EVIDENCE-GUIDED EXPERT EXPANSION FOR DOMAIN-AGNOSTIC INCREMENTAL AUDIO CLASSIFICATION

Technical Report

Peihong Zhang, Yiqiang Cai, Shengchen Li

Xi'an Jiaotong-Liverpool University
{Peihong.Zhang20, Yiqiang.Cai21}@student.xjtlu.edu.cn, Shengchen.Li@xjtlu.edu.cn

ABSTRACT

Domain-agnostic incremental audio classification is not only a forgetting problem: after new-domain experts have been learned, the system must still decide which frozen expert is reliable for an unlabeled-domain test sample. This report studies this routing problem under the DCASE 2026 Task 7 constraints: no external data, no pretrained audio model, no replay of previous-domain audio, no evaluation-set calibration, and no test-time adaptation. We ask whether compact aggregate evidence from frozen experts can replace the unavailable domain label at inference time. Our final system freezes the provided D1 expert, expands compact D2 and D3 experts, and routes each sample using batch-normalization evidence, class-conditional prototype distance, entropy, margin, and a calibrated D2/D3 energy score. A five-model P3a ensemble obtains 70.82% D2 accuracy, 62.41% D3 accuracy, and 66.62% D2/D3 average accuracy on the development split. Relative to the same top-five ensemble with the clean router, energy evidence reduces D3-to-D2 routing from 24.69% to 18.24%. Negative results are also informative: hard routers, tiny learned meta-routers, D3-oriented model mixing, and fixed TTA each improved one diagnostic but reduced robustness or D2 accuracy.

***Index Terms*—** audio classification, routing, DCASE

1. INTRODUCTION

DCASE 2026 Task 7 studies domain-agnostic incremental learning for audio classification [1]. It extends the DCASE tradition of reproducible acoustic-scene and sound-event evaluation [2] to a setting in which a system learns new domains without revisiting previous-domain training data. Recent audio continual-learning work has considered class-incremental multi-label sound classification [3], online domain-incremental acoustic scenes [4], and domain-incremental audio classification [5]. Task 7 adds a particularly sharp inference-time difficulty: the domain identity is hidden. A strong D2 or D3 expert is useful only if the router can recognize when that expert is compatible with the current sample.

The central question of this report is therefore: *can compact, aggregate evidence from frozen experts serve as a reliable domain-agnostic routing signal without replay or evaluation-time adaptation?* This differs from standard continual learning, where the dominant concern is often weight-level forgetting and catastrophic interference [6, 7, 8, 9, 10, 11, 12]. Here, the main visible error mode was D3 samples being assigned to D2 even when the D3 expert itself remained useful under forced-domain evaluation. The scientific claim is modest: aggregate evidence does not solve domain-agnostic incremental learning, but it gives an auditable way to expose and reduce a specific routing bottleneck.

The contribution is threefold. First, we construct a no-replay expert-expansion system that freezes the organizer-provided D1 expert and adds compact D2/D3 experts using the challenge CNN/BN backbone, related in layout to CNN14 from PANNs [13] but trained without pretrained PANNs weights. Second, we define a compact evidence memory combining frozen batch-normalization statistics, class-conditional prototypes, entropy, margin, and relative energy scores. Third, we report negative results that are useful for future systems: reducing D3-to-D2 alone is insufficient if D2 collapses, and TTA or learned meta-routing can disturb the evidence distribution under small visible splits.

2. EVIDENCE-GUIDED EXPERT EXPANSION

2.1. Frozen experts and aggregate memory

Let $e \in \{1, 2, 3\}$ index the domain experts. The D1 expert is the unchanged organizer checkpoint. D2 and D3 are expanded by updating selected convolutional blocks and the classifier head; the final P3a family updates D2 blocks 4–6 and D3 blocks 2–6 with cosine classifier heads. Audio is resampled to 32 kHz and represented by 64-band log-mel features using a 1024-sample window, 320-sample hop, and a 50–14000 Hz range. Experts are trained on 4-s clips using Adam, learning rate 10^{-4}, batch size 32, 120 epochs, weight decay 10^{-5}, label smoothing 0.05, and cosine learning-rate decay. Unlike distillation-style preservation [14, 9] or exemplar replay [10], the final system does not train against stored previous-domain samples or sample-level embeddings. At submission time all expert weights are frozen.

For each expert, the memory contains aggregate statistics rather than stored audio or sample-level embeddings. For evidence layer ℓ, class c, and expert e, we store training-split counts, means $\mu_{e,\ell,c}$, and variances $\sigma^2_{e,\ell,c}$ for selected activations. For a test sample x, expert e predicts $\hat{y}_e = \arg\max_y p_e(y|x)$ and we compute a class-conditional prototype distance

$$d_e(x) = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \left\langle \frac{(f_{e,\ell}(x) - \mu_{e,\ell,\hat{y}_e})^2}{\sigma^2_{e,\ell,\hat{y}_e} + \epsilon} \right\rangle. \quad (1)$$

Using the expert top-1 class makes this a compatibility score, not a proof of the class label. This distinction matters: hard prototype routers were fragile when the top-1 class was wrong.

Batch-normalization evidence uses frozen running statistics [15]. Conceptually, for BN layer ℓ we measure the normalized mismatch between the current activation statistics and the expert's

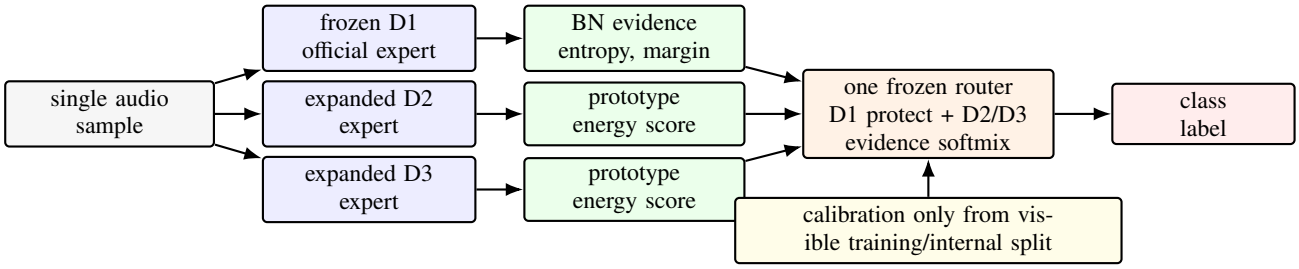

Figure 1: Frozen inference pipeline. Evaluation samples are processed independently by all experts. The router uses compact evidence, not hidden-domain metadata or evaluation-set statistics.

stored BN mean and variance, then average over the evidence layers. Lower BN distance indicates stronger domain compatibility. These statistics are never refreshed on development or evaluation data.

2.2. Router and energy compatibility

The router first protects D1: if D1 has the strongest BN evidence and remains competitive by entropy and margin, D1 is selected. Otherwise, D2 and D3 are compared. For a scalar evidence quantity q_e, $r(q_e)$ is its rank among the D2/D3 experts, with rank 1 better. The clean D2/D3 evidence score is

$$v_e = r(d_e) + 0.5r(b_e) + 0.3r(h_e) - 0.3r(m_e), \quad e \in \{2, 3\}, \quad (2)$$

where b_e, h_e, and m_e denote BN distance, normalized entropy, and probability margin. The final posterior is a soft expert mixture,

$$p(y|x) = \sum_{e\in\{2,3\}} \alpha_e p_e(y|x), \quad \alpha_e = \frac{\exp(-v_e/\tau)}{\sum_j \exp(-v_j/\tau)}. \quad (3)$$

Energy is used as a relative expert-compatibility score, not as a calibrated probability. This is related to confidence and OOD detection, where maximum-softmax, calibration, ODIN, and energy scores have been studied for deciding whether a prediction should be trusted [16, 17, 18, 19]. Our use is narrower: the score only compares D2 and D3 experts after D1 protection, and it is never temperature-tuned on evaluation data. We compute

$$E_e(x;T) = -T \log \sum_c \exp(z_{e,c}(x)/T). \quad (4)$$

For $e \in \{2, 3\}$, μ_e^E and σ_e^E are estimated only on the visible training/internal split, and $s_e(x) = -(E_e(x;T) - \mu_e^E)/(\sigma_e^E + \epsilon)$. The selected D3_ENERGY router uses $T = 2.0$ and $v_e \leftarrow v_e - 0.6s_e$. No D1 energy normalization is used. For ensembles, same-expert logits and evidence are averaged across checkpoints before this single router is applied, following the variance-reduction motivation of deep ensembles [20].

3. DEVELOPMENT ANALYSIS

Table 2 is organized around the routing gap rather than only final submissions. The forced-domain columns estimate expert capacity when the domain is known. The top-five clean ensemble reaches 75.62% forced-D2 and 63.59% forced-D3, but its domain-agnostic D3 score is only 60.61% because 24.69% of D3 samples are routed to D2. D3_ENERGY keeps the same expert capacity while reducing D3-to-D2 to 18.24%, raising the D2/D3 average by 0.73 points over the top-five clean router. The remaining gap to forced-domain performance shows that routing ambiguity, not only classifier capacity, is still the dominant bottleneck.

Table 1: Frozen configuration used for the submitted systems. Calibration never uses development or evaluation-set statistics.

Component	Setting
Experts	D1 official checkpoint frozen; P3a D2 updates blocks 4–6; P3a D3 updates blocks 2–6; cosine heads.
Evidence	BN mismatch, class-conditional prototype distance, entropy, margin, and D2/D3 energy.
Router	Systems 1–3 use D1 protect first, then one D2/D3 soft evidence mixture; system 4 uses a frozen output-level vote.
Energy	$T = 2.0$, weight 0.6, D2/D3 z-normalization from visible training/internal split only.
Ensemble	Systems 1–3 average same-expert logits and evidence before routing; system 4 votes over five frozen variant outputs.
Inference	One frozen prediction per sample; no eval statistics, BN refresh, adaptation, or metadata use.

This comparison also explains the final ranking of submissions. Relative to the single clean P3a system, STABLE_MAIN gains 2.06 points on D3 and 0.85 points in average accuracy while losing 0.36 points on D2. The D2 loss is not ignored; it is the price paid for reducing a systematic D3 failure. D1_SAFE is therefore kept as a separate conservative submission rather than replacing the main system. Its visible D2/D3 average is lower, but it selects D1 more often under rank-1 D1 evidence and provides a hedge against hidden-D1 routing risk.

The fourth submitted system, MAJORITY5, is a fixed output-level majority vote over five frozen trained variants: boundary-margin, pair-boundary, STABLE_MAIN, class-balanced pair-boundary, and D1-distilled pair-boundary. On the development split it obtains 71.29% D2, 60.15% D3, and 65.72% average classwise accuracy. Because it votes over final class outputs rather than selecting one D2/D3 expert, the routing and forced-expert columns are not defined for this row. The vote table and its member order are fixed without using evaluation labels or evaluation-set statistics.

Figure 2 shows the mechanism behind this trade-off. Energy does not change the D1 gate in STABLE_MAIN; D1 selection on D2/D3 remains 1.88%/1.49%. Its main effect is a D2/D3 boundary shift: D3 samples assigned to D3 increase from 73.82% to 80.27%, while D2 samples assigned to D3 also increase from 28.95% to 36.93%. Thus the improvement is not a free calibration gain. It is a controlled reallocation of ambiguous samples toward the D3 expert, justified because the previous clean router had a large D3-to-D2 bias.

Per-class audits were used only as engineering checks, not for per-class threshold tuning. They found no label-map, class-order,

Table 2: Routing gap and final systems on the development split. Forced columns use the correct D2/D3 expert and upper-bound the router.

System	D2	D3	Avg.	D3→D2	F-D2	F-D3
single clean	71.19	60.35	65.77	24.81	74.92	63.90
top5 clean	71.16	60.61	65.89	24.69	75.62	63.59
top5 STABBN	71.34	60.91	66.13	24.57	75.62	63.59
STABLE_MAIN	70.82	62.41	66.62	18.24	75.62	63.59
D1_SAFE	70.42	60.72	65.57	22.46	75.20	63.37
D3_ORIENTED	70.40	62.11	66.26	18.24	74.92	63.90
MAJORITY5	71.29	60.15	65.72	–	–	–

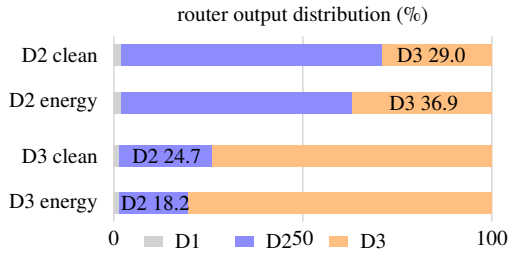

Figure 2: Clean vs. energy routing for the P3a top-five ensemble. Energy shifts D3 samples away from D2 while also moving some D2 samples toward D3, exposing the main trade-off.

checkpoint-loading, or ensemble-weighting error. The most visible hard class in the development diagnostics was *fire*: STABLE_MAIN reached 32.43% on D3 fire and 41.67% on D2 fire, with substantial cross-routing in both domains. This is consistent with acoustic overlap rather than an output-index bug, and it argues against class-specific manual correction under the challenge rules.

4. NEGATIVE RESULTS AND ROBUSTNESS

The rejected experiments are useful because they identify failure modes rather than only weaker scores. Table 3 summarizes the main lessons, while Table 4 gives the larger development trajectory. Aggressive hard routing can minimize D3-to-D2 but does not optimize the task objective: the best strict diagnostic reached 9.68% D3-to-D2, yet D2 fell to 68.12%. This indicates that D2/D3 overlap is asymmetric; rescuing D3 can damage D2 if the router is too discontinuous. A tiny logistic meta-router trained on evidence features was also insufficient, reaching 64.51% D2/D3 average accuracy. The evidence features are valuable, but the visible split was too small for a learned router to replace the rule-based compatibility score robustly.

We also evaluated fixed test-time aggregation because TTA is often tempting for audio; however, adapting or updating parameters at test time, as in test-time adaptation methods such as entropy minimization [21], is outside the frozen evaluation protocol. Even fixed non-adaptive TTA did not consistently help. Window-level routing blurred the file-level evidence and increased D3 confusion in several candidates, so no-TTA was kept for STABLE_MAIN and D3_ORIENTED. D1_SAFE remains a conservative hedge for hidden D1 uncertainty rather than a D2/D3 optimum.

The broader lesson from these failures is methodological. In this task, the router is evaluated through the downstream class prediction, not through domain identification accuracy. A router that is excellent at avoiding one off-diagonal cell can still be worse if it sends acoustically ambiguous D2 samples to D3 or if it changes the evidence distribution seen by the classifier. For future work, this suggests reporting forced-domain expert scores, routing matrices, and final class accuracy together; without all three, it is difficult to distinguish weak experts from weak domain evidence.

Table 3: Ablations and negative results.

Variant	Observation	Lesson
Hard proto	D3→D2 9.68%, but D2 68.12%.	Do not optimize one routing error alone.
Meta-router	Logistic evidence router: 64.51% Avg.	Small splits favor auditable rules over learned gates.
D3-mix	D3-oriented members improved D3 evidence but pushed D2 below 70%.	Expert diversity can shift boundaries.
Fixed TTA	STABLE_MAIN hop1 mean-logit TTA: 66.22 vs. 66.62 no-TTA.	Window averaging changes evidence.
Pruning	Each top-five leave-one-out reduced Avg. by 0.26–0.37.	The ensemble gain was not from one member.

Table 4: Development trajectory. Rejected systems shaped the final interpretation.

Stage	Main lesson
BN-only	Stable and compliant, but D2/D3 separation was too weak without class evidence.
Adapter-DIL	Small updates were attractive, but final D3 capacity lagged behind expanded experts.
X1/X2	Block search identified late-block D2 and deeper D3 updates as a useful capacity pattern.
Class proto	Improved interpretability; hard nearest-prototype routing was brittle when top-1 labels were wrong.
P3a expert	Forced-D3 reached 63.90% single-model and 63.59% top-five, exposing routing as the larger bottleneck.
Energy	Reduced top-five D3→D2 from 24.69% to 18.24%, becoming the final router after audits.

5. FINAL SUBMISSION SYSTEMS

The final package contains four frozen systems rather than a single tuned model. This is intentional: the challenge hides D1 during final evaluation, so the submitted systems cover different risk profiles while preserving the same scientific mechanism. STABLE_MAIN is the primary system, using the P3a top-five ensemble and D3_ENERGY without TTA. D1_SAFE uses a conservative five-model ensemble and 4-s hop-1 mean-logit aggregation because its role is not to maximize visible D2/D3 average, but to reduce risk when D1-like hidden samples are present. D3_ORIENTED uses the same D3_ENERGY router on the best single P3a checkpoint; it has less ensemble complexity and nearly the same D3 routing behavior. MAJORITY5 is a frozen majority vote over five pre-trained variants, with fixed member order for rare ties, and is included as an auditable consensus risk profile rather than as an evaluation-adapted system.

No candidate is selected by inspecting evaluation audio or evaluation output distributions. All router profiles, ensemble members, TTA choices, and energy statistics are frozen before evaluation release. This distinction is important for reproducibility: the final systems differ only in pre-declared risk profiles, not in evaluation-day adaptation.

6. COMPLIANCE, LIMITATIONS, AND CONCLUSION

All calibration is frozen before evaluation release. Energy means and variances are estimated only from the visible training/internal

split; ensemble energy statistics are recomputed from ensemble logits on that same split. During evaluation, each sample is processed independently. The system uses no external data, no external pretrained audio representation, no replay of old-domain audio, no sample-level old-domain embedding cache, no evaluation-set statistics, no BN refresh, no test-time adaptation, and no hidden metadata.

The main limitation is that hidden D1 behavior cannot be fully assessed from visible D2/D3 diagnostics. The second limitation is unresolved D2/D3 overlap: STABLE_MAIN still routes 18.24% of D3 development samples to D2 and routes more D2 samples to D3 than the clean router. These errors suggest that future domain-agnostic audio DIL should study uncertainty and evidence calibration specifically for overlapping acoustic domains rather than treating domain routing as a standard closed-set classification problem.

Two design principles follow from the experiments. First, evidence should be interpreted as compatibility with an expert, not as a separate domain classifier. This is why the final router combines prototype distance, BN mismatch, entropy, margin, and energy instead of assigning a domain from one score. Each signal fails differently: BN evidence is stable but coarse, prototypes are class-aware but depend on top-1 correctness, entropy and margin are sensitive to overconfidence, and energy is helpful mainly for relative D2/D3 separation. Their combination is less elegant than a learned router, but it is auditable and can be frozen before evaluation release.

Second, submission systems should preserve different risk profiles instead of over-compressing all diagnostics into one leaderboard-oriented model. STABLE_MAIN is selected because it improves the D2/D3 average and repairs the main D3-to-D2 error mode. D1_SAFE is not selected by the visible average; it is kept because hidden D1 is the least observable part of the task. D3_ORIENTED is a simpler high-D3 fallback, and MAJORITY5 is a fixed consensus over five frozen variants. This separation matters scientifically: it avoids using hidden evaluation behavior to decide which risk should be preferred after release.

The results also indicate what not to over-interpret. A lower D3-to-D2 rate is useful only when the corresponding D2 loss remains controlled. A stronger expert is useful only when the router can expose it to the right samples. A smoother TTA prediction is useful only when the evidence used for routing remains coherent at the file level. These observations are mundane from an engineering perspective, but they are central for domain-agnostic incremental audio learning because the final error is the product of expert capacity, evidence reliability, and routing policy.

Future work should therefore study evidence calibration under domain overlap as a first-class problem. Promising directions include uncertainty scores that are calibrated across experts without evaluation statistics, reliability-aware layer selection for BN evidence, and routers that can be trained on visible internal splits while remaining interpretable enough to audit. Under Task 7 rules, such methods would still need to avoid old-domain audio replay, sample-level embedding memory, evaluation-set calibration, and any form of test-time parameter update.

For reproducibility, the final experiments were frozen as configuration files rather than as interactive scripts with manual choices. Each candidate records its checkpoint list, router profile, TTA setting, calibration source, and development diagnostics. The inference code averages same-expert logits and evidence before applying one router, which avoids a common but subtle ensemble error: routing each model independently and then voting. Energy statistics for ensembles are recomputed from the ensemble logits on the visible split, rather than reusing single-checkpoint statistics. These details are mundane but important because small implementation differences can change the D2/D3 routing matrix even when the expert checkpoints are identical.

The report also separates diagnostic and submission evidence. Development labels are used to audit errors, compute routing matrices, and write this analysis. They are not used to tune evaluation-day parameters. Hidden evaluation files are treated as independent samples: the system does not infer domains from file names, directory structure, ordering, batch composition, file counts, or aggregate output distributions. This protocol is stricter than many ordinary audio-classification pipelines, but it is necessary for the task's domain-agnostic setting.

Finally, the negative results should be read as constraints on future designs rather than as failed attempts to be ignored. Learned meta-routing may become viable with stronger visible validation protocols, but in this setting it overfit the small split. Hard routing may become useful if prototype evidence is made robust to wrong top-1 classes, but in our experiments it damaged D2. Fixed TTA may help a pure classifier, but for an evidence-routed ensemble it changes the statistics seen by the router. These are practical research questions exposed by the challenge, and they are at least as informative as the small gain of the final system.

This also suggests a useful reporting template for future Task 7 systems. Development accuracy alone should be accompanied by forced-domain expert accuracy, the routing confusion matrix, and an explicit statement of which statistics were estimated before evaluation release. A method can otherwise appear strong for the wrong reason: it may have a better expert but a worse router, or a better router that merely transfers errors from D3 to D2. In our experiments, the most interpretable comparisons were those that separated expert capacity from routing policy. This separation made it possible to reject attractive but unstable variants without relying on hidden evaluation behavior. For a challenge where the domain label is unavailable at inference time, such diagnostics are not bookkeeping details; they are part of the scientific object being measured.

The same separation also bounds the conclusion. The proposed system should not be read as evidence that expert expansion alone solves continual audio learning, or that energy scores are generally calibrated across acoustic domains. Instead, the result is narrower and more reproducible: when the old expert is frozen, replay is disallowed, and the hidden domain label is unavailable, compact aggregate evidence can reveal whether a failure comes from expert capacity or from expert selection. This is why the report emphasizes routing matrices and forced-domain scores alongside accuracy. The final improvement is modest in absolute value, but it changes the dominant D3 error mode while preserving a clean audit trail from visible split statistics to frozen evaluation inference.

This report supports three takeaways. First, under no-replay and no-evaluation-adaptation constraints, domain-agnostic audio DIL is substantially a routing problem after expert expansion. Second, compact aggregate evidence can reduce D3 misrouting without storing old audio, but energy should be interpreted as a relative compatibility heuristic, not a probability. Third, intuitive fixes such as hard routing, tiny meta-routers, D3-heavy ensembles, and fixed TTA can improve a single diagnostic while hurting the robust submission criterion.

7. REFERENCES

[1] "DCASE 2026 challenge task 7: Domain-agnostic incremental learning for audio classification," https://dcase.community/challenge2026/task-domain-agnostic-incremental-learning-for-audio-classification, accessed: 2026-05-31.

[2] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018.

[3] M. Mulimani and A. Mesaros, "Class-incremental learning for multi-label audio classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[4] ——, "Online domain-incremental learning approach to classify acoustic scenes in all locations," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2024.

[5] ——, "Domain-incremental learning for audio classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[6] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of Learning and Motivation*. Academic Press, 1989, vol. 24, pp. 109–165.

[7] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[8] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proc. International Conference on Machine Learning (ICML)*, 2017, pp. 3987–3995.

[9] Z. Li and D. Hoiem, "Learning without forgetting," in *Proc. European Conference on Computer Vision (ECCV)*, 2016, pp. 614–629.

[10] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2001–2010.

[11] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[12] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3366–3385, 2022.

[13] 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.

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

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. International Conference on Machine Learning (ICML)*, 2015, pp. 448–456.

[16] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. International Conference on Machine Learning (ICML)*, 2017, pp. 1321–1330.

[18] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[19] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 21 464–21 475.

[20] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[21] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.