An Entropy-Guided Curriculum Learning Strategy for Data-Efficient Acoustic Scene Classification under Domain Shift

Peihong Zhang, Yuxuan Liu*, Zhixin Li, Rui Sang, Yiqiang Cai, Yizhou Tan, Shengchen Li

School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China
{Peihong.Zhang20, Yuxuan.Liu2204, Zhixin.Li22, Rui.Sang22,
Yiqiang.Cai21, Yizhou.Tan22}@student.xjtlu.edu.cn
Shengchen.Li@xjtlu.edu.cn

Abstract—Acoustic Scene Classification (ASC) faces challenges in generalizing across recording devices, particularly when labeled data is limited. The DCASE 2024 Challenge Task 1 highlights this issue by requiring models to learn from small labeled subsets recorded on a few devices. These models need to then generalize to recordings from previously unseen devices under strict complexity constraints. While techniques such as data augmentation and the use of pre-trained models are well-established for improving model generalization, optimizing the training strategy represents a complementary yet less-explored path that introduces no additional architectural complexity or inference overhead. Among various training strategies, curriculum learning offers a promising paradigm by structuring the learning process from easier to harder examples. In this work, we propose an entropy-guided curriculum learning strategy to address the domain shift problem in data-efficient ASC. Specifically, we quantify the uncertainty of device domain predictions for each training sample by computing the Shannon entropy of the device posterior probabilities estimated by an auxiliary domain classifier. Using entropy as a proxy for domain invariance, the curriculum begins with high-entropy samples and gradually incorporates low-entropy, domain-specific ones to facilitate the learning of generalizable representations. Experimental results on multiple DCASE 2024 ASC baselines demonstrate that our strategy effectively mitigates domain shift, particularly under limited labeled data conditions. Our strategy is architecture-agnostic and introduces no additional inference cost, making it easily integrable into existing ASC baselines and offering a practical solution to domain shift.

Index Terms—Acoustic Scene Classification, Curriculum Learning, Domain Generalization, Domain Shift

1. INTRODUCTION

Acoustic Scene Classification (ASC) aims to recognize the environment from an audio segment and serves as a fundamental task in computational auditory analysis [1], [2]. A major challenge in real-world ASC systems is domain shift caused by variations in recording devices, which can significantly degrade model generalization to unseen devices. The challenge is further exacerbated when only limited labeled training data is available. The DCASE 2024 Challenge Task 1 [3] explicitly targets this scenario by requiring systems to be trained on small fractions (as low as 5%) of labeled audio from a limited set of devices and evaluated on recordings from previously unseen devices. In addition, participating models must adhere to strict complexity constraints, including a maximum of 128 kB of parameters and 30 MMACs per second of audio.

Domain shift becomes particularly challenging when labeled training data is limited, as models struggle to learn domain-invariant representations from limited supervision [4]. Figure 1 illustrates this phenomenon in ASC, where spectrograms of the same scene recorded by different devices exhibit noticeable discrepancies.

To address such challenges, existing approaches often turn to data augmentation strategies—such as Freq-MixStyle [5] and device impulse response simulation [6]—or leverage pretrained models [7] to enhance model generalization across recording devices. These
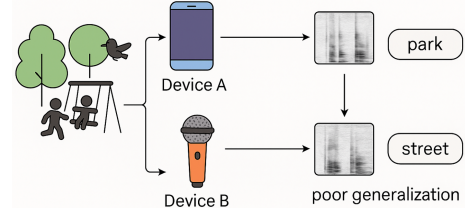
Fig. 1: Example of domain shift in ASC caused by device variability. Although the two spectrograms represent the same scene, device-specific characteristics cause noticeable differences. As a result, a model trained on one device may fail to correctly classify inputs from another, illustrating poor generalization.

techniques improve feature diversity and representation quality but may also introduce additional complexity [8], [9], such as increased model size, reliance on external resources, or inference overhead.

While these approaches are effective, they focus on modifying the data or the initial state of the model. The training curriculum itself—the very order in which data is presented to the model—remains a relatively underexplored dimension for enhancing domain generalization [10]. This observation motivates our central question: Can we devise a lightweight, model-agnostic training strategy that optimizes the learning trajectory itself to build domain-invariant representations with limited supervision?

To address this challenge, we draw inspiration from Curriculum Learning [11], a training paradigm that organizes the presentation of training data from easy to hard, thereby optimizing the learning trajectory. This approach is particularly well-suited for data-limited domain shift scenarios, as it facilitates the progressive acquisition of generalizable representations [12]. By introducing less confounded, domain-invariant samples early in training, the model can establish a robust feature foundation before gradually incorporating harder, domain-specific examples. Such staged learning has been shown to enhance generalization in a variety of tasks [13], [14]. Building on this idea, we propose an entropy-guided curriculum learning strategy that complements existing approaches without modifying the ASC model architecture. Specifically, we employ an auxiliary domain classifier to estimate the device posterior probabilities for each training sample. The Shannon entropy of this distribution serves as a proxy for domain invariance, based on the intuition that samples with high entropy—those that confuse the device classifier—are more likely to be device-agnostic and thus easier for learning generalizable features [15]. The curriculum therefore begins with these domain-invariant samples to establish a robust feature foundation, and gradually introduces harder, domain-specific examples to refine decision boundaries.

To assess the effectiveness of our approach, we conduct evaluations across multiple ASC systems submitted to the DCASE 2024 Challenge Task 1, with diverse model architectures and training setups. Experimental results show that our method consistently improves cross-

device performance under low-resource conditions. These findings demonstrate the potential of entropy-guided curriculum learning as a lightweight, architecture-agnostic, and easily integrable strategy for improving domain generalization in ASC.

2. BACKGROUND

2.1. Domain Shift in Data-Efficient ASC

Domain shift, particularly induced by device variability, is a well-established challenge in Acoustic Scene Classification (ASC) [16]. Differences in microphone frequency response and spatial positioning introduce systematic shifts in audio characteristics across devices [17]. These discrepancies can significantly impair cross-device generalization, as models trained on one device distribution often fail to transfer effectively to others [18]. The impact is especially pronounced under limited labeled data, where insufficient device diversity prevents the model from learning robust, domain-invariant representations [3].

When training data covers a diverse range of scenes and recording devices, domain-invariant features can emerge naturally, as the model learns to disentangle scene-relevant cues from device-specific artifacts [5]. In contrast, under data-efficient settings, such coverage is lacking, making the learning process prone to overfitting device characteristics. For instance, if recordings corresponding to a specific scene are predominantly collected using a single device, the model may implicitly associate the spectral properties of that device with the scene label [19]. Such spurious correlations introduce domain leakage and compromise robustness [16], [20], as the model relies on incidental device-dependent patterns rather than genuine scene-discriminative features. Furthermore, early exposure to domain-specific samples may bias the internal representation space of the model, steering the learning process toward device-dependent features [21].

The above findings suggest that, under limited supervision, it may be beneficial to control the order in which training samples are presented. By prioritizing domain-invariant examples early in training, the model may better avoid overfitting to device-specific cues. This intuition motivates the use of a curriculum-based learning schedule to improve generalization without altering model architecture.

2.2. Entropy-Guided Curriculum Learning

Curriculum Learning (CL) is a training paradigm inspired by the human learning process, in which knowledge is acquired progressively—from simple concepts to more complex ones [11]. By exposing the model to easier examples in the early stages of training and gradually increasing sample difficulty, CL has been shown to improve optimization stability and generalization, particularly in settings with limited labeled data [21], [22].

A key component of curriculum learning is the definition of sample difficulty, which determines the order in which training examples are presented. Existing curricula commonly rely on heuristics such as loss magnitude [22], prediction confidence [23], classification margin [21], or uncertainty estimates derived from Bayesian inference [24]. While these strategies are effective in many supervised learning tasks, their core assumption—that task performance is a valid proxy for sample difficulty—becomes untenable in the context of domain shift. The reason is that task performance is an inherently ambiguous metric: a sample with low classification loss may be either genuinely simple and domain-invariant, or deceptively easy due to spurious, domain-specific cues. This ambiguity renders such a metric an unsound foundation for any curriculum whose objective is to improve generalization.

To address this limitation, we propose to define sample difficulty in terms of domain uncertainty rather than task performance. The core idea is to prioritize training on examples that are minimally influenced

by domain-specific features, and are therefore more likely to generalize across domains. To quantify domain uncertainty, we employ an auxiliary domain classifier to estimate the device posterior distribution for each sample. The Shannon entropy of this distribution serves as a proxy for domain invariance: samples with higher entropy indicate greater ambiguity in device identity, suggesting a lower reliance on device-dependent features [15]. By organizing data according to this entropy-based measure, we construct a curriculum that gradually transitions from domain-invariant to domain-specific examples, thereby facilitating the learning of more generalizable representations. While curriculum learning has been explored in domain generalization for tasks like semantic segmentation [13] and fault diagnosis [14], its application to acoustic scene classification remains largely unexplored.

3. PROPOSED METHOD: ENTROPY-GUIDED CURRICULUM LEARNING

3.1. Overview and Problem Setup

We propose an entropy-guided curriculum learning strategy to enhance the generalization capability of ASC models under domain shift. The strategy comprises three stages: (1) quantifying domain uncertainty using an auxiliary device classifier, (2) constructing a training curriculum by ranking samples based on entropy, (3) performing staged model training from domain-invariant to domain-specific samples.

Each ASC model consists of a feature extractor f_{feat} and a scene classifier f_{cls}. To construct the curriculum, we substitute f_{cls} with a lightweight device classifier f_{dom}, which is trained to predict device identities. For each training sample x, the device classifier produces a posterior probability distribution over device labels. We compute the Shannon entropy $H(x)$ of this distribution to measure domain ambiguity. A higher entropy indicates greater uncertainty in device identity, implying stronger domain invariance.

We rank all samples by entropy and split them at the median: the top 50% form the domain-invariant subset $\mathcal{X}_{\text{inv}}$, while the remaining form the domain-specific subset $\mathcal{X}_{\text{spec}}$. Training begins exclusively on $\mathcal{X}_{\text{inv}}$ and progressively incorporates $\mathcal{X}_{\text{spec}}$ through mixed mini-batches. This method requires no changes to model architecture or hyperparameters, ensuring seamless integration into existing ASC models.

3.2. Entropy-Based Curriculum Construction

To estimate domain uncertainty, we first freeze the feature extractor f_{feat} and train an auxiliary device classifier f_{dom} based on it. Given an sample x, the classifier produces a posterior distribution over device:

$$\mathbf{p}_d(x) = \text{softmax}(f_{\text{dom}}(f_{\text{feat}}(x))) \tag{1}$$

The classifier is trained using device labels and the standard cross-entropy loss. After training, we compute the Shannon entropy of the predicted distribution for each sample:

$$H(x) = -\sum_{i=1}^{D} p_d^{(i)}(x) \log p_d^{(i)}(x) \tag{2}$$

where D is the number of recording devices. Higher entropy indicates greater ambiguity in device identity, which implies that the sample is less influenced by device-specific characteristics and therefore exhibits stronger domain invariance.

Let $\{x_i\}_{i=1}^{N}$ denote the training set. We sort all samples based on their entropy values and partition them into two curriculum subsets:

$$\mathcal{X}_{\text{inv}} = \{x_i \in \mathcal{D}_{\text{train}} \mid \text{rank}(H(x_i)) \leq \lfloor 0.5N \rfloor\} \tag{3}$$

$$\mathcal{X}_{\text{spec}} = \mathcal{D}_{\text{train}} \setminus \mathcal{X}_{\text{inv}} \tag{4}$$
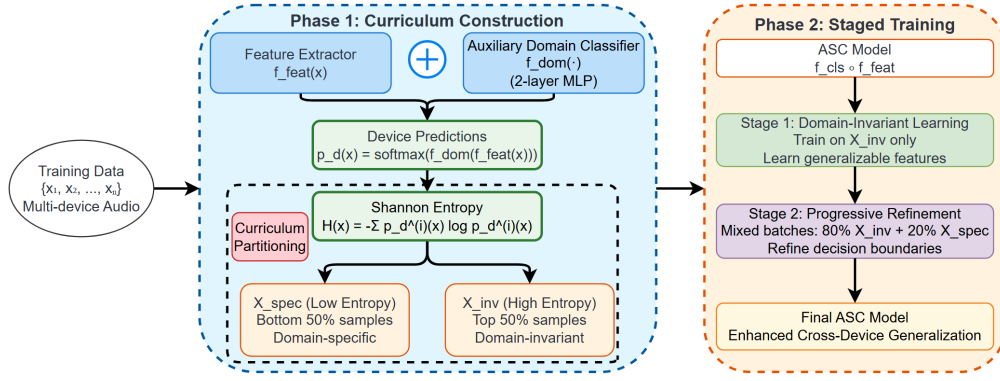
Fig. 2: Overview of the proposed entropy-guided curriculum learning framework. In Phase 1, an auxiliary domain classifier estimates the device posterior distribution for each training sample, from which entropy is computed to quantify domain uncertainty. Based on entropy values, samples are partitioned into domain-invariant (X_{inv}) and domain-specific (X_{spec}) subsets. In Phase 2, the ASC model is initially trained on X_{inv} to promote domain-invariant feature learning, followed by a staged refinement phase using mixed batches composed of both X_{inv} and X_{spec}, thereby enhancing robustness to domain shift.

We adopt a median-based entropy split as a simple yet effective proxy for curriculum ordering, selecting the top 50% high-entropy samples as domain-invariant. This empirical threshold ensures a balanced curriculum distribution between domain-invariant and domain-specific examples. While this strategy is effective, we acknowledge that sample difficulty may vary continuously. Future work will explore adaptive or weighted curriculum scheduling schemes.

In the second training stage, training mini-batches $\mathcal{B} \subset \mathcal{D}_{\mathrm{train}}$ are constructed using a fixed-ratio sampling strategy:

$$\mathcal{B} = \mathcal{B}_{\mathrm{inv}} \cup \mathcal{B}_{\mathrm{spec}}, \quad |\mathcal{B}_{\mathrm{inv}}| = \lfloor 0.8B \rfloor, \quad |\mathcal{B}_{\mathrm{spec}}| = B - |\mathcal{B}_{\mathrm{inv}}| \quad (5)$$

The gradual incorporation of domain-specific samples facilitates stable convergence and improves generalization across unseen devices.

3.3. Curriculum-Guided Model Training

After curriculum construction, we reinstate the original scene classifier f_{cls} and train the full model $f_{\mathrm{cls}} \circ f_{\mathrm{feat}}$.

Training is performed using the standard cross-entropy loss:

$$\mathcal{L}_{\mathrm{ASC}} = -\sum_{i=1}^{C} y^{(i)} \log \hat{y}^{(i)}, \quad \hat{y} = \mathrm{softmax}(f_{\mathrm{cls}}(f_{\mathrm{feat}}(x))) \quad (6)$$

where C is the number of scene classes and y is the one-hot label.

The model is trained in two stages:

- **Stage 1:** The model is first trained exclusively on $\mathcal{X}_{\mathrm{inv}}$ to facilitate the learning of domain-invariant representations.
- **Stage 2:** To enhance cross-device generalization, the model is progressively exposed to domain-specific inputs by constructing each mini-batch with a fixed 80:20 ratio of samples from $\mathcal{X}_{\mathrm{inv}}$ and $\mathcal{X}_{\mathrm{spec}}$, respectively. This controlled inclusion enables the model to adapt to domain-specific cues while preserving the domain-invariant features learned during Stage 1.

The transition from Stage 1 to Stage 2 is triggered when the validation loss on $\mathcal{X}_{\mathrm{inv}}$ stops improving beyond a predefined threshold for several consecutive epochs. This reflects the core principle of curriculum learning: once the model has sufficiently learned from easier, domain-invariant samples, it is gradually exposed to more challenging, domain-specific data. Throughout both stages, we strictly follow the original training configurations of each ASC baseline, including the optimizer, learning rate schedule, and batch size. This ensures fair comparison and highlights the plug-and-play compatibility of our method. By progressively shifting the training focus, the staged curriculum promotes robust representation learning and improves generalization to unseen devices.

4. EXPERIMENT

4.1. Experimental Setup

4.1.1. Datasets: Our experimental evaluation is performed using the DCASE 2024 Task 1 dataset [25], a benchmark derived from the TAU Urban Acoustic Scenes 2022 Mobile development set for the Acoustic Scene Classification (ASC) task. The dataset consists of one-second audio clips representing ten distinct acoustic scenes, recorded across 12 European cities with multiple real and simulated mobile devices. In total, the development set contains 230,350 audio segments, amounting to approximately 64 hours of audio.

As detailed in Table 1, the dataset is partitioned into training and testing sets with a specific device distribution. The training set includes labeled data from real devices (A, B, C) and simulated devices (S1, S2, S3). A significant characteristic of the training data is its imbalance, with Device A contributing a substantial majority of the segments. The test partition comprises data from all training devices and, critically, introduces three previously unseen simulated devices (S4, S5, S6) to assess the generalization capabilities of the models.

Table 1: Distribution of DCASE 2024 Task 1 Development Dataset

Devices	Type	Total segments	Train segments	Test segments
A	Real	144,000	102,150	3,300
B	Real	10,780	7,490	3,290
C	Real	10,770	7,480	3,290
S1, S2, S3	Simulated	3 × 10,800	3 × 7,500	3 × 3,300
S4, S5, S6	Simulated	3 × 10,800	-	3 × 3,300
Total		**230,350**	**139,620**	**29,680**

To investigate performance under data-limited conditions, we strictly adhere to the official low-resource protocol of the DCASE 2024 Challenge. This protocol provides five predefined training subsets drawn from the multi-device data (A, B, C, S1, S2, S3), which contain 5%, 10%, 25%, 50%, and 100% of the total labeled data, respectively. All models in our study are trained exclusively on these specified low-resource subsets.

4.1.2. Evaluation Metrics: Model performance is primarily evaluated using the official DCASE 2024 Task 1 metric: class-wise average accuracy. This metric computes the mean of per-class

Table 2: Classification accuracy (%) on the DCASE2024 Task 1 evaluation set across different training data splits. Each system is compared with and without our proposed curriculum learning strategy. The last two columns report accuracy on seen and unseen devices under the 5% training condition.

System	5%	10%	25%	50%	100%	Seen (5%)	Unseen (5%)
DCASE2024 Baseline [26]	44.00	46.95	51.47	54.40	56.84	45.3	42.4
+ our strategy	**46.30**	**49.10**	**52.80**	**55.20**	**57.00**	**47.5**	**44.0**
Cai_XJTLU [7]	48.91	53.16	58.09	59.47	62.12	50.6	46.7
+ our strategy	**51.50**	**55.10**	**59.50**	**61.25**	**63.20**	**52.3**	**49.3**
Han_SJTUTHU [27]	54.35	56.69	59.09	60.38	61.82	55.6	52.7
+ our strategy	**56.60**	**58.25**	**60.40**	**61.65**	**62.20**	**57.3**	**55.2**

accuracies, ensuring a balanced performance assessment across all scene categories. It is defined as:

$$\text{Accuracy}_{\text{avg}} = \frac{1}{C} \sum_{c=1}^{C} \frac{N_c^{\text{correct}}}{N_c} \qquad (7)$$

where C denotes the total number of scene classes, N_c represents the total number of samples belonging to class c, and N_c^{correct} is the count of correctly classified samples for that specific class.

4.1.3. Baseline Systems: To validate the general applicability of our entropy-guided curriculum strategy, we evaluate three representative baseline systems from DCASE2024:

1) **DCASE2024 Official Baseline** [26]: employs CP-Mobile, a CNN with residual inverted bottleneck blocks. It utilizes knowledge distillation from a teacher ensemble (PaSST and CP-ResNet), with augmentations like Freq-MixStyle and device impulse response simulation.

2) **Cai_XJTLU** [7] (ranked 4th): adopts a teacher–student approach with a pretrained PaSST teacher and a TF-SepNet-64 student trained via knowledge distillation. Training includes SpecAugment, Mixup, noise addition, and pseudo-labeling.

3) **Han_SJTUTHU** [27] (ranked 1st): uses a student–teacher framework with SSCP-Mobile (a spatially separable CP-Mobile variant) as the student, and a PaSST teacher ensemble. Techniques include knowledge distillation, SpecAugment, Freq-MixStyle, and model pruning.

4.1.4. Implementation Details: Our entropy-guided curriculum learning strategy is implemented in an architecture-agnostic manner. Accordingly, we follow each baseline system's original training protocol, including optimizer, learning rate schedule, batch size, and other hyperparameters.

- **Auxiliary Domain Classifier:** A lightweight two-layer Multi-Layer Perceptron (MLP) is used to predict device identities from features extracted by the frozen primary encoder f_{feat}. The MLP comprises one hidden layer with ReLU activation and a softmax output layer, and is optimized using Adam with a learning rate of 1×10^{-4}.

- **Curriculum-Guided Training:** The ASC model is trained in two stages as described in Section 3.3. The transition to Stage 2 is based on validation performance on $\mathcal{X}_{\text{inv}}$, and training follows the fixed 80:20 sampling strategy. All models are trained under their respective official settings to ensure fair comparison.

4.2. Results and Analysis

Table 2 shows the classification accuracy on the DCASE2024 Task 1 evaluation set across varying proportions of labeled training data. Performance improvements introduced by our strategy are particularly notable under data-limited conditions (e.g., 5%–25% of training data), where models are more susceptible to domain shift.

To better assess the impact of our method on domain generalization, we decompose performance under the 5% training condition into
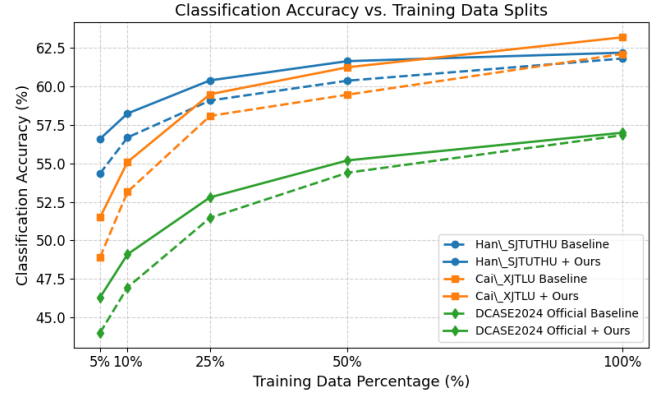
Fig. 3: Classification accuracy across varying training data splits for different baseline systems, with and without our proposed strategy.

seen-device and unseen-device subsets, where domain shift is most pronounced. As shown in the additional columns of Table 2, our strategy consistently improves both subsets, with notably larger gains on unseen devices. For instance, in the Cai_XJTLU, accuracy on unseen devices increases from 46.7% to 49.3% (+2.6%), whereas the improvement on seen devices is comparatively modest (+1.7%).

We further illustrate these improvements in Fig. 3, which presents accuracy curves across varying data splits. The curves clearly show that our curriculum strategy consistently enhances model generalization across different data regimes. Specifically, improvements gradually diminish as more training data becomes available, which is consistent with the expectation that abundant labeled data already facilitates effective learning of domain-invariant features, thereby reducing the relative benefits of training strategies.

Overall, these results confirm the efficacy of our entropy-guided curriculum learning strategy in improving ASC model generalization under domain shift, particularly in scenarios characterized by limited labeled training data. Importantly, the proposed method does not introduce additional inference overhead or architectural complexity, making it a practical solution for ASC tasks under domain shift.

5. CONCLUSION

We proposed an entropy-guided curriculum learning strategy to address domain shift in ASC under limited labeled data. By employing an auxiliary domain classifier to estimate domain uncertainty via entropy, the training process is structured to progress from domain-invariant to domain-specific samples, facilitating the acquisition of robust and generalizable representations. Experimental results on multiple DCASE2024 Task 1 baseline systems demonstrate consistent performance gains across training regimes, with particularly notable improvements in low-resource scenarios. The proposed strategy is architecture-agnostic, incurs no additional inference overhead, and integrates seamlessly with existing ASC models.

ACKNOWLEDGEMENTS

REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[2] Y. Cai, P. Zhang, and S. Li, "Tf-sepnet: An efficient 1d kernel design in cnns for low-complexity acoustic scene classification," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 821–825.

[3] F. Schmid, P. Primus, T. Heittola, A. Mesaros, I. Martín-Morató, K. Koutini, and G. Widmer, "Data-efficient low-complexity acoustic scene classification in the dcase 2024 challenge," *arXiv preprint arXiv:2405.10018*, 2024.

[4] Y. Tan, H. Ai, S. Li, and G. Fazekas, "Domain adaptation of few-shot bioacoustic event detection in different environments," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[5] F. Schmid, S. Masoudian, K. Koutini, and G. Widmer, "Cp-jku submission to dcase22: Distilling knowledge for low-complexity convolutional neural networks from a patchout audio transformer," *Tech. Rep., Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2022.

[6] Y. Cai, M. Lin, C. Zhu, S. Li, and X. Shao, "Dcase2023 task1 submission: Device simulation and time-frequency separable convolution for acoustic scene classification," *Tech. Rep., Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2023.

[7] Y. Cai, M. Lin, S. Li, and X. Shao, "Dcase2024 task1 submission: Data-efficient acoustic scene classification with self-supervised teachers," DCASE Challenge, Tech. Rep, Tech. Rep., 2024.

[8] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[9] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[10] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 8, pp. 8052–8072, 2022.

[11] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[12] J. Liu, Y. Chen, H. Liu, H. Zhang, and Y. Zhang, "From less to more: Progressive generalized zero-shot detection with curriculum learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19016–19029, 2022.

[13] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2020–2030.

[14] Y. Wang, J. Gao, W. Wang, X. Yang, and J. Du, "Curriculum learning-based domain generalization for cross-domain fault diagnosis with category shift," *Mechanical Systems and Signal Processing*, vol. 212, p. 111295, 2024.

[15] A. Arsenos, D. Kollias, E. Petrongonas, C. Skliros, and S. Kollias, "Uncertainty-guided contrastive learning for single source domain generalisation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6935–6939.

[16] Y. Tan, H. Ai, S. Li, and M. D. Plumbley, "Acoustic scene classification across cities and devices via feature disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1286–1297, 2024.

[17] J. Bai, M. Wang, H. Liu, H. Yin, Y. Jia, S. Huang, Y. Du, D. Zhang, D. Shi, W.-S. Gan *et al.*, "Description on ieee icme 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift," *arXiv preprint arXiv:2402.02694*, 2024.

[18] T. Morocutti, F. Schmid, K. Koutini, and G. Widmer, "Device-robust acoustic scene classification via impulse response augmentation," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 176–180.

[19] Y. Yan, W. Liu, Y. Chai, and Y. Ren, "Semi-supervised acoustic scene classification under domain shift with mixmatch and information bottleneck optimization," in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 2024, pp. 1–4.

[20] Y. Liu, P. Zhang, R. Sang, Z. Li, and S. Li, "Maia: An inpainting-based approach for music adversarial attacks," 2025. [Online]. Available: https://arxiv.org/abs/2509.04980

[21] G. Hacohen and D. Weinshall, "On the power of curriculum learning in training deep networks," in *International conference on machine learning*. PMLR, 2019, pp. 2535–2544.

[22] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *International conference on machine learning*. PMLR, 2018, pp. 2304–2313.

[23] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 322–330.

[24] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[25] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, pp. 56–60. [Online]. Available: https://arxiv.org/abs/2005.14623

[26] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, "Distilling the knowledge of transformers and CNNs with CP-mobile," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, 2023, pp. 161–165.

[27] H. Bing, H. Wen, C. Zhengyang, J. Anbai, C. Xie, F. Pingyi, L. Cheng, L. Zhiqiang, L. Jia, Z. Wei-Qiang, and Q. Yanmin, "Data-efficient acoustic scene classification via ensemble teachers distillation and pruning," DCASE2024 Challenge, Tech. Rep., May 2024.