RELIABILITY-GUIDED DISTILLATION FOR CROSS-DOMAIN MOSQUITO SPECIES CLASSIFICATION

Technical Report

Shuanglin Li

XiangJiang Laboratory
Changsha, China
slay575@163.com

Ruxiao Qian

XiangJiang Laboratory
Changsha, China
ruxiaoqian@gmail.com

ABSTRACT

This report describes our BioDCASE 2026 Task 5 system for cross-domain mosquito species classification (CD-MSC). The development data are highly imbalanced across species-domain cells, with D5 dominating the training distribution, while the official evaluation emphasizes balanced accuracy under unseen-domain shift. We view the CD-MSC task as challenging along four axes: weak acoustic separability in narrow-band, low-SNR wingbeat recordings; cross-domain distribution shift; confounded and imbalanced species–domain coverage; and model-selection uncertainty under a constrained development protocol. To expose domain and coverage risks during development, we first construct a species–domain held-out development split. Under this protocol, our submitted systems combine a compact CQT–MTRCNN backbone, gradient-reversal-based domain regularization, reliability-guided distillation of calibrated teacher logits, and guarded prediction-level ensembles. Because official evaluation labels are hidden, we use development-protocol evidence primarily for risk control rather than as a claim of final evaluation performance. The submitted single-model and ensemble candidates are selected to favor unseen-domain robustness and prediction diversity, with guarded ensembles used to reduce overfitting to the observed species–domain structure in the development data.

***Index Terms*—** Mosquito Classification, Domain Shift, Knowledge Distillation

1. INTRODUCTION

BioDCASE 2026 Task 5, Cross-Domain Mosquito Species Classification (CD-MSC), aims to classify mosquito recordings into one of nine closed-set species under domain shift [1]. The official evaluation set contains both seen-domain and unseen-domain samples, while the unseen domains are still drawn from the five known recording domains. Participants submit only the predicted species label for each evaluation clip; domain labels are hidden and used by the organizers to compute seen-domain and unseen-domain metrics.

The central difficulty of CD-MSC is that the development data are not only imbalanced across species, but also highly confounded across species-domain cells. In particular, D5 dominates the training distribution, whereas the official ranking emphasizes balanced accuracy on unseen domains [1]. This creates a mismatch between the objective that a standard classifier can easily optimize and the objective that matters for evaluation. A model may perform well by exploiting recording-domain artifacts or background conditions, while failing to learn species-discriminative mosquito signatures that transfer across domains [2].

We therefore treat CD-MSC as a domain generalization and reliability problem. Our development protocol first constructs an official-like species-domain re-split to expose unseen-domain behavior before the final evaluation stage. Building on the official CD-MSC baseline [1], which adopts the MTRCNN architecture originally introduced in [3], we replace the log-mel frontend with CQT features. We further apply gradient-reversal-based domain regularization [4] to discourage the representation from encoding recording-domain information. However, adversarial domain regularization may also weaken species-discriminative information when it is entangled with domain-specific recording conditions [5, 6]. To counterbalance this trade-off, we add soft species-level supervision via knowledge distillation [7]. Calibrated teacher logits provide soft species-level targets beyond the hard labels [7, 8], helping the student retain inter-species discrimination while learning a more domain-robust representation. Since teacher predictions may become less reliable under recording-domain shift [9], we weight the distillation signal according to teacher reliability, reducing the influence of uncertain or domain-sensitive soft targets [10, 11].

Since the official evaluation labels are hidden, final system selection is based on development unseen-domain accuracy, domain-gap diagnostics, and prediction-diversity checks rather than a single validation score. Our final submission includes four prediction-level ensemble systems that follow the same overall recipe but represent different risk profiles: an aggressive peak-performance ensemble, a stable ensemble, a pre-selected guarded logit ensemble, and a low-domain-gap ensemble. This design balances peak development performance with robustness to hidden species-domain shifts.

2. METHOD

2.1. Species-Domain Holdout Split and Metrics

The official development release provides labeled `TrainVal` data, but no labels for the final evaluation set. Since the evaluation protocol is closed-set in species but includes species-specific unseen domains within the known domain set, random validation splits can mainly measure seen-domain fitting rather than robustness to held-out species–domain combinations.

For each species, one available domain is held out from the training subset, as shown in Table 1. Clips in the correspond-

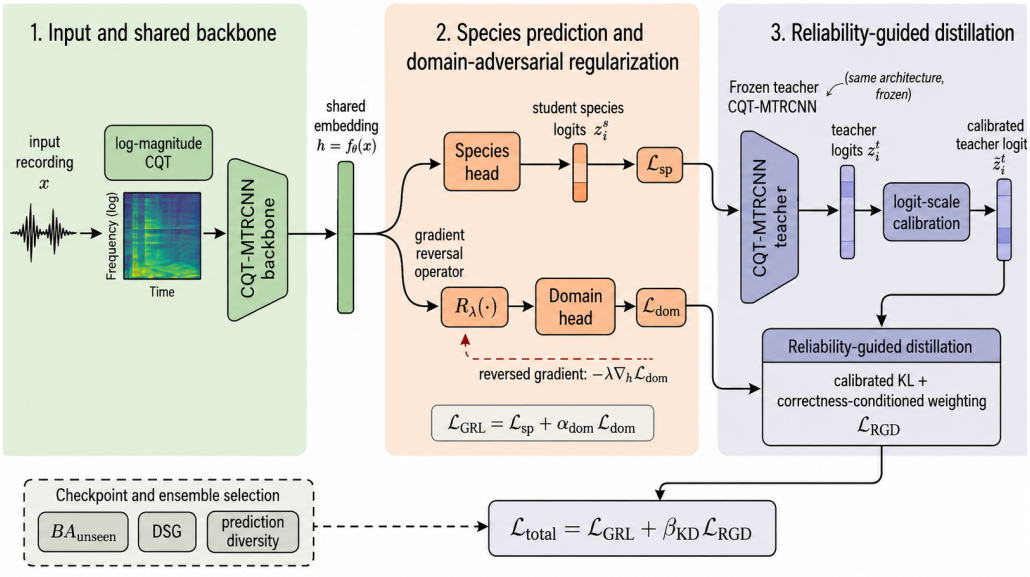

Figure 1: Overview of the challenge solutions. The student uses log-magnitude CQT features and a CQT-MTRCNN backbone. Domain-adversarial regularization reduces domain-predictive information through a gradient reversal branch, while reliability-guided distillation transfers calibrated teacher logits with correctness-conditioned weighting. Final checkpoints and prediction-level ensembles are selected using $\mathrm{BA}_{\mathrm{unseen}}$, DSG, and prediction diversity.

ing species–domain cell are used only as unseen-domain samples in validation and test, while the remaining cells form the seen-domain training/evaluation data. This produces 240,460 training clips, 15,461 validation clips, and 15,459 test clips, with 2,102 and 2,100 unseen-domain clips in validation and test, respectively. The validation and test partitions are constructed with nearly identical species–domain matrices to support stable model comparison. Although D5 still dominates the seen-domain data, the held-out cells force checkpoint and ensemble selection to account for non-D5 and rare-domain generalization.

We follow the official CD-MSC metric definitions [1]. We report SpeciesBA over all samples, $\mathrm{BA}_{\mathrm{seen}}$ and $\mathrm{BA}_{\mathrm{unseen}}$ on seen- and unseen-domain subsets, and DSG as the absolute seen–unseen gap. Model selection prioritizes $\mathrm{BA}_{\mathrm{unseen}}$, while DSG, per-species behavior, and prediction diversity are used as guard diagnostics against split-specific overfitting.

2.2. CQT–MTRCNN Backbone

We keep the classifier close to the official MTRCNN-style baseline [3] to avoid confounding our analysis with unconstrained architecture search. This provides a controlled basis for evaluating gradient reversal, reliability-guided distillation, and ensemble selection, consistent with prior domain generalization studies emphasizing training objectives, regularization, and model selection over architecture changes [12].

Mosquito wingbeat recordings are short, narrow-band, and quasi-periodic, with species-related information often reflected in localized frequency and harmonic structure. We therefore treat acoustic feature extraction as a task-specific design choice rather than the main contribution of this work. We considered STFT spectrograms, log-mel spectrograms, PCEN features, LEAF features [13], and CQT features as candidate representations.

STFT spectrograms preserve fine spectral detail, but their fixed analysis window imposes a single time–frequency resolution trade-off. Log-mel features are compact, but frequency-band pooling can obscure narrow harmonics and subtle wingbeat-frequency shifts. PCEN and LEAF features add adaptive compression or learnable filtering, but also introduce extra frontend choices that are harder to select reliably under species–domain imbalance. We therefore use log-compressed magnitude CQT features, whose logarithmic frequency spacing and approximately constant relative bandwidth are well aligned with the harmonic and pitch-like structure of mosquito wingbeats while keeping the frontend fixed.

We replace the log-mel feature in the official MTRCNN-style pipeline with log-compressed magnitude CQT features. To keep the model interface unchanged, we use 64 CQT bins, matching the 64-bin baseline frontend. The resulting time-frequency maps can therefore be processed by the same convolutional backbone and temporal pooling layers without architecture changes.

2.3. Domain-Adversarial Regularization

The major difficulty of the challenge is closed-set species recognition under domain shift [1]. In the development data, species labels and recording domains are strongly confounded: several species are concentrated in only a few domains, with D5 dominating the seen-domain distribution. As a result, a species classifier trained only with cross-entropy may learn domain shortcuts [14]: it can obtain favorable seen-domain performance by exploiting recording-domain-specific acoustic patterns that are correlated with species labels, but these patterns may not remain predictive when the same species appears in a held-out domain. We therefore apply gradient-reversal-based domain regularization [15] to reduce the accessibility of recording-domain information from the shared embedding and encourage more domain-robust species representations.

Table 1: Species-domain clip distribution and held-out cells in the development split. Rows correspond to domains and columns to species; bold entries are excluded from training and used as unseen-domain validation/test samples.

Dataset	Ae.aeg	Ae.alb	Cx.qui	An.gam	An.ara	An.dir	Cx.pip	An.min	An.ste	Total
D1	123	79	**672**	**818**	**1820**	87	66	200	200	4065
D2	0	**419**	0	0	0	0	66	**99**	200	784
D3	**192**	19	0	0	0	0	**68**	200	200	679
D4	22	0	13	0	0	**40**	0	51	**74**	200
D5	81250	18000	71371	46180	19297	0	29554	0	0	265652
Total	81587	18517	72056	46998	21117	127	29754	550	674	271380

We implement this regularization by attaching an auxiliary recording-domain classifier to the shared backbone embedding. The species branch predicts mosquito species, while the auxiliary branch predicts the recording domain. Following domain-adversarial training [15], a gradient reversal layer is inserted between the shared encoder and the domain classifier. The layer acts as the identity function in the forward pass, but reverses the gradient from the domain-classification loss before it reaches the shared encoder. Thus, the domain classifier is optimized to recognize recording domains, whereas the shared encoder is optimized to make domain information less accessible. The domain branch acts as a training-time regularizer; final predictions are produced from the species logits alone.

Formally, let $h = f_\theta(x)$ denote the shared embedding of an input recording x. The species classifier produces $p_\phi(y \mid h)$, and the domain classifier produces $q_\psi(d \mid \mathcal{R}(h))$, where $\mathcal{R}(\cdot)$ denotes the gradient reversal operator. The operator returns its input unchanged in the forward pass and multiplies the backward gradient by -1. The implemented training loss is

$$\mathcal{L}_{\mathrm{GRL}} = \mathcal{L}_{\mathrm{sp}} + \mathcal{L}_{\mathrm{dom}}, \tag{1}$$

where $\mathcal{L}_{\mathrm{sp}}$ is the species cross-entropy loss and $\mathcal{L}_{\mathrm{dom}}$ is the domain cross-entropy loss. Although the domain loss is added to the scalar loss, its gradient reaches the shared encoder through $\mathcal{R}(\cdot)$ and is therefore reversed.

This regularizer reduces access to domain-predictive structure and discourages domain-correlated shortcuts, but species–domain confounding can make some domain-related patterns species-discriminative. This robustness-separability trade-off motivates the reliability-guided distillation module described next.

2.4. Reliability-Guided Distillation

Domain-adversarial regularization reduces reliance on recording-domain information, but CD-MSC does not cleanly separate domain and species variation. Under strong species-domain confounding, suppressing domain-related patterns may also weaken fine-grained species separability. We therefore use knowledge distillation [7] to counterbalance this robustness-separability trade-off with soft species-level supervision.

The teacher uses the same CQT-MTRCNN backbone as the student, but is trained only for species classification, selected under the held-out development protocol, and frozen during student training. Its logits provide soft species-level targets that capture inter-species similarity beyond the hard labels. Since the teacher is learned from the same imbalanced and species-domain-confounded data, its predictions may also reflect domain-conditioned biases; we therefore treat the distillation signal as informative but reliability-dependent.

A vanilla distillation objective would minimize the KL divergence between softened teacher and student predictions [7]. In this task, directly transferring raw teacher logits is risky for two reasons. First, the teacher and student may operate at different logit scales, so teacher confidence can dominate the update even when the class ranking is useful. Second, because the teacher is trained on the same imbalanced species–domain data, its errors can be structured by domain bias rather than random noise.

We therefore use a reliability-guided distillation loss with two controls. First, teacher logits are calibrated to the current student logit scale before distillation. For sample i, let z_i^t and z_i^s denote teacher and student logits. We compute:

$$\tilde{z}_i^t = \rho \left[\frac{z_i^t - \mu(z_i^t)}{s(z_i^t)} s(z_i^s) + \mu(z_i^s) \right] + (1-\rho) z_i^t, \tag{2}$$

where $\mu(\cdot)$ and $s(\cdot)$ are the mean and standard deviation over species classes, and ρ controls the calibration strength. This preserves the teacher's class ranking while aligning its confidence scale with the student.

Second, we condition the distillation signal on teacher reliability. Let:

$$m_i = \mathbf{1}\left[\arg\max_c z_{i,c}^t = y_i \right], \tag{3}$$

where $m_i = 1$ indicates that the teacher predicts the ground-truth species. For teacher-correct samples, we use the full calibrated soft target. For teacher-wrong samples, we suppress direct target-class imitation and retain only a down-weighted non-target dark-knowledge term:

$$\ell_i^{\mathrm{RGD}} = m_i \, \mathrm{KL}_\tau(\tilde{z}_i^t, z_i^s) + (1-m_i)\gamma \, \mathrm{NCKD}_\tau(\tilde{z}_i^t, z_i^s, y_i), \tag{4}$$

where KL_τ is temperature-scaled KL distillation, NCKD_τ denotes non-target-class distillation after excluding the ground-truth class [16], and γ controls how much information is retained from teacher-wrong samples.

We refer to this calibrated and correctness-conditioned objective as reliability-guided distillation (RGD). At the batch level, we average the per-sample RGD loss with optional reliability weights:

$$\mathcal{L}_{\mathrm{RGD}} = \frac{\sum_{i=1}^B r_i \ell_i^{\mathrm{RGD}}}{\sum_{i=1}^B r_i}, \tag{5}$$

where r_i denotes an additional sample weight, used for confidence filtering or domain-level reliability weighting. The final student objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{GRL}} + \beta_{\mathrm{KD}} \mathcal{L}_{\mathrm{RGD}}, \tag{6}$$

where β_{KD} controls the distillation strength. In our final systems, the distillation term is activated after a short warm-up. This

Table 2: Single-model development results and controls on the internal held-out test split. Bold numbers mark the best semantic-teacher GRL–distillation result in each column. The permuted-teacher row is a negative control and is excluded from highlighting.

System	GRL	Logit calib.	Correct.-cond.	SpeciesBA	$\mathrm{BA}_{\mathrm{unseen}}$	DSG
MTRCNN baseline	–	–	–	0.5411	0.1751	0.7055
Teacher-only classifier	–	–	–	0.6287	0.2445	0.5658
No-GRL reliability distillation	–	✓	✓	0.6337	0.2355	0.5803
Vanilla logit distillation	✓	–	–	0.5508	0.3353	0.2959
Student-scale calibrated KL	✓	✓	–	0.5616	0.3209	0.3183
Calibration + correctness conditioning	✓	✓	✓	**0.5872**	**0.3513**	0.3434
Domain-reweighted reliability KD	✓	✓	✓	0.5776	0.3491	0.3290
Cross-teacher reliability check	✓	✓	✓	0.5568	0.3456	**0.2881**
Permuted-teacher control†	✓	✓	✓	0.5540	0.3553	0.2691

makes distillation a species-structure-preserving complement to GRL while limiting the transfer of teacher overconfidence and domain-biased errors.

2.5. Checkpoint and Ensemble Selection

Checkpoints are selected on the internal species–domain held-out split using $\mathrm{BA}_{\mathrm{unseen}}$ as the primary criterion, with SpeciesBA and DSG as guard diagnostics because overall and seen-domain accuracy are dominated by seen-domain/D5 samples. Final systems are small prediction-level ensembles formed by averaging logits or probabilities from selected GRL-distillation models. To limit overfitting to the fixed development split, we avoid stacking and extensive weight optimization, and filter ensembles by $\mathrm{BA}_{\mathrm{unseen}}$, DSG, and prediction diversity so that submitted systems are both strong and non-redundant.

3. DEVELOPMENT RESULTS AND FINAL SYSTEMS

3.1. Training Configurations

We train with AdamW [17] for up to 100 epochs using batch size 64, learning rate 10^{-3}, weight decay 10^{-4}, dropout 0.2, and early stopping after epoch 10 with patience 5. Audio is resampled to 8 kHz and normalized; training uses random 2-second crops, while all validation/test/evaluation runs use deterministic full-clip inference with valid-length masking. Unless stated otherwise, $\beta_{\mathrm{KD}} = 0.003$, $T = 2$, and distillation starts after 10 warm-up epochs. Checkpoints are selected by $\mathrm{BA}_{\mathrm{unseen}}$ with DSG as a guard.

3.2. Single-Model Development Study

Table 2 reports representative single-model results and controls. The MTRCNN baseline obtains favorable seen-domain diagnostics but transfers poorly to held-out species–domain cells, illustrating the species–domain shortcut problem. The teacher-only and no-GRL distillation controls further show that teacher supervision alone does not solve the domain-shift problem. This supports the use of GRL as the main representation-level regularizer.

Among GRL-based distillation variants, vanilla logit distillation already improves $\mathrm{BA}_{\mathrm{unseen}}$ over the baseline, while the best semantic-teacher single model combines logit calibration with correctness-conditioned transfer. Domain-reweighted and cross-teacher variants give comparable unseen-domain performance, indicating sensitivity to domain composition and teacher choice. The permuted-teacher negative control achieves strong $\mathrm{BA}_{\mathrm{unseen}}$, showing that part of the gain comes from regularization and loss shaping rather than purely semantic teacher dark knowledge.

Table 3: Prediction-level ensemble progression results.

System family	Members	$\mathrm{BA}_{\mathrm{unseen}}$	DSG
Best Single Teacher	1	0.3513	0.3434
Early guarded pair	2	0.4450	0.1255
Final guarded range	2–3	0.4944–0.4972	0.0969–0.1233

Table 4: Final submitted systems results.

System	Role	Members	$\mathrm{BA}_{\mathrm{unseen}}$	DSG
task5_1	Diversity	3	0.4947	0.1233
task5_2	Stable	2	0.4972	0.1089
task5_3	Guarded	3	0.4969	0.1046
task5_4	Low DSG	2	0.4944	0.0969

3.3. Prediction-Level Ensemble Study

Because single models are sensitive to seed, teacher choice, and rare species–domain cells, final systems use small prediction-level ensembles. We restrict ensembling to logit/probability averaging and avoid stacking or large member counts to reduce validation-set memorization.

Table 3 summarizes the development progression. Ensembling gives the largest practical gain in the final stage, improving $\mathrm{BA}_{\mathrm{unseen}}$ while substantially reducing DSG. Because the internal split contains a fixed species–domain matrix, we treat peak ensemble scores cautiously and use prediction diversity, DSG, and simple averaging behavior as guard diagnostics.

Table 4 lists the four submitted systems. They use different ensemble compositions to cover different risk profiles. `task5_2` and `task5_3` are the most conservative high-performing systems, `task5_4` prioritizes the smallest domain-shift gap, and `task5_1` is retained as a diversity-oriented guard against hidden species–domain shifts. On the unlabeled official evaluation, pairwise disagreement among the four submitted prediction files ranges from 5.1% to 9.2%, while 87.3% of clips receive the same prediction from all four systems. This indicates that the submitted systems share a stable prediction core while still providing non-identical risk profiles.

4. REFERENCES

[1] Y. Hou, V. Zdravkovic, M. Sinka, Y. Li, W. Wang, M. D. Plumbley, K. Willis, and S. Roberts, "Biodcase 2026 challenge baseline for cross-domain mosquito species classification," 2026. [Online]. Available: https://arxiv.org/abs/2603.20118

[2] Y. Hou, Z. Liu, X. Shen, and S. Roberts, "Learning domain-robust bioacoustic representations for mosquito species classification with contrastive learning and distribution alignment," in *ICASSP*, 2026, pp. 15 207–15 211.

[3] Y. Hou, Q. Ren, W. Wang, and D. Botteldooren, "Sound-based recognition of touch gestures and emotions for enhanced human-robot interaction," in *ICASSP*, 2025, pp. 1–5.

[4] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*. PMLR, 2015, pp. 1180–1189.

[5] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon, "On learning invariant representations for domain adaptation," in *ICML*. PMLR, 2019, pp. 7523–7532.

[6] M. Long, Z. Cao, J. Wang, and M. Jordan, "Conditional adversarial domain adaptation," *NeurIPS*, vol. 31, 2018.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 1321–1330.

[9] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *ICCV*, 2019, pp. 4794–4802.

[10] 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 *ICML*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 2304–2313.

[11] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 4334–4343.

[12] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *ICLR*, 2021.

[13] N. Zeghidour, O. Teboul, F. D. C. Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," *arXiv preprint arXiv:2101.08596*, 2021.

[14] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, no. 59, pp. 1–35, 2016.

[16] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *CVPR*, 2022, pp. 11 953–11 962.

[17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.