Lijian GAO, Qing ZHU, Yaxin SHEN, Qirong MAO, Yongzhao ZHAN, 2025. Dynamic prompting class distribution optimization for semi-supervised sound event detection. *Frontiers of Information Technology & Electronic Engineering*, 26(4):556-567. https://doi.org/10.1631/FITEE.2400061

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

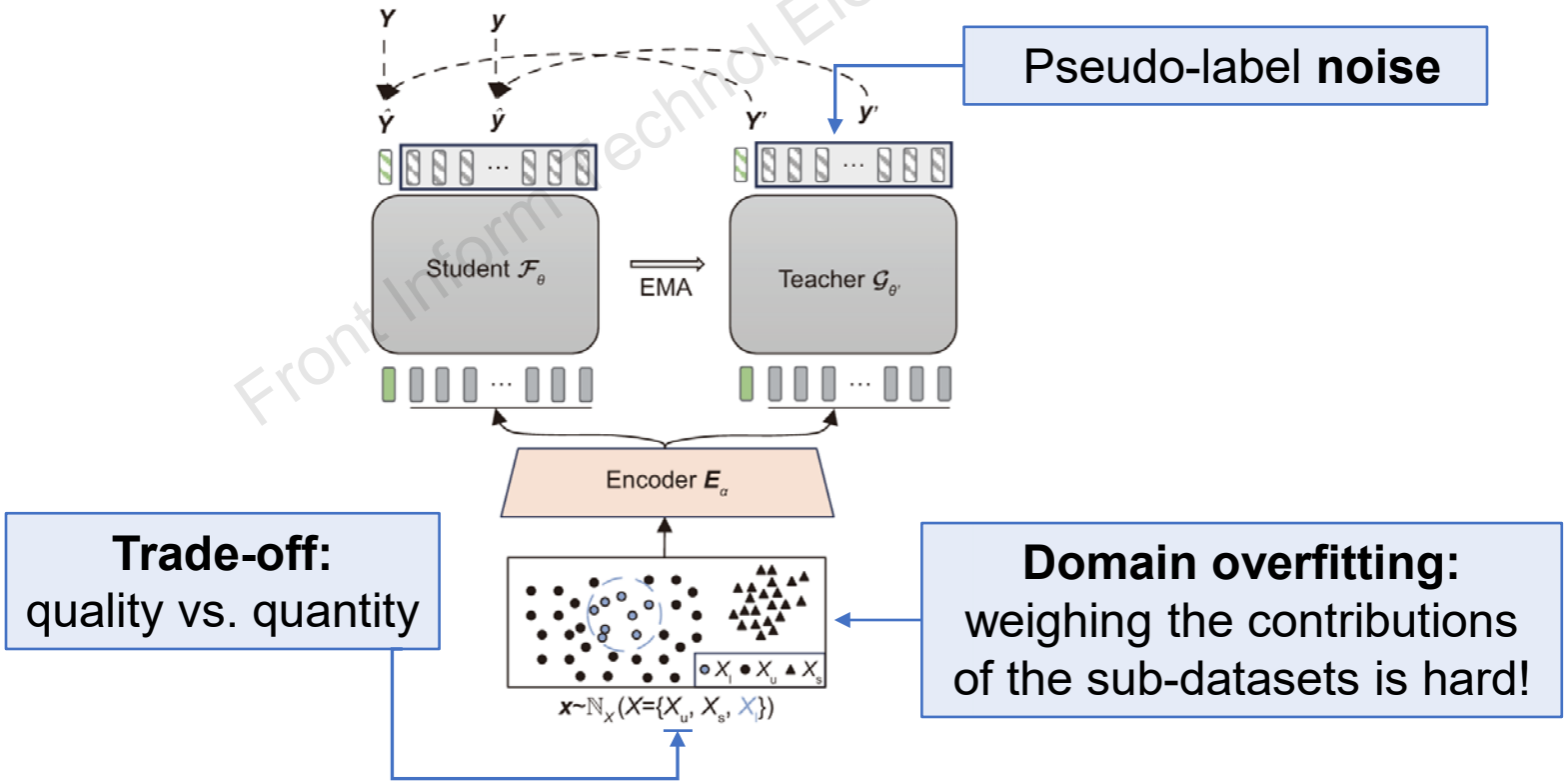
Key words: Prompt tuning; Class distribution learning; Semi-supervised learning; Sound event detection

Qirong MAO
E-mail: mao_qr@ujs.edu.cn
ORCID: https://orcid.org/0000-0002-0616-4431

Lijian GAO
E-mail: ljgao@ujs.edu.cn
ORCID: https://orcid.org/0000-0002-6458-0660

Challenges

- **Noisy interference in class distribution learning**: pseudo-label noise and domain knowledge bias;
- **Trade-off between quality and quantity**: low efficiency in the utilization of unlabeled data;
- **Domain overfitting with increased algorithm complexity**: domain adaptation algorithms require careful design.



Motivation & novelties

- By optimizing only a small portion of the parameters (i.e., prompt tokens), **prompt tuning** (PT) effectively tailors class distribution information for specific downstream tasks while retaining the generalization capacity of pre-trained models.
- 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.

- ✓ Propose an advanced **semi-supervised class distribution learning** framework (i.e., 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.
- ✓ Avoid the decline in generalization capability
  - PADO models prior knowledge as prompt information during only supervised training to dynamically guide the learning of class distribution.

Methodology: PADO framework

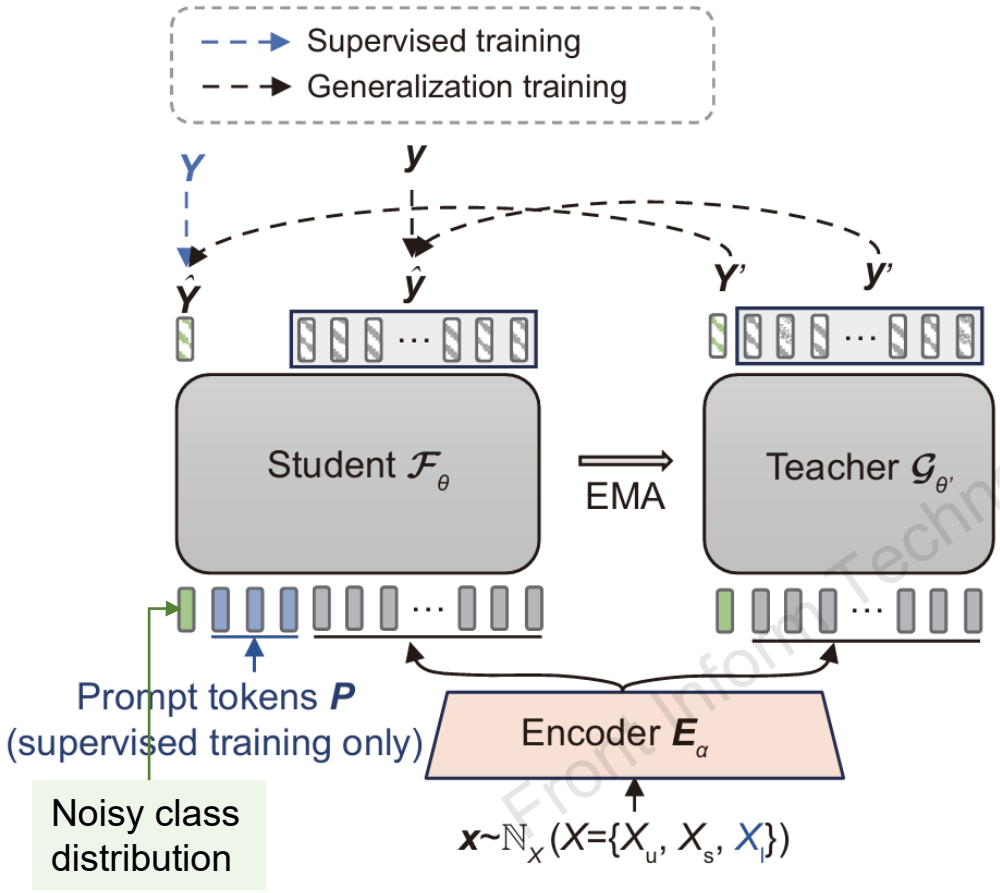


Fig. 2 Framework of the PADO-based SSED, containing supervised training and generalization training

➢ **Generalization training**

→ Learning from pseudo-labels:

1. Get $\boldsymbol{x}$ from $\{X_u, X_s\}: \boldsymbol{x} \sim \mathbb{N}_X(X_u, X_s)$
2. Generate pseudo-labels:

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

3. Minimize the consistency error.

Generalization performance is improved, but the **class distribution is noisy**!

➢ **Supervised training**

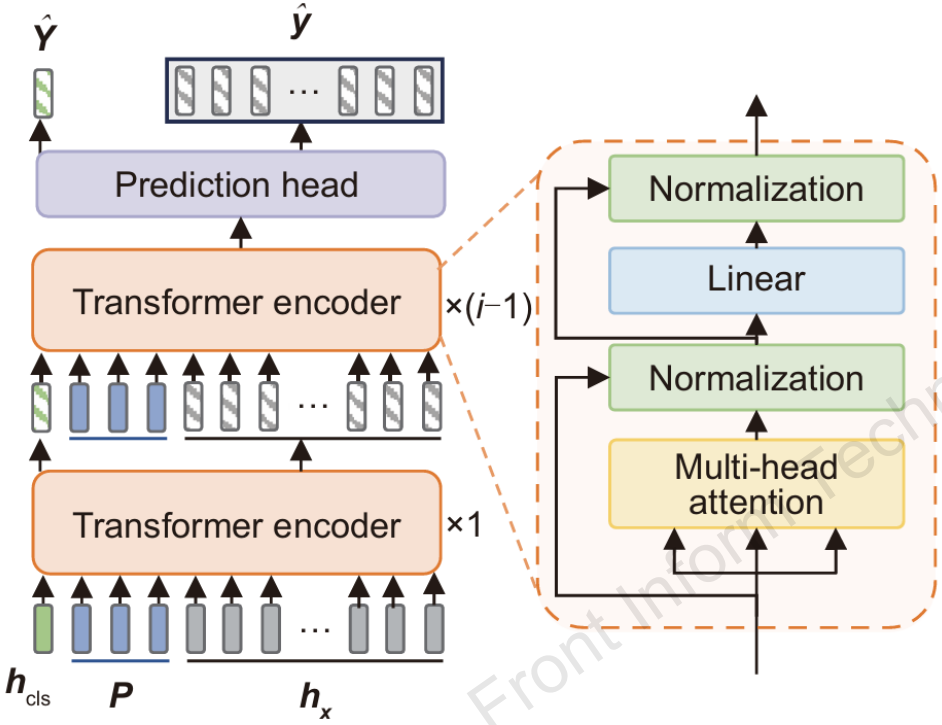
→ Optimizing distribution by real labeled data X_l :

1. Build a real distribution by prompts ***P***;
2. Interact with the noisy distribution dynamically to optimize joint loss.

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

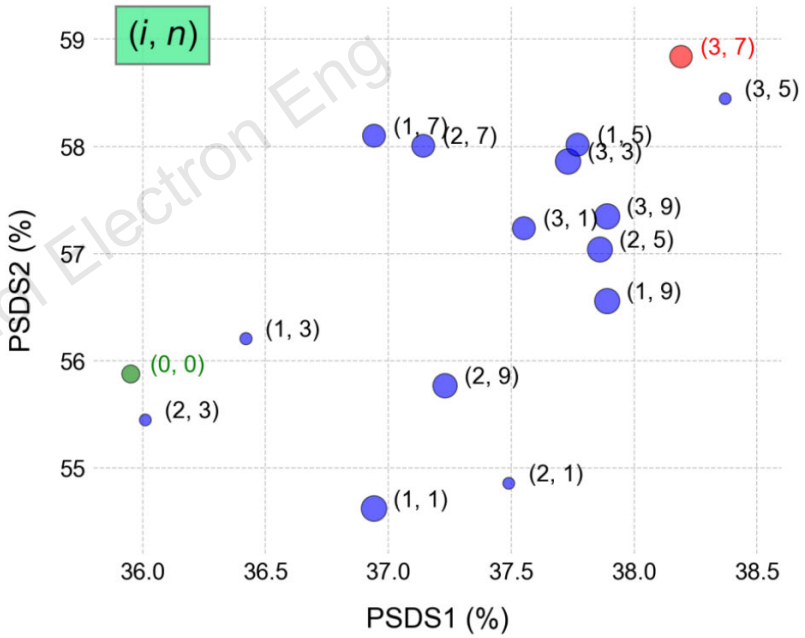
Maintaining the generalization ability while optimizing the noise distribution!

Methodology: model ablations



Details of the backbone models in PADO

- Grid search for the optimal setting in PADO



The best setting for the number of encoder layers (*i*) and the number of prompt tokens (*n*) is *i*=3 and *n*=7 (red point), indicating that deep prompt tuning performs better.

Results

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

- **Comparisons with MT-based baseline models**
- ✓ The proposed PADO-based semi-supervised learning framework achieves significant performance improvements for the aforementioned advanced MT-based SSED models.

Results

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**

- **Comparisons with SOTA methods**
- ✓ Our PADO-based methods achieve remarkable performance on all metrics in SSED tasks, outperforming the SOTA models.
- ✓ The PADO-Joint-Former achieves a new SOTA performance in SSED tasks on all three benchmark datasets.

Results

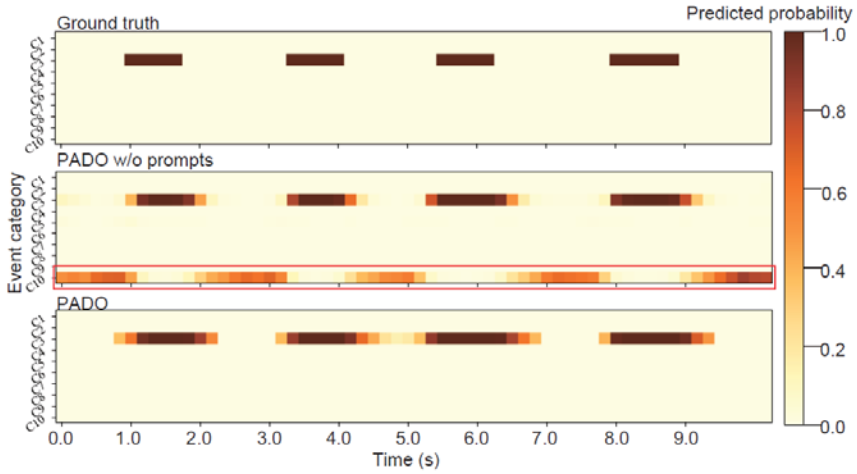


Fig. 4 Visualization of event localization for a test sample (1LKP1ZyHgVg_0_10.wav) from DCASE 2020 evaluation set, where the sound event "Cat" is active

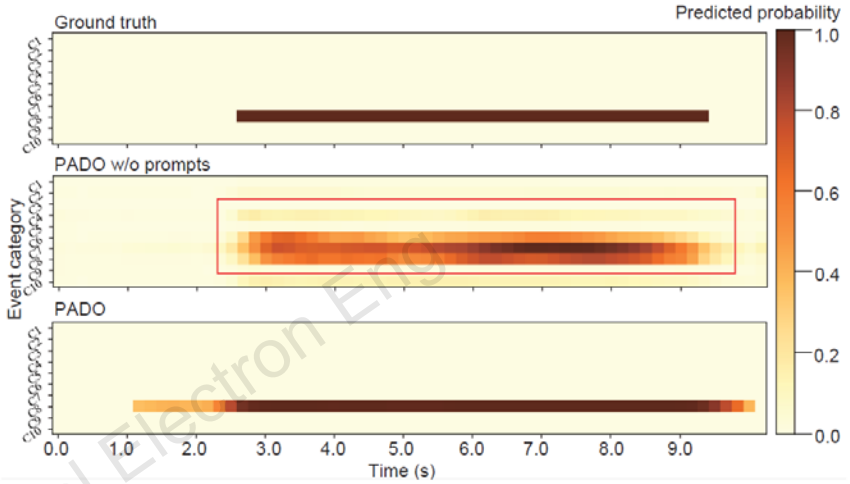


Fig. 5 Visualization of event localization for a test sample (dfRtayqQAls_14_24.wav) from DCASE 2020 evaluation set, where the sound event "Running water" is active

- ➢ **Qualitative visualization of localization**
- ✓ After embedding the prompt token, the **false detection rate** is significantly reduced.
- ✓ After embedding the prompt token, the **confusion between classes** is significantly reduced.

PADO effectively improves the efficiency of class distribution learning!

Lijian Gao received the M.S. and Ph.D. degrees in computer science and technology from Jiangsu University, Zhenjiang, China, in 2019 and 2024, respectively. He is currently a Lecturer at the School of Computer Science and Communication Engineering, Jiangsu University. His research interests include audio and speech signal processing, specifically sound event detection.

Qirong Mao received the M.S. and Ph.D. degrees in computer application technology from Jiangsu University, Zhenjiang, China, in 2002 and 2009, respectively. She is currently a Professor and Dean of the School of Computer Science and Communication Engineering at Jiangsu University. Her research interests include affective computing, pattern recognition, and multimedia analysis.

Yongzhao Zhan received the B.S. degree in computer science and technology from Fuzhou University, Fujian, China, in 1984, and the Ph.D. degree in computer science and technology from Nanjing University, Nanjing, China, in 2000. He is currently a Professor with the School of Computer Science and Communication Engineering, Jiangsu University. His research interests include multimedia analysis and pattern recognition.