

ENERGY RING: STABILIZING FEATURE ENERGY FOR RELIABLE TEST-TIME DISCOVERY

Tianle Liu¹, Fan Lyu^{*2}, Tingliang Feng³, Chenggong Ni⁴, Tianqi Wang¹, Zhang Zhang⁵, Fuyuan Hu^{*1}, Rui Yao⁶, Tao Zhou⁷

¹School of Electronics and Information Engineering, Suzhou University of Science and Technology

²Computer Vision Center, Universitat Autònoma de Barcelona

³College of Computer Science, Northwest University

⁴School of Intelligence Science and Technology, Nanjing University

⁵New Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

⁶School of Computer Science and Technology, China University of Mining and Technology

⁷School of Computer Science Engineering, North Minzu University

tianleliu@post.usts.edu.cn

fanlyu@cvc.uab.cat

fuyuanhu@mail.usts.edu.cn

ABSTRACT

Test-Time Discovery (TTD) aims to recognize known classes and detect novel ones while the model is already deployed, operating directly on a continual unlabeled test-time stream. Existing TTD methods are overly conservative, making known and unknown features collapse together and difficult to separate during test-time adaptation. In this paper, we revisit TTD from an energy perspective and find that the natural energy gap between known and unknown classes is unstable and easily collapses. We therefore propose Energy Ring to stabilize this gap by anchoring the two types of classes on separate energy rings. A lightweight adapter enlarges the norm-based gap while preserving known-class directions, and a perturbation consistency module forms compact clusters for unknown classes. During testing, energy-guided confidence weighting improves prediction, and a stability-based rectification mechanism corrects unreliable updates. Experimental results show that Energy Ring significantly improves TTD performance against existing TTD methods. Our code will be available at <https://github.com/LeTianL-TT/ER-for-TTD>.

1 INTRODUCTION

Novel Class Discovery (NCD) (Pu et al., 2023; Wu et al., 2023) aims to identify new categories from unlabeled data using knowledge from known classes. However, it separates the testing and discovery phases, assuming a fixed unlabeled set available in advance. Such a static setting is unsupported in open environments, where new classes appear continuously after deployment. Recently, Test-Time Discovery (TTD) (Lyu et al., 2025b) was proposed to detect and learn novel classes while the model is already deployed. Unlike traditional NCD, TTD operates in a continual, unlabeled stream where both known and unknown samples arrive sequentially under test-time constraints, while discovering and testing in real time. This capability is essential for real-world applications where new concepts may emerge unexpectedly after model deployment (Tan et al., 2026a; Lyu et al., 2026a;b). Existing TTD methods usually take a conservative strategy to avoid false discovery of new classes and to protect performance on known ones. As shown in Fig. 1, HM (Lyu et al., 2025b) freezes the backbone so that pretrained features stay fixed. This prevents drift on known classes but also blocks useful updates for new classes. DAA (Liu et al., 2025) adds a learnable adapter that allows updates, yet forces all feature vectors to have the same norm. This improves stability but removes a key source of variation, which may lead to crowding and confusion across unknown and known classes. In this paper, we revisit the TTD problem from an energy view. Prior studies (Sodano et al., 2024; Gu et al., 2025; Park et al., 2023; Ding et al., 2026) suggest that feature energy, measured by the norm of a feature, reflects class-specific information. In many cases, known classes have higher energy and

^{*}Corresponding authors.

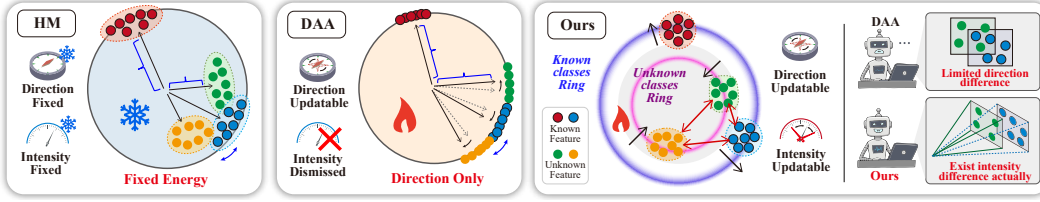


Figure 1: Energy comparisons among existing TTD methods and our method. HM keeps the backbone frozen and classifies based on fixed spatial energy. DAA collapses the original energy intensity onto a single hypersphere. Energy Ring compresses known and unknown classes onto different energy rings, yielding spatial-energy difference and retaining separability.

unknown classes have lower energy. This natural gap can be used as a cue to separate old and new classes at test time, reducing interference and improving discovery. *However, the energy of unknown classes is easy to turn chaotic, and the boundary between high and low energy erodes, yielding the unreliability of the whole system.*

Motivated by this, we aim to *amplify and stabilize* the energy difference between known and unknown classes. Features are expected to be distributed at varying heights during TTD just like concentric rings, with each category forming distinct clusters, thereby facilitating easier discrimination and discovery. In this paper, we therefore propose **Energy Ring**, a TTD framework that can anchor known classes on a high-energy ring and unknown classes on a low-energy ring. Specifically, we introduce a lightweight energy ring adapter that enlarges separation while preserving feature directions of known classes to retain prior knowledge. In addition, we enforce consistent perturbations to encourage compact and stable clustering of unknown samples along the energy ring. During testing, the learned energy-ring structure enables query classification via energy-weighted confidence, while a high-perturbation priority rectification mechanism identifies unstable predictions and corrects them, further improving the accuracy and stability of the energy rings. Our method improves agreement performance by 3.33% compared to DAA (Liu et al., 2025). In summary, our contributions are:

- (1) We propose Energy Ring, a TTD framework that stabilizes the inherently fragile energy gap by anchoring known and unknown classes on separate energy rings.
- (2) We design a lightweight energy ring adapter and a perturbation-consistency module that jointly enlarge the energy gap, preserve known-class directions, and produce compact, stable clusters for unknown classes.
- (3) Extensive experiments demonstrate the effectiveness of Energy Ring, showing clear improvements over prior TTD methods.

2 RELATED WORK

Novel Category Discovery (NCD) (Han et al., 2020; Zhang et al., 2023; Zhong et al., 2021) tackles the problem of recognizing classes that never appear in the training labels. Standard NCD pipelines first learn a transferable representation from the labelled data, then cluster the unlabelled stream to reveal new groups. DTC (Han et al., 2019) exploits semantic anchors from related classes to reduce clustering ambiguity; GCD (Vaze et al., 2022) and PromptCCD (Cendra et al., 2024) replace the traditional k-means stage with contrastive learning prompting for purer prototypes. On-the-fly Category Discovery (OCD) (Du et al., 2023; Zheng et al., 2024) pushes NCD toward real-time operation by hashing incoming images and assigning them to the nearest Hamming bucket. OCD, however, is designed only for discovery whose backbone remains frozen, no fine-tuning on the new classes is performed. NCD methods treat discovery and adaptation as two separate stages.

Test-Time Training (TTT) (Liang et al., 2023; Sun et al., 2020; Tan et al., 2026b; Lyu et al., 2025a) is a paradigm that adapts a pre-trained model to distribution shifts while the model is already deployed. Instead of relying on a frozen model, TTT injects self-supervised objectives and uses the test stream itself to update the backbone on the fly. Representative methods such as TTT++ (Liu et al., 2021b) and TTT-MAE (Gandelsman et al., 2022) show that this on-line alignment significantly improves accuracy under covariate shift (Lin et al., 2023; Mei et al., 2026; Tan et al., 2026c). Yet the vast majority of TTT work assumes the model is only asked to preserve performance on the known classes, leaving it blind to categories that emerge after deployment.

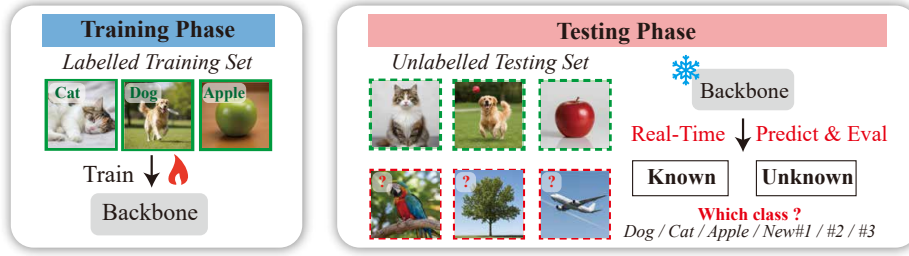


Figure 2: Task setting of TTD. In TTD, a model is pretrained on data containing only known classes. During deployment, the test data may include both known and unknown classes, requiring the model to predict known classes and discover novel ones.

Test-Time Discovery (TTD) unifies the two desiderata: retain accuracy on old classes while instantly detecting and learning new ones during inference. The first TTD implementation, pioneered by Lyu et al. (Lyu et al., 2025b), keeps the network frozen and instead maintains an external memory built with Locality-Sensitive Hashing. By avoiding any gradient update, the method sidesteps catastrophic forgetting, but it also forfeits the possibility of refining the feature space when old and new classes overlap. DAA retains the adaptation spirit of TTT while embracing the discovery mandate of NCD: it updates only a small set of adapters in real time, thereby sharpening the decision boundary between known and emerging classes without erasing previous knowledge.

3 METHODS

3.1 TTD DEFINITION AND ENERGY-BASED DISCOVERY

As shown in Fig. 2, TTD aims to recognize known classes and discover unknown ones in a continuous unlabeled stream after model deployment. Formally, let $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^N$ denote the training set, where $x_i \in \mathcal{X}$ is input image and $y_i \in \mathcal{Y}_{\text{kn}}$ is its label from the known class set $\mathcal{Y}_{\text{kn}}$. During testing, the model encounters a test stream $\mathcal{D}_{\text{test}} = \mathcal{D}_{\text{kn}} \cup \mathcal{D}_{\text{un}}$, where $\mathcal{D}_{\text{un}}$ may contain unknown classes that were never seen during training. TTD requires model to satisfy two objectives simultaneously. First, *novel class discovery and learning*, where inputs $x \in \mathcal{D}_{\text{test}}^{\text{un}}$ must be identified as belonging to unseen classes and assigned new labels to distinguish them from known-class samples. Second, *unified classification*, where all inputs $x \in \mathcal{D}_{\text{test}}$ are accurately classified into either a known class $y \in \mathcal{Y}_{\text{kn}}$ or a novel class $y \in \mathcal{Y}_{\text{seen}}$, with $\mathcal{Y}_{\text{seen}}$ denoting the set of dynamically discovered novel classes.

In this paper, we study TTD from an energy view. During testing, the model receives a mixture of known and unknown samples and must make decisions based on the learned representation $z = F(x)$. For known classes, the supervised objective explicitly optimizes the dependency between visual features and their corresponding class identifiers. In contrast, unknown classes are absent from training and their visual information is not explicitly organized or aligned with the learned class-discriminative space. This distinction often manifests in the magnitude of learned representations. Discriminative training tends to amplify feature responses along directions that are useful for separating known classes, producing more confident and higher-energy representations for in-distribution samples. This motivation is consistent with prior OOD detection and novelty discovery studies showing that known samples often exhibit stronger activation energy, while unknown samples tend to produce lower feature responses (Gu et al., 2025; Sodano et al., 2024; Park et al., 2023; Ni et al., 2025). Thus, we use the feature norm $\|z\|_2$ as a soft indicator of representation confidence, rather than as a strict separation criterion between known and unknown classes. This preliminary creates a natural difference in energy, reflected in both the norm and direction of features. As illustrated in Fig. 3, known classes usually appear with higher energy and unknown ones with lower energy. Simulated features have lower average norms and therefore provide an appropriate low-energy initialization

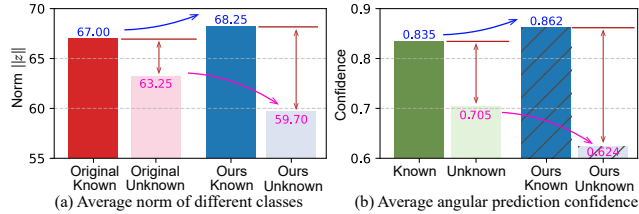


Figure 3: Comparison of norm and angular prediction for test samples. Our method further expands the differences of energy in both intensity and direction.

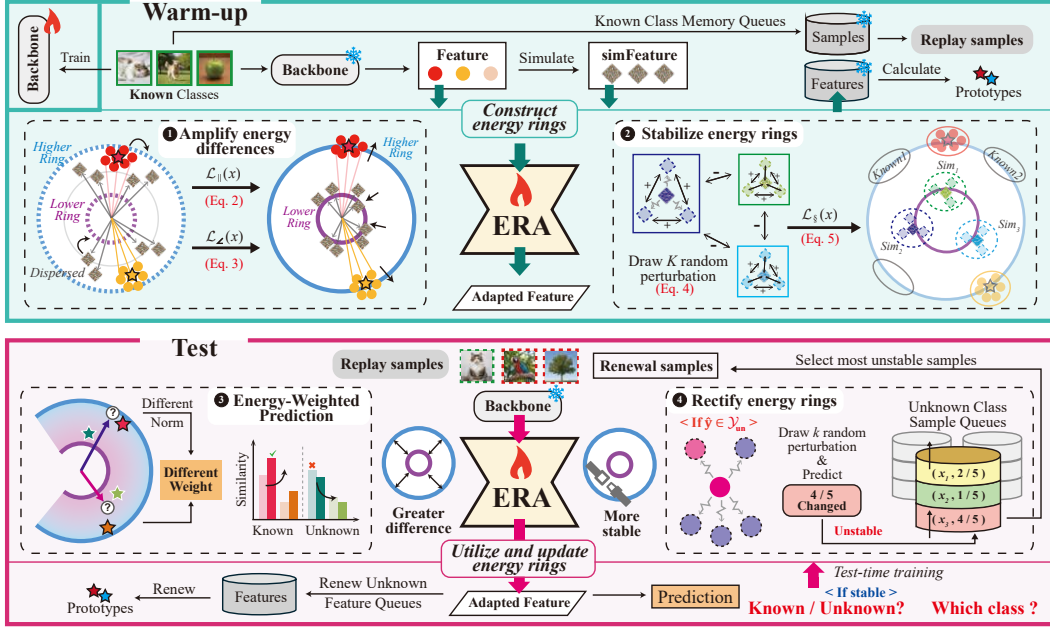


Figure 4: Method Schema. ERA projects features onto concentric energy rings: high-energy ring for known classes and low-energy ring for unknowns, together with perturbation-consistency during training. During testing, we conduct energy-weighted prediction, while selectively rectifying high-perturbation samples to refresh memory queues and update ERA to rectify energy rings.

cue. However, this separation is unstable as their purpose is to expose ERA to ambiguous boundary regions rather than to imitate the semantics of future categories. The norms of different classes overlap, the angular differences are weak, and as more novel classes arrive, these patterns may drift and cause confusion between old and new categories. These observations motivate us to design a method that enlarges and stabilizes the energy difference so that known and unknown classes remain distinguishable during test-time discovery.

We therefore explicitly construct two non-overlapping energy rings, keeping known classes at higher energy ring and unknown classes at lower energy ring. We introduce a lightweight **Energy Ring Adapter (ERA)** and (i) construct energy rings through warm up after training the backbone model to achieve the effect of expanding the energy difference among known and unknown classes while maintaining the energy direction of the learned old classes; (ii) enhance the stability of the unknown classes energy ring through perturbations and prepare for rectification. During testing, we (iii) utilize the constructed energy rings for energy weighted prediction to improve the accuracy of discovery and classification and (iv) conduct test-time training to correct and optimize the energy rings based on the predicted results and high-perturbation priority rectification mechanism.

3.2 PRECONSTRUCT ENERGY RINGS THROUGH WARM-UP

Following DAA (Liu et al., 2025), we also adopt a warm-up strategy to enlarge the gap between known and unknown classes before testing. Unlike DAA, which collapses all features onto a single hypersphere and removes energy differences, our goal is to preserve and further widen the natural gap observed earlier. Since known and unknown classes already show different energy scales, warm-up encourages them to lie on different energy levels, making later discovery more reliable. To achieve this, we introduce an adapter ERA inserted into the pre-trained backbone, which transforms the original feature $\mathbf{r}(x) = f(x)$ into an adapted representation $\mathbf{z} = \text{ERA}(\mathbf{r}(x))$. Specifically, we simulate unknown samples during warm-up by mixing pairs of features with a coefficient λ drawn from a Beta distribution and adding Gaussian noise, forming a synthetic feature $\tilde{\mathbf{r}}(x) = \lambda \mathbf{r}(x) + (1 - \lambda) \mathbf{r}(x') + \mathbf{n}$. These mixed features serve as proxies for unknown classes (Liu et al., 2025; Zhang et al., 2018) and help stabilize and enlarge the energy separation before test-time discovery. Below, we construct the energy rings during warm-up through **three aspects**: 1) *constraining known and unknown classes onto*

different energy tracks (norm constraint), 2) maintaining separability for known classes (direction constraint), and 3) enhancing the stability of the energy rings under perturbation.

First, we enforce a norm constraint that pushes known samples onto an outer ring and unknown samples onto an inner ring, creating a clear spatial gap between them:

$$\mathcal{L}_{||}(x) = \left(\|\text{ERA}(\mathbf{r}(x))\|_2 - (\mathbf{1}_{\text{kn}}(x) r_{\text{kn}} + \mathbf{1}_{\text{un}}(x) r_{\text{un}}) \right)^2, \quad (1)$$

where $\mathbf{1}(x)$ is an indicator function, if x is a known class, $\mathbf{1}_{\text{kn}} = 1$, $\mathbf{1}_{\text{un}} = 0$ and vice versa. r_{kn} and r_{un} denote known and unknown classes energy ring.

Second, to preserve previously learned knowledge, ERA must not alter the semantic direction of known-class features. We maintain feature direction by maximizing cosine similarity between the original and adapted features:

$$\mathcal{L}_{\angle}(x) = 1 - \cos(\mathbf{r}(x), \text{ERA}(\mathbf{r}(x))). \quad (2)$$

Third, even with spatial separation and direction preservation, the simulated unknown samples generated by mixup and noise remain random and unstable. To enhance the consistency of unknown classes, we introduce a perturbation-based constraint. For each adapted unknown feature $\tilde{\mathbf{z}} = \text{ERA}(\tilde{\mathbf{r}}(x))$, we create K norm-preserving perturbations by sampling random directions $\delta \sim \mathcal{N}(0, \mathbf{I})$, projecting them onto the tangent space of $\tilde{\mathbf{z}}$, and scaling them to an angular radius ε . The perturbed features therefore remain on the same energy ring:

$$\mathbf{z}_b = \tilde{\mathbf{z}} + \varepsilon \frac{\delta - (\delta^\top \tilde{\mathbf{z}}) \tilde{\mathbf{z}} / \|\tilde{\mathbf{z}}\|^2}{\|\delta - (\delta^\top \tilde{\mathbf{z}}) \tilde{\mathbf{z}} / \|\tilde{\mathbf{z}}\|^2\|}. \quad (3)$$

We treat the K perturbations of the same feature as soft positives $\mathcal{P}_b = \{\mathbf{z}_b^{(i)} \mid i = 1, \dots, K\}$ and perturbations from different features as soft negatives $\mathcal{N}_b = \{\mathbf{z}_{b'}^{(i)} \mid b' \neq b, i = 1, \dots, K\}$. This encourages perturbations of the same unknown sample to stay close, while pushing those from different samples apart, improving the stability and separability of unknown classes without altering norm information:

$$\mathcal{L}_{\S} = \frac{1}{K} \sum_b \max(0, \|\mathbf{z}_b^{(0)} - \bar{\mathcal{P}}_b\| - \|\mathbf{z}_b^{(0)} - \bar{\mathcal{N}}_b\|). \quad (4)$$

The final warm-up loss combines all three terms:

$$\mathcal{L}_{\text{warmup}} = \mathcal{L}_{||} + \mathcal{L}_{\angle} + \mathcal{L}_{\S}. \quad (5)$$

3.3 ENERGY RING GUIDED PREDICTION AND DISCOVERY

Warm-up establishes the initial energy rings for known and unknown classes. We now move to the testing phase and describe how these energy rings guide prediction and discovery. To provide energy information, similar to DAA, we maintain memory queues for both known and undiscovered classes. For a test image x , its ERA representation $\mathbf{z} = \text{ERA}(\mathbf{r}(x))$. We first compute an energy coefficient ρ based on its position between the two energy rings:

$$\rho = \frac{\|\mathbf{z}\|_2 - r_{\text{un}}}{r_{\text{kn}} - r_{\text{un}}} \in [0, 1]. \quad (6)$$

Let μ_c be the prototype of class c . Features closer to the outer ring should receive slightly higher similarity toward known classes. Given a small amplitude $\alpha \in (0, 1)$, we adjust cosine similarities within a narrow range $[1 - \alpha, 1 + \alpha]$, so the energy prior nudges but does not dominate the prediction:

$$\tilde{s}_c = \begin{cases} \cos(\mathbf{z}, \mu_c) \cdot [1 + \alpha(2\rho - 1)], & c \in \mathcal{Y}_{\text{kn}} \\ \cos(\mathbf{z}, \mu_c) \cdot [1 - \alpha(2\rho - 1)], & c \in \mathcal{Y}_{\text{seen}}. \end{cases} \quad (7)$$

This weighting increases confidence for known classes on the outer ring and decreases confidence for unknown classes on the inner ring. We then use a confidence threshold γ to decide whether a test sample belongs to an existing class or a new one:

$$\hat{y} = \begin{cases} \arg \max_{c \in (\mathcal{Y}_{\text{kn}} \cup \mathcal{Y}_{\text{seen}})} (\tilde{s}_c), & \text{if } \max_c \tilde{s}_c > \gamma, \\ \text{new discovered class}, & \text{otherwise.} \end{cases} \quad (8)$$

To ensure the reliability of an unknown prediction, we adopt a stability recheck strategy. For a sample predicted as an unknown class, we reuse the perturbation mechanism used in warm-up. We draw k random directions $\delta_k \sim \mathcal{N}(0, \mathbf{I})$, project them at the feature $\mathbf{z}_{\hat{y}}$, and form perturbed features $\mathbf{z}^{(k)} = \mathbf{z} + \delta_k$. For each perturbed feature we recompute the similarity scores $\tilde{s}_c^{(k)}$ and obtain a new prediction. The stability of the original prediction is then measured by

$$S = \frac{1}{k} \sum_{i=1}^k \mathbf{1} \left[\arg \max_c \tilde{s}_c^{(i)} = \hat{y} \right] \quad (9)$$

where $\mathbf{1}(\cdot)$ is the indicator function. For each unknown-class sample, we store the tuple (x_c^i, S^i) in the queue. Samples with the lowest stability are selected for rectification and added to the test-time training, helping reduce prototype drift and improving the accuracy of class representations. And the model will treat this sample as the first instance of this new seen class. The image feature $\mathbf{z}$ will be used as the initial prototype for this class, denoted as $\mu_{\hat{y}}$. Subsequent samples will be compared to all known and seen prototypes, including this new $\mu_{\hat{y}}$.

3.4 TEST-TIME TRAINING TO STABILIZE ENERGY RINGS

During testing phase, ERA not only produces energy-aware features but also continues to learn from the streaming batch. The process requires real-time updating of the adapter, strengthening the dual ring spatial structure, while preventing the forgetting of old knowledge, and performing memory correction to selectively update the most unstable unknown samples to refine the class margin. For known classes, we continue to use lower energy ring loss and direction-preservation loss, which penalizes deviations between the original and adapted features and constrains known class features on the larger energy ring to ensure that the ERA does not alter the feature representations significantly. Contrary to known classes, stable unknown classes that have been evaluated are processed using larger energy ring loss and contrastive loss, which encourages ERA to better adapt to new knowledge of unknown classes. Specifically, given a test batch data $\mathcal{B}$, the test-time training with ERA can be represented by two kinds of loss functions. The first kind of loss is the self-supervised learning loss via the pseudo labels:

$$\mathcal{L}_{\text{kn}}(x) = \mathbb{E}_{\hat{y} \in \mathcal{Y}_{\text{kn}}} \left[\mathcal{L}_{||}(x | \hat{y}) + \mathcal{L}_{\angle}(x | \hat{y}) \right], \quad \mathcal{L}_{\text{un}}(x) = \mathbb{E}_{\substack{\hat{y} \in \mathcal{Y}_{\text{un}} \\ S > \kappa}} \left[\mathcal{L}_{||}(x | \hat{y}, S) + \mathcal{L}_{\text{cont}}(x | \hat{y}, S) \right], \quad (10)$$

where $\mathcal{L}_{\text{cont}}$ represents Contrastive Loss (Mao et al., 2023). The second kind represents the replay for known and the renewal for seen classes via retrain samples in memory. Specifically, for every discovered class p we retrieve the samples with the lowest S_p^i (highest instability) regularly, and these samples would be reprocessed, re-evaluated, and update ERA again:

$$\mathcal{L}_{\text{kn}}^{\text{replay}}(x) = \mathbb{E} \mathcal{L}_{||}(x) + \mathcal{L}_{\angle}(x), \quad \mathcal{L}_{\text{un}}^{\text{renewal}}(x) = \mathbb{E} \mathcal{L}_{||}(x) + \mathcal{L}_{\text{cont}}(x). \quad (11)$$

The overall testing phase loss is shown below:

$$\mathcal{L}_{\text{test}}(x) = \lambda_1 \cdot \mathbb{E}_{\hat{y} \in \mathcal{Y}_{\text{kn}}} \left(\mathcal{L}_{\text{kn}}(x | \hat{y}) + \mathcal{L}_{\text{kn}}^{\text{replay}}(x) \right) + \lambda_2 \cdot \mathbb{E}_{\hat{y} \in \mathcal{Y}_{\text{seen}}} \left(\mathcal{L}_{\text{un}}(x | \hat{y}) + \mathcal{L}_{\text{un}}^{\text{renewal}}(x | \hat{y}) \right), \quad (12)$$

where λ_1 and λ_2 are hyperparameters that balance the contributions of each loss term.

4 EXPERIMENT

4.1 EXPERIMENTAL DETAILS

Datasets. We conduct our experiments based on four benchmark datasets. CIFAR100 (C100) (Krizhevsky et al., 2009) is a commonly used classic dataset. Tiny ImageNet (Le & Yang, 2015) is a subset of ImageNet (Deng et al., 2009). Caltech-UCSD Birds-200-2011 (CUB) (Wah et al., 2011) and FGVC-Aircraft (Maji et al., 2013) are complex fine-grained datasets. All these datasets are split into known and unknown classes (7:3). The model is trained on the known training set and tested on the mixture of known and unknown test sets. Because the dataset is used for discovery, we follow the names of four transformed datasets from HM (Lyu et al., 2025b): CIFAR100D, CUB-200D, Tiny-ImageNetD, and AircraftD.

Table 1: Comparisons on four datasets. Real-time evaluation reflects the accumulated performance across all test batches, while post evaluation reassesses all test samples after testing. **Bold** and underlining denote the best and second-best results.

Method		Real-time Evaluation			Post Evaluation			
		KA↑	TA↑	CA↑	KA↑	TA↑	CA↑	KF↓
CIFAR100D	L2P (Wang et al., 2022b)	59.93±2.15	8.57±2.49	43.10±4.30	50.53±7.25	9.60±1.50	27.39±2.11	27.85±7.21
	DP (Wang et al., 2022a)	66.09±1.01	8.80±1.69	48.34±6.78	56.19±2.00	8.68±2.06	28.93±2.25	29.06±1.99
	GMP (Cendra et al., 2024)	72.77±1.20	7.37±0.88	42.26±4.54	67.21±2.53	13.04±2.59	27.10±1.77	17.69±2.53
	PHE (Zheng et al., 2024)	72.35±1.12	16.74±1.05	37.34±1.31	70.12±1.07	14.55± 1.01	24.65±1.35	2.12 ±0.92
	HM (Lyu et al., 2025b)	79.17± 0.13	21.13± 0.62	56.37 ±1.42	80.73 ±1.59	31.03±1.24	34.81± <u>1.22</u>	<u>3.41</u> ±1.49
	DAA (Liu et al., 2025)	80.81±0.44	31.94±0.98	45.38± 0.88	80.27 ± 0.60	35.63±1.11	37.31 ± 1.20	4.10±0.60
	Ours	81.79 ± <u>0.42</u>	33.31 ± <u>0.77</u>	47.42±1.04	80.40 ± <u>0.66</u>	37.23 ±1.18	38.25 ±1.34	4.00± 0.50
CUB200D	L2P (Wang et al., 2022b)	46.22±1.53	9.01± <u>0.87</u>	55.37±7.79	31.97±3.35	4.75± <u>0.73</u>	24.48±1.65	42.29±3.14
	DP (Wang et al., 2022a)	53.69±1.24	43.68±2.20	45.68±5.88	63.37±3.27	45.94±0.91	7.92±1.10	5.85±3.11
	GMP (Cendra et al., 2024)	62.97±1.33	46.44±1.87	47.99±4.34	58.11±3.00	48.02±1.20	10.31±1.45	5.46±2.77
	PHE (Zheng et al., 2024)	57.29±1.23	32.68±1.66	57.22±2.75	54.77±1.75	27.92± 0.43	21.36±1.36	2.18 ±0.71
	HM (Lyu et al., 2025b)	66.20±0.55	58.30 ±2.37	43.33±6.10	64.42± <u>0.65</u>	65.28 ±1.78	37.25±4.90	4.07±0.47
	DAA (Liu et al., 2025)	68.09 ± 0.33	54.49±1.03	58.34 ± 1.35	<u>66.26</u> ± 0.39	64.13±2.23	<u>44.65</u> ± 0.88	3.60 ± 0.39
	Ours	67.10±1.48	<u>57.63</u> ± 0.47	<u>58.19</u> ± <u>1.42</u>	67.38 ±1.82	64.23±1.78	45.53 ±1.05	3.00±0.52
F-ImageNetD	L2P (Wang et al., 2022b)	46.25±1.41	7.79±2.92	53.55±6.47	29.50±3.77	10.38±2.42	23.28± 0.79	47.97±3.77
	DP (Wang et al., 2022a)	46.51± 0.58	6.41±0.93	58.27±6.10	28.53±3.33	9.63±2.10	26.80±1.38	47.57±3.32
	GMP (Cendra et al., 2024)	62.47±1.40	6.25±1.72	58.02±4.29	63.95±2.04	15.64±2.63	26.31±2.33	16.86±2.04
	PHE (Zheng et al., 2024)	63.28±1.29	13.54±1.20	70.33±3.80	58.39± 1.14	12.80± 1.05	19.42±1.45	1.88± 0.35
	HM (Lyu et al., 2025b)	75.31±1.31	16.04± <u>0.76</u>	73.81 ±2.67	74.94±2.20	16.23± <u>1.24</u>	37.43±1.30	1.15 ±2.18
	DAA (Liu et al., 2025)	76.38±0.82	24.10±0.80	71.70± <u>2.02</u>	75.50 ±1.96	23.17±1.35	36.41±1.12	2.39±1.56
	Ours	76.72 ±1.25	25.23 ± 0.71	<u>72.60</u> ± 1.43	<u>75.28</u> ±1.98	24.33 ±1.38	37.93 ±1.49	<u>1.50</u> ± <u>0.95</u>
AircraftD	L2P (Wang et al., 2022b)	43.21±1.08	20.11±1.20	35.52±1.93	40.17±2.34	21.00±1.21	27.21± 1.33	15.04±2.10
	DP (Wang et al., 2022a)	40.18±1.10	21.88±1.23	34.80±1.79	42.53± 1.46	20.22± 0.96	28.15±1.74	12.68±1.74
	GMP (Cendra et al., 2024)	41.26± <u>1.04</u>	21.12± <u>0.67</u>	35.10±2.50	32.56±1.75	21.25± <u>1.02</u>	26.34± <u>1.48</u>	22.65±2.25
	PHE (Zheng et al., 2024)	45.82± 1.02	22.56±1.05	41.72±2.13	47.28±2.88	25.09±1.76	24.38±1.67	7.93±1.13
	HM (Lyu et al., 2025b)	<u>50.35</u> ±1.08	28.01±0.85	40.23±2.22	49.77 ±1.94	29.80 ±1.25	28.74±1.53	5.44±1.56
	DAA (Liu et al., 2025)	49.37±1.46	27.79± 0.31	40.51± 1.05	48.87±1.73	28.37±1.69	28.99±1.69	5.61± <u>1.04</u>
	Ours	50.88 ±1.42	28.49 ±0.76	41.90 ±1.91	<u>49.42</u> ± <u>1.66</u>	<u>29.79</u> ±1.50	30.79 ±1.55	5.20 ± 0.96

Implementation details. In our implementation, we build our method on the prompt-based method L2P (Wang et al., 2022b), which employs a ViT-B/16 backbone (Khan et al., 2022) following the pretraining procedure of NCD and GMP work. We employed the contrastive loss of the GCD literature when we fine-tune the pretrained model on the known classes, using SGD optimizer and cosine decay learning rate scheduler with an initial learning rate of 0.1 and minimum learning rate of 0.0001, and weight decay of 0.00005. All input images are resized to 224×224 and augmented to match the pretrained backbone settings.

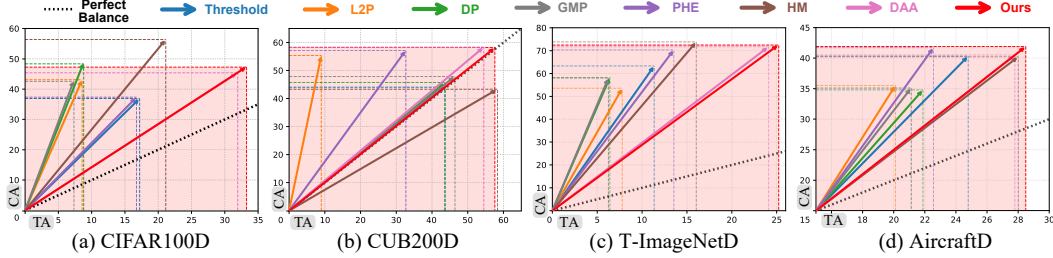
Evaluation metrics. Following HM, we use the following metrics. For unknown classes, we use two agreement metrics: (1) *True-label Agreement ratio (TA)*. This metric measures the maximum proportion of samples from a given true class that are predicted as the same class; and (2) *Cluster Agreement ratio (CA)*. This metric measures the maximum proportion of samples from a given predicted cluster that have the same true label. We also provide some clustering metrics that traditional NCD methods use, including Hungarian Cluster Accuracy (HCA) (Meilă, 2003), Adjusted Rand Index (ARI) (Rand, 1971), Normalized Mutual Information (NMI) (McDaid et al., 2011) and V-Measure (Rosenberg & Hirschberg, 2007). For known classes, we also employ two key metrics to comprehensively assess the model’s performance: *Known Accuracy (KA)* and *Known Forgetting (KF)*. KA measures the traditional classification accuracy of the model on known classes while KF quantifies the degree of performance degradation on known classes over time.

4.2 MAJOR COMPARISONS

In this paper, we first compare our methods with naive thresholding-based training-free methods. When exceeding the threshold, the naive method will be considered to have discovered a new class. And we compare with some training-required methods including L2P (Wang et al., 2022b), DP (Wang et al., 2022a), and GMP (Cendra et al., 2024) as these methods update prompts like TTA and CL methods. We also compare with the recent PHE (Zheng et al., 2024) method for OCD and HM (Lyu et al., 2025b), DAA (Liu et al., 2025) for TTD.

Table 2: Comparisons of clustering metrics of post evaluation on four datasets.

Method	CIFAR100D				CUB200D				T-ImageNetD				AircraftD			
	HCA	ARI	NMI	VM	HCA	ARI	NMI	VM	HCA	ARI	NMI	VM	HCA	ARI	NMI	VM
HM (Lyu et al., 2025b)	63.69	48.35	73.08	73.08	50.46	27.81	70.25	70.25	56.64	40.35	72.70	72.70	33.30	21.33	46.46	46.46
DAA (Liu et al., 2025)	65.62	53.96	75.20	75.21	58.19	32.43	75.29	75.29	58.91	42.50	73.53	73.53	32.08	21.20	46.44	46.44
Ours	66.38	53.54	75.36	75.36	58.34	33.70	75.53	75.53	58.25	43.26	73.36	73.36	33.21	22.13	47.02	47.02

Figure 5: Comparison on balance of Real-time TA and CA. Perfect Balance represents the most balanced TA and CA ($TA = CA$).

Comparisons on TA & CA. The analysis of our experimental results, as illustrated in Table 1, offers key insights into the performance of various methods in the context of Test-Time Discovery (TTD). Training-based methods, which update model parameters upon encountering new classes, tend to suffer from performance degradation. This decline is largely due to their immediate adaptation to novel classes, resulting in lower TA and CA, as well as an increased risk of catastrophic forgetting. On the other hand, training-free methods, such as HM, avoid parameter updates but has restricted capacity to refine its representations in response to novel data. While DAA incorporates a lightweight module that can be updated which only focuses on energy direction and the amplification of inter-class differences remains limited. Our proposed method strikes a more balanced effectiveness among these challenges, achieving more consistent performance across all four datasets. These results demonstrate that our approach enhances the network’s adaptability in discovering novel classes without compromising the retention of previously acquired knowledge.

Comparison of compatibility metrics. TA and CA directly reflect online discovery quality and class consistency as the primary metrics, while the four clustering metrics mainly serve compatibility with post-hoc NCD evaluation. The results in Table 2 demonstrate that our method also has better clustering performance on post evaluation.

Balance of TA & CA. TA and CA are two complementary indicators, and a high individual indicator does not necessarily indicate a good overall classification performance. Therefore, higher values of both metrics, along with greater balance between them, indicate better classification performance. We compared the balance in Fig. 5 and found that some methods have higher values but deviate far from the Perfect Balance line. And our method is more balanced and effective (larger red areas) on different datasets.

4.3 ANALYSIS ON ENERGY RING

Comparison on norm and confidence. Fig. 3 shows that there is indeed a certain norm difference between the features of known and unknown classes obtained by the backbone network, as stated in (Sodano et al., 2024; Lyu et al., 2025b). The average norm of unknown classes is slightly lower than that of known classes. Our method trains simulated classes as unknown classes and significantly reduce the norm of the simulation classes, which successfully simulate unknown classes to a certain extent, thus reducing the norm of the unknown classes. And our strategy makes their average confidence differences larger, which means that unknown classes are easier to be discovered.

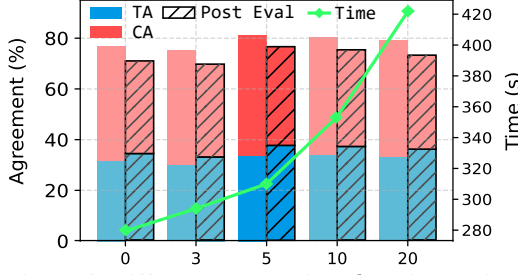
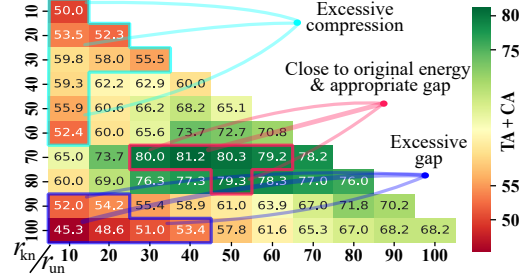
Energy weight. We exploit the spatial variation in feature responses to re-weight categorical similarity, using a norm-based modulation that remains deliberately light (Sec.3.3). The feature-to-category similarities already lie in $[0, 1]$ and the overlap zone clusters around 0.7. Consequently, as Table 3 shows, only a small modulation factor is required, once the exponent becomes large, the relative reliabilities are overturned and the norm effectively seizes control of the decision; in the limiting case of a purely norm-governed weight, the accuracy collapses.

Table 3: Different energy weight during testing.

Energy weight α	Real-time Eval			Post Eval			
	TA	CA	T+C	TA	CA	T+C	KF
0 (No norm)	31.81	44.59	76.40	35.25	36.99	72.24	4.30
0.01	32.56	45.58	78.14	36.42	37.43	73.85	4.07
0.05	33.52	47.66	81.18	37.74	38.92	76.66	3.65
0.10	29.25	48.23	77.48	34.92	40.95	75.87	3.83
0.50	25.70	49.48	75.18	27.25	36.51	63.76	8.49
(Norm-based)	12.87	27.16	40.03	20.92	24.79	45.71	12.90

Table 4: Ablation study. T: Test-time Training, W: Energy-weighted prediction, P: Perturbation rectification, T+C: TA+CA.

T	W	P	Real-time Eval			Post Eval			
			TA	CA	T+C	TA	CA	T+C	KF
✓			30.81	42.13	72.94	34.91	38.53	73.44	2.27
	✓		24.68	53.74	78.42	24.89	42.68	67.57	6.84
	✓		29.61	45.54	75.15	32.44	36.45	68.89	3.05
✓	✓		31.60	44.99	76.59	34.52	36.52	71.04	4.22
✓	✓	✓	33.52	47.66	81.18	37.74	38.92	76.66	3.65

Figure 6: Different perturbations k during testing.Figure 7: Different r_{kn} & r_{un} during Warm-Up.

Ablation study. Table 4 summarizes the ablation results. Test-time Training (T) alone lifts CA from 42.13 to 53.74, but results in larger KF and lowers post-TA, revealing instability. Energy-weighted prediction (W) without T already outperforms the fixed model, pushing real-time TA+CA from 72.94 to 75.15, confirming the effectiveness of energy-weight assistance. Injecting perturbation consistency rectification (P) into the full pipeline helps rectify unstable samples, attaining the best overall trade-off (TA+CA 81.18 real-time and 76.66 post). These results demonstrate that W and P can effectively enable positive test-time update without catastrophic forgetting.

Different Perturbation. We mainly enhance the stability among unknown classes through perturbations, and during testing, we make corrections based on the results of perturbations. In Fig. 6, insufficient perturbations makes the judgment of stability reckless and coincidental changes or no changes can be considered unstable or very stable, so the effect cannot be effectively improved, while excessive perturbations does not significantly improve the effect and introduces greater time cost.

Energy rings during warm-up. We treat the energy ring r_{kn} and r_{un} as a coupled hyper-parameter pair that determines the spatial energy layout of feature. Fig. 7 shows when r_{un} is too small the inner ring is pushed close to the origin with a large gap between r_{kn} , making classification difficult and yielding unstable clusters. The experiment of Fig 3 finds that the norm of backbone features for known classes centers around 70. Therefore, we fix the known class feature norm at 70 and handle unknown classes for better results.

Post visualization using T-SNE. In Fig. 8, we employ t-SNE (Van der Maaten & Hinton, 2008) to visualize the distribution of test samples. Methods that directly update the base model parameters often risk disrupting the model’s existing structure, leading to destructive confusion between classes. In contrast, HM methods cannot update the model, leaving the feature distribution static. Meanwhile, DAA dismisses the spatial difference of feature, limiting its angular amplification effect. In contrast, our method strikes a balance by not only distinguishing features based on angles but also accounting for the energy intensity between the inner and outer energy rings, and by learning updates that separate features across different classes and maintain clearer spatial boundaries than other methods.

Comparison of unknown class label matching. Fig. 9 shows the matching between predictions and ground truth, where the base-model updating method easily leads to chaotic predictions and misclassification. In the fixed-energy method like HM, lots of old samples are classified into unknown clusters. DAA has improved its ability to discover new classes, but there are still some categories that fail to be discovered. Our method can discover a larger number of new classes and make the matching more balanced while classifying fewer known classes as unknown.

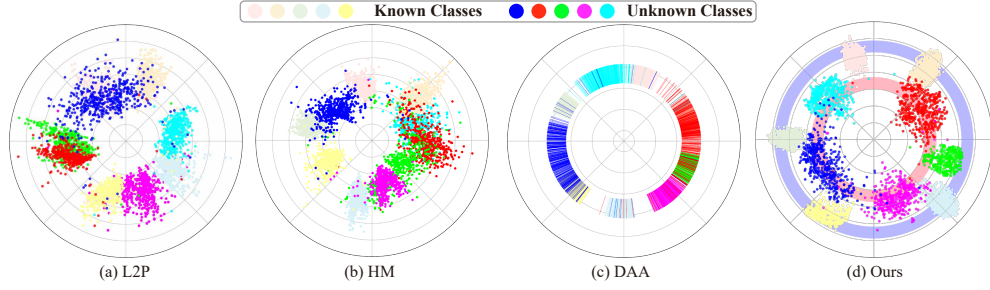


Figure 8: T-SNE visualization in angular space on CIFAR100D with 5 known and 5 unknown classes.

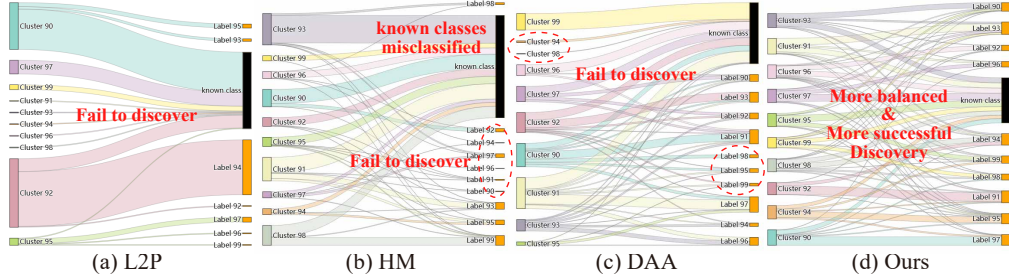


Figure 9: Matching comparison between unknown class predictions and ground truth across methods on CIFAR-100D (90+10). Black rectangles indicate old class samples misclassified as unknowns.

5 CONCLUSION

In this paper, we introduced the Energy Ring framework to address the challenging Test-Time Discovery task. Our method enables instant recognition of unknown classes while safeguarding the accuracy of previously learned ones. The lightweight energy ring adapter compresses feature onto concentric energy rings, preserves known-class directions, enlarges the energy gap and produces compact, stable clusters for unknown classes. Comprehensive experiments on four datasets demonstrate that Energy Ring outperforms in both adaptability and stability. Although energy ring adds marginal computation during warm-up and test-time training, the overhead is negligible compared with the gains in discovery performance. However, the current energy cycle still needs to rely on pre-set energy tracks. Future work will explore more efficient adapter designs, noise-robust update rules, and extensions to multi-modal and continual learning scenarios. And we will explore how to dynamically set energy tracks to achieve wider applicability in the open world.

REFERENCES

- Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- Dhanajit Brahma and Piyush Rai. A probabilistic framework for lifelong test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Chentao Cao, Zhun Zhong, Zhanke Zhou, Tongliang Liu, Yang Liu, Kun Zhang, and Bo Han. Noisy test-time adaptation in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Fernando Julio Cendra, Bingchen Zhao, and Kai Han. Promptccd: Learning gaussian mixture prompt pool for continual category discovery. In *European Conference on Computer Vision*, 2024.
- Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. Improved test-time adaptation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.
- Luqiong Ding, Jiayao Tan, Chenggong Ni, Fuyuan Hu, and Fan Lyu. Vd-lora: Adaptive reuse of low-rank directions for continual learning. In *European Conference on Computer Vision*, 2026.
- Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Ruoyi Du, Dongliang Chang, Kongming Liang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. On-the-fly category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 2022.
- Zhengqing Gao, Xu-Yao Zhang, and Cheng-Lin Liu. Unified entropy optimization for open-set test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- Jiawei Gu, Ziyue Qiao, and Zechao Li. Activation shape matters: Ood detection with norm-entropy fusion. In *Proceedings of ACM International Conference on Multimedia*, 2025.
- Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- Kai Han, Sylvester-Alvise Rebuffi, Sébastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *International Conference on Learning Representations*, 2020.
- X. Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. 2022.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, 2019.
- Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 2022.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 2020.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 2017.
- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 2015.
- Jian Liang, R. He, and Tien-Ping Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 2023.
- Huiwei Lin, Baoquan Zhang, Shanshan Feng, Xutao Li, and Yunming Ye. Pcr: Proxy-based contrastive replay for online class-incremental continual learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

- Jiashuo Liu, Zheyang Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021a.
- Tianle Liu, Fan Lyu, Chenggong Ni, Zhang Zhang, Fuyuan Hu, and Liang Wang. Daa: Amplifying unknown discrepancy for test-time discovery. *Advances in Neural Information Processing Systems*, 2025.
- Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 2021b.
- Fan Lyu, Gongbo Cheng, Daofeng Liu, Linglan Zhao, Zhang Zhang, Fuyuan Hu, and Liang Wang. Mitigating catastrophic forgetting in online continual learning with dual-margin contrastive replay. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a.
- Fan Lyu, Tianle Liu, Zhang Zhang, Fuyuan Hu, and Liang Wang. Test-time discovery via hashing memory. *arXiv preprint arXiv:2503.10699*, 2025b.
- Fan Lyu, Gongbo Cheng, Daofeng Liu, Linglan Zhao, Zhang Zhang, Fuyuan Hu, and Liang Wang. Mitigating catastrophic forgetting in online continual learning with dual-margin contrastive replay. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(4):4433–4449, 2026a. doi: 10.1109/TCSVT.2025.3639309.
- Fan Lyu, Linglan Zhao, Chengyan Liu, Yinying Mei, Zhang Zhang, Baoqing Yu, Tao Zhang, Fuyuan Hu, and Liang Wang. Gain: Global-atomic interaction graph for few-shot class-incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(7):10019–10035, 2026b. doi: 10.1109/TCSVT.2026.3673928.
- Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International Conference on Machine Learning*, 2023.
- Aaron F McDaid, Derek Greene, and Neil Hurley. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*, 2011.
- Yinying Mei, Linglan Zhao, Fan Lyu, Tianle Liu, Fenglei Xu, and Fuyuan Hu. Expansive geometry stabilization for exemplar-free continual learning with pre-trained models. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2026.
- Marina Meilă. Comparing clusterings by the variation of information. In *Proceedings of Learning Theory and Kernel Machines*, 2003.
- Chenggong Ni, Fan Lyu, Jiayao Tan, Fuyuan Hu, Rui Yao, and Tao Zhou. Maintaining consistent inter-class topology in continual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofu Chen, Shijian Zheng, Peilin Zhao, and Minghui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, 2022.
- Jaewoo Park, Jacky Chen Long Chai, Jaeho Yoon, and Andrew Beng Jin Teoh. Understanding the feature norm for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Daiqing Qi, Handong Zhao, and Sheng Li. Better generative replay for continual federated learning. *International Conference on Learning Representations*, 2023.

- William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 1971.
- Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007.
- Matteo Sodano, Federico Magistri, Lucas Nunes, Jens Behley, and Cyrill Stachniss. Open-world semantic segmentation including class similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, 2020.
- Jiayao Tan, Tianle Liu, Fuyuan Hu, Wei Feng, and Liang Wan. E-logic prompt: Unified energy-logic framework for continual visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026a.
- Jiayao Tan, Fan Lyu, Tianle Liu, Fuyuan Hu, and Wei Feng. Towards dynamic modality alignment in multimodal continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.
- Jiayao Tan, Fan Lyu, Chenggong Ni, Fuyuan Hu, Wei Feng, and Rui Yao. Exposing mixture and annotating confusion for active universal test-time adaptation. In *International Conference on Learning Representations*, 2026c.
- Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.
- Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.
- Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.
- Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, 2022a.
- Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.
- Yanan Wu, Zhixiang Chi, Yang Wang, and Songhe Feng. Metagcd: Learning to continually learn in generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, pp. 5635–5662, 2024.
- Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. 2018.
- Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Prototypical hash encoding for on-the-fly fine-grained category discovery. *Advances in Neural Information Processing Systems*, 2024.

Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Lei Zhu, Yanyu Xu, Yong Liu, Rick Siow Mong Goh, and Xinxing Xu. Ad 2 mix: Adversarial and adaptive mixup for unsupervised domain adaptation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.