challenges of SSED tasks.

4 Experimental evaluation

4.1 Experimental setup

4.1.1 Datasets

Here, we describe the widely-adopted DCASE challenge datasets in SSED, including DCASE 2019 (Turpault et al., 2019), DCASE 2020 (Turpault et al., 2020) and DCASE 2021 (Wisdom et al., 2021) (task 4) datasets, in which the training sets contain real-recorded weakly-labeled and unlabeled data with some synthetic strongly labeled data. The validation set is available to contestants for performance evaluation during the challenges, whereas the evaluation set is used for the official ranking. Except for the synthetic data, others in all three DCASE challenge datasets are sub-sets of Audio Set (Gemmeke et al., 2017), a large-scale real-recorded dataset for audio classification. Details of the datasets are listed in Table 1. Additionally, the datasets cover 10 target sound events in domestic scenarios, including "Speech," "Dog," "Alarm," "Dishes," "Frying," "Blender," "Running water," "Vacuum cleaner," "Cat," and "Electric shaver."

4.1.2 Model comparison

As illustrated in Table 2, we first choose three advanced self-attention-based models as baselines to build the PADO framework: (1) Transformer, a three-layer model as shown in Fig. 2b, (2) ConformerSED (Miyazaki et al., 2020a), the winner of the DCASE 2020 challenge, and (3) Joint-Former (Gao LJ et al., 2023) (our previous work), one of the SOTA methods in SSED. For fair comparisons, the ConformerSED (https://github.com/m-koichi/ConformerSED.git) and Joint-Former (https://github.com/mastergofujs/Joint-Former.git) models are reproduced using the open-source code provided in the respective literature. Based on the baseline models, we build our PADO-based models referred to as PADO-Transformer, PADO-Conformer and PADO-Joint-Former in Table 2. Also, we extensively compare the proposed methods with other SOTA models on DCASE 2019,

Table 1 Detailed structures of DCASE 2019, 2020, and 2021 challenge datasets

Dataset	No. of audios in the training set			No. of audios in the Val. set	No. of audios in the Eval. set	Duration (s)	Category No.
	Weakly labeled	Unlabeled	Synthetic				
DCASE 2019	1578	14 412	2045	1168	692	10	10
DCASE 2020	1578	14 412	2584	1168	692	10	10
DCASE 2021	1578	14 412	10 000	1168	692	10	10

No.: number; Val.: validation; Eval.: evaluation

2020, and 2021 challenge datasets, including guided learning (GL) (Lin et al., 2020) (the winner of DCASE 2019), SparseTrans (Guan et al., 2022), and SAN (Wakayama and Saito, 2022).

All the compared models mentioned above are built in an MT framework, without optimizing the noisy interference in class distribution learning. To this end, we consider three advanced methods for alleviating the noisy interference for comparison, as shown in Table 2, including SCT, CNMF, and MMT. Specifically, SCT adopts a two-stage shift consistency training strategy for PL. CNMF proposes a two-stage convolutive non-negative matrix factorization method for PL. In contrast, MMT focuses on DA.

4.1.3 Data preprocessing

Following ConformerSED, we down-sampled all data in the datasets to 16 kHz and extracted 64-dimensional Mel spectrogram features as the raw input. The size of the Hanning window in the short-time Fourier transform is 1024, with a hop size of 323 samples. Ultimately, we obtained 496 frames of 64-dimensional Mel spectrogram features for each sound signal. The settings of data preprocessing remained consistent throughout all experiments in this paper.

4.1.4 Metrics

We followed the settings of DCASE challenges which adopt the event-based F1 (EB-F1) score (Mesaros et al., 2016) and the polyphonic sound detection score (PSDS) (Bilen et al., 2020) as official metrics for ranking. The PSDS metric can be further classified into two sub-metrics for two testing scenarios: PSDS1 emphasizes the performance of audio localization, whereas PSDS2 focuses more on audio tagging performance. DCASE challenges use different metrics to evaluate the performance of SSED systems, for example, EB-F1 on DCASE 2019 and 2020, and PSDS1 and PSDS2 on DCASE 2021.

Table 2 Summary of the compared methods

Model	Framework	Strategy
Transformer	MT	–
ConformerSED (Miyazaki et al., 2020a)	MT	–
Joint-Former (Gao LJ et al., 2023)	MT	–
GL (Lin et al., 2020)	MT	–
SparseTrans (Guan et al., 2022)	MT	–
SAN (Wakayama and Saito, 2022)	MT	–
SCT (Koh et al., 2021)	Two-stage	PL
CNMF (Chan and Chin, 2021)	Two-stage	PL
MMT (Zheng et al., 2021a)	MT	DA
PADO-Transformer	PADO	PT
PADO-Conformer	PADO	PT
PADO-Joint-Former	PADO	PT

CNMF: convolutive non-negative matrix factorization; GL: guided learning; SAN: self-attention network; SCT: shift consistency training; MMT: mutual mean teacher; DA: domain adaptation; MT: mean teacher; PADO: prompting class distribution optimization; PL: pseudo-labeling; PT: prompt tuning; SED: sound event detection. "–" denotes that relevant data are not available

In contrast, we applied the three metrics on all the datasets to enable comprehensive comparisons in this study.

4.1.5 Training settings

All experiments were conducted on PyTorch 1.2.0. The compute unified device architecture (CUDA) we used for training was NVIDIA RTX 3090, with a batch size of 32 during training. The training was set to run 30 000 epochs. The hyper-parameter $\omega(n)$ in Eq. (3) ramps up from 1 to a maximum value of 2.0 within 6000 steps.

4.2 Ablation studies

To search for the optimal settings of prompt tokens in the proposed PADO, we designed a series of ablation experiments for comparison, using

ConformerSED as the baseline model. The most important components in PADO are the number of prompt tokens and the number of encoder layers with embedded prompt tokens.

As shown in Fig. 3, the ablation results are visualized in a scatter plot, where the horizontal axis represents PSDS1, the vertical axis represents PSDS2, and the size of the data points denotes the EB-F1 score. Each data point's coordinates (i, n) represent the number of encoder layers i with embedded prompt tokens and the number of prompt tokens n. We evaluated n in PADO through a grid search from {1, 3, 5, 7, 9} and i from {1, 2, 3}. Clearly, scatters located closer to the upper-right region and with larger data point areas indicate better overall performance.

Upon thorough analysis, it is evident that PADO achieves the best performance (red point in Fig. 3) when i=3 and n=7. Therefore, the prompt tokens are set as $3 \times 7 \times 128$ in PADO, where 128 denotes the dimension of a token. In this configuration, the EB-F1, PSDS1, and PSDS2 scores are 49.90%, 38.19%, and 58.84%, respectively. Note that when (i=0, n=0), PADO reverts to the basic MT framework, resulting in 47.90%, 35.95%, and 55.88% performance on EB-F1, PSDS1, and PSDS2, respectively. Consequently, PADO exhibits a significant improvement over the original MT framework, with performance gains of 2.00 percentage points (pp) on EB-F1, 2.24 pp on PSDS1, and 2.96 pp on PSDS2, respectively.

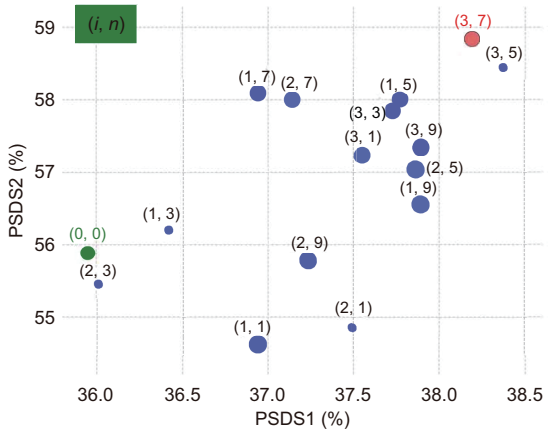


Fig. 3 Grid search for optimal settings of prompt tokens in prompting class distribution optimization (PADO). The red point represents the setting that achieved the optimal performance, and the green point is the basic mean teacher (MT) framework. References to color refer to the online version of this figure

4.3 Performance comparison with SOTA models

4.3.1 Comparisons with MT-based baseline models

To thoroughly evaluate the effectiveness and generality of the proposed PADO, we chose three advanced MT-based methods as baselines for comparison on the validation and evaluation sets of DCASE 2019, DCASE 2020, and DCASE 2021 datasets. These methods include a basic Transformer and two advanced Transformer-based SSED models (ConformerSED and Joint-Former). Performance of PADO-based SSED methods and MT-based baselines on DCASE 2019, 2020, and 2021 validation sets is shown in Table 3. Since the results of baseline models on the validation set were not publicly accessible, we reproduced them for comparison. Experimental results indicate that the proposed PADO-based semi-supervised learning framework achieves significant performance improvements for the aforementioned advanced MT-based SSED models. Specifically, compared to the Transformer, PADO-Transformer shows comprehensive improvements in all metrics across all datasets. For instance, the performance improvement ranges from 0.3 pp to 1.2 pp on the DCASE 2019 validation set, from 0.1 pp to 0.8 pp on the DCASE 2020 validation set, and from 0.1 pp to 2.1 pp on the DCASE 2021 validation set. Compared to ConformerSED, PADO-Conformer shows improvements of 0.4 pp to 1.6 pp on the DCASE 2019 validation set, 1.2 pp to 1.9 pp on the DCASE 2020 validation set, and 0.4 pp to 1.9 pp on the DCASE 2021 validation set. However, compared to Joint-Former, our PADO-Joint-Former does not outperform Joint-Former in EB-F1 on DCASE 2019 and 2021 validation sets. This is because Joint-Former requires careful adjustment of a set of hyperparameters, such as the weights of reconstruction loss. When adopting the Joint-Former within our PADO framework to reproduce its results, we do not perform an additional grid search of the hyper-parameters. Nevertheless, the proposed PADO-Joint-Former achieves significant performance improvements on all other metrics. Specifically, PADO-Joint-Former improves PSDS1 and PSDS2 by 0.7 pp and 0.6 pp on the DCASE 2019 validation set, respectively. On the

Table 3 Performance of PADO-based SSED methods and MT-based baselines on DCASE 2019, 2020, and 2021 validation sets (%)

Model	DCASE 2019 Val. set			DCASE 2020 Val. set			DCASE 2021 Val. set		
	EB-F1	PSDS1	PSDS2	EB-F1	PSDS1	PSDS2	EB-F1	PSDS1	PSDS2
Transformer	41.0	25.5	40.0	42.1	27.7	43.5	38.7	24.4	42.1
PADO-Transformer	**41.3**	**26.6**	**41.2**	**42.3**	**27.8**	**44.3**	**40.8**	**25.0**	**42.2**
ConformerSED (Miyazaki et al., 2020a)	41.4	25.9	44.3	41.7	27.5	46.6	40.3	24.7	43.4
PADO-Conformer	**41.8**	**27.5**	**45.2**	**43.6**	**29.1**	**47.8**	**40.7**	**26.0**	**45.3**
Joint-Former (Gao LJ et al., 2023)	**42.9**	27.2	43.6	43.2	27.8	44.6	**42.1**	26.9	45.8
PADO-Joint-Former	42.3	**27.9**	**44.2**	**44.4**	**29.7**	**48.0**	41.9	**27.7**	**46.6**

EB-F1: event-based F1; Val.: validation; MT: mean teacher; PADO: prompting class distribution optimization; PSDS: polyphonic sound detection score; SSED: semi-supervised sound event detection. The methods underlined indicate reproduced results. The better performing results between our methods and the reproduced related SOTA methods are in bold, respectively

DCASE 2020 validation set, all metrics are improved by 1.2 pp to 3.4 pp, and on the DCASE 2021 validation set, PSDS1 and PSDS2 are improved by 0.8 pp and 0.8 pp, respectively.

4.3.2 Comparisons with MT-based SOTA models

Here, we compare our PADO-based models with the SOTA models on DCASE 2019, 2020, and 2021 sets, including GL, ConformerSED, SparseTrans, Joint-Former, and SAN. Since the SOTA models reported their experimental results only on the public DCASE challenge evaluation sets, we compare the evaluation performance on the evaluation sets with the SOTA models, as shown in Table 4. As a result, our PADO-based methods achieve remarkable performance on all metrics in SSED tasks, outperforming the SOTA models. Notably, the PADO-Joint-Former reaches a new SOTA performance in SSED tasks on all three benchmark datasets.

4.3.3 Comparisons with advanced optimization methods

To comprehensively evaluate the effectiveness of the proposed PADO, we compare it with three advanced methods aimed at optimizing class distribution for SSED tasks, namely SCT, CNMF, and MMF, as shown in Table 4. The results show that PADO outperforms these advanced methods in class distribution optimization for SSED by 0.5 pp to 5.8 pp on EB-F1 on the DCASE 2020 evaluation set. Notably, PADO achieves dynamic class distribution optimization with only a lightweight set of parameters, i.e., prompt tokens. In contrast, the aforementioned advanced methods require the assistance of additional models or complex training strategies, such as two-stage pre-training in SCT and CNMF. The experimental results demonstrate the superior performance of PADO.

4.4 Qualitative visualization of localization

Finally, we visualize and compare the audio localization performance on the DCASE 2020 evaluation set to discuss the significance of prompt tokens in PADO. Specifically, we use the trained PADO-Joint-Former to visualize the localization results of two exemplar sound recordings. Then, we remove prompt tokens from our model to observe the difference in localization performance. The localization performance is visualized in Figs. 4 and 5, which consists of three rows of subplots, illustrating the visualization of ground truth, localization results after removing prompt tokens (denoted as PADO w/o prompts), and localization results with prompt tokens retained (denoted as PADO), respectively. The horizontal axis represents the time, and the vertical axis represents the event category, denoted as C1 to C10 for {"Alarm," "Blender," "Cat," "Dishes," "Dog," "Electric shaver," "Fry," "Running water," "Speech," "Vacuum cleaner"}. Clearly, the PADO-Joint-Former with retained prompt tokens significantly improves localization performance.

Specifically, Fig. 4 illustrates a sound signal with the true label "Cat," where the PADO-Joint-Former with retained prompt tokens exhibits excellent localization performance (as shown in the

Table 4 Performance of PADO-based SSED methods and SOTA methods on DCASE 2019, 2020, and 2021 evaluation sets (%)

Model	DCASE 2019 Eval. set			DCASE 2020 Eval. set			DCASE 2021 Eval. set		
	EB-F1	PSDS1	PSDS2	EB-F1	PSDS1	PSDS2	EB-F1	PSDS1	PSDS2
GL (Lin et al., 2020)	42.7	–	–	–	–	–	–	–	–
ConformerSED (Miyazaki et al., 2020a)	–	–	–	46.0	–	–	–	–	–
SparseTrans (Guan et al., 2022)	–	–	–	47.6	–	–	–	–	–
Joint-Former (Gao LJ et al., 2023)	51.3	–	–	49.5	–	–	–	33.9	55.1
SAN (Wakayama and Saito, 2022)	–	–	–	–	–	–	–	29.2	55.0
SCT (Koh et al., 2021)	–	–	–	45.1	–	–	–	–	–
CNMF (Chan and Chin, 2021)	–	–	–	46.3	–	–	–	–	–
MMT (Zheng et al., 2021a)	–	–	–	49.4	–	–	–	–	–
Transformer	46.6	35.3	53.1	46.4	38.1	55.6	42.4	32.9	52.1
PADO-Transformer	**47.9**	**36.9**	**54.0**	**50.9**	**39.6**	**57.8**	**44.9**	**33.2**	**54.0**
ConformerSED (Miyazaki et al., 2020a)	47.9	36.0	55.9	46.7	36.8	58.2	42.9	32.9	56.2
PADO-Conformer	**49.9**	**38.2**	**58.8**	**49.9**	**40.1**	**61.6**	**44.2**	**33.8**	**56.4**
Joint-Former (Gao LJ et al., 2023)	50.9	40.0	60.1	50.7	39.3	59.3	44.5	34.4	56.9
PADO-Joint-Former	**51.4**	**42.1**	**61.2**	**50.9**	**41.8**	**61.9**	**46.7**	**36.9**	**57.0**

CNMF: convolutive non-negative matrix factorization; Eval.: evaluation; EB-F1: event-based F1; PADO: prompting class distribution optimization; PSDS: polyphonic sound detection score; SOTA: state-of-the-art; SSED: semi-supervised sound event detection. "–" indicates that the corresponding metrics are not reported. The methods underlined indicate reproduced results. The better performing results between our methods and the reproduced related SOTA methods are in bold, respectively

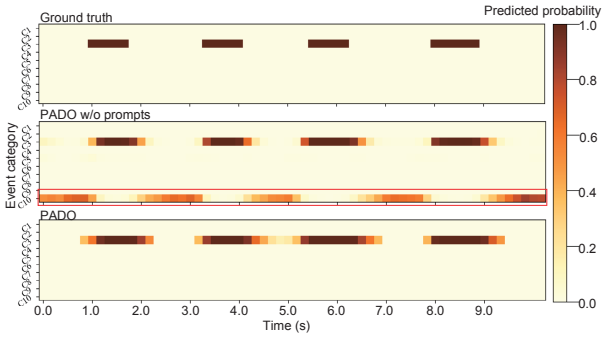


Fig. 4 Visualization of event localization for a test sample (1LKP1ZyHgVg_0_10.wav) from DCASE 2020 evaluation set, where the sound event "Cat" is active

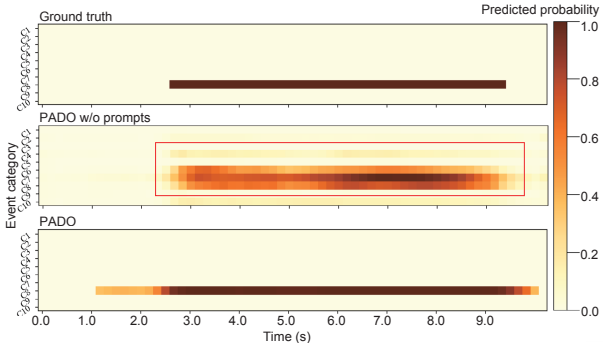


Fig. 5 Visualization of event localization for a test sample (dfRtayqQAls_14_24.wav) from DCASE 2020 evaluation set, where the sound event "Running water" is active

third-row subplot) with minimal false positive predictions. However, when removing prompt tokens from PADO, the audio localization performance significantly declines, as indicated by the red box in the second-row subplot, where the model incorrectly identifies the sound event as a "Vacuum cleaner." In Fig. 5, there is confusion among "Running water," "Speech," and "Fry" for PADO without prompt tokens, leading to inaccurate classification of the specific sound events.

In summary, prompt tokens in PADO play a

pivotal role in precisely identifying and localizing specific sound events. This is due to their capacity to model the real class distribution information in semi-supervised learning, thereby aiding in the optimization of noisy-class distributions for SSED tasks (as illustrated in the second subplot of Fig. 5).

5 Conclusions

Addressing the challenges of noisy interference from pseudo-labels of unlabeled data and domain knowledge gaps in SSED tasks, this paper proposes the PADO method for optimizing class distribution learning through dynamic prompt tuning. PADO leverages a set of independent prompt tokens to model the prior information of the true distribution, aiding in the optimization of noisy-class distributions. Experimental results on prominent datasets (i.e., DCASE 2019, DCASE 2020, and DCASE 2021) demonstrate the effectiveness and generality of PADO.

Contributors

Lijian GAO designed the research. Qing ZHU and Yaxin SHEN improved the experiments. Lijian GAO drafted the paper. Yongzhao ZHAN helped organize the paper. Lijian GAO and Qirong MAO revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are openly available in several public repositories: DCASE 2019 at https://github.com/turpaultn/DCASE2019_task4/, DCASE 2020 at https://github.com/turpaultn/DESED, and DCASE 2021 at https://github.com/DCASE-REPO/DESED_task/tree/master/recipes/dcase2021_task4_baseline.

References

Bilen Ç, Ferroni G, Tuveri F, et al., 2020. A framework for the robust evaluation of sound event detection. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.61-65. https://doi.org/10.1109/ICASSP40776.2020.9052995

Brown TB, Mann B, Ryder N, et al., 2020. Language models are few-shot learners. Proc 34th Int Conf on Neural Information Processing Systems, Article 159.

Chan TK, Chin CS, 2021. Detecting sound events using convolutional macaron net with pseudo strong labels. Proc IEEE 23rd Int Workshop on Multimedia Signal Processing, p.1-6. https://doi.org/10.1109/MMSP53017.2021.9733668

Crocco M, Cristani M, Trucco A, et al., 2016. Audio surveillance: a systematic review. *ACM Comput Surv*, 48(4):52. https://doi.org/10.1145/2871183

Dinkel H, Wu MY, Yu K, 2021. Towards duration robust weakly supervised sound event detection. *IEEE/ACM Trans Audio Speech Lang Process*, 29:887-900. https://doi.org/10.1109/TASLP.2021.3054313

Fu YW, Xu KL, Mi HB, et al., 2019. A mobile application for sound event detection. Proc 28th Int Joint Conf on Artificial Intelligence, p.1-7. https://doi.org/10.24963/ijcai.2019/941

Gao LJ, Mao QR, Dong M, et al., 2019. On learning disentangled representation for acoustic event detection. Proc 27th ACM Int Conf on Multimedia, p.2006-2014. https://doi.org/10.1145/3343031.3351086

Gao LJ, Zhou L, Mao QR, et al., 2022. Adaptive hierarchical pooling for weakly-supervised sound event detection. Proc 30th ACM Int Conf on Multimedia, p.1779-1787. https://doi.org/10.1145/3503161.3548097

Gao LJ, Mao QR, Dong M, 2023. Joint-Former: jointly regularized and locally down-sampled Conformer for semi-supervised sound event detection. Proc 24th Annual Conf of the Int Speech Communication Association, p.2753-2757. https://doi.org/10.21437/Interspeech.2023-344

Gao LJ, Mao QR, Dong M, 2024. On local temporal embedding for semi-supervised sound event detection. *IEEE/ACM Trans Audio Speech Lang Process*, 32:1687-1698. https://doi.org/10.1109/TASLP.2024.3369529

Gao TY, Fisch A, Chen DQ, 2021. Making pre-trained language models better few-shot learners. Proc 59th Annual Meeting of the Association for Computational Linguistics and 11th Int Joint Conf on Natural Language Processing, p.3816-3830. https://doi.org/10.18653/v1/2021.acl-long.295

Gemmeke JF, Ellis DPW, Freedman D, et al., 2017. Audio Set: an ontology and human-labeled dataset for audio events. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.776-780. https://doi.org/10.1109/ICASSP.2017.7952261

Gu YX, Han X, Liu ZY, et al., 2022. PPT: pre-trained prompt tuning for few-shot learning. Proc 60th Annual Meeting of the Association for Computational Linguistics, p.8410-8423. https://doi.org/10.18653/v1/2022.acl-long.576

Gu ZD, He KJ, 2024. Affective prompt-tuning-based language model for semantic-based emotional text generation. *Int J Semantic Web Inform Syst*, 20(1):1-19. https://doi.org/10.4018/IJSWIS.339187

Guan YD, Xue JB, Zheng GB, et al., 2022. Sparse self-attention for semi-supervised sound event detection. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.821-825. https://doi.org/10.1109/ICASSP43922.2022.9747834

Gulati A, Qin J, Chiu CC, et al., 2020. Conformer: convolution-augmented Transformer for speech recognition. Proc 21st Annual Conf of the Int Speech Communication Association, p.5036-5040. https://doi.org/10.21437/Interspeech.2020-3015

Imoto K, Tonami N, Koizumi Y, et al., 2020. Sound event detection by multitask learning of sound events and scenes with soft scene labels. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.621-625. https://doi.org/10.1109/ICASSP40776.2020.9053912

Jia ML, Tang LM, Chen BC, et al., 2022. Visual prompt tuning. Proc 17th European Conf on Computer Vision, p.709-727. https://doi.org/10.1007/978-3-031-19827-4_41

Koh CY, Chen YS, Liu YW, et al., 2021. Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.376-380. https://doi.org/10.1109/ICASSP39728.2021.9414350

Kong QQ, Xu Y, Wang WW, et al., 2020. Sound event detection of weakly labelled data with CNN-Transformer and automatic threshold optimization. *IEEE/ACM Trans Audio Speech Lang Process*, 28:2450-2460. https://doi.org/10.1109/TASLP.2020.3014737

Li YX, Liu ML, Drossos K, et al., 2020. Sound event detection via dilated convolutional recurrent neural networks. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.286-290. https://doi.org/10.1109/ICASSP40776.2020.9054433

Lin LW, Wang XD, Liu H, et al., 2020. Guided learning for weakly-labeled semi-supervised sound event detection. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.626-630. https://doi.org/10.1109/ICASSP40776.2020.9053584

Mesaros A, Heittola T, Virtanen T, 2016. Metrics for polyphonic sound event detection. *Appl Sci*, 6(6):162. https://doi.org/10.3390/app6060162

Mesaros A, Heittola T, Virtanen T, et al., 2021. Sound event detection: a tutorial. *IEEE Signal Process Mag*, 38(5):67-83. https://doi.org/10.1109/MSP.2021.3090678

Miyazaki K, Komatsu T, Hayashi T, et al., 2020a. Conformer-based sound event detection with semi-supervised learning and data augmentation. Proc 5th Workshop on Detection and Classification of Acoustic Scenes and Events, p.100-104.

Miyazaki K, Komatsu T, Hayashi T, et al., 2020b. Weakly-supervised sound event detection with self-attention. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.66-70. https://doi.org/10.1109/ICASSP40776.2020.9053609

Murugesan B, Hussain R, Bhattacharya R, et al., 2024. Prompting classes: exploring the power of prompt class learning in weakly supervised semantic segmentation. Proc IEEE/CVF Winter Conf on Applications of Computer Vision, p.290-301. https://doi.org/10.1109/WACV57701.2024.00036

Park JS, Kim SH, 2020. Sound learning-based event detection for acoustic surveillance sensors. *Multimed Tools Appl*, 79(23-24):16127-16139. https://doi.org/10.1007/s11042-019-7547-y

Serizel R, Turpault N, Shah A, et al., 2020. Sound event detection in synthetic domestic environments. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.86-90. https://doi.org/10.1109/ICASSP40776.2020.9054478

Singhal K, Azizi S, Tu T, et al., 2023. Large language models encode clinical knowledge. *Nature*, 620:172-180. https://doi.org/10.1038/s41586-023-06291-2

Sohn K, Chang H, Lezama J, et al., 2023. Visual prompt tuning for generative transfer learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.19840-19851. https://doi.org/10.1109/CVPR52729.2023.01900

Tarvainen A, Valpola H, 2017. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. 31st Int Conf on Neural Information Processing Systems, p.1195-1204.

Turpault N, Serizel R, Shah AP, et al., 2019. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. Workshop on Detection and Classification of Acoustic Scenes and Events, p.253-257.

Turpault N, Wisdom S, Erdogan H, et al., 2020. Improving sound event detection in domestic environments using sound separation. 5th Workshop on Detection and Classification of Acoustic Scenes and Events, p.205-209.

Wakayama K, Saito S, 2022. CNN-Transformer with self-attention network for sound event detection. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.806-810. https://doi.org/10.1109/ICASSP43922.2022.9747762

Wang YH, Chauhan J, Wang W, et al., 2023. Universality and limitations of prompt tuning. 37th Int Conf on Neural Information Processing Systems, Article 3305.

Wisdom S, Erdogan H, Ellis DPW, et al., 2021. What's all the fuss about free universal sound separation data? Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.186-190. https://doi.org/10.1109/ICASSP39728.2021.9414774

Xu H, Xie HT, Tan QF, et al., 2023. Meta semi-supervised medical image segmentation with label hierarchy. *Health Inform Sci Syst*, 11(1):26. https://doi.org/10.1007/s13755-023-00222-1

Yan J, Song Y, Dai LR, et al., 2020. Task-aware mean teacher method for large scale weakly labeled semi-supervised sound event detection. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.326-330. https://doi.org/10.1109/ICASSP40776.2020.9053073

Zheng X, Song Y, Dai LR, et al., 2021a. An effective mutual mean teaching based domain adaptation method for sound event detection. Proc 22nd Annual Conf of the Int Speech Communication Association, p.556-560. https://doi.org/10.21437/Interspeech.2021-281

Zheng X, Song Y, McLoughlin I, et al., 2021b. An improved mean teacher based method for large scale weakly labeled semi-supervised sound event detection. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.356-360. https://doi.org/10.1109/ICASSP39728.2021.9414931