Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37840

Cross-DATA: Cross-Data Attention Distillation for Lightweight Sound Classification with Limited Data

Fengkai Qu

School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China
18853577160@163.com

Abstract. Using the UrbanSound8K dataset as a representative case, this study investigates how lightweight acoustic models can be trained with limited labeled data in resource-constrained environments. Because conventional knowledge distillation confines knowledge transfer to a single target dataset, it leaves the rich prior knowledge in external audio sources largely unused. We therefore propose Cross-DATA, a framework that treats a large-scale, general audio dataset such as FSD50K as an auxiliary knowledge source and integrates a newly designed Cross-data Attention Distillation mechanism. With mixed data batches as input, a powerful teacher model establishes inter-feature correlation patterns that the compact student network is encouraged to reproduce. In this way, the teacher's generic acoustic representations and discriminative capabilities can be transferred beyond the distribution of the target data, rather than being reduced to data augmentation or feature imitation. Experiments on UrbanSound8K show that the resulting student model surpasses traditional single-dataset distillation approaches in classification accuracy while using only about 4% of the teacher model's parameters and computational cost. Cross-DATA consequently alleviates the representational capacity deficit of lightweight models in small-data scenarios.

Keywords: Cross-data attention distillation, Knowledge distillation, Lightweight sound classification, Edge computing

1. Introduction

Edge computing has become a major platform for deploying artificial intelligence models on mobile devices, embedded systems, and wearable devices as the Internet of Things (IoT) and smart devices have rapidly evolved [1, 2]. In applications such as smart homes, industrial monitoring, autonomous driving, and environmental sensing, audio classification supports intelligent decision-making and safety monitoring by recognizing acoustic events in real time [3]. Directly deploying large-scale deep neural networks, however, is generally impractical because edge devices provide limited computational power, memory capacity, and energy resources [4]. These constraints make the development of lightweight, high-accuracy audio classification models suitable for edge deployment an important research challenge [5, 6].

Limited labeled data remains a major obstacle even when lightweight neural network architectures are available for audio classification [7]. Practical tasks, including industrial anomaly

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37840

detection and specialized environmental sound recognition, are often supported by relatively small datasets [8]; UrbanSound8K, for instance, includes only 8,732 audio samples. Such conditions can cause large models to overfit, while lightweight models may not possess enough representational capacity to learn complex acoustic patterns [9]. Lightweight audio classification must therefore address two constraints simultaneously: strict model complexity requirements and inadequate training data.

Existing approaches also make insufficient use of external acoustic knowledge [10]. Traditional training and knowledge distillation typically operate only on the target dataset, even though large-scale datasets such as AudioSet and FSD50K contain diverse acoustic information that could strengthen model generalization. Transferring this generalized knowledge effectively is thus essential to compensate for the limited representational capacity of lightweight models and improve their performance in small-data scenarios.

Cross-DATA (Cross-data Attention Distillation) is proposed to resolve these limitations by using a large-scale external audio dataset as an auxiliary knowledge source for knowledge distillation. The framework forms mixed batches from target and auxiliary data, allowing the student model to acquire generalized acoustic representations from the teacher model. Rather than merely aligning feature values or output predictions, a cross-data attention distillation loss transfers inter-feature correlation patterns. On UrbanSound8K, Cross-DATA yields substantially better classification performance than conventional single-dataset distillation and other representative baselines, although the lightweight model requires only approximately 4% of the teacher model's parameters and computational cost.

2. Method

2.1. The cross-data framework

Cross-DATA transfers generalized acoustic knowledge from a powerful teacher model to a lightweight student model when target-domain data are limited. As shown in Figure 1, an auxiliary large-scale audio dataset broadens the teacher's knowledge, whereas the student remains dedicated to the target classification task. Throughout training, the fixed teacher supplies both output-level and feature-level supervision for student optimization. Let $\mathcal{D}_{target} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_t}$ represent the labeled target dataset and $\mathcal{D}_{aux} = \{\mathbf{x}_j\}_{j=1}^{N_a}$ the auxiliary dataset. At every iteration, samples from the two datasets form the mixed batch $\mathbf{X}_{mix} = Concat(\mathbf{X}_{target}, \mathbf{X}_{aux})$, with the mixing ratio γ controlling the share of auxiliary samples. By processing this mixed batch, the teacher produces feature representations, prediction logits, and attention maps that encode acoustic correlations across target and auxiliary data.

Only target samples are presented to the lightweight student model. From these inputs, the student generates task predictions and intermediate feature representations used to construct its attention map. This asymmetric training strategy lets the teacher exploit external-data diversity without diverting the student from target-domain classification. Knowledge is then conveyed through softened output distributions together with attention-based feature correlations.

Three complementary loss functions jointly guide student optimization. Ground-truth target labels first provide direct supervision through the classification loss $L_{cls} = CE(\mathbf{p}_s, \mathbf{y}_{target})$. Output-level knowledge is then transferred by minimizing the discrepancy between the teacher's and student's softened prediction distributions, giving $L_{kd} = \tau^2 KL\left(\mathbf{p}_s^{\tau}, \mathbf{p}_t^{\tau}[:m]\right)$. Here, τ is the distillation temperature, and alignment uses only teacher predictions associated with target samples.

DOI: 10.54254/2755-2721/2026.37840

The cross-data attention distillation loss constitutes the central element of Cross-DATA. From mixed target and auxiliary data, the teacher derives attention representations that express feature correlation patterns across diverse acoustic samples. The student attention map is aligned with the target-related portion of the teacher attention map through $L_{cdad} = \frac{1}{m^2} \|\mathbf{A}_s - \mathbf{A}_t[:m, :m]\|_F^2$. The full training objective is $L_{total} = \alpha L_{cls} + \beta L_{kd} + \lambda L_{cdad}$, where α, β and λ are weighting coefficients. Jointly minimizing these objectives enables the student to solve the target classification task while inheriting prediction knowledge and cross-dataset feature correlation patterns from the teacher.

Training begins by pre-training or fine-tuning the teacher model so that it develops generalized acoustic representations and task-specific discriminative capability. Its parameters are subsequently fixed, while the randomly initialized student model is optimized over multiple training epochs.

At each iteration, the teacher processes a mixed batch assembled from target and auxiliary samples, producing soft predictions and attention representations from the combined data. The student simultaneously processes only the target samples and returns its corresponding predictions and attention representations. The classification loss, knowledge distillation loss, and cross-data attention distillation loss are combined into the overall objective. Backpropagation updates only student parameters; the teacher remains unchanged. Once training ends, deployment retains solely the optimized lightweight student model.

2.2. Complexity analysis and practical considerations

Let $B = m + n$ be the mixed batch size, and let T_t and T_s denote the computational costs of the teacher and student forward passes. Per-iteration computational complexity is approximated by $O(T_t + T_s + C_{att})$, in which C_{att} represents attention computation and alignment. With direct pairwise attention, this additional cost is usually quadratic in batch size, namely $O(B^2)$; efficient attention mechanisms can reduce the overhead further.

Model parameters, intermediate feature activations, and attention representations account for most memory consumption. Because the teacher is fixed, student optimization does not require storage of teacher-parameter gradients, thereby lowering the training memory requirement. Relative to conventional knowledge distillation, Cross-DATA adds teacher computation for auxiliary samples and attention alignment, but both operations are confined to training.

Neither the auxiliary dataset nor the teacher model is needed at inference time. Edge devices deploy only the lightweight student model after training. The framework can therefore strengthen the student's representation and generalization capabilities without adding computational or memory overhead to practical inference.

3. Experiment

3.1. Datasets and preprocessing

UrbanSound8K, a widely used benchmark for urban sound classification, is used to evaluate Cross-DATA. The dataset provides 8,732 labeled sound clips across 10 classes, and each clip lasts approximately 4 seconds on average. Following the standard evaluation protocol, we use the official 10-fold cross-validation split and report mean classification accuracy (%) together with the standard deviation across folds. FSD50K serves as the external auxiliary knowledge source and contains 51,197 audio clips from 200 sound-event categories. Machine sounds, human activities, and natural

DOI: 10.54254/2755-2721/2026.37840

backgrounds contribute diverse acoustic patterns that enrich the teacher's prior knowledge, although FSD50K labels are never used as target labels. More specifically, the official 200-class FSD50K taxonomy does not overlap with the 10 UrbanSound8K classes (air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren, and street_music). Auxiliary samples are never regarded as labeled instances of an UrbanSound8K class, preventing label leakage between the datasets.

All audio files are resampled to 16 kHz and converted to mono-channel before feature extraction. Using a Hann window of 2048 samples (128 ms) and a hop length of 512 samples (32 ms), we derive 64-dimensional log-Mel spectrograms that produce a 64 × 128 time-frequency representation for a nominal 4-second clip. Clips below 4 s are zero-padded, while longer recordings are centrally truncated to maintain a consistent input size. Random time stretching (20%) and pitch shifting are additionally applied during training to improve robustness and generalization.

3.2. Model architectures

The teacher model is PANNs (Pre-trained Audio Neural Networks), specifically the Wavegram-Logmel-CNN14 variant pre-trained on AudioSet. It comprises 14 convolutional blocks, an attention pooling head, and 2048-dimensional embeddings; the final classification layer is replaced for target-task adaptation. Fine-tuning proceeds for 20 epochs on FSD50K with 200-way cross-entropy and then for 15 epochs on the UrbanSound8K training split. The resulting teacher is frozen and supplies only soft labels and attention representations during distillation. It contains approximately 81M parameters and requires about 16.0 GFLOPs per clip. The lightweight student couples a narrowed MobileNetV2 backbone (width multiplier 0.35) with a two-layer BiGRU for temporal modeling. MobileNetV2 processes the input $64 times 128$ Mel-spectrogram; its feature maps are average-pooled over frequency and passed to the BiGRU with 128 hidden units per direction, followed by a linear classifier and softmax layer.

With approximately 3.2×10^6 parameters and 0.62 GFLOPs per clip, the student uses about 4% of the teacher's parameters and is appropriate for edge deployment. In every training iteration, B_t labeled target clips from UrbanSound8K, whose labels are $y \in \{0, \ldots, 9\}$, are concatenated with B_a auxiliary FSD50K clips to form $\mathbf{X}_{mix} = [\mathbf{X}_t; \mathbf{X}_a]$. The teacher processes this combination to build attention maps under a richer cross-data context, whereas the student receives only $\mathbf{X}_t$. The classification loss $L_{cls} = CE(\mathbf{p}_s, \mathbf{y})$ is evaluated exclusively on the B_t target samples, and auxiliary positions are excluded through $\texttt{ignore_index} = -100$. Consequently, FSD50K's 200-class labels never enter the 10-class classification loss. This masking design prevents label leakage while retaining the benefit of generalized acoustic knowledge learned by the teacher from both datasets.

3.3. Hyperparameter settings and training details

The principal hyperparameters for Cross-DATA and the baseline methods are summarized below. Unless noted otherwise, settings were selected by grid search on a held-out validation fold (Fold 1 of UrbanSound8K) and then fixed for all methods to support a fair comparison. AdamW optimizes the student for at most 50 epochs with an initial learning rate of 3×10^{-4}, weight decay of 1×10^{-4}, cosine annealing, and early stopping with patience 1. Each mixed batch contains 32 samples: 16 target and 16 auxiliary samples, corresponding to $\gamma = 0.5$. Knowledge distillation uses $\tau = 3$, and the classification, knowledge distillation, and cross-data attention distillation loss weights are $(\alpha, \beta, \lambda) = (1.0, 0.5, 1.5)$. Each coefficient was selected from $\{0.1, 0.5, 1.0, 1.5, 2.0\}$, while τ was

DOI: 10.54254/2755-2721/2026.37840

chosen from $\{1,2,3,4,5\}$ according to validation accuracy. Scaled dot-product self-attention with four heads is used in the attention module. For the student, the MobileNetV2 width multiplier is 0.35, the two-layer BiGRU has 128 hidden units per direction, and the classifier head uses dropout 0.2. Teacher adaptation consists of 20 FSD50K fine-tuning epochs with Adam at 1×10^{-4}, followed by 15 UrbanSound8K fine-tuning epochs under the same optimizer and learning rate. The teacher is then frozen for knowledge generation during student training.

3.4. Computational resources and inference efficiency

All experiments ran on a server with two NVIDIA RTX 3090 GPUs (24GB VRAM each), 64GB RAM, and an Intel Xeon Gold 6226R CPU. Teacher adaptation took approximately 8h in Stage 1 and 1.5h in Stage 2. Per fold, Cross-DATA student training required about 3h, compared with approximately 2h for vanilla knowledge distillation, because mixed-batch teacher forward passes and attention-map computation increase training cost. These operations disappear at inference, when the execution graph contains only MobileNetV2+BiGRU. The student consumes approximately 0.62 GFLOPs and 28MB peak memory, with measured latency of about 8.3ms per clip on an RTX 3090 and 45.2ms per clip on a Raspberry Pi-4B equipped with an ARM Cortex-A72 processor. These results indicate that deployment remains feasible on resource-constrained edge devices.

3.5. Reproducibility and baseline implementations

For an apple-to-apple comparison, the baselines share Cross-DATA's student input features, data augmentation, 10-fold cross-validation protocol, and optimizer settings unless a method inherently requires another configuration. Student (Scratch) retains the same lightweight backbone, augmentation strategy, 50-epoch cosine-annealing schedule, and early stopping, but eliminates the teacher and auxiliary data and minimizes only L_{cls}; it therefore provides the lower-bound reference. Vanilla KD freezes the teacher and optimizes $L_{cls} + \beta L_{kd}$ using target-only batches with $\gamma = 0$ and $\tau = 3$. FSD50K Pre-train + Fine-tune trains the student for 20 epochs on FSD50K using 200-way cross-entropy before CE-only fine-tuning on UrbanSound8K, without teacher guidance. MixUp + Mixed-data linearly interpolates Mel-spectrogram pairs sampled from target and auxiliary datasets using $\alpha_{mix} = 0.2$, then trains the student with CE and no explicit teacher supervision.

Contrastive Distillation maps teacher and student embeddings onto a 128-dimensional unit sphere through a two-layer projection head. It combines standard CE with a symmetrised InfoNCE objective at $T = 0.1$ and applies stop-gradient to the teacher branch. Teacher and student representations of the same clip form positive pairs, whereas the other clips in the mixed batch serve as negatives. MAML uses 5-way 5-shot tasks drawn from UrbanSound8K and tunes the inner-loop learning rate for every fold. EfficientNet-B0 is also evaluated as an architecture reference with the same input and feature pipeline and a comparable computational budget, but it is trained from scratch on UrbanSound8K. This comparison tests whether Cross-DATA's gain originates from cross-data distillation rather than merely from adopting a newer backbone.

3.6. Results

Table 1 compares the high-capacity teacher with lightweight students trained under different strategies. PANNs provides the 92.1% accuracy reference but requires 81.0M parameters and 16.0G FLOPs per clip. Training the 3.2M-parameter, 0.62G-FLOP student from scratch lowers accuracy to 77.3%, revealing a 14.8-point compression deficit even though the deployed model uses

DOI: 10.54254/2755-2721/2026.37840

approximately 4% of the teacher's parameters and less than 4% of its listed FLOPs. Additional supervision recovers this loss to different degrees: Vanilla KD reaches 81.0%, mixed-data pre-training 80.5%, Contrastive Distillation 80.8%, and MAML 79.5%. With the same 3.2M-parameter scale and 0.62G computational budget, Cross-DATA obtains 84.7%. This corresponds to gains of 7.4 points over Student (Scratch), 3.7 over Vanilla KD, 4.2 over mixed-data pre-training, 3.9 over contrastive supervision, and 5.2 over MAML, thereby recovering exactly half of the teacher--scratch gap. Because the deployed network size and FLOPs remain unchanged and auxiliary samples are not treated as UrbanSound8K labels, the improvement supports relational attention supervision rather than increased inference capacity or label leakage. Cross-DATA thus offers the strongest accuracy among the listed compact-model strategies under the stated 10-fold protocol, although PANNs retains the highest absolute accuracy and Table 1 does not report standard deviations.

Table 1. Comparison with state-of-the-art and baselines on UrbanSound8K (10-fold CV). All methods use log-Mel features except where noted

Method	Parames (M)	FLOPs (G)	Mem (MB)	Time (RTX 3090)	Time (RPi 4B)	Ace (%)
Teacher(PANNS)	81.0	16.0	320	48.0	-	92.1
Student(scratch)	3.2	0.62	28	8.3	45.2	77.3
Vanilla KD	3.2	0.62	28	8.3	45.2	81.0
MixUp+Pre-train	3.2	0.62	28	8.3	45.2	80.5
Contrastive (w/o label)	3.2	0.62	28	9.1	46.5	80.8
MAML	3.2	0.62	28	8.5	45.5	79.5
Cross-DATA (ours)	3.2	0.62	28	8.3	45.2	84.7

Table 2 shows that Cross-DATA's accuracy improvement does not enlarge the deployed model. Student (Scratch), Vanilla KD, and Cross-DATA each use 3.2M parameters, 0.62G FLOPs per clip, approximately 28MB peak working memory, 8.3ms latency on the RTX 3090, and 45.2ms latency on the Raspberry Pi-4B. This identical inference profile follows directly from the training design: teacher forward passes, mixed batches, and attention-map matching are discarded after checkpoint selection, leaving only the MobileNetV2+BiGRU student. By comparison, the teacher achieves 92.1% accuracy with 81M parameters, 16.0G FLOPs, approximately 320MB peak memory, and 48ms RTX latency; no Raspberry Pi result is reported. Relative to this teacher, the Cross-DATA student reduces parameter count and FLOPs by roughly 96%, peak memory by approximately 91%, and RTX latency by about 5.8 times, at an accuracy cost of 7.4 points. Cross-DATA nevertheless raises compact-model accuracy from 77.3% for Student (Scratch) and 81.0% for Vanilla KD to 84.7%. Its additional computation is therefore concentrated in offline training rather than continuous edge inference. Table 2 does not imply shorter training time; instead, it confirms that the method preserves the baseline student's compact inference profile while producing substantially higher accuracy.

DOI: 10.54254/2755-2721/2026.37840

Table 2. Efficiency profile (inference). Memory = peak working set during a single 4-s clip forward pass

Method	Parames (M)	FLOPs (G)	Mem (MB)	Time (RTX 3090)	Time (RPi 4B)	Ace (%)
Teacher(PANNS)	81.0	16.0	320	48.0	-	92.1
Student(scratch)	3.2	0.62	28	8.3	45.2	77.3
Vanilla KD	3.2	0.62	28	8.3	45.2	81.0
Cross-DATA (ours)	3.2	0.62	28	8.3	45.2	84.7

Table 3 separates the contributions that produce the full Cross-DATA accuracy of 84.7%. In A1, removing L_{cdad} while retaining cross-data processing and teacher adaptation reduces accuracy to 81.0%, a 3.7-point decline. Because teacher supervision and mixed target--auxiliary batches remain available, this result indicates that output-level knowledge transfer does not replace explicit attention-relation matching. A2 retains attention distillation and the adapted teacher but removes cross-data mixing; accuracy falls to 80.2%, the largest decrease at 4.5 points. The teacher target is therefore most effective when formed in a richer mixed-data context, and auxiliary data influence the transferred signal rather than merely supplying more examples. A3 preserves cross-data processing and attention loss but removes two-stage teacher adaptation, yielding 82.9% and a 1.8-point drop. Broad acoustic exposure and target-task discrimination consequently improve teacher knowledge beyond the two central mechanisms. A4 substitutes feature-value MSE for relational attention matching and reaches 82.1%, 2.6 points below the complete model. Since cross-data and teacher adaptation remain unchanged, this comparison favors matching inter-feature relations over directly equating raw feature values across different architectures and capacities. Overall, cross-data mixing enriches the teacher context, teacher adaptation makes its knowledge more relevant, and attention-relation matching transfers that context to the student. All four ablations move consistently below 84.7%; however, because Table 3 provides no confidence intervals or per-class scores, significance across folds would require further statistical analysis.

Table 3. Ablation study on UrbanSound8K (10-fold CV). √=included; ×=removed/replaced. Drops ↓ are w.r.t. the full model

Setup	L_{edad}	Cross-data	Teacher Adapt.	Ace (%)
Cross-DATA (ours)	√	√	√	84.7
A1: w/o Attn. Dist (λ=0)	×	√	√	81.0(↓3.7)
A2: w/o Cross-data.	√	×	√	80.2(↓4.5)
A3: w/o Teacher Adapt.	√	√	×	82.9(↓1.8)
A4: Feature-MSE.	(MSE)	√	√	82.1(↓2.6)

DOI: 10.54254/2755-2721/2026.37840

4. Conclusion

This study introduces Cross-DATA to improve lightweight urban sound classification with limited labeled data on resource-constrained edge devices. In contrast to conventional knowledge distillation centered on a single target dataset, the framework draws generalized acoustic knowledge from a large-scale auxiliary audio dataset such as FSD50K. Its cross-data attention distillation mechanism guides the student model to learn inter-feature correlation patterns generated by the teacher from mixed target--auxiliary data, extending supervision beyond output predictions or individual-feature imitation. UrbanSound8K experiments show that the lightweight student outperforms the evaluated baselines while retaining approximately 4% of the teacher's parameters and computational cost. Ablation and visualization studies further substantiate the roles of cross-data learning, attention distillation, and two-stage teacher adaptation. Cross-DATA nevertheless depends on teacher quality and auxiliary-data diversity and incurs extra attention-computation overhead during training. Future research will examine automated auxiliary-data selection, more efficient attention mechanisms, cross-modal knowledge distillation, and privacy-preserving distributed distillation for resource- and privacy-sensitive applications.

References

[1] Hu, R., Chu, X., Dou, D., et al. (2025). Toward real-world applicability: Lightweight underwater acoustic localization model through knowledge distillation. *IEEE Journal of Oceanic Engineering, 50*(2), 1429–1442.

[2] Huang, C., Chen, Y., Zhou, B., et al. (2025). LightEdu-Net: Noise-resilient multimodal edge intelligence for student-state monitoring in resource-limited environments. *Sensors, 25*(24), 7529.

[3] You, M., Ren, B., Han, X., et al. (2024). Cross-factory polarizer sheet surface defect inspection system based on multiteacher knowledge amalgamation. *IEEE Transactions on Instrumentation and Measurement, 73*, 1–16.

[4] Deng, W., Zhao, C., & Zhang, Z. (2026). Generalizing driver behavior recognition in complex driving scenarios via hierarchical attention network with hybrid-split strategy. *Journal of Transportation Engineering, Part A: Systems, 152*(3), 04025150.

[5] Bellafkir, H., Vogelbacher, M., Gottwald, J., et al. (2022). Bat echolocation call detection and species recognition by transformers with self-attention. In *International Conference on Intelligent Systems and Pattern Recognition* (pp. 189–203). Springer International Publishing.

[6] Jiang, B., Fei, J., Luo, F., et al. (2025). Federated learning for internet of underwater things based on lightweight distillation and data refinement. *IEEE Internet of Things Journal*.

[7] Wu, W., Lyu, Y., Zheng, W., et al. (2025). TerraDA: A domain-adaptive framework with dynamic multimodal remote sensing data fusion and consistency learning for cross-city land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

[8] Liu, K., Huang, Z., Wang, C. D., et al. (2024). Fine-grained learning behavior-oriented knowledge distillation for graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 36*(5), 9422–9436.

[9] Shen, Z., Shi, C., Shen, R., et al. (2023). Spatial downscaling of near-surface air temperature based on deep learning cross-attention mechanism. *Remote Sensing, 15*(21), 5084.

[10] Ibrahim, M. R., Benavente, R., Ponsa, D., et al. (2025). HyDA-Net: A hybrid dense attention network for remote sensing multi-image super-resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18*, 7592–7614.