OR-KDL: ORTHOGONAL KNOWLEDGE DISTILLED LORA FOR DOMAIN-AGNOSTIC INCREMENTAL LEARNING IN DCASE 2026 TASK 7

Technical Report

*Hanseul Kim, Eojin Kim, Jihyuk Lee, Junho Bae, Donghyeok Park, and Chanjun Chun**

Department of Computer Engineering, Chosun University, Gwangju, Republic of Korea
{khs020512, jjj333, dlwlgur02, rfvq56, ehdgur4277, cjchun}@chosun.ac.kr

ABSTRACT

We present OR-KDL (Orthogonal Knowledge distilled LoRA), a continual adaptation system for the domain-incremental acoustic classification setting of DCASE 2026 Task 7. A CNN14 backbone pretrained on D1 is kept frozen throughout, with adaptation to D2 and D3 performed solely via domain-specific low-rank adapters, preserving the pretrained representation while minimizing trainable parameters. Knowledge distillation (KD) is applied at each stage: logit distillation for D1→D2, and logit combined with cosine feature distillation for D2→D3 to mitigate representational drift. To prevent catastrophic forgetting without access to prior-domain data, Weight-based Orthogonal Gradient Projection (OGP) constructs protected subspaces via SVD of previous domain weights and adapter updates, projecting gradients onto their orthogonal complement. We compare standard LoRA and a CoLoRA-based adapter under this framework, and select the ensemble size N of a stratified 5-fold Top-N soft-voting ensemble on a held-out validation set. The final system CoLoRA with uniform augmentation and a Top-20 ensemble achieves D2@D3 of 79.73%, D3@D3 of 66.55%, Accuracy of 73.14%, and Forgetting of 4.04%p.

***Index Terms*—** DCASE 2026, domain-incremental learning, LoRA, CoLoRA, knowledge distillation, weight-based orthogonal gradient projection, K-Fold, class-selective augmentation

1. INTRODUCTION

This report describes the systems we submitted for the DCASE 2026 Challenge Task 7 [1]. The task is formulated as a domain-agnostic incremental learning problem, a domain-incremental learning (DIL) [2] setting in which a model must adapt to sequentially changing domains while preserving performance on previously seen domains, under a fixed label space and without access to domain identity at inference time. It comprises three acoustic domains, D1, D2, and D3, but imposes an additional constraint: D1 is provided only as a pretrained checkpoint, without any accompanying training data. Consequently, adaptation to D2 and D3 must acquire new-domain capacity while avoiding interference with the representation space already encoded in the D1 model.

To address this challenge, we propose Orthogonal Knowledge distilled LoRA (OR-KDL), a continual adaptation framework built on a frozen CNN14 [3] backbone. OR-KDL combines three components: domain-specific low-rank adapters for D2 and D3, knowledge distillation (KD) [4, 5] for semantic consistency across adaptation stages, and Weight-based Orthogonal Gradient Projection (OGP) [6, 7, 8] based on singular-vector subspaces extracted from prior-domain weights. The key design principle is modularity: the low-rank adapter is decoupled from OGP and distillation, allowing different adapter formulations to be evaluated under the same continual learning framework.

Within OR-KDL, we compare two low-rank adapter formulations: standard LoRA [9] as the baseline, and an additive shared-basis variant adapted from CoLoRA [10], which decomposes the update into a layer-wise structural term and a block-shared basis term. We further analyze stratified K-fold ensembling and augmentation strategy to examine their effects on the stability–plasticity trade-off. Results are reported in Section 3.

2. PROPOSED METHOD

The proposed system operates on a frozen CNN14 backbone with three independent components – domain-specific LoRA adapters, Weight-based Orthogonal Gradient Projection (OGP), and KD (Figure 1) – so that the adapter design can be evaluated separately from OGP and distillation.

The model is built on a CNN14-style convolutional backbone loaded from the pretrained D1 checkpoint. Since no D1 training data is provided to reconstruct its representation space, the backbone is kept entirely frozen throughout training, and adaptation to subsequent domains is delegated exclusively to lightweight low-rank adapters attached to the frozen backbone. For each domain d, the adapter induces a weight update at layer l of the general form

$$\Delta W_{d,l} = f_d(A_{d,l}, B_{d,l}), \tag{1}$$

where $f_d(\cdot)$ is the adapter function for domain d and $A_{d,l}, B_{d,l}$ are its low-rank factors. At inference time, the domain-specific updates are additively composed onto the frozen backbone:

$$W_{\text{eff},l}^{(d)} = W_{D1,l} + \sum_{i=2}^{d} \Delta W_{i,l}, \quad d \in \{2, 3\}. \tag{2}$$

The D3 model, obtained after the final adaptation stage, is evaluated on both the D2 and D3 test sets under the same configuration and without access to domain labels.

2.1. Low-Rank Adaptation (LoRA)

We compare two low-rank schemes for the adapter function $f_d(\cdot)$: standard LoRA and an additive shared-basis variant adapted from CoLoRA. Both keep the frozen backbone unchanged and parameterize the domain-specific update with a small number of trainable low-rank factors.

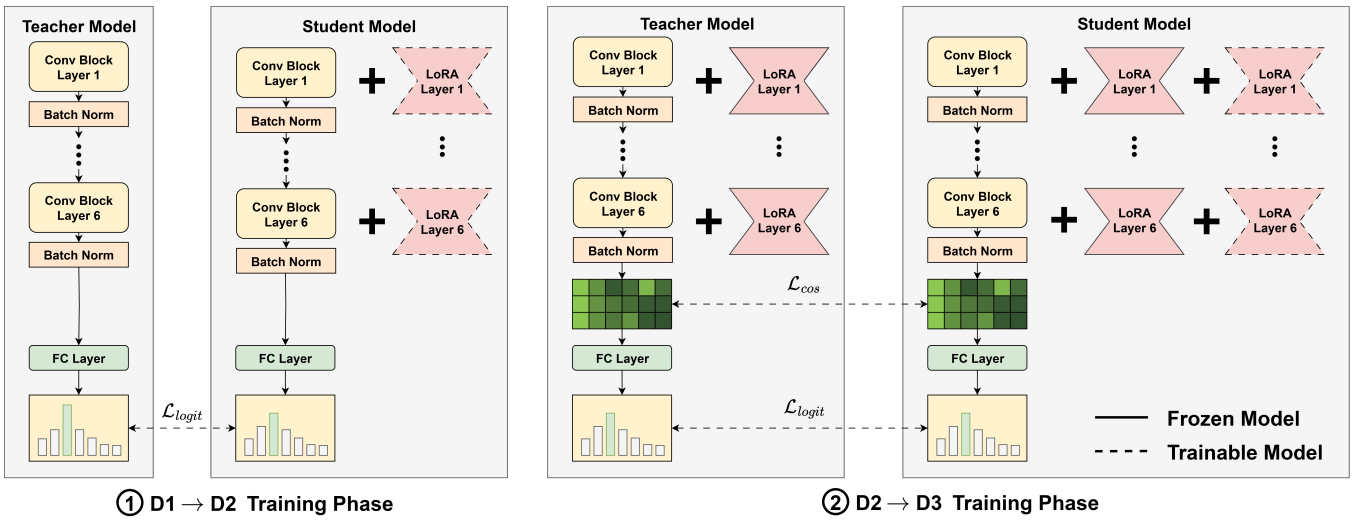

Figure 1: Overall architecture of the proposed OR-KDL framework. (Left) D1→D2 phase: a domain-specific adapter is trained with logit-based distillation from the D1 teacher. (Right) D2→D3 phase: a second adapter is added with both logit and cosine feature distillation from the D2 teacher. Weight-based OGP constrains adapter updates to directions orthogonal to the prior-domain subspace in both phases.

LoRA approximates the weight update by a single low-rank product, factoring it into a down-projection and an up-projection. For domain d at layer l:

$$\Delta W_{d,l} = \frac{\alpha}{r}(B_{d,l}A_{d,l}) \tag{3}$$

where $A_{d,l} \in \mathbb{R}^{r\times c_{\text{in}}}$ and $B_{d,l} \in \mathbb{R}^{c_{\text{out}}\times r}$ are the low-rank factors, $r \ll \min(c_{\text{in}}, c_{\text{out}})$ is the rank, and α/r controls the update magnitude relative to the frozen weights. Composed onto the frozen path via Eq. (3), only $A_{d,l}$ and $B_{d,l}$ are trainable, while $W_{D1,l}$ receives no gradient.

Each domain maintains its own factors $\{A_{d,l}, B_{d,l}\}$, with $B_{d,l}$ zero-initialized so that $\Delta W_{d,l} = 0$ at the start of training; the adapter then acquires only the target-domain representations as training proceeds.

CoLoRA extends the layer-wise LoRA term with an additive block-shared term, distributing a common basis across the two convolutions within a residual block. For each convolutional layer $l \in \{1, 2\}$,

$$\Delta W_{d,l} = s_A(B_l A_l) + s_B(B_s A_{s,l}) \tag{4}$$

where A_l, B_l are layer-specific factors, $A_{s,l}$ is a layer-specific projection, B_s is a block-shared factor common to both convolutions, and s_A, s_B are scaling coefficients. This formulation increases cross-layer capacity under the same rank budget while preserving the OGP constraint on all input-side factors.

2.2. Knowledge Distillation

While gradient projection constrains update directions, it does not explicitly enforce semantic consistency with the teacher domain. We therefore apply knowledge distillation at each adaptation stage, independently of the adapter architecture.

For D1→D2, logit-based distillation minimizes the Kullback-Leibler (KL) divergence between temperature-scaled softmax outputs of the D1 teacher and the student. For D2→D3, cosine feature distillation is added to logit distillation:

$$\mathcal{L}_{\text{KD}} = \lambda_{logit} \cdot \mathcal{L}_{\text{logit}} + \lambda_{cos} \cdot \underbrace{\big(1 - \cos(f_{\text{student}}, f_{\text{teacher}})\big)}_{\mathcal{L}_{\text{cos}}} \tag{5}$$

The cosine term improved both D2 retention and D3 accuracy in our experiments, suppressing representational drift that logit-level alignment alone cannot capture.

2.3. Weight-based Orthogonal Gradient Projection

Existing OGP methods [6, 7, 8] require previous task data to construct the projection subspace. Since no D1 training data is available, we propose a data-free variant, Weight-based OGP, which extracts the projection basis directly from prior-domain weight matrices rather than from task-specific activations. For D2 training, each D1 convolutional kernel is reshaped into $W_{D1,l} \in \mathbb{R}^{c_{\text{out}}\times(c_{\text{in}}\,k^2)}$ and decomposed via singular value decomposition (SVD) [11]:

$$W_{D1,l} = U_l \Sigma_l V_l^\top \tag{6}$$

The right singular vectors V_l span the input-side subspace of D1 and are already orthonormal as a direct consequence of SVD, requiring no additional orthonormalization at this stage. For D3 training, the basis is extended by decomposing $\Delta W_{2,l}$ via SVD to obtain $V_{2,l}^{\Delta}$, which is concatenated with the D1 basis V_l. Although V_l and $V_{2,l}^{\Delta}$ are each individually orthonormal, their concatenation $[V_l \mid V_{2,l}^{\Delta}]$ is not guaranteed to be orthonormal in general, since the column spaces of the two bases may be non-orthogonal to each other. QR decomposition is therefore applied to re-orthonormalize the concatenated basis while preserving its column space, yielding a single orthonormal basis that jointly spans the D1 and D2 subspaces. The orthonormal basis Q_l is defined as:

$$Q_l = \begin{cases} V_l & \text{(D2 phase)} \\ \text{QR}([V_l \mid V_{2,l}^{\Delta}]) & \text{(D3 phase)} \end{cases} \tag{7}$$

where $\mathrm{QR}(\cdot)$ denotes the orthonormal factor returned by QR decomposition. The projection is then applied to the input-side and output-side gradients of the adapter at layer l:

$$G'_{A,l} = G_{A,l} - G_{A,l}Q_lQ_l^\top, \quad G'_{B,l} = G_{B,l} - Q_lQ_l^\top G_{B,l} \quad (8)$$

3. PERFORMANCE EVALUATION

3.1. Training Strategy

The following settings are shared across all experiments: LoRA rank $r_{D2} = 128$ and $r_{D3} = 64$, optimizer AdamW [12] with learning rate 3×10^{-4} and cosine annealing, Focal Loss [13] with $\gamma = 2$, and $K = 5$ for the stratified K-fold protocol, with all models trained for 150 epochs per domain. For knowledge distillation, the logit distillation weight and temperature are set to $\lambda_{logit} = 0.1$, $\tau = 2.0$ for D1→D2, and $\lambda_{logit} = 0.5$, $\tau = 1.5$ for D2→D3. Cosine feature distillation is applied only during D2→D3 with $\lambda_{cos} = 0.3$; it is disabled for D1→D2 ($\lambda_{cos} = 0.0$).

To improve generalization and reduce sensitivity to data splits, we apply gain scaling, mixup [14], and SpecAugment [15] as data augmentation, and use balanced sampling with focal loss ($\gamma = 2$) to address class imbalance.

A key observation during training is that applying augmentation uniformly across all classes is suboptimal: the phone and baby classes exhibit distinctive, compact spectro-temporal signatures that are easily corrupted by aggressive augmentation. We therefore adopt a class-selective augmentation strategy, disabling gain scaling, mixup, and SpecAugment for these classes while keeping them active for all others, which consistently improves their per-class accuracy without degrading overall performance.

For the ensemble, we apply a stratified K-fold protocol, where K is the number of folds and N is the number of top-ranked models aggregated in the final soft-voting ensemble. For D2, we fix $K = 5$ and train five D2 fold adapters independently; each is then paired with five D3 folds, yielding 25 D3 models in total. These models are ranked by accuracy on a validation split held out from the training data, and a soft-voting ensemble is formed by averaging the class posterior probabilities of the Top-N models. N is selected solely on this validation split, without any access to the test set. Once fixed, the validation split is merged back into the training data and the final submission models are retrained on the full data to avoid wasting labeled data.

We report macro accuracy across five metrics: D2@D2, D2@D3, D3@D3, Acc, and Fr, where Fr measures the degree of forgetting on D2 induced by D3 adaptation.

3.2. Results

Table 1 reports our four challenge submission systems, combining two adapter types (LoRA and CoLoRA) with two augmentation strategies (uniform and class-selective), all evaluated under the validation-selected Top-20 ensemble. CoLoRA with uniform augmentation attains the highest overall accuracy (73.14% Acc) and is selected as our primary submission. Notably, CoLoRA yields consistently higher D2@D2 than LoRA under both augmentation strategies (83.77/83.16 vs. 81.24/82.05%), suggesting that its block-shared basis better preserves source-domain representations. The overall-accuracy ranking, however, depends on the augmentation strategy: CoLoRA leads under uniform (73.14 vs. 72.97% Acc) while LoRA leads under class-selective (73.06 vs. 72.88% Acc),

Table 1: Performance of our four challenge submission systems, compared by LoRA type and augmentation strategy. All systems use the final models retrained on the full data with the validation-selected Top-20 soft-voting ensemble.

LoRA	Aug.	D2@D2	D2@D3	D3@D3	Acc	Fr
LoRA	Uniform	81.24	77.17	68.78	72.97	4.07
	Class-sel.	82.05	80.12	66.00	73.06	1.93
CoLoRA	Uniform	83.77	79.73	66.55	73.14	4.04
	Class-sel.	83.16	79.14	66.61	72.88	4.02

Table 2: Top-N soft-voting ensemble accuracy across ensemble sizes. The Dataset column indicates the evaluation dataset: Valid refers to the held-out validation dataset used to select N (highlighted: N=20), whereas Test refers to the official test dataset evaluated after retraining on the full training data with the selected N fixed.

Dataset	N=5	N=10	N=15	N=20	N=25
Valid	73.21	72.89	73.17	73.68	72.55
Test	72.92	73.32	72.81	73.14	72.87

so adapter choice interacts with augmentation strategy. The augmentation effect is likewise adapter-dependent: for LoRA, class-selective augmentation cuts forgetting from 4.07%p to 1.93%p while slightly improving accuracy (72.97 → 73.06%), trading D3@D3 (68.78 → 66.00%) for stronger D2 retention (D2@D3 77.17 → 80.12%); for CoLoRA it leaves forgetting essentially unchanged (4.04 → 4.02%p) and lowers accuracy. Thus class-selective augmentation is attractive specifically for LoRA when stability is prioritized over plasticity.

Table 2 presents an ablation on the ensemble size N, selected solely on the validation split without any access to the test set. Among the candidates, N=20 attains the highest validation accuracy (73.68%) and is chosen; the validation split is then merged back and the final models retrained on the full data. The validation-selected N=20 remains near-optimal on the test set (73.14% Acc), 0.18%p below the test-set optimum (N=10, 73.32% Acc), confirming that validation-based selection generalizes to the test set.

Overall, OR-KDL balances stability and plasticity across domains, with the Top-N ensemble providing consistent gains under a test-agnostic selection of N.

4. CONCLUSION

We presented OR-KDL, a continual adaptation framework for domain-incremental learning that combines Weight-based OGP, knowledge distillation, and modular low-rank adapters on a frozen CNN14 backbone. Across four submission systems, neither adapter dominated across augmentation strategies; class-selective augmentation substantially reduced forgetting for LoRA and appears to come with a marginal accuracy gain as well, though alongside reduced D3@D3 retention. CoLoRA appears to consistently retain source-domain performance better than LoRA, suggesting that its block-shared basis aids representation preservation. We select CoLoRA with uniform augmentation and the validation-selected Top-20 soft-voting ensemble (73.14% Acc, 4.04%p forgetting) as our primary submission, with the ensemble size N chosen on a held-out validation split without test-set access.

5. ACKNOWLEDGMENTS

This work was supported in part by the Innopolis Foundation funded by the Ministry of Science and ICT (2022-DD-UP-0312, Science and Technology Project Opens the Future of the Region), and in part by the Commercialization Promotion Agency for R&D Outcomes (COMPA) funded by the Ministry of Science and ICT (MSIT) (RS-2025-02634220, Development of a reinforcement learning-based hyper-personalized on-device AI agent and system for application to future cars).

6. REFERENCES

[1] R. Casciotti, M. Mulimani, M. Harju, J. R. Jensen, and A. Mesaros, "Domain-agnostic incremental learning for sound classification. a DCASE 2026 Challenge task," arXiv preprint:2606.02173, 2026.

[2] M. Mulimani and A. Mesaros, "Domain-incremental learning for audio classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[3] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proceedings of the Deep Learning and Representation Learning Workshop (NIPS)*, 2015.

[5] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[6] X. Wang, T. Chen, Q. Ge, H. Xia, R. Bao, R. Zheng, Q. Zhang, T. Gui, and X. Huang, "Orthogonal subspace learning for language model continual learning," in *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

[7] M. Farajtabar, N. Azizan, A. Mott, and A. Li, "Orthogonal gradient descent for continual learning," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[8] S. Wang, X. Li, J. Sun, and Z. Xu, "Training networks in null space of feature covariance for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[10] W. Ran, W. Zhang, S. Pang, Z. Zhu, J. Liu, J. Liu, X. Cao, Q. Li, Y. Yan, and C. Ma, "Correlated low-rank adaptation for convnets," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[11] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische Mathematik*, vol. 14, no. 5, pp. 403–420, 1970.

[12] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[14] H. Zhang, M. Cissé, Y. N. Dauphin, and D. López-Paz, "mixup: Beyond empirical risk minimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of Interspeech*, 2019.