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

Prompting class distribution optimization dynamically 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: mao_qr@ujs.edu.cn

Received Jan. 27, 2024; Revision accepted June 27, 2024; Crosschecked

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 associated with noise interference introduced by pseudo-labels or domain knowledge gaps. To alleviate noise 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 (SOTA) methods on the DCASE 2019, 2020, and 2021 challenge SSED datasets, PADO demonstrates significant performance improvements. Furthermore, it is ready to be extended 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:**

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 are required to recognize the categories of events and mark the onset and offset times for each event in a mixed audio recording, which generally contains 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, that is, 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; Li et al., 2020; Serizel et al., 2020; Gao LJ et al., 2022, 2024). However, the fine-grained manual annotation at the frame level is extremely time-consuming and results

‡ Corresponding author

* Project supported by the National Nature Science Foundation of China (No. 62176106), 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 (No. KYCX22_3668 & KYCX21_3373), the Key Project of National Nature Science Foundation of China (No. U1836220), and the Jiangsu key research and development plan (industry foresight and key core technology, No. BE2020036)

ORCID: Lijian GAO, https://orcid.org/0000-0002-6458-0660

in a severe shortage of annotated samples and 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 or synthetic data for generalization learning to effectively mitigate overfitting on the limited labeled data. Recently, the most popular methods in SSED commonly build upon the consistency regularization assumption (Tarvainen and Valpola, 2017) to establish a teacher-student framework (Yan et al., 2020; Koh et al., 2021; Zheng et al., 2021b; Gao LJ et al., 2023). In the teacher-student framework, the teacher model provides pseudo-labels to guide the student model for generalization training on unlabeled or 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, the noise interference introduced by pseudo-labels or domain knowledge bias of synthetic data in class distribution learning is often under-considered, resulting in sub-optimal performance in SED tasks. Therefore, one of the most challenging tasks in SSED is alleviating the noise interference in class distribution learning. To this end, some recent work achieves performance gains through effective pseudo-labeling (PL) strategies (Chan and Chin, 2021; Koh et al., 2021), or tries to transfer the domain knowledge of the synthetic data domain into the 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. Consequently, this often comes with an increase in 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 fundamental 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 task, 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 of the real labeled data. Then, the prior information learned by the 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 pioneer the integration of 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 unlabeled or synthetic data, and effectively alleviates the noise interference for semi-supervised learning.

2. PADO avoids directly optimizing the noisy distribution information modeled by class tokens, thus preventing the decline in generalization capability. Instead, it models prior knowledge as prompt information only during 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 the recent literature related to our work in this study.

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. In recent years, the convolutional recurrent neural network (CRNN) is often adopted as the backbone model in SSED tasks, which employs a convolutional neural network (CNN) as the front-end network for the RNN model. With its remarkable spatial and temporal feature modeling capability, 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 the sequential spectrogram. Then, to jointly model local spatial and global context information, some recent work has attempted to combine a CNN and Transformer (Kong et al., 2020; Miyazaki et al., 2020a; Wakayama and Saito, 2022; Gao LJ et al., 2023). One of the most representative work 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, a more recent work, named Joint-Former (Gao LJ et al., 2023), combines a Masked Auto-encoder with the Conformer model, further improving the performance of SSED.

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

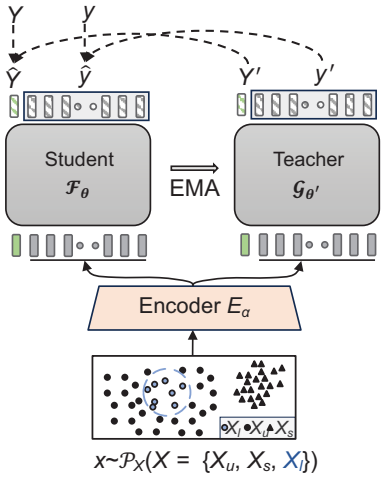


Fig. 1 Framework of MT-based SSED. MT: mean teacher; SSED: semi-supervised sound event detection

2.2 Class distribution optimization for SSED

In alleviating noise interference during class distribution learning for SSED tasks, recent efforts have been directed toward two main aspects to optimize the noised class distribution information: pseudo-label learning and DA. For PL, Koh et al. (2021) introduced a two-stage PL strategy through a pre-trained pseudo-label generation model for unlabeled data; Chan and Chin (2021) performed a convolutive non-negative matrix factorization (CNMF) algorithm to annotate the pseudo-labels. For DA, Chan and Chin (2021) proposed the use of two different MT models to guide student training, combining an adversarial trained domain classifier with MT models.

3 Methodology

In this section, we will 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 small real labeled data X_l. As illustrated in Fig. 1, the training set can be defined as $X = \{X_u, X_s, X_l\}$, where $x \in X$ are spectrograms of audio signals. The real labeled data $X_l \in X$ only contains clip-level ground truth, denoted as weakly labeled data. Typically, the basic Mean-Teacher (MT) framework for SSED tasks contains three components as shown in Fig. 2, including an encoder E_α for raw input modeling, a student model $\mathcal{F}_\theta$ and a teacher model $\mathcal{G}_{\theta'}$, where α, θ and θ' denotes the parameters corresponding to each module respectively. The detailed descriptions of MT-based SSED architecture is described in detail as follows.

Assuming that the length of each sequential data $x \in X$ is T, the dimension of x is D, that is $x \in \mathbb{R}^{d\times T}$, and the number of event categories in the dataset is K. For a randomly sampled spectrogram feature x from the training set X (denoted as $x \sim P_X$), the Encoder E_α is required to down-sample the long-time sequential input for reducing computation costs, and learns the latent features h for x (the gray block sequence in Fig. 1), that is, $h = E_\alpha(x)$, where $h \in \mathbb{R}^{d'\times T'}$, d' represents the dimension of h and T' is sequence length after down-sampling. Next, the latent features h will be concatenated with the independently learnable class token $h_{\text{cls}} \in \mathbb{R}^{d'\times 1}$ (the green block in Fig. 1) and fed into both the student model $\mathcal{F}_\theta$ and the teacher model $\mathcal{G}_{\theta'}$ to get the SED predictions $(\hat{Y}, \hat{y})$ and $(\hat{Y}', \hat{y}')$, respectively,

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

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

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

According to the consistency assumption, the student model $\mathcal{F}_\theta$ 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, the weight $\omega(n)$ is initialized close to 0 (i.e., $\omega(0) = \mathrm{e}^{-5} \approx 0.007$) and gradually increases 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)), and both of them are the sum of tagging error ℓ_{tag} and localization error ℓ_{tag}, where (Y, y) are the real-labeled ground truth, while (Y', y') are the predictions of the teacher model.

Following the MT framework setup, the teacher model is an ensemble of the historical student models and performs the exponential moving average (EMA) algorithm to update the parameters θ' of the teacher model as Eq. (6), where t is the current training epoch and α is the smoothing coefficient, typically set to 0.999.

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

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

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

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

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 finish the training of the MT-based SSED framework. However, the noise interference from the pseudo-labels and the knowledge gaps of the synthetic data domain in the class distribution learning process are underconsidered. To this end, we will discuss our proposed solution as follows, i.e., prompting class distribution optimization (PADO) for SSED.

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 (generalization & supervised training) run synchronously until convergence.) The generalization training is responsible for improving the generalization performance of the model 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 distribution introduced during the generalization training stage. The details are as follows.

3.2.1 Generalization training process in PADO

First, given the unlabeled and synthetic data (i.e., $x \sim P_{\{X_u,X_s\}}$), consistent with the basic MT framework, PADO optimizes the consistency loss $\mathcal{L}_{\text{con}}$ as Eq. (5) for generalization training. In this case, the generalization performance of the model is improved, reducing the over-fitting on the small amount of real labeled data. However, the noise interference from the pseudo-label and domain bias of synthetic data are still under-considered, resulting in a noisy class distribution. So, 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 at alleviating the potential noisy interference in class distribution information. Here, PADO ingeniously addresses this challenge by introducing prompt tokens $\mathbf{P}$, a group of independent learnable parameters, during supervised training on real labeled data (i.e., $x \sim P_{X_l}$). Specifically, as illustrated in Fig. 2b, the prompt tokens $\mathbf{P}$ are embedded into the latent feature sequences $h_x = E_\alpha(x)$ of each layer of the Transformer encoders. Note that the prompt tokens are randomly initialized at each layer. Then, efficient interactions are performed among prompt tokens $\mathbf{P}$, class token h_{cls} and latent features h_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 $x \in \{X_l\}$).

$$\hat{Y}, \hat{y} = \begin{cases} \mathcal{F}_\theta(\mathcal{C}[h_{\text{cls}}, h_x]), & \text{if } x \in \{X_u, X_s\}, \\ \mathcal{F}_\theta(\mathcal{C}[h_{\text{cls}}, \mathbf{P}, h_x]), & \text{if } x \in \{X_l\}. \end{cases} \tag{7}$$

More specifically, taking the i^{th} layer encoder Att^i as an example, Eqs. (8)–(10) define the interaction process when embedding prompt tokens $\mathbf{P}$ through the self-attention mechanism. $PE_{\text{sinusoidal}}$ denotes sinusoidal position encoding, where the positional indices of the latent feature h_x should remain consistent during both generalization training and supervised training. For example, assuming that the length of prompt tokens is 3 and the sequence length of h_x is 10, then during generalization training, the positional indices should be $\{0, 4, 5, ..., 13\}$, while during the supervised training, the positional indices are set to $\{0, 1, 2, 3, 4, 5, ..., 13\}$ because the prompt tokens are embedded. Therefore, the positional index of the class token is set to 0, the positional indices of prompt tokens are $\{1, 2, 3\}$, and the positional indices of spectrogram features remain fixed at $\{4, 5, 6, ..., 13\}$. Additionally, Q, K, and V in the equations represent query, key, and value vectors in the attention mechanism, respectively. MultiHead 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 the joint constraints $\mathcal{L}_{\text{joint}}$ (Eq. (3)), only the student model embedded with prompt tokens (as shown in Fig. 2b) is required for inference.

$$h^i = \mathcal{C}[h^i_{\text{cls}}, \mathbf{P}^i, h^i_x] + PE_{\text{sinusoidal}} \tag{8}$$

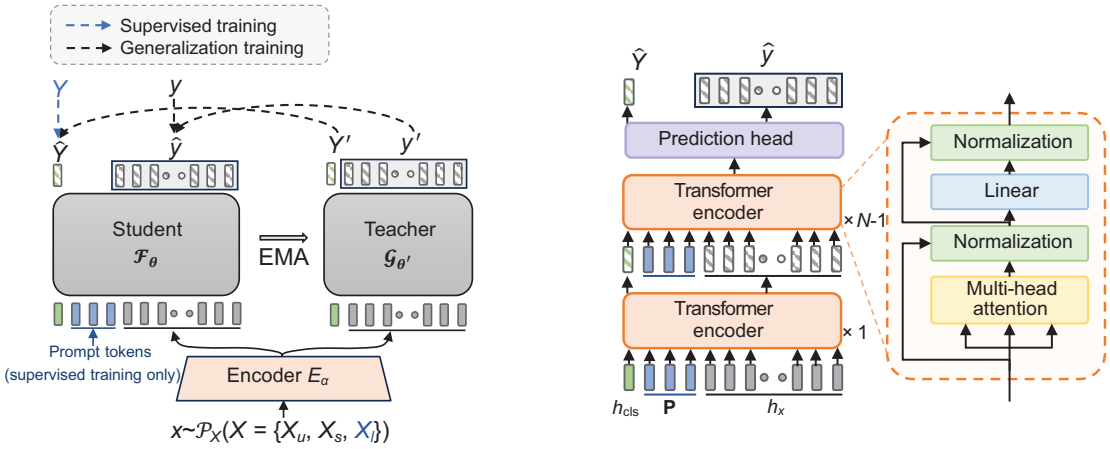
$$Q^i, K^i, V^i = h^i W^i_Q, h^i W^i_K, h^i W^i_V \tag{9}$$

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

$$\hat{Y}, \hat{y} = \text{Pred}(h^I_{\text{cls}}), \text{Pred}(h^I_x) \tag{11}$$

Based on the simple yet powerful training strategy offered by PADO, we can easily train a 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 label 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 the generation learning on the unlabeled data. In this case, the prior information about real label data 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



(a) (b)

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

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

Dataset	Training set			Validation set	Evaluation set	Duration (s)	Category number
	Weakly labeled	Unlabeled	Synthetic				
DCASE 2019	1578	14412	2045	1168	692		
DCASE 2020	1578	14412	2584	1168	692	10	10
DCASE 2021	1578	14412	10,000	1168	692		

(Eq. (4)). Notably, PADO avoids directly optimizing the noisy 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 Experiment setup

4.1.1 Dataset

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 public to contestants for performance evaluation during the challenges, whereas the evaluation set was used for the official ranking. Except for the synthetic data, the others in all three DCASE challenge datasets are sub-sets of AudioSet (Gemmeke et al., 2017), a large-scale real-recorded dataset for audio classification. Details of the datasets are listed in Table 1. In addition, the datasets cover 10 target sound events in domestic scenarios, including "Speech", "Dog", "Alarm", "Dishes", "Frying", "Blender", "Running water", "Vacuum cleaner", and "Electric shaver".

4.1.2 Models comparison

We first choose three advanced self-attention-based models as the baseline to build our PADO framework. As illustrated in Table 2, the baseline models are (1) a three-layers Transformer 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-

Table 2 Summary of the compared methods

ID	Model	Framework	Strategy
1	Transformer	MT	–
2	ConformerSED (Miyazaki et al., 2020a)	MT	–
3	Joint-Former (Gao et al., 2023)	MT	–
4	GL (Lin et al., 2020)	MT	–
5	SparseTrans (Guan et al., 2022)	MT	–
6	SAN (Wakayama and Saito, 2022)	MT	–
7	SCT (Koh et al., 2021)	Two-stage	PL
8	CNMF (Chan and Chin, 2021)	Two-stage	PL
9	MMT (Zheng et al., 2021a)	MT	DA
10	PDAO-Transformer	PADO	PT
11	PDAO-ConformerSED	PADO	PT
12	PDAO-Joint-Former	PADO	PT

CNMF: convolutive non-negative matrix factorization; DA: domain adaptation; MT: mean teacher; PADO: prompting class distribution optimization; PL: pseudo-labeling; PT: prompt tuning

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 (10) PADO-Transformer, (11) PADO-Conformer and (12) PADO-JointFormer in Table 2. Also, we extensively compared the proposed methods with the other SOTA models on DCASE 2019, 2020, and 2021 challenge datasets, including (4) GL (Lin et al., 2020) (the winner of DCASE 2019), (5) SparseTrans (Guan et al., 2022), and (6) 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 also consider four advanced methods for alleviating the noise interference for comparison. As shown in Table 2(7)-(10), they are SCT (Koh et al., 2021), CNMF (Chan and Chin, 2021) and MMT (Zheng et al., 2021a). Specifically, SCT adopts a two-stage shift consistency training strategy for PL. CNMF proposed a convolutive nonnegative matrix factorization method for PL. In contrast, MMT focused on DA.

4.1.3 Data preprocessing

Following with ConformerSED (Miyazaki et al., 2020a), we down-sampled all data in 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 get 496 frames of 64-dimensional Mel spectrogram features for each sound signal. The settings of data preprocessing remain consistent throughout all experiments in this paper.

4.1.4 Metrics

We follow the settings of DCASE challenges which adopt the Event-Based F1 score (EB-F1) (Mesaros et al., 2016) and the Polyphonic Sound Detection Score (PSDS) (Bilen et al., 2020) as the 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. In contrast, we perform the three metrics on all the datasets for comprehensive comparisons in this study.

4.1.5 Training settings

All experiments were developed on PyTorch 1.2.0. The CUDA devices we used for training were NVIDIA RTX 3090, with a batch size 32 during training. The training iterations were set to 30,000 epochs. The hyperparameter $\omega(n)$ in Eq. (3) ramps up from 1 to 6,000 steps with a maximum of 2.0.

4.2 Ablation studies

First of all, to search the optimal settings for the prompt tokens in the proposed PADO, we designed a group 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 layers of the encoders embedded with prompt tokens. Therefore, we evaluate the

Table 3 Performance comparisons on DCASE 2019, 2020, and 2021 validation sets: PADO-based SSED methods vs. MT-based baselines

Models	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-ConformerSED	**41.8**	**27.5**	**45.2**	**43.6**	**29.1**	**47.8**	**40.7**	**26.0**	**45.3**
Joint-Former (Gao et al., 2023)	**42.9**	27.2	43.6	43.2	27.8	44.6	**42.1**	26.9	45.8
PADO-JointFormer	42.3	**27.9**	**44.2**	**44.4**	**29.7**	**48.0**	41.9	**27.7**	**46.6**

EB-F1: event-based F1; MT: mean teacher; PADO: prompting class distribution optimization; PSDS: polyphonic sound detection score; SSED: semi-supervised sound event detection. The underlined methods indicate replicated results. The best-performing results are in bold

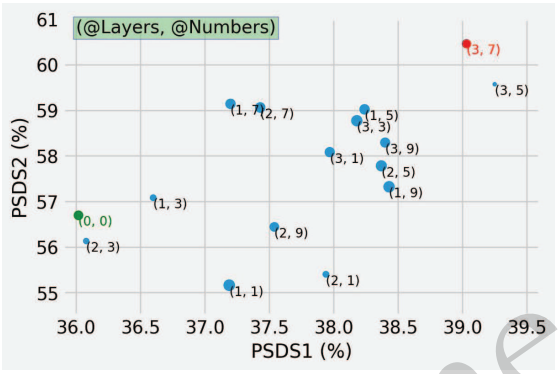


Fig. 3 Grid search for optimal settings of prompt tokens in PADO, where the coordinates of the data points represent (layers, numbers), the red point represents the settings that achieved optimal performance, and the green point is the basic MT framework. MT: mean teacher; PADO: prompting class distribution optimization

impact of the number of prompt tokens in PADO through grid-search from {1, 3, 5, 7, 9}, and the layers of encoders embedded with prompt tokens from {1, 2, 3}.

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. The coordinates of the data points represent layers of the encoder embedded with prompt tokens, and the number of prompt tokens, that is (Layers, Numbers). Clearly, scatters that are located closer to the upper-right region and have 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 the layer of embedded encoders is 3 (each layer of the encoder embeds prompt tokens), and the number of tokens is 7. Finally, the prompt tokens are set as $3 \times 7 \times 128$ in PADO. In this configuration (i.e., (layers=3, number=7)), the EB-F1, PSDS1, and PSDS2 scores are 49.90%, 38.19%, and 58.84%, respectively. Note that when (layers=0, number=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% on EB-F1, 2.24% on PSDS1, and 2.96% on PSDS2, respectively.

4.3 Performance comparison with SOTAs

4.3.1 Comparisons with MT-based baseline models

To thoroughly evaluate the effectiveness and generality of the proposed PADO, here we choose three advanced MT-based methods as baselines for comparison on the validation and evaluation set in DCASE 2019, DCASE 2020 and DCASE 2021 datasets. They are a basic Transformer and two advanced Transformer-based SSED models (ConformerSED (Miyazaki et al., 2020a) and Joint-Former (Gao et al., 2023)).

Performance comparisons on the validation set of DCASE 2019, 2020 and 2021 datasets are shown in Table 3. Because the baseline models did not report the results on the validation set, we reproduced the models (underlined text) using the same settings they used for comparison. Experimental results shown in Table 3 indicate that the proposed PADO-based semi-supervised learning frame-

Table 4 Performance comparison on DCASE 2019, 2020, and 2021 evaluation sets: PADO-based SSED methods vs. SOTA methods

Models	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 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-ConformerSED	**49.9**	**38.2**	**58.8**	**49.9**	**40.1**	**61.6**	**44.2**	**33.8**	**56.4**
Joint-Former (Gao et al., 2023)	50.9	40.0	60.1	50.7	39.3	59.3	44.5	34.4	56.9
PADO-JointFormer	**51.4**	**42.1**	**61.2**	**50.9**	**41.8**	**61.9**	**46.7**	**36.9**	**57.0**

CNMF: convolutive non-negative matrix factorization; 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. The solid line (-) indicates that the corresponding metrics were not reported; The methods marked with underlined text indicate replicated results; The best-performing results are in bold

work 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, on the DCASE 2019 validation set, the performance improvement ranges from 0.3% to 1.2%. On the DCASE 2020 validation set, the improvement ranges from 0.1% to 0.8%, and on the DCASE 2021 validation set, PADO-Transformer brings a performance improvement ranging from 0.1% to 2.1%. Compared to ConformerSED (Miyazaki et al., 2020a), on the DCASE 2019 validation set, PADO-ConformerSED improves by 0.4% to 1.6%; on the DCASE 2020 validation set, the improvement is from 0.8% to 1.9%; and on the DCASE 2021 validation set, except for the EB-F1 metric, PSDS1 and PSDS2 improve by 0.4% and 1.3%, respectively. Finally, compared to Joint-Former (Gao et al., 2023), our PADO-JointFormer does not outperform the vanilla Joint-Former in EB-F1 on DCASE 2019 and 2021 validation sets.. This is because there is a set of hyper-parameters that need to be carefully adjusted in Joint-Former (Gao et al., 2023), for example, weights of reconstruction loss. However, when we adopt the Joint-Former in our PADO framework to reproduce its results, we do not perform an additional grid search of the hyper-parameters. Nevertheless, the proposed PADO-JointFormer achieved significant performance improvements on all other metrics. Specifically, on the DCASE 2019 validation set, PADO-JointFormer, except for the EB-F1 metric, improves PSDS1 and PSDS2 by 0.7% and 0.8%, respectively. On the DCASE 2020 validation set, all metrics improve by 0.9% to 1.4%, and on the DCASE 2021 validation set, except for the EB-F1 metric, PSDS1 and PSDS2 improve by 0.8% and 0.8%, respectively.

4.3.2 Comparisons with MT-based SOTA models

Here we compare our PADO-based models with the SOTAs on DCASE 2019, 2020, and 2021 sets, that is, GL, ConformerSED, SparseTrans, Joint-

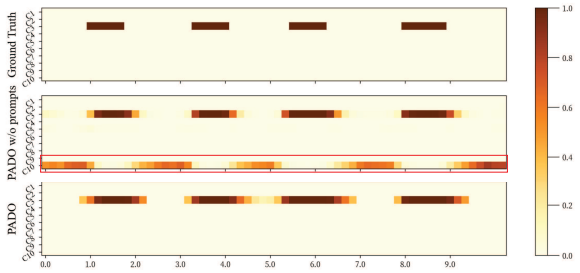


Fig. 4 Visualization of event localization for a test sample (1LKP1ZyHgVg_0_10.wav), where the sound event "Cat" is active

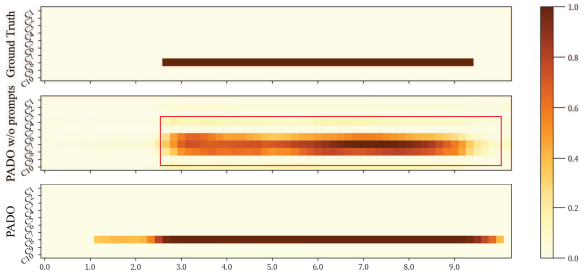


Fig. 5 Visualization of event localization for a test sample (dfRtayqQAls_14_24.wav), where the sound event "Running Water" is active

Former, and SAN. Because the SOTAs only reported the experimental results on the public evaluation sets on DCASE challenge datasets, we compare the evaluation performance on the evaluation sets with the SOTAs as shown in Table 4. As a result, our PADO-based methods achieves remarkable performance on all the metrics in SSED tasks, outperforming the SOTAs. Notably, the PADO-JointFormer 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, here we compare the performance with three advanced methods aimed at optimizing class distribution for SSED tasks, that is, SCT, CNMF, and MMF as shown in Table 4. The results show that PADO outperforms the advanced methods in class distribution optimization for SSED by 0.5% to 5.8% on EB-F1. Notably, PADO achieves class distribution optimization dynamically with only a lightweight set of parameters, that is, prompt tokens. In contrast, the aforementioned advanced methods require the assistance of additional models or complex training strategies, for example, two-stage pre-training in SCT (Koh et al., 2021) and CNMF (Chan and Chin, 2021). Clearly, 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 the prompt tokens in PADO. Specifically, we use the trained PADO-JointFormer to visualize the localization results of two exemplar sound recordings. Then, we remove the prompt tokens from our model to observe the difference in the localization performance. The localization performance is visualized in Fig. 4 and Fig. 5, which consists of three rows of subplots, illustrating the visualization of ground truth, localization results after removing prompt token (denoted as "PADO w/o prompts"), and localization results with prompt tokens retained (denoted as "PADO"), respectively. The horizontal axis represents time, and the vertical axis represents event categories, denoted as $C1$ to $C10$ for {"Alarm," "Blender," "Cat," "Dishes," "Dog," "Electric shaver," "Fry," "Running water," "Speech," "Vacuum cleaner"}. Clearly, the PADO-JointFormer with retained prompt tokens significantly improves localization performance.

Specifically, Fig. 4 illustrates a sound signal with the true label "Cat," where the PADO-JointFormer with retained prompt tokens exhibits excellent localization performance (as shown in the third-row subplot) with minimal false positive predictions. However, when removing prompt tokens in 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 between "Running water," "Speech" and "Fry" of 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 attributed 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 in the case of the second subplot in Fig. 5).

5 Conclusions

Addressing the noise 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 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. Li-jian GAO and Qirong MAO revised and finalized the paper.

Conflict of interest

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

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. Joint Conf of the 59th Annual Meeting of the Association for Computational Linguistics and the 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 Inf 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. 5th the 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(7972):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 the 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 the 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. WhatâĂŹs 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 Inf 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 22rd 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