DURA: DUAL-STAGE UNKNOWN-AWARE REPRESENTATION ADAPTATION FOR UNDERWATER ON-THE-FLY CATEGORY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

On-the-Fly Category Discovery (OCD) aims to recognize known categories and discover new ones from an unlabeled data stream. In underwater scenarios, fine-grained inter-class similarity and substantial appearance variations make this task particularly challenging. Existing methods either keep offline representations fixed during inference or adapt representations learned mainly from known categories. To address this issue, we propose DURA, a Dual Stage Unknown Aware Representation Adaptation framework for OCD. In the offline training stage, DURA constructs pseudo unknown samples and explicitly separates them from known category prototypes, preparing the representation for novel category discovery. In the online inference stage, DURA dynamically expands the prototype memory and adapts both the encoder and category prototypes using the observed query stream, allowing the model to adjust to emerging category structures. For systematic evaluation, we establish UOCDBench, to our knowledge the first benchmark dedicated to underwater OCD. It combines our newly introduced AquaOCD343, containing 343 categories and 67,651 images, with four existing underwater datasets. Experiments on UOCDBench and standard OCD benchmarks demonstrate that DURA consistently achieves state-of-the-art performance, particularly improving novel category discovery while maintaining strong known category recognition.

1 INTRODUCTION

Conventional visual recognition systems are commonly developed under a closed world assumption, where the category space is fixed during training and remains unchanged at test time (Geng et al., 2020; Rawat & Wang, 2017; He et al., 2016). This assumption is restrictive in real world environments, where previously unseen categories may appear after deployment. To address this limitation, recent studies have introduced category discovery as an open world learning paradigm that transfers knowledge from labeled categories to uncover novel categories in unlabeled data. Representative formulations include Novel Category Discovery (NCD) and Generalized Category Discovery (GCD) (Han et al., 2020; Zhao & Han, 2021; Fini et al., 2021; Caselli et al., 2026; Rathore et al., 2026; Rastegar et al., 2023). NCD assumes that the unlabeled data contain only novel categories, whereas GCD allows known and novel categories to coexist. However, both settings typically rely on a predefined unlabeled query set and perform category discovery offline. On-the-Fly Category Discovery (OCD) further extends category discovery to sequential test streams, where models must recognize known categories and discover novel ones as unlabeled samples arrive (Du et al., 2023; Wu et al., 2026; Li et al., 2026c; Zhang et al., 2026).

In underwater environments, deployed recognition systems may encounter categories absent from their training data, making OCD particularly relevant to underwater applications. However, underwater OCD presents challenges beyond those considered in conventional benchmarks. Aquatic categories often exhibit high inter class similarity, with category specific differences appearing only in subtle variations of shape, texture, color patterns, or local structures (Rost et al., 2026; Fuad et al., 2026b; Cong et al., 2025). Meanwhile, changes in water conditions, illumination, imaging environments, and acquisition devices can produce substantial appearance variations across underwater scenes (Li et al., 2026a;b). Representations pretrained on natural images, such as CLIP (Radford

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2021) and DINO (Caron et al., 2021), may therefore transfer imperfectly to underwater imagery. These factors make it difficult to preserve fine grained category discrimination while accommodating novel category structures after deployment. Despite its practical relevance, underwater OCD remains largely unexplored, as existing studies have focused predominantly on generic visual benchmarks.

Existing OCD methods mainly improve novel category discovery by optimizing the offline training stage, including enhancing feature discriminability (Du et al., 2023), improving category prototype modeling (Zheng et al., 2024), and constructing virtual or pseudo unknown samples to simulate potential novel categories (Liu et al., 2025). However, these methods typically keep the learned representation fixed during inference and cannot adapt to the category structure revealed by newly arrived samples. TALON (Wu et al., 2026) addresses this limitation through online adaptation, but its offline representation is still learned mainly from known categories. Thus, offline preparation for potential unknowns and online adaptation to emerging categories remain insufficiently coordinated.

Although this problem is particularly challenging in underwater imagery, the underlying requirement is general to OCD: the representation should be prepared for unseen categories before deployment and remain adaptable as novel category structures emerge. To address this issue, we propose DURA, a Dual Stage Unknown Aware Representation Adaptation framework. DURA does not rely on underwater-specific architectures or physical priors. Instead, it connects offline pseudo unknown separation with online representation and prototype adaptation. In the offline stage, pseudo unknown samples are separated from known category prototypes to prepare the representation for novel category discovery. In the online stage, the representation and prototype memory are further adapted using the observed query stream.

We study DURA primarily in underwater OCD, where fine-grained category similarity and substantial appearance variations make online discovery particularly challenging. To support systematic evaluation in this domain, we introduce AquaOCD343, a large-scale fine-grained underwater dataset containing 343 categories and 67,651 images, and establish UOCDBench by integrating AquaOCD343 with four existing underwater datasets. Experiments on both UOCDBench and standard OCD benchmarks demonstrate the effectiveness of DURA across underwater and conventional visual scenarios.

Our main contributions are summarized as follows:

- We propose DURA, a Dual Stage Unknown Aware Representation Adaptation framework for OCD. DURA combines pseudo unknown separation in offline training with representation adaptation and prototype refinement at test time. It separates pseudo unknown samples from known category prototypes through maximum similarity repulsion and updates online prototypes according to assignment confidence and the number of supporting samples. Extensive experiments show that DURA achieves state-of-the-art (SOTA) performance on both underwater and standard OCD benchmarks, with particularly strong improvements in novel category discovery.
- We introduce AquaOCD343, a large scale fine grained underwater dataset containing 343 categories and 67,651 images. Its broad category coverage and challenging visual variations provide a testbed for evaluating known category recognition and novel category discovery in underwater imagery.
- We establish UOCDBench, to our knowledge the first benchmark dedicated to OCD in underwater imagery. It integrates AquaOCD343 and four existing underwater datasets under a unified evaluation protocol, enabling systematic comparison of OCD methods across diverse underwater settings.

2 Related Work

2.1 Open Set Recognition

Open Set Recognition (OSR) aims to recognize samples from known categories while identifying samples that belong to previously unseen categories at test time (Scheirer et al., 2012). Unlike conventional closed set recognition, OSR explicitly considers the incompleteness of the training label

space. Early OSR methods mainly rely on generative modeling or discriminative learning to characterize unknown regions and establish effective decision boundaries between known and unknown samples (Ge et al., 2017; Neal et al., 2018; Bendale & Boult, 2016; Guo et al., 2021). Recent studies have increasingly investigated OSR from the perspective of representation learning and unknown simulation. SpHOR (Bahavan et al., 2026) revisits OSR through representation learning and improves the transferability of learned features to unseen categories. ViPER (Wang et al., 2026) generates near OOD samples by disrupting positional information in Vision Transformers, providing a data free way to expose the model to potential unknown regions. Recent work also studies adaptation under open set test distributions, including COSTA (Zhang & Li, 2026) with robust prototype learning and DOCO (Yang et al., 2026) under domain and semantic shifts. However, OSR mainly identifies unknown samples, whereas OCD must further organize them into novel categories as unlabeled samples arrive sequentially.

2.2 Category Discovery

Category discovery aims to discover semantic categories from unlabeled data by transferring knowledge learned from labeled categories. Early NCD methods learn transferable representations from labeled categories and use clustering or pseudo labeling to organize unlabeled samples into novel categories (Han et al., 2019; 2020). GCD relaxes this assumption by allowing known and novel categories to coexist in the unlabeled data (Vaze et al., 2022). RAPL (Liu et al., 2024) focuses on fine-grained category discovery, while Hilo (Wang et al., 2025) studies category discovery under domain shift. However, GCD still assumes that a fixed unlabeled set is available before discovery, which differs from practical settings where test samples arrive sequentially and require immediate predictions.

OCD removes this restriction by replacing the fixed unlabeled set with a sequential test stream, where models recognize known categories and discover novel ones as samples arrive (Du et al., 2023). SMILE (Du et al., 2023) represents categories with hash-based descriptors, while PHE (Zheng et al., 2024) improves category representation through prototypical hash encoding. DiffGRE (Liu et al., 2025) and LTC (Zhang et al., 2026) improve offline learning through synthesized novel or pseudo unknown samples, whereas TALON (Wu et al., 2026) introduces test-time adaptation for online category discovery. DURA combines these two directions by coupling offline unknown aware learning with online representation adaptation. Different from existing methods, DURA explicitly separates pseudo unknowns from known prototypes and refines online prototypes according to both assignment confidence and sample support.

3 Methodology

3.1 Problem Definition.

We follow the standard OCD setting. Let $\mathcal{D}_S = \{(x_i, y_i)\}_{i=1}^{N_S} \subset \mathcal{X} \times \mathcal{Y}_S$ denote a labeled support set available for offline training, where $\mathcal{Y}_S$ contains the known categories. At test time, the model receives an unlabeled query stream $\mathcal{D}_Q = \{x_t\}_{t=1}^{N_Q}$ sequentially, whose underlying label space $\mathcal{Y}_Q$ contains both known and novel categories, with $\mathcal{Y}_S \subset \mathcal{Y}_Q$. The novel category set is therefore defined as $\mathcal{Y}_N = \mathcal{Y}_Q \setminus \mathcal{Y}_S$. Labels of query samples are unavailable during inference and are used only for evaluation. For each arriving sample x_t, the model must immediately recognize a known category or discover a novel one using only the support set and previously observed query samples, without access to future samples.

3.2 Method Overview

Figure 1 presents the DURA framework, which consists of an offline training stage and an online inference stage. In the offline training stage, DURA learns discriminative representations from labeled known samples and constructs pseudo unknown samples to explicitly separate potential unknowns from the known prototype set. This stage provides an unknown-aware representation for subsequent category discovery.

In the online inference stage, known-category prototypes initialize a dynamic prototype memory. Incoming queries are matched against the current memory to recognize existing categories or create

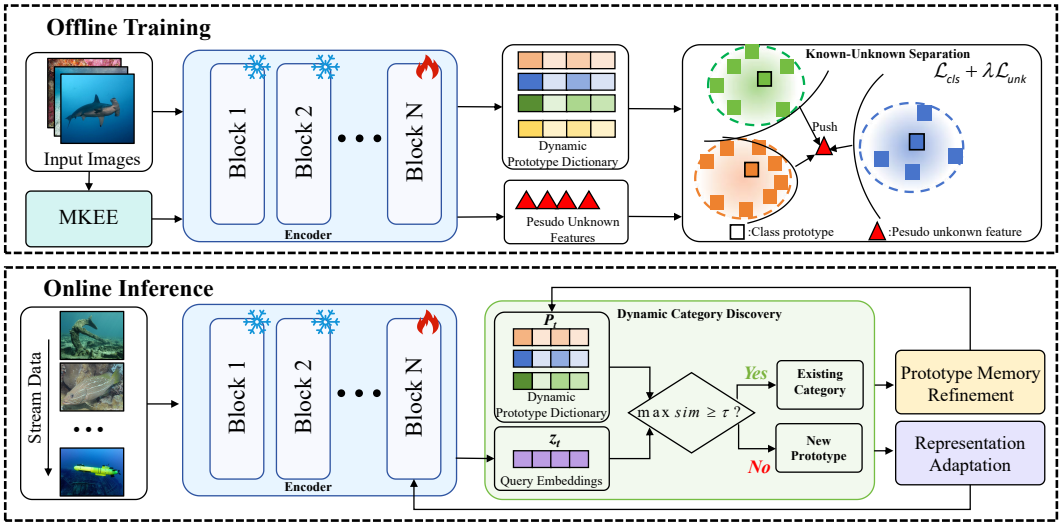


Figure 1: Overview of the proposed DURA framework. During offline training, pseudo unknown samples are generated to improve the separation between known and unknown categories while learning discriminative category representations. In online inference stage, each query is encoded into an embedding and matched against a dynamic prototype memory to recognize an existing category or create a new prototype. The representation and prototype memory are subsequently adapted using the current test data for processing future samples.

new prototypes. The observed category assignments are then used to adapt the encoder and refine the prototype memory, with prototype updates controlled by assignment confidence and sample support.

3.3 Offline Training

In the offline training stage, DURA learns discriminative representations from labeled known samples and constructs pseudo unknown samples to explicitly separate potential unknowns from the known prototype set. The two types of samples share the same visual encoder.

Representation Learning. Given a sample x_i, the visual encoder extracts an ℓ_2-normalized representation $\mathbf{z}_i = f_\theta(x_i)/\|f_\theta(x_i)\|_2$. Each known-category prototype is initialized as the normalized mean representation of its labeled samples. Representations and prototypes are normalized before each similarity computation; hence, $\mathbf{z}_i^\top \mathbf{p}_k$ denotes cosine similarity. The prototypes and encoder parameters are then jointly optimized.

OCD aims to distinguish novel categories that may be semantically close to the known ones. A representation optimized only for known-category recognition may provide insufficient separation between semantically similar categories, making novel samples difficult to distinguish from existing categories. To improve category separation, we apply an angular margin into prototype-based classification, enlarging inter-class margins and improving intra-class compactness. Let $\theta_{i,k}$ denote the angle between the representation $\mathbf{z}_i$ and prototype $\mathbf{p}_k$. The classification logits are defined as

$$\ell_{i,k} = \begin{cases} s\cos(\theta_{i,y_i} + m), & k = y_i, \\ s\cos\theta_{i,k}, & k \neq y_i, \end{cases} \tag{1}$$

where m is the angular margin and s is the scaling factor. The corresponding classification objective is

$$\mathcal{L}_{\text{cls}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\ell_{i,y_i})}{\sum_{k=1}^{K}\exp(\ell_{i,k})}. \tag{2}$$

The angular margin imposes a stricter classification criterion, requiring each sample to achieve a higher similarity to its own prototype than to competing prototypes. This pulls samples toward their corresponding prototypes while enlarging the angular gaps between different known categories.

Such compact and separated category representations make semantically similar categories easier to distinguish and leave clearer space for separating potential unknown categories in the following stage.

Pseudo Unknown Generation. Offline training exposes the model only to labeled known categories, whereas test time discovery requires it to distinguish previously unseen ones. This mismatch leaves the model without direct supervision for separating unknown categories from known ones. Therefore, we construct pseudo unknown samples from the labeled training data to provide such supervision during offline training.

To obtain informative pseudo unknown samples, we first generate interpolated samples using Mixup (Zhang et al., 2017). Specifically, given two randomly paired samples x_i and x_j, we construct a mixed sample as $x^{\mathrm{mix}} = \lambda x_i + (1-\lambda)x_j$, where $\lambda \sim \mathrm{Beta}(\alpha_{\mathrm{mix}}, \alpha_{\mathrm{mix}})$ and α_{mix} controls the interpolation ratio. Such interpolation produces structured samples between known categories, but the mixed samples may still remain within high confidence regions of the known distribution.

Following LTC (Zhang et al., 2026), we adopt Minimizing Kernel Energy and Maximizing Entropy (MKEE) to improve the quality of pseudo unknown samples. MKEE perturbs Mixup samples toward regions with high prediction uncertainty and low feature density. Let $\mathcal{H}(x)$ denote the prediction entropy over known categories and $\rho(x)$ denote the kernel density estimated from known representations. The density is estimated from the distances between the generated sample and a reference set of known representations using an Gaussian radial basis function (RBF) kernel. MKEE jointly considers prediction entropy and feature density:

$$\mathcal{J}(x) = \mathcal{H}(x) - \lambda_\rho \rho(x), \qquad x^u = x^{\mathrm{mix}} + \epsilon \frac{\nabla_{x^{\mathrm{mix}}} \mathcal{J}(x^{\mathrm{mix}})}{\|\nabla_{x^{\mathrm{mix}}} \mathcal{J}(x^{\mathrm{mix}})\|_2}, \tag{3}$$

where λ_ρ controls the contribution of the density term, and ϵ determines the perturbation magnitude. The entropy term increases prediction uncertainty over the known categories, preventing the generated sample from being strongly associated with a single known category. Meanwhile, the density term drives the sample away from densely populated regions of known representations. As a result, the generated pseudo unknowns occupy more uncertain and sparsely populated regions than ordinary interpolated samples. These samples are subsequently used in the category separation objective to reduce their similarity to known prototypes and provide direct supervision for unknown category separation.

Pseudo Unknown Separation. Given the generated pseudo unknown sample x_i^u, the same visual encoder extracts its representation $\mathbf{z}_i^u$. A pseudo unknown sample should have low similarity to every known category. We therefore compare $\mathbf{z}_i^u$ with all known prototypes and use the maximum cosine similarity to identify its closest known category:

$$s_i^u = \max_k \mathbf{z}_i^{u\top} \mathrm{sg}(\mathbf{p}_k), \tag{4}$$

where $\mathrm{sg}(\cdot)$ denotes the stop gradient operation, which keeps the known prototypes fixed when optimizing $\mathcal{L}_{\mathrm{unk}}$. Thus, $\mathcal{L}_{\mathrm{unk}}$ updates the encoder while treating the known prototypes as fixed references. The maximum similarity operation focuses the supervision on the known category with which each pseudo unknown is most easily confused, while stop gradient prevents the generated samples from shifting the learned known category anchors. DURA therefore reduces unknown to known affinity by adapting the encoder without disrupting the prototype structure established from labeled samples. We penalize samples whose maximum similarity exceeds a predefined margin m_u:

$$\mathcal{L}_{\mathrm{unk}} = \frac{1}{B_u} \sum_{i=1}^{B_u} [s_i^u - m_u]_+, \qquad \mathcal{L}_{\mathrm{off}} = \mathcal{L}_{\mathrm{cls}} + \lambda_u \mathcal{L}_{\mathrm{unk}}, \tag{5}$$

where B_u is the number of pseudo unknown samples, $[a]_+ = \max(0, a)$, and λ_u controls the contribution of the unknown separation loss. As a result, this loss pushes pseudo unknown representations away from the known prototype set instead of moving the known prototypes toward the pseudo unknown samples. Together, $\mathcal{L}_{\mathrm{cls}}$ promotes compact and well-separated known categories, while $\mathcal{L}_{\mathrm{unk}}$ reduces the association between pseudo unknown samples and existing known categories. The resulting unknown-aware representation provides the initialization for online discovery, connecting explicit offline unknown separation with adaptation to the novel category structure observed in the query stream.

3.4 ONLINE INFERENCE

Inspired by TALON (Wu et al., 2026), we incorporate online representation adaptation into the inference stage, allowing both the encoder and the dynamic prototype memory to exploit the category structure revealed by incoming queries. Each query is matched against the current prototypes to recognize an existing category or discover a new one. The resulting category assignments are subsequently used to adapt the representation and refine the prototype memory as the query stream evolves. In DURA, this online procedure starts from the unknown-aware representation learned through offline pseudo unknown separation.

Dynamic Category Discovery. The prototype memory is initialized with known category prototypes computed from the labeled samples using the trained encoder. As novel categories are discovered, new prototypes are progressively added to the memory.

Given an incoming query x_t, the encoder extracts its representation $\mathbf{z}_t$. The query representation and the prototypes are normalized before matching, such that their inner products correspond to cosine similarity. We identify the most similar prototype as

$$s_t^{\max} = \max_k \mathbf{z}_t^\top \mathbf{p}_k, \qquad k_t^* = \arg\max_k \mathbf{z}_t^\top \mathbf{p}_k. \tag{6}$$

If $s_t^{\max} \geq \tau$, the query is assigned to the existing category represented by $\mathbf{p}_{k_t^*}$. Otherwise, the query is regarded as belonging to a new category, and its representation initializes a new prototype that is appended to the memory. The prototype memory therefore contains both the original known categories and the novel categories discovered from the query stream.

Representation Adaptation. Query predictions are generated sequentially before any batch-level adaptation is performed. After a query batch has been fully processed, the predictions and representations of these already observed queries are used for one adaptation step, and the updated encoder is applied only to subsequent queries. Offline training prepares the representation for potential unknowns, but cannot capture the category structure that actually emerges in the query stream. DURA uses the predicted assignments of the current query batch to adapt the encoder to the observed category structure while keeping the prototype memory fixed.

For each query representation $\mathbf{z}_i$, we compute its prediction distribution over the current prototype memory from cosine similarities:

$$q_{i,k} = \frac{\exp\left(\gamma\, \mathbf{z}_i^\top \mathbf{p}_k\right)}{\sum_j \exp\left(\gamma\, \mathbf{z}_i^\top \mathbf{p}_j\right)}, \qquad \mathcal{L}_{\text{ent}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_k q_{i,k} \log q_{i,k}, \tag{7}$$

where γ is a scaling factor and B is the batch size. Minimizing $\mathcal{L}_{\text{ent}}$ encourages confident predictions over the current categories.

We further exploit the predicted assignments to preserve the category structure of the current batch. Let $\mathcal{S}_c$ contain the queries assigned to category c, and let $\mathcal{C}_B$ denote the categories appearing in the batch. Their mean representations are aligned with the corresponding prototypes:

$$\bar{\mathbf{z}}_c = \text{norm}\left(\frac{1}{|\mathcal{S}_c|}\sum_{i\in\mathcal{S}_c}\mathbf{z}_i\right), \qquad \mathcal{L}_{\text{align}} = -\frac{1}{|\mathcal{C}_B|}\sum_{c\in\mathcal{C}_B}\bar{\mathbf{z}}_c^\top \mathbf{p}_c. \tag{8}$$

This term maintains consistency between each query group and its assigned prototype.

Meanwhile, different category groups should remain distinguishable as the representation adapts. We minimize their pairwise cosine similarity:

$$\mathcal{L}_{\text{sep}} = \frac{1}{|\mathcal{C}_B|(|\mathcal{C}_B|-1)}\sum_{\substack{c,c'\in\mathcal{C}_B \\ c\neq c'}}\bar{\mathbf{z}}_c^\top \bar{\mathbf{z}}_{c'}. \tag{9}$$

This term preserves separation between the category groups observed in the current batch. When only one category is present, $\mathcal{L}_{\text{sep}}$ is omitted.

The representation adaptation objective is

$$\mathcal{L}_{\text{adapt}} = \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{sep}}. \tag{10}$$

The prototype memory is fixed in this optimization. Only the affine parameters of the normalization layers in the final encoder block are updated to limit representation drift.

Prototype Memory Refinement. To keep the prototype memory representative of the categories observed in the query stream, DURA refines each prototype using its assigned query representations. For prototype $\mathbf{p}_j$, let $\mathcal{S}_j$ denote the set of assigned representations, $n_j = |\mathcal{S}_j|$, and $\bar{\mathbf{z}}_j$ their mean representation. We estimate the assignment confidence by the average cosine similarity between the assigned representations and the current prototype:

$$\mathrm{conf}_j = \frac{1}{n_j} \sum_{\mathbf{z}_i \in \mathcal{S}_j} \mathbf{z}_i^\top \mathbf{p}_j. \tag{11}$$

The prototype is then refined using an adaptive update coefficient:

$$\alpha_j = \eta_j \, \mathrm{conf}_j \frac{n_j}{n_j + \kappa_j}, \qquad \mathbf{p}_j \leftarrow \mathrm{norm}\left((1-\alpha_j)\mathbf{p}_j + \alpha_j \bar{\mathbf{z}}_j\right). \tag{12}$$

The confidence term reduces the influence of uncertain assignments, while $n_j/(n_j+\kappa_j)$ suppresses large updates when only a few supporting queries are available. For known prototypes, $\eta_j = \eta_{\mathrm{known}}$ and $\kappa_j = \kappa_{\mathrm{known}}$; for newly discovered prototypes, $\eta_j = \eta_{\mathrm{novel}}$ and $\kappa_j = \kappa_{\mathrm{novel}}$. We use a smaller update rate for known prototypes to preserve their learned structure, while newly discovered prototypes are updated more rapidly as additional queries are observed. Prototypes assigned only a single query are kept unchanged.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate DURA in both underwater and conventional visual scenarios. For underwater scenes, we constructed the UOCDBench with five datasets: AquaOCD343, AQUA20 (Fuad et al., 2026a), AquaOV255 (Li et al., 2025), MAS3K (Li et al., 2022), and USIS16K (Hong et al., 2025). AquaOCD343 is a large scale fine grained underwater dataset introduced in this work, containing 343 categories and 67,651 images, while the other four datasets are used to further evaluate the generalization of DURA across different underwater data. Beyond underwater datasets, we also conduct experiments on five commonly used visual benchmarks, including Oxford Pets (Parkhi et al., 2012), CIFAR-10, CIFAR-100 (Krizhevsky, 2009), CUB-200-2011 (Wah et al., 2011), and ImageNet-100 (Russakovsky et al., 2015), to evaluate the generality of DURA under standard OCD settings. (More detailed dataset settings are provided in the supplementary material.)

Evaluation Metrics. Following the standard evaluation protocol in (Du et al., 2023), we evaluate performance using clustering accuracy (ACC) under both Greedy–Hungarian and Strict–Hungarian.

$$\mathrm{ACC} = \max_{\pi \in \mathcal{P}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[y_i = \pi(\hat{y}_i)\right], \tag{13}$$

where y_i and $\hat{y}_i$ denote the ground truth and predicted labels, respectively, and $\mathcal{P}$ denotes all possible label permutations. We evaluated the performance of all methods using accuracy on Old, New, and All categories.

Implementation Details. We compare DURA with six representative OCD methods under the standard evaluation protocols. For standard OCD benchmarks, we use the results reported in the original publications when available. Since previous methods do not report results on UOCDBench, we rerun them using their publicly available implementations and published settings. Detailed method configurations, result sources, pretrained backbones, and optimization hyperparameters are provided in Appendix A.2.

4.2 Results on UOCDBench

Table 1 compares DURA with existing OCD methods on five underwater datasets. DURA-DINOv2 achieves the best novel category accuracy on all five datasets under both evaluation protocols. DURA variants also achieve the best overall accuracy across the benchmark, with DURA-DINO performing best on AQUA20 under Strict–Hungarian. These consistent improvements demonstrate the effectiveness of DURA in discovering novel categories across diverse underwater environments.

Table 1: Comparison with SOTA methods on UOCDBench. The best and second best results are highlighted in **bold** and underlined, respectively.

	Method	AquaOCD343			AQUA20			AquaOV255			MAS3K			USIS16K		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Greedy–Hungarian	SLC	29.1	30.8	27.8	39.0	30.7	47.9	52.8	50.3	55.1	39.3	44.4	36.5	49.1	56.3	42.0
	RankStat	29.2	37.6	22.3	38.4	32.6	44.7	45.0	44.8	45.1	49.0	71.1	36.7	58.1	60.8	55.5
	SMILE	37.3	43.3	32.3	63.8	79.3	47.2	29.3	35.6	23.6	43.8	72.0	28.1	37.0	48.7	25.5
	PHE	61.6	82.8	29.7	66.9	89.3	42.9	52.5	93.0	29.0	66.1	72.9	28.0	47.6	78.1	41.8
	LTC	66.4	88.0	33.9	72.6	89.7	54.2	51.3	84.7	31.9	69.3	75.1	36.8	46.3	**87.0**	39.9
	TALON	75.0	94.5	45.8	74.8	85.9	63.0	61.3	91.7	43.7	70.0	76.1	36.0	48.3	78.6	42.6
	DURA-DINO	75.3	94.7	46.2	83.8	90.2	76.9	61.8	93.1	44.9	66.7	72.0	36.8	49.2	78.8	43.6
	DURA-DINOv2	**79.4**	**95.3**	**55.5**	**89.2**	**93.1**	**85.0**	**75.7**	**96.6**	**63.5**	**83.6**	**89.9**	**51.5**	**64.1**	86.0	**59.9**
	DURA-CLIP	73.2	92.2	44.7	76.8	90.9	61.8	62.0	93.3	43.8	72.2	77.7	41.4	47.5	**87.0**	40.0
Strict–Hungarian	SLC	26.9	28.0	26.0	37.9	28.8	47.6	49.3	46.8	51.6	37.8	41.7	35.4	49.3	51.6	47.0
	RankStat	25.9	35.0	18.3	37.5	31.5	43.9	40.9	41.7	40.2	44.5	65.4	31.9	52.7	55.1	50.4
	SMILE	26.7	31.9	22.4	43.5	70.1	14.9	21.8	31.7	12.9	36.6	65.6	19.0	28.5	42.3	15.0
	PHE	53.1	74.0	21.8	52.5	87.9	14.5	45.3	84.3	22.6	35.3	72.9	12.6	42.0	53.1	39.9
	LTC	60.0	**86.9**	19.6	53.3	56.4	49.9	47.0	80.0	27.8	32.7	74.2	7.5	40.8	60.8	37.0
	TALON	65.6	84.8	36.9	56.5	81.3	29.9	53.8	82.5	37.2	34.4	**76.1**	9.2	41.8	54.1	39.4
	DURA-DINO	65.3	84.7	36.2	**82.1**	**88.0**	75.8	57.8	86.8	41.0	29.3	71.5	3.8	42.7	54.4	40.5
	DURA-DINOv2	**70.1**	85.2	**47.4**	71.5	59.3	**84.6**	**71.4**	**89.2**	**61.1**	**63.3**	69.0	**59.8**	**62.8**	**66.6**	**59.1**
	DURA-CLIP	64.1	83.4	35.0	57.5	57.1	57.9	54.3	83.2	37.5	39.2	74.3	18.0	39.5	54.1	36.7

Table 2: Comparison with existing OCD methods on standard benchmarks. Results for the comparison methods are taken from their original publications. The best and second-best reported results are highlighted in **bold** and underlined, respectively. "–" indicates that the corresponding result was not reported in the original paper.

	Method	CIFAR-10			CIFAR-100			ImageNet-100			CUB-200-2011			Oxford Pets		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Greedy–Hungarian	SLC	65.9	96.5	50.9	46.9	62.1	16.6	34.2	86.6	7.1	30.2	46.5	22.1	–	–	–
	RankStat	56.5	81.1	44.2	36.9	45.7	19.3	33.1	74.2	12.4	22.7	30.0	19.1	–	–	–
	SMILE	78.2	**99.3**	67.6	61.3	70.7	42.5	39.9	87.1	16.2	41.1	67.6	27.8	54.1	66.1	47.8
	PHE	83.0	98.0	75.5	64.8	78.8	36.9	53.1	83.5	38.1	46.9	76.0	32.4	63.3	91.3	48.6
	LTC	80.5	98.1	71.7	66.8	81.2	37.8	54.0	90.5	35.7	51.9	82.9	36.3	68.6	92.5	56.0
	TALON	86.2	95.4	79.3	72.5	85.2	47.0	84.1	**94.3**	63.4	52.6	83.3	37.2	81.0	92.4	75.1
	DURA-DINO	91.9	92.0	91.9	71.7	83.5	48.1	79.1	83.0	77.7	53.8	84.2	39.7	82.2	**96.9**	74.4
	DURA-DINOv2	**96.0**	96.3	**95.8**	**80.6**	**89.7**	**62.4**	**86.1**	88.3	**81.7**	**72.8**	**92.4**	**62.9**	**82.6**	95.8	**75.7**
	DURA-CLIP	93.4	92.8	93.9	67.2	75.0	51.6	67.2	75.5	50.5	61.3	87.9	48.0	75.2	92.0	66.4
Strict–Hungarian	SLC	41.5	58.3	33.3	44.4	59.0	15.1	32.9	86.6	5.2	28.6	44.0	20.9	35.5	41.3	33.1
	RankStat	42.1	49.3	38.6	35.0	44.0	17.0	31.1	73.3	9.8	21.2	26.9	18.4	33.2	42.3	28.4
	SMILE	49.9	39.9	54.9	51.6	61.6	31.7	33.8	74.2	13.5	32.2	50.9	22.9	42.9	38.7	45.1
	PHE	53.1	19.3	70.0	56.0	70.1	27.8	39.2	49.3	34.1	36.4	55.8	27.0	48.3	53.8	45.4
	LTC	53.9	19.5	71.1	57.7	75.1	22.8	41.0	61.5	30.7	35.9	52.4	27.6	52.3	52.2	52.4
	TALON	65.0	46.1	79.3	64.7	77.4	39.3	82.6	**92.0**	63.4	43.6	63.5	33.6	69.2	58.5	74.8
	DURA-DINO	71.1	43.4	91.9	61.3	75.0	33.9	80.3	81.9	77.3	39.2	55.1	31.2	70.6	66.4	72.8
	DURA-DINOv2	**96.0**	**96.3**	**95.8**	**72.4**	**80.7**	**55.6**	**85.0**	87.1	**81.0**	**61.3**	**63.7**	**60.1**	**77.5**	**81.1**	**75.7**
	DURA-CLIP	65.9	28.6	93.9	55.8	62.3	42.6	60.5	73.2	47.8	48.1	58.4	42.9	68.0	77.1	63.3

The improvements are particularly pronounced on AQUA20 and MAS3K. On AQUA20, DURA-DINOv2 achieves novel-category accuracies of 85.0% and 84.6%, surpassing the best baselines by 22.0% and 34.7% under Greedy–Hungarian and Strict–Hungarian, respectively. On MAS3K, it achieves 51.5% and 59.8%, improving the corresponding best baseline results by 14.7% and 24.4%. The larger improvements under Strict–Hungarian indicate that DURA better preserves a coherent category structure when known and novel categories are matched jointly.

DURA also performs consistently across different pretrained backbones. With the same DINO backbone, DURA-DINO outperforms TALON in overall accuracy on four of the five datasets under Greedy–Hungarian and three under Strict–Hungarian. For example, on AQUA20, its overall accuracy reaches 83.8% and 82.1%, compared with 74.8% and 56.5% for TALON. DURA-CLIP also achieves competitive results and surpasses TALON in novel-category accuracy on AQUA20 and AquaOV255 under Strict–Hungarian. These results show that the improvements of DURA are consistent across different pretrained representations.

4.3 Results on Standard OCD Datasets

Table 2 compares DURA with existing OCD methods on five standard datasets. DURA achieves the best overall and novel category accuracy on all five datasets under both evaluation protocols. Com-

pared with TALON, it improves overall accuracy by an average of 8.3% and 13.4% under Greedy–Hungarian and Strict–Hungarian, respectively. The improvements in novel category accuracy are even more pronounced, reaching 15.3% and 15.6% on average under the two protocols.

In particular, DURA-DINOv2 substantially outperforms TALON on CUB-200-2011, improving overall accuracy by 20.2% and 17.7% under Greedy–Hungarian and Strict–Hungarian, respectively. Its novel category accuracy reaches 62.9% and 60.1%, compared with 37.2% and 33.6% for TALON, yielding improvements of 25.7% and 26.5%. These results highlight the effectiveness of DURA in distinguishing visually similar fine-grained categories. On CIFAR-10, DURA-DINOv2 achieves 96.0% overall accuracy under Strict–Hungarian, surpassing TALON by 31.0%.

The benefits also extend to other pretrained backbones. Under Greedy–Hungarian, both DURA-DINO and DURA-CLIP outperform TALON in novel-category accuracy on CIFAR-10, CIFAR-100, and CUB-200-2011. Together, these results demonstrate that DURA improves category discovery across different pretrained backbones and extends effectively from underwater imagery to standard visual benchmarks.

4.4 Ablation Studies

Effects of Components. Table 3 evaluates the main components of DURA. The Mixup-only variant retains $\mathcal{L}_{\text{unk}}$ but directly uses mixed samples as pseudo unknowns without MKEE perturbation. Compared with full DURA, its novel-category accuracy decreases by 1.2% and 1.7% under Greedy–Hungarian and Strict–Hungarian, respectively, demonstrating the benefit of generating pseudo unknowns in uncertain and low-density regions. Disabling online adaptation reduces novel-category accuracy by 1.6% and 1.1% under the two protocols. Removing either representation adaptation or prototype refinement also degrades performance. These results show that MKEE provides more informative offline unknown supervision, while encoder adaptation and prototype refinement both contribute to online category discovery.

Table 3: Ablation study of the main components of DURA on AquaOCD343.

	Greedy			Strict		
Variant	All	Old	New	All	Old	New
Mixup only	79.0	94.8	54.3	69.4	85.1	45.7
w/o Online Adapt.	78.6	95.1	53.9	69.5	84.9	46.3
w/o Repr. Adapt.	79.0	**95.3**	54.5	69.8	**85.7**	45.8
w/o Proto. Refine.	78.9	95.1	54.5	68.9	84.8	46.8
DURA	**79.4**	**95.3**	**55.5**	**70.1**	85.2	**47.4**

Effects of Loss Functions. Table 4 evaluates the effects of different training objectives in DURA. Removing $\mathcal{L}_{\text{cls}}$ produces the highest performance degradation, which confirms its importance for learning discriminative representations of known categories. Furthermore, removing $\mathcal{L}_{\text{unk}}$ significantly reduces the accuracy for new categories, indicating that the separation of unknown samples is crucial for the novel category discovery task.. In addition, we found that introducing supervised contrastive learning resulted in a slight decline in performance. This suggests that highly compact representations of known categories may limit the ability to distinguish between newly added unknown categories.

Table 4: Ablation study of training objectives on AquaOCD343.

	Greedy			Strict		
Variant	All	Old	New	All	Old	New
w/o $\mathcal{L}_{\text{cls}}$	64.3	76.2	46.3	59.0	72.4	38.8
w/o $\mathcal{L}_{\text{unk}}$	77.3	93.9	52.4	67.3	**89.0**	34.8
+ $\mathcal{L}_{\text{sup}}$	78.0	93.6	54.5	68.7	83.1	47.1
DURA	**79.4**	**95.3**	**55.5**	**70.1**	85.2	**47.4**

5 Conclusion

In this paper, we propose DURA, a Dual-Stage Unknown-Aware Representation Adaptation framework for On-the-Fly Category Discovery in underwater imagery. By combining pseudo-unknown supervision in offline training with representation and prototype adaptation at test time, DURA better prepares for unseen categories and adapts to emerging category structures. We further introduce AquaOCD343 and establish UOCDBench for systematic evaluation of underwater OCD. Extensive experiments on underwater and standard OCD benchmarks demonstrate that DURA consistently outperforms existing methods, particularly in novel-category discovery.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript translation and editing. All AI assisted content was reviewed and verified by the authors. Generative AI was not used to generate experimental results or replace empirical evaluation. The authors take full responsibility for the final content of this work.

REFERENCES

Thiru Thillai Nadarasar Bahavan, Sachith Seneviratne, and Saman Halgamuge. Sphor: A representation learning perspective on open-set recognition for identifying unknown classes in deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6901–6910, 2026.

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, 2016.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640. IEEE, 10 2021.

Lorenzo Caselli, Marco Mistretta, Simone Magistri, and Andrew Bagdanov. Spectralgcd: Spectral concept selection and cross-modal representation learning for generalized category discovery. In *International Conference on Learning Representations*, volume 2026, pp. 18162–18182, 2026.

Runmin Cong, Zongji Yu, Hao Fang, Haoyan Sun, and Sam Kwong. Uis-mamba: Exploring mamba for underwater instance segmentation via dynamic tree scan and hidden state weaken. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 343–352, 2025.

Ruoyi Du, Dongliang Chang, Kongming Liang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. On-the-fly category discovery. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11691–11700. IEEE, 2023.

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9264–9272. IEEE, 2021.

Taufikur Rahman Fuad, Sabbir Ahmed, and Shahriar Ivan. AQUA20: A Benchmark Dataset for Underwater Species Classification under Challenging Conditions. *Arabian Journal for Science and Engineering*, March 2026a. ISSN 2191-4281. doi: 10.1007/s13369-026-11231-z. URL https://doi.org/10.1007/s13369-026-11231-z.

Taufikur Rahman Fuad, Sabbir Ahmed, and Shahriar Ivan. AQUA20: A benchmark dataset for underwater species classification under challenging conditions. *Arabian Journal for Science and Engineering*, pp. 1–22, 2026b. ISSN 2193-567X. doi: 10.1007/s13369-026-11231-z.

ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017.

Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020.

Yunrui Guo, Guglielmo Camporese, Wenjing Yang, Alessandro Sperduti, and Lamberto Ballan. Conditional variational capsule network for open set recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 103–111. IEEE, 2021.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8400–8408. IEEE, 2019.

Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. *arXiv preprint arXiv:2002.05714*, 2020.

John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Lin Hong, Xin Wang, Yihao Li, and Xia Wang. USIS16K: High-quality dataset for underwater salient instance segmentation. *arXiv preprint arXiv:2506.19472*, 2025.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario, Canada, 2009.

Bingyu Li, Tao Huo, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Exploring the underwater world segmentation without extra training. *arXiv preprint arXiv:2511.07923*, 2025.

Bingyu Li, Tao Huo, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Exploring the underwater world segmentation without extra training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39879–39889, 2026a.

Bingyu Li, Feiyu Wang, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Maris: Marine open-vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24939–24949, 2026b.

Chunming Li, Shidong Wang, and Haofeng Zhang. Adaptive gaussian expansion for on-the-fly category discovery. In *The Fourteenth International Conference on Learning Representations*, 2026c.

Lin Li, Bo Dong, Eric Rigall, Tao Zhou, Junyu Dong, and Geng Chen. Marine animal segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:2303–2314, 2022. doi: 10.1109/TCSVT.2021.3094383.

Xiao Liu, Nan Pu, Haiyang Zheng, Wenjing Li, Nicu Sebe, and Zhun Zhong. Generate, refine, and encode: Leveraging synthesized novel samples for on-the-fly fine-grained category discovery. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1078–1087. IEEE, 2025.

Yu Liu, Yaqi Cai, Qi Jia, Binglin Qiu, Weimin Wang, and Nan Pu. Novel class discovery for ultra-fine-grained visual categorization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17679–17688. IEEE, 2024.

Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *European conference on computer vision*, pp. 620–635. Springer, 2018.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505. IEEE, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Ganti, Girish Sastry, Amanda Askell, Pamela Agarwal, et al., and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 7 2021.

Sarah Rastegar, Hazel Doughty, and Cees Snoek. Learn to categorize or categorize to learn? self-coding for generalized category discovery. *Advances in Neural Information Processing Systems*, 36:72794–72818, 2023.

Vaibhav Rathore, Divyam Gupta, Moloud Abdar, Subhasis Chaudhuri, and Biplab Banerjee. Focus: Bridging fine-grained recognition and open-world discovery across domains. *arXiv preprint arXiv:2603.14240*, 2026.

Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9):2352–2449, 2017.

Thomas Manuel Rost, Martina Figlia, F. Morgado-Dias, and Marko Radeta. Label-efficient underwater image classification with logistic regression on frozen foundation model embeddings, 2026. URL https://arxiv.org/abs/2604.00313.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772, 2012.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. *arXiv preprint arXiv:2201.02609*, 2022.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, Pasadena, CA, USA, 2011.

Hongjun Wang, Sagar Vaze, and Kai Han. Hilo: A learning framework for generalized category discovery robust to domain shifts. In *International Conference on Learning Representations*, volume 2025, pp. 80686–80727, 2025.

Yu Wang, Jiabo Xie, Yucan Zhou, Junxian Mu, Qinghua Hu, and Pengfei Zhu. Disrupting positional encoding for effective open set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6633–6642, 2026.

Yanan Wu, Yuhan Yan, Tailai Chen, Zhixiang Chi, ZiZhang Wu, Yi Jin, Yang Wang, and Li Zhenbo. Talon: Test-time adaptive learning for on-the-fly category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yingkai Yang, Chaoqi Chen, and Hui Huang. Back to source: Open-set continual test-time adaptation via domain compensation. *arXiv preprint arXiv:2604.21772*, 2026.

Bohan Zhang, Weidong Tang, Zhixiang Chi, Yi Jin, Zhenbo Li, Yang Wang, and Yanan Wu. Learning through creation: A hash-free framework for on-the-fly category discovery. *ArXiv*, abs/2603.13858, 2026. URL https://api.semanticscholar.org/CorpusID:286568460.

Can Zhang and Ruirui Li. Costa: Collaborative open-set test-time adaptation through robust prototype learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6154–6163, 2026.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *Advances in Neural Information Processing Systems*, 34:22982–22994, 2021.

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*, 37:101428–101455, 2024.

A APPENDIX

A.1 DATASET AND BENCHMARK DETAILS

We evaluate DURA on two groups of datasets: five underwater datasets included in UOCDBench and five standard datasets commonly used for OCD evaluation. Let $\mathcal{Y}_S$ denote the set of known categories available during offline training and $\mathcal{Y}_Q$ denote the complete category space of the query stream, which contains both known and novel categories. Accordingly, the novel category set is defined as $\mathcal{Y}_N = \mathcal{Y}_Q \setminus \mathcal{Y}_S$. The labeled support set and the unlabeled query stream are denoted by $\mathcal{D}_S$ and $\mathcal{D}_Q$, respectively.

Underwater Datasets. UOCDBench consists of AquaOCD343, AQUA20 (Fuad et al., 2026a), AquaOV255 (Li et al., 2025), MAS3K (Li et al., 2022), and USIS16K (Hong et al., 2025). These datasets cover diverse underwater categories and imaging conditions, enabling evaluation across different category scales and levels of visual granularity. For each dataset, the categories are divided into known and novel subsets. For the known categories, 50% of the images are used to construct the labeled support set $\mathcal{D}_S$. The remaining known category images are combined with all images from the novel categories to form the unlabeled query stream $\mathcal{D}_Q$. Therefore, $\mathcal{D}_Q$ contains samples from both known and novel categories. The resulting dataset statistics are summarized in Table 5.

Table 5: Statistics of the underwater datasets included in UOCDBench.

	AQUA20	AquaOV255	MAS3K	USIS16K	AquaOCD343
$\lvert\mathcal{Y}_S\rvert$	10	117	20	80	156
$\lvert\mathcal{Y}_Q\rvert$	20	255	37	160	343
$\lvert\mathcal{D}_S\rvert$	2,778	5,566	797	5,344	25,376
$\lvert\mathcal{D}_Q\rvert$	5,393	15,157	2,113	10,807	42,275

AquaOCD343. AquaOCD343 contains 67,651 images covering 343 categories. It combines images from AquaOV255, USIS10K, and DYB-PlanktonNet with images collected from online websites and underwater plankton images captured by our team. Its category vocabulary covers fish, crustaceans, corals, plankton, underwater objects, and particle categories. By bringing these data sources together, AquaOCD343 supports the evaluation of category discovery across a broad range of underwater targets, including fine grained biological categories and other objects encountered in underwater imagery.

AquaOCD343 exhibits a pronounced long tailed class distribution. The number of images per category ranges from 2 to 7,760, with a mean of 197.2 and a median of 75. In particular, 19 categories contain fewer than 20 images, whereas 9 categories contain at least 1,000 images. This variation requires discovery methods to handle both frequently observed categories and rare categories represented by only a few samples.

We partition AquaOCD343 into 156 known categories and 187 novel categories. Under the support/query split reported in Table 5, novel categories account for 54.5% of the categories in the query stream but only 40.0% of its images. The model must therefore discover a larger set of unseen categories from fewer observations per category on average. This setting introduces additional challenges associated with rare category discovery and unequal category frequencies.

UOCDBench. We establish UOCDBench, to our knowledge the first benchmark dedicated to on-the-fly category discovery in underwater imagery. It brings together AquaOCD343, AQUA20, AquaOV255, MAS3K, and USIS16K under a common evaluation framework. Each dataset is evaluated independently using the category and support/query partitions summarized in Table 5. All methods learn from labeled known category samples and process an unlabeled query stream containing both known and novel categories, without access to future queries. We report All, Old, and New accuracy under both Greedy–Hungarian and Strict–Hungarian matching, enabling systematic comparison across different underwater category spaces and sample distributions.

Standard Datasets. We additionally evaluate DURA on five standard OCD benchmarks: CIFAR-10, CIFAR-100 (Krizhevsky, 2009), ImageNet-100 (Russakovsky et al., 2015), CUB-200-

2011 (Wah et al., 2011), and Oxford Pets (Parkhi et al., 2012). CIFAR-10, CIFAR-100, and ImageNet-100 mainly evaluate category discovery in general object recognition, whereas CUB-200-2011 and Oxford Pets provide more fine grained category distinctions. Following TALON (Wu et al., 2026), we use the same known/novel category partitions and support/query splits for these datasets. Specifically, 50% of the images from the known categories form the labeled support set, while the remaining known category images and all novel category images form the unlabeled query stream. The detailed statistics are reported in Table 6.

Table 6: Statistics of the standard datasets used in our experiments.

	CIFAR-10	CIFAR-100	ImageNet-100	CUB-200	Pets
$\lvert\mathcal{Y}_S\rvert$	5	80	80	100	19
$\lvert\mathcal{Y}_Q\rvert$	10	100	100	200	38
$\lvert\mathcal{D}_S\rvert$	12.5K	20.0K	31.9K	1.5K	0.9K
$\lvert\mathcal{D}_Q\rvert$	37.5K	30.0K	95.3K	4.5K	2.7K

A.2 Implementation Details

Compared Methods and Result Sources. We compare DURA with six representative OCD methods, including SLC (Hartigan, 1975), RankStat, SMILE (Du et al., 2023), PHE (Zheng et al., 2024), LTC (Zhang et al., 2026), and TALON (Wu et al., 2026).

For the standard OCD benchmarks, we report the results published in the original papers under their respective experimental settings. Results that were not evaluated or reported in the corresponding papers are denoted by "–".

Since previous methods do not report results on UOCDBench, we evaluate them using their publicly available implementations and published configurations, including the pretrained backbones adopted in their original papers. All methods use the support/query partitions and evaluation protocols defined by UOCDBench. The detailed backbone configurations and result sources are summarized in Table **??**.

For DURA, we report results using CLIP-pretrained ViT-B/16, DINO-pretrained ViT-B/16, and DINOv2-pretrained ViT-B/14. These variants are denoted as DURA-CLIP, DURA-DINO, and DURA-DINOv2, respectively.

Offline Optimization. We train all DURA variants for 100 epochs with a batch size of 16. We use AdamW with a weight decay of 1×10^{-4}. The learning rates for the trainable backbone parameters and category prototypes are set to 1×10^{-5} and 5×10^{-4}, respectively. We use a cosine learning rate schedule with a 10-epoch warm-up initialized at 1×10^{-6}. The scale factor s in $\mathcal{L}_{\text{cls}}$ is set to 30, and the angular margin m is set to 0.4.

Pseudo Unknown Generation. We set $\alpha_{\text{mix}} = 0.4$, $\epsilon = 0.05$, and $\lambda_\rho = 0.1$. For each labeled batch, one pseudo unknown sample is generated for each input sample, resulting in the same number of pseudo unknowns as labeled samples. The reference representations for density estimation are extracted from the current labeled batch. The Gaussian radial basis function (RBF) kernel bandwidth is determined by the median pairwise distance between the generated and reference representations.

Sequential Inference and Online Adaptation. Query samples are processed sequentially within each query batch. Each arriving sample is first encoded and matched against the current prototype memory. It is assigned to an existing category if its maximum similarity satisfies the discovery threshold; otherwise, a new prototype is created immediately and becomes available to subsequent queries. Predictions are recorded before online adaptation and are not revised retrospectively.

After all samples in the current query batch are assigned, DURA performs one gradient step for representation adaptation, followed by prototype refinement. Only queries in the current batch are used for these updates. The updated encoder and prototype memory are then used for subsequent query batches. The final incomplete query batch, if any, is processed in the same manner.

Online Representation Adaptation. Only the affine parameters of the normalization layers in the final Transformer block are updated using SGD, while the prototype memory is held fixed during encoder optimization. The learning rate is set to 0.1 for CLIP and 0.01 for DINO and DINOv2. The scaling factor γ is set to 30, and the gradient norm is clipped at 1.0. The category assignments obtained before adaptation are fixed when computing the adaptation losses. When the current query batch contains only one predicted category, $\mathcal{L}_{\text{sep}}$ is set to zero.

Prototype Refinement and Discovery Threshold. Representation adaptation is performed before prototype refinement. Prototype refinement uses the normalized query representations extracted before the encoder update. For each prototype $\mathbf{p}_j$, the assigned set $\mathcal{S}_j$ contains the queries in the current batch assigned to category j, and $n_j = |\mathcal{S}_j|$ denotes their number. Prototypes with no assigned queries remain unchanged, and those supported by only one query are not refined.

The prototype update rates are $\eta_{\text{known}} = 0.02$ and $\eta_{\text{novel}} = 0.1$. The corresponding damping parameters are $\kappa_{\text{known}} = 32$ and $\kappa_{\text{novel}} = 8$. The discovery threshold is estimated from the labeled support set as the 0.001 quantile of the maximum cosine similarities to the known prototypes and is capped at $\tau_{\max} = 0.65$.

All experiments conducted in this work use a single NVIDIA RTX 5090 GPU.

A.3 Comparison of Estimated Category Numbers

Table 7 compares the total numbers of categories estimated by different methods on UOCDBench. The ground truth count includes both known and novel categories in the query stream. An estimate closer to the ground truth indicates more accurate category count estimation. This comparison complements clustering accuracy by revealing overestimation and underestimation of the category space.

Table 7: Comparison of estimated total category numbers on UOCDBench. Ground truth denotes the actual number of known and novel categories combined. The estimates with the smallest and second smallest differences from the ground truth are highlighted in **bold** and underlined, respectively.

Method	AQUA20	AquaOV255	MAS3K	USIS16K	AquaOCD343
Ground truth	20	255	37	160	343
SLC	196	2,437	846	698	13548
AutoNovel	71	852	174	599	3,562
SMILE	174	2,922	536	468	4,584
PHE	63	852	79	289	1,094
LTC	45	726	<u>50</u>	499	1,656
TALON	29	<u>607</u>	115	217	981
DURA-DINO	<u>24</u>	1,226	20	**117**	**703**
DURA-DINOv2	**23**	768	**31**	<u>116</u>	<u>710</u>
DURA-CLIP	34	**548**	107	104	759

Different DURA variants produce the closest estimates on different datasets. DURA-DINOv2 estimates 23 and 31 categories on AQUA20 and MAS3K, respectively, compared with the ground-truth counts of 20 and 37. DURA-DINO provides the closest estimates on USIS16K and AquaOCD343, while DURA-CLIP performs best on AquaOV255. On AquaOCD343, DURA-DINO and DURA-DINOv2 estimate 703 and 710 categories, respectively, compared with 981 for TALON, 1,094 for PHE, and 1,656 for LTC. These results show that DURA can reduce excessive category estimation, although the effectiveness varies across backbones and datasets.

Category count estimation nevertheless remains challenging. All evaluated methods overestimate the number of categories on AquaOV255 and AquaOCD343, whereas all DURA variants underestimate the count on USIS16K. Moreover, a close category count does not necessarily imply accurate

Table 8: Dual-stage ablation on AquaOCD343. All variants retain offline classification training with $\mathcal{L}_{\text{cls}}$.

Offline unknown supervision	Online adaptation	Greedy–Hungarian			Strict–Hungarian		
		All	Old	New	All	Old	New
Off	Off	73.0	87.4	51.2	63.8	81.9	36.7
On	Off	78.6	95.1	53.9	69.5	84.9	46.3
Off	On	77.3	93.9	52.4	67.3	**89.0**	34.8
On	On	**79.4**	**95.3**	**55.5**	**70.1**	85.2	**47.4**

sample assignments, since incorrect splitting and merging can offset each other. The estimated counts should therefore be considered together with the clustering accuracies reported in Table 1.

A.4 Visualization of Online Category Discovery

To illustrate how novel categories emerge during online discovery, we visualize a sequence of t-SNE maps on AquaOCD343 using DURA-DINOv2, as shown in Figure 2. Panel (a) shows the initial known categories, while panels (b–f) present successive stages of discovery as additional query samples are processed. Known category samples are shown in gray, and different colors indicate different discovered novel categories. As the query stream progresses, additional novel category clusters appear alongside the known categories. This visualization illustrates how DURA progressively expands its category space using incoming observations.

A.5 Dual Stage Ablation

Table 8 examines the contributions of offline unknown supervision and online adaptation on AquaOCD343. All four variants retain offline classification training with $\mathcal{L}_{\text{cls}}$ and dynamic category discovery, ensuring that the comparison isolates the effects of the two components.

Enabling offline unknown supervision alone improves overall accuracy by 5.6% and 5.7% under Greedy–Hungarian and Strict–Hungarian, respectively. The corresponding improvements in novel category accuracy reach 2.7% and 9.6%, indicating that explicitly separating pseudo unknowns from known prototypes provides a stronger foundation for discovery. Online adaptation alone also improves overall accuracy under both protocols. However, its novel category accuracy under Strict–Hungarian decreases from 36.7% to 34.8%, showing that adapting representations trained only on known categories does not consistently benefit novel category discovery.

Combining both components achieves the highest overall and novel category accuracy among the four variants under both protocols. Compared with the baseline without either component, full DURA improves overall accuracy by 6.4% and 6.3%, and novel category accuracy by 4.3% and 10.7%, respectively. Furthermore, adding online adaptation to the model trained with unknown supervision increases novel category accuracy by another 1.6% and 1.1%. These results support the complementary roles of the two components: offline unknown supervision prepares the representation for unseen categories, while online adaptation further improves discovery using the observed query stream.

A.6 Hyperparameter Analysis

We conduct hyperparameter analysis on AquaOCD343 using DINOv2 as the backbone. Unless otherwise specified, all other settings are kept unchanged.

Effects of the Unknown Separation Margin. Table 9 studies the effect of the unknown separation margin m_u. Setting $m_u = 0.2$ achieves the best overall and novel category accuracy under both evaluation protocols. When the margin is either smaller or larger, the accuracy on novel categories decreases, while the accuracy on known categories remains relatively stable. This result shows that a moderate separation margin better separates pseudo unknowns from known prototypes without disrupting the learned known category structure.

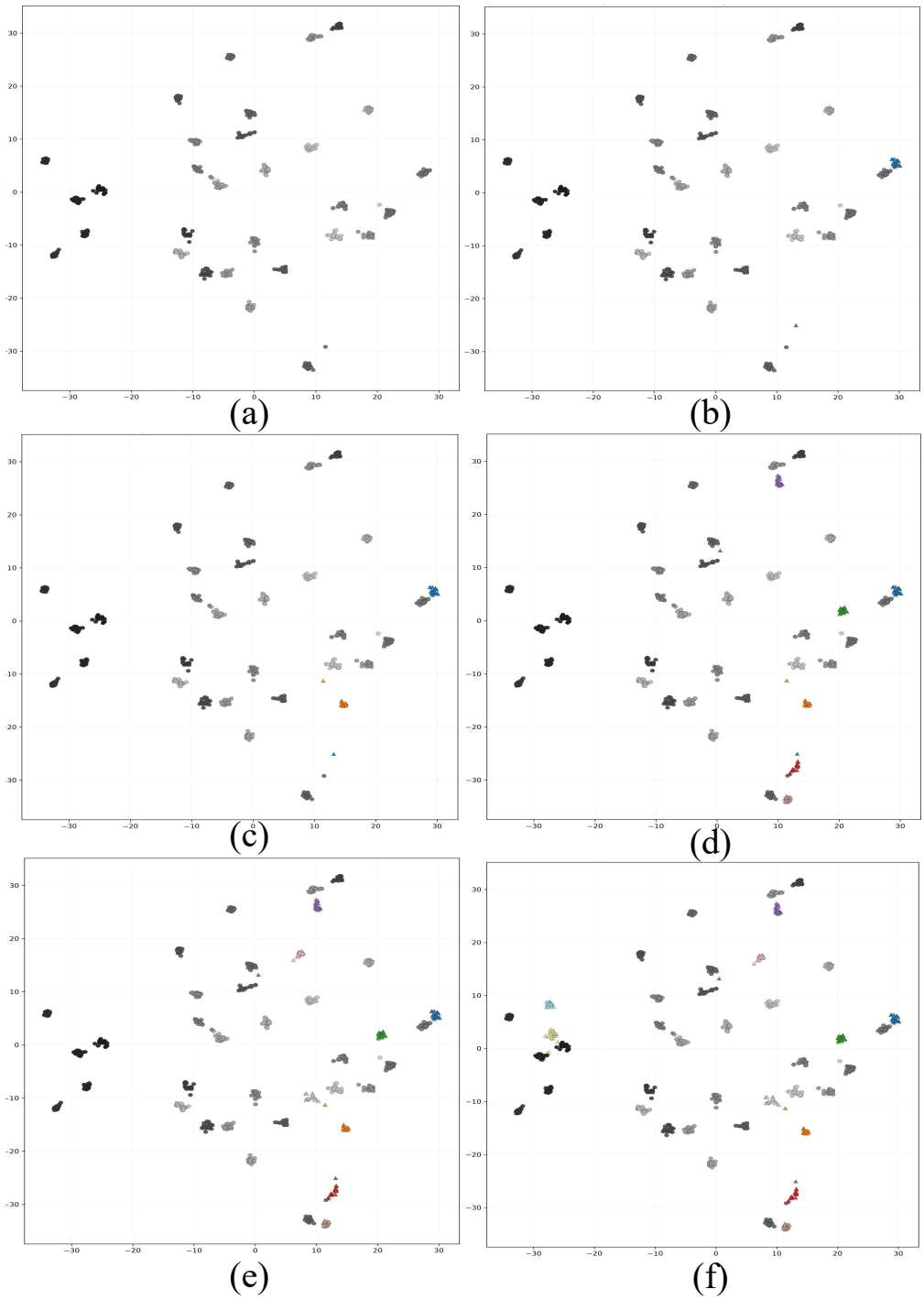


Figure 2: Progressive category discovery on AquaOCD343 using DURA-DINOv2. (a) Initial known categories. (b–f) Successive discovery stages showing the gradual emergence of novel categories as the query stream is processed. Gray points represent known category samples, while different colors indicate different discovered novel categories.

Table 9: Sensitivity analysis of the unknown separation margin m_u on AquaOCD343.

	Greedy			Strict		
m_u	All	Old	New	All	Old	New
0.0	77.5	95.7	50.1	68.4	85.2	43.2
0.1	77.2	95.7	49.4	69.4	**86.4**	43.9
0.2	**79.4**	95.3	**55.5**	**70.1**	85.2	**47.4**
0.3	77.8	95.7	51.0	69.0	85.5	44.1
0.4	77.7	**95.8**	50.4	68.4	85.5	42.6

Table 10: Sensitivity analysis of the unknown loss weight λ_u on AquaOCD343.

	Greedy			Strict		
λ_u	All	Old	New	All	Old	New
0.01	77.6	**95.8**	50.3	68.4	**91.9**	33.1
0.05	77.2	95.6	49.6	68.1	85.2	42.3
0.10	**79.4**	95.3	**55.5**	**70.1**	85.2	**47.4**
0.20	77.5	95.6	50.2	68.4	85.0	43.6
0.50	77.2	95.6	49.5	68.1	91.6	33.0

Effects of the Unknown Loss Weight. Table 10 evaluates the effect of the unknown separation weight λ_u. DURA achieves the highest overall and novel category accuracy at $\lambda_u = 0.1$ under both evaluation protocols. Reducing or increasing the weight leads to lower accuracy on novel categories, with a particularly clear degradation under Strict–Hungarian. These results show that an appropriate contribution of $\mathcal{L}_{\text{unk}}$ is important for balancing known category learning and unknown category separation.

Effects of the Discovery Threshold. Table 11 examines the effect of the discovery threshold cap $\tau_{\max}$. The best overall and novel category accuracy is obtained at $\tau_{\max} = 0.65$ under both evaluation protocols. Increasing the threshold beyond this value slightly improves the accuracy on known categories but reduces the accuracy on novel categories. This result shows that the discovery threshold controls the trade off between assigning queries to existing categories and creating new category prototypes.

Table 11: Sensitivity analysis of the discovery threshold cap $\tau_{\max}$ on AquaOCD343.

	Greedy			Strict		
$\tau_{\max}$	All	Old	New	All	Old	New
0.55	77.3	95.7	50.1	68.2	84.8	43.2
0.60	77.2	95.7	49.5	68.0	85.4	41.8
0.65	**79.4**	95.3	**55.5**	**70.1**	85.2	**47.4**
0.70	77.8	**95.8**	50.9	69.0	**86.0**	43.4
0.75	78.4	**95.8**	52.2	69.4	85.9	44.7

Effects of the Novel Prototype Update Rate. Table 12 evaluates the effect of the novel prototype update rate η_{novel}. Setting $\eta_{\text{novel}} = 0.1$ achieves the best performance across all metrics under both evaluation protocols. Both smaller and larger update rates lead to clear performance degradation, especially in the accuracy on novel categories. The results show that a moderate update rate is important for adapting novel prototypes to incoming observations while maintaining stable category representations.

A.7 Comparison under Different Pretrained Backbones

Table 13 compares DURA with existing OCD methods under different pretrained backbones. For DINOv2, we directly compare DURA with TALON using the same pretrained backbone. We addi-

Table 12: Sensitivity analysis of the novel prototype update rate η_{novel} on AquaOCD343.

	Greedy			Strict		
η_{novel}	All	Old	New	All	Old	New
0.02	74.9	94.4	45.7	65.2	85.1	35.4
0.05	73.3	93.0	43.7	63.5	83.0	34.2
0.10	**79.4**	**95.3**	**55.5**	**70.1**	**85.2**	**47.4**
0.15	73.2	93.3	43.0	63.4	83.2	33.6
0.20	70.4	89.9	41.1	60.2	79.6	31.1

Table 13: Comparison of OCD methods using different pretrained backbones on standard benchmarks. TALON denotes its DINOv2-ViT-Base configuration as "DINO" in its reported results; we identify it as DINOv2 according to its implementation details. Results for SMILE, PHE, LTC, and TALON are taken from their original publications, while the DURA results are obtained in this work. The best result within each backbone group and evaluation protocol is highlighted in **bold**.

			CIFAR-10			CIFAR-100			ImageNet-100			CUB-200-2011			Oxford Pets		
Protocol	Backbone	Method	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Greedy–Hungarian	DINO	SMILE	78.2	**99.3**	67.6	61.3	70.7	42.5	39.9	87.1	16.2	41.1	67.6	27.8	54.1	66.1	47.8
		PHE	83.0	98.0	75.5	64.8	78.8	36.9	53.1	83.5	38.1	46.9	76.0	32.4	63.3	91.3	48.6
		LTC	80.5	98.1	71.7	66.8	81.2	37.8	54.0	**90.5**	35.7	51.9	82.9	36.3	68.6	92.5	56.0
		DURA-DINO	**91.9**	92.0	**91.9**	**71.7**	**83.5**	**48.1**	**79.1**	83.0	**77.7**	**53.8**	**84.2**	**39.7**	**82.2**	**96.9**	**74.4**
	DINOv2	TALON	86.2	95.4	79.3	72.5	85.2	47.0	84.1	**94.3**	63.4	52.6	83.3	37.2	81.0	92.4	75.1
		DURA-DINOv2	**96.0**	**96.3**	**95.8**	**80.6**	**89.7**	**62.4**	**86.1**	88.3	**81.7**	**72.8**	**92.4**	**62.9**	**82.6**	**95.8**	**75.7**
	CLIP	SMILE	82.4	97.4	74.9	56.4	64.6	40.0	47.5	71.0	35.7	43.7	69.7	30.8	58.2	77.5	48.1
		PHE	79.3	97.0	70.4	66.1	80.3	37.5	52.9	87.8	35.5	44.2	70.3	31.1	64.1	86.2	52.4
		LTC	88.6	**98.1**	83.8	**70.7**	81.5	49.3	55.6	87.7	39.5	57.8	83.9	44.8	73.0	92.6	62.7
		TALON	84.7	96.0	76.3	69.9	**82.8**	44.2	**83.6**	**96.1**	**58.2**	58.9	87.3	44.7	74.9	**94.6**	64.6
		DURA-CLIP	**93.4**	92.8	**93.9**	67.2	75.0	**51.6**	67.2	75.5	50.5	**61.3**	**87.9**	**48.0**	**75.2**	92.0	**66.4**
Strict–Hungarian	DINO	SMILE	49.9	39.9	54.9	51.6	61.6	31.7	33.8	74.2	13.5	32.2	50.9	22.9	42.9	38.7	45.1
		PHE	53.1	19.3	70.0	56.0	70.1	27.8	39.2	49.3	34.1	36.4	**55.8**	27.0	48.3	53.8	45.4
		LTC	53.9	19.5	71.1	57.7	**75.1**	22.8	41.0	61.5	30.7	35.9	52.4	27.6	52.3	52.2	52.4
		DURA-DINO	**71.1**	**43.4**	**91.9**	**61.3**	75.0	**33.9**	**80.3**	**81.9**	**77.3**	**39.2**	55.1	**31.2**	**70.6**	**66.4**	**72.8**
	DINOv2	TALON	65.0	46.1	79.3	64.7	77.4	39.3	82.6	**92.0**	63.4	43.6	63.5	33.6	69.2	58.5	74.8
		DURA-DINOv2	**96.0**	**96.3**	**95.8**	**72.4**	**80.7**	**55.6**	**85.0**	87.1	**81.0**	**61.3**	**63.7**	**60.1**	**77.5**	**81.1**	**75.7**
	CLIP	SMILE	51.9	19.7	68.0	46.7	55.3	29.5	35.7	41.4	32.8	34.7	55.2	24.5	40.3	37.4	41.8
		PHE	52.4	18.3	69.5	56.8	71.9	26.5	39.2	60.7	28.4	35.1	54.5	25.4	52.0	52.3	51.9
		LTC	54.6	19.3	72.3	60.0	73.5	32.8	45.9	67.9	35.0	42.5	51.7	37.8	58.9	59.5	58.5
		TALON	56.9	**31.1**	76.3	**61.6**	**75.0**	34.9	**80.9**	**93.6**	**54.9**	45.5	**60.7**	37.8	64.0	65.4	**63.3**
		DURA-CLIP	**65.9**	28.6	**93.9**	55.8	62.3	**42.6**	60.5	73.2	47.8	**48.1**	58.4	**42.9**	**68.0**	**77.1**	**63.3**

tionally report the DINO results of SMILE, PHE, and LTC as references, together with the CLIP-based results of all methods.

With DINOv2, DURA achieves the best overall and novel category accuracy on all five datasets under both evaluation protocols. Under Greedy–Hungarian, DURA improves overall accuracy over TALON by 9.8, 8.1, 2.0, 20.2, and 1.6 percentage points on CIFAR-10, CIFAR-100, ImageNet-100, CUB-200-2011, and Oxford Pets, respectively. The corresponding improvements in novel-category accuracy are 16.5, 15.4, 18.3, 25.7, and 0.6 percentage points. The improvement is particularly clear on CUB-200-2011, where the novel-category accuracy increases from 37.2% to 62.9%.

The improvements remain substantial under Strict–Hungarian. Compared with TALON, DURA improves overall accuracy by 31.0, 7.7, 2.4, 17.7, and 8.3 percentage points across the five datasets. The corresponding gains in novel-category accuracy are 16.5, 16.3, 17.6, 26.5, and 0.9 percentage points. These results show that DURA provides consistent improvements over TALON when using the same DINOv2 backbone.

The CLIP-based comparison further evaluates the effect of the proposed method under a different pretrained representation. Under Greedy–Hungarian, DURA achieves the best novel-category accuracy on four of the five datasets and the best overall accuracy on CIFAR-10, CUB-200-2011, and Oxford Pets. Under Strict–Hungarian, DURA achieves the best novel-category accuracy on

CIFAR-10, CIFAR-100, and CUB-200-2011, and ties with TALON on Oxford Pets. These results demonstrate that the effectiveness of DURA is not restricted to a particular pretrained backbone.