Learning through Creation: A Hash-Free Framework for On-the-Fly Category Discovery

Supplementary Material

Appendix Overview

In this appendix, we provide additional details and experimental results for the main paper. Section A presents a comparative study of CLIP and DINO as backbones for the On-the-Fly Category Discovery (OCD) task. Section C discusses the effectiveness of the AutoTau mechanism and other results under varying hyperparameters. Section D offers a theoretical analysis of the Minimizing Kernel Energy & Maximizing Entropy (MKEE) approach and its impact on known-class accuracy. Finally, Section E addresses the limitations of our proposed method and suggests potential future improvements.

A. Backbone Control Study

To further validate that CLIP serves as a more effective backbone than DINO for On-the-Fly Category Discovery (OCD), we perform additional experiments under two algorithmic variants: the standard Greedy-Hungarian (our main evaluation protocol), and a stricter variant Strict-Hungarian, where pseudo-unknown samples are only used to model novel categories without influencing predictions for known classes. Results are reported in Table 5.

Across both settings, switching from DINO to CLIP consistently improves performance for SMILE, PHE, and our proposed LTC—especially on the four fine-grained datasets. This trend aligns with findings in Liu et al. [30], which compares CLIP and DINO using identical architectures (ViT-B/16) and training data (ImageNet-style corpus of 100M images). Their study shows that CLIP yields stronger results in fine-grained classification due to its vision-language pretraining, which helps the model capture high-level semantic distinctions between visually similar yet categorically different classes. This capability is particularly well-suited to OCD, where recognizing subtle differences among previously unseen categories is essential.

We therefore suggest building OCD-related tasks upon a CLIP-pretrained backbone, in order to fully leverage its semantic representation capabilities and achieve stronger performance. Future work should also explore model designs that are better aligned with CLIP-pretrained ViT-B/16, aiming to further exploit the rich high-level semantics it encodes and enhance novel category recognition in open-world scenarios.

B. Implementation Details

B.1. Dataset Details

As outlined in Table 6, our method is evaluated across multiple benchmarks. Following the protocol established in prior OCD works [11, 63], the categories within each dataset are divided into subsets of seen and unseen categories. Specifically, 50% of the samples from the seen categories are used to form the labeled training set $\mathcal{D}_S$, while the remainder forms the unlabeled set $\mathcal{D}_Q$ for on-the-fly testing.

Table 6. Statistics of datasets used in our experiments.

	CUB	Scars	Pets	Food	CIFAR10	CIFAR100	ImageNet100
$\lvert Y_Q\rvert$	200	196	38	101	10	100	100
$\lvert Y_S\rvert$	100	98	19	51	5	50	50
$\lvert\mathcal{D}_S\rvert$	1.5K	2.0K	0.9K	19.1K	12.5K	12.5K	32.5K
$\lvert\mathcal{D}_Q\rvert$	4.5K	6.1K	2.7K	56.6K	37.5K	37.5K	97.5K

B.2. Training Details

We use the AdamW optimizer with a learning rate of 1×10^{-2} for both the backbone and projection head, and apply a weight decay of 0.05. All models are trained for 100 epochs with a consistent batch size of 128 across all datasets to ensure fair comparison with prior methods. For MKEE, we set $\varepsilon = 0.05, \sigma_0 = 1, \lambda_\rho = 0.1$. The warm-up period for MKEE is set to 1 epoch, meaning the perturbation is activated starting from the second epoch. All experiments were run on NVIDIA RTX 4090 GPUs.

B.3. Compared Methods Details

Following the experimental setup in SMILE [11] and PHE [63], we evaluate our method against several competitive baselines. Below we provide brief descriptions of each:
Ranking Statistics (RankStat). [14] AutoNovel employs ranking-based heuristics, using the top-3 indices from feature embeddings to encode each category. This approach is naturally compatible with OCD due to its lightweight descriptor and nonparametric clustering behavior. To ensure fairness, we use the same DINO-ViT-B-16 backbone and discard auxiliary training stages that require access to unlabeled unknowns. The embedding is projected to 32 dimensions, yielding a total prediction space of $C_{32}^3 = 4,906$, which is on par with SMILE ($2^{12} = 4096$) when using 12-bit binary codes.
Winner-Take-All (WTA). [21] As an alternative to RankStat, WTA mitigates reliance on global embedding order by

Table 5. Comparison of variants.

	Method	Backbone	CIFAR10 (%)			CIFAR100 (%)			ImageNet-100 (%)			CUB-200-2011 (%)			Stanford Cars (%)			Oxford Pets (%)			Food101 (%)		
			All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Greedy–Hungarian	SMILE	DINO	78.2	99.3	67.6	61.3	70.7	42.5	39.9	87.1	16.2	41.1	67.6	27.8	33.4	58.4	21.3	54.1	66.1	47.8	34.4	64.0	19.4
	SMILE	CLIP	82.4	97.4	74.9	56.4	64.6	40.0	47.5	71.0	35.7	43.7	69.7	30.8	36.7	57.2	26.8	58.2	77.5	48.1	40.5	70.4	25.2
	PHE	DINO	83.0	98.0	75.5	64.8	78.8	36.9	53.1	83.5	38.1	46.9	76.0	32.4	46.3	78.3	30.8	63.3	91.3	48.6	50.0	89.3	30.0
	PHE	CLIP	79.3	97.0	70.4	66.1	80.3	37.5	52.9	87.8	35.5	44.2	70.3	31.1	46.4	78.1	31.1	64.1	86.2	52.4	47.8	88.4	27.0
	LTC	DINO	80.5	98.1	71.7	66.8	81.2	37.8	54.0	90.5	35.7	51.9	82.9	36.3	42.3	79.5	24.4	68.6	92.5	56.0	43.0	82.2	23.0
	LTC	CLIP	88.6	98.1	83.8	70.7	81.5	49.3	55.6	87.7	39.5	57.8	83.9	44.8	56.6	90.3	40.4	73.0	92.6	62.7	54.7	90.7	36.4
Strict–Hungarian	SMILE	DINO	49.9	39.9	54.9	51.6	61.6	31.7	33.8	74.2	13.5	32.2	50.9	22.9	26.2	46.6	16.3	42.9	38.7	45.1	24.2	54.3	8.8
	SMILE	CLIP	51.9	19.7	68.0	46.7	55.3	29.5	35.7	41.4	32.8	34.7	55.2	24.5	32.4	46.2	25.7	40.3	37.4	41.8	33.3	56.3	21.5
	PHE	DINO	53.1	19.3	70.0	56.0	70.1	27.8	39.2	49.3	34.1	36.4	55.8	27.0	31.3	61.9	16.8	48.3	53.8	45.4	29.1	64.7	11.1
	PHE	CLIP	52.4	18.3	69.5	56.8	71.9	26.5	39.2	60.7	28.4	35.1	54.5	25.4	36.2	54.2	27.4	52.0	52.3	51.9	33.5	58.6	20.6
	LTC	DINO	53.9	19.5	71.1	57.7	75.1	22.8	41.0	61.5	30.7	35.9	52.4	27.6	32.6	59.8	19.4	52.3	52.2	52.4	29.7	66.6	10.9
	LTC	CLIP	54.6	19.3	72.3	60.0	73.5	32.8	45.9	67.9	35.0	42.5	51.7	37.8	49.3	74.0	37.4	58.9	59.5	58.5	37.6	72.3	36.7

selecting index maxima within local feature groups. We divide the 48-dimension embedding into three parts and extract the index of the maximum value in each, forming a descriptor of length 3. The resulting prediction space is $16^3 = 4096$, aligning with other methods for consistent comparison.

Sequential Leader Clustering (SLC). [15] SLC is a classic online clustering method for streaming data. We train the encoder using only labeled support data and apply SLC on extracted features during test time. Hyperparameters are tuned on CUB and fixed across all datasets to ensure comparability.

Meta-Learning for Domain Generalization (MLDG). [26] Unlike conventional NCD approaches that leverage both support and query sets, OCD restricts training to known-class samples, posing a generalization challenge. To address this, we adapt MLDG for the OCD setting, treating different classes as meta-train and meta-test domains in each iteration. MLDG is then applied on top of our baseline model to promote domain-robust feature learning across class boundaries.

C. Additional Results under Varying Hyperparameters

C.1. Effectiveness of AutoTau

We further investigated the effectiveness of the AutoTau mechanism on the CUB dataset. Using both the Strict Hungarian and Greedy Hungarian metrics, as shown in Fig. 6, AutoTau significantly mitigates the impact of suboptimal initial thresholds (τ_{init}) and improves model performance. Notably, the enhancement is more pronounced when evaluated using the Strict Hungarian metric. As demonstrated in Fig. 7, the improvement reaches around 5% for certain suboptimal initial values of τ_{init}.

C.2. Additional Hyperparameter Sensitivity

To assess the sensitivity of the model to hyperparameters, we conducted experiments with varying values of α and γ_{mm}. Specifically, we performed parameter sensitivity analysis with $\alpha = 0.7$ and $\gamma_{mm} = 0.5$. The results, as shown in Fig. 8, demonstrate that selecting the appropriate parameters can influence the model's performance by approximately 2%.

D. Theoretical Analysis of MKEE

This appendix analyzes why MKEE (*Minimizing Kernel Energy & Maximizing Entropy*) improves novel-category discovery while slightly lowering known-class accuracy in On-the-Fly Category Discovery (OCD). The analysis explain the trade-offs observed in experiments, such as the significant gains in novel class accuracy accompanied by a minor decline in known class performance.

D.1. Rationale for One-Step Gradient Approximation Perturbation

The objective of MKEE is to generate pseudo-unknown samples x_{pus} by maximizing entropy and minimizing kernel density to encourage model uncertainty:

$$\mathcal{J}(x) = H(p(y|x)) - \lambda_1 \rho(x),$$

where $H(p(y|x)) = -\sum_c p_c \log p_c$ is the predictive entropy, and $\rho(x)$ estimates the kernel density of the sample in the known feature space. The Mixup sample $x_{\text{mix}} = \lambda x_i + (1-\lambda) x_j$ is perturbed via gradient ascent:

$$x_{\text{pus}} = x_{\text{mix}} + \varepsilon \cdot \frac{\nabla_x \mathcal{J}(x)}{\|\nabla_x \mathcal{J}(x)\|_2}.$$

These samples are used in training through a max-margin loss (as described in Section 3.3 of the main text) to jointly optimize known and unknown categories. To justify the one-step rule, we apply Taylor's theorem with Lagrange remainder: for any small Δx,

$$\mathcal{J}(x + \Delta x) = \mathcal{J}(x) + \nabla_x \mathcal{J}(x)^\top \Delta x + \tfrac{1}{2} \Delta x^\top H_x(\xi) \Delta x,$$

where $H_x(\xi)$ is the Hessian at some point ξ. If $\|H_x(\xi)\|_2 \leq L$ locally, then

$$\mathcal{J}(x + \Delta x) \geq \mathcal{J}(x) + \nabla_x \mathcal{J}(x)^\top \Delta x - \tfrac{L}{2} \|\Delta x\|_2^2.$$

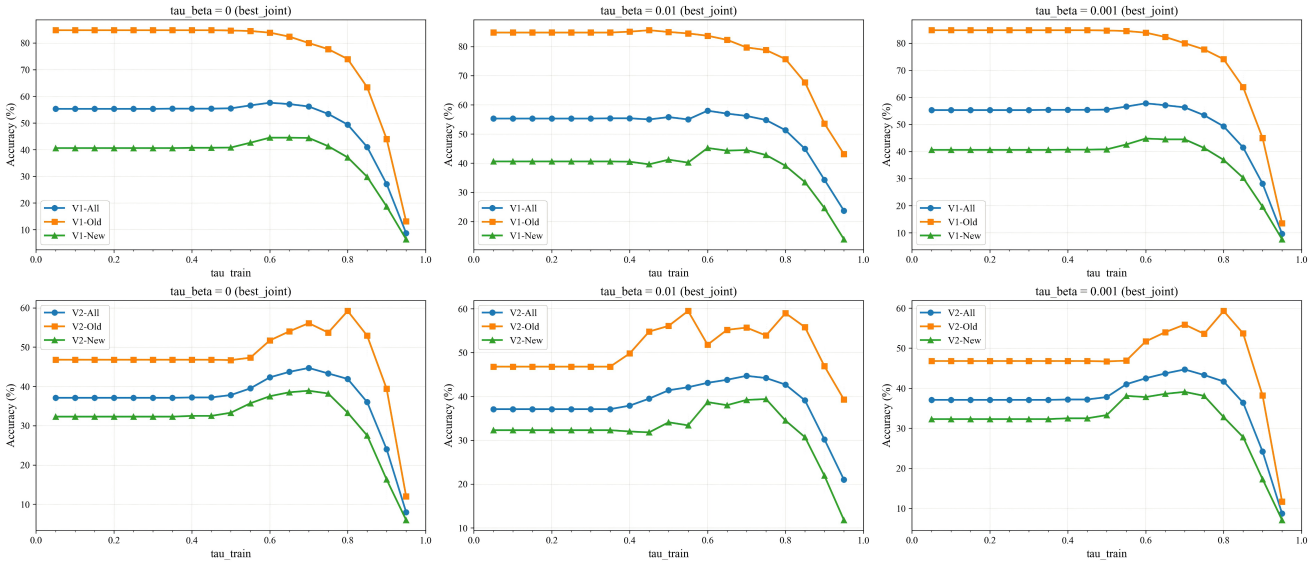

Figure 6. Performance comparison of AutoTau against fixed τ_{init} values using the Strict Hungarian and Greedy Hungarian metrics. AutoTau improves model performance by mitigating the impact of suboptimal initial thresholds.

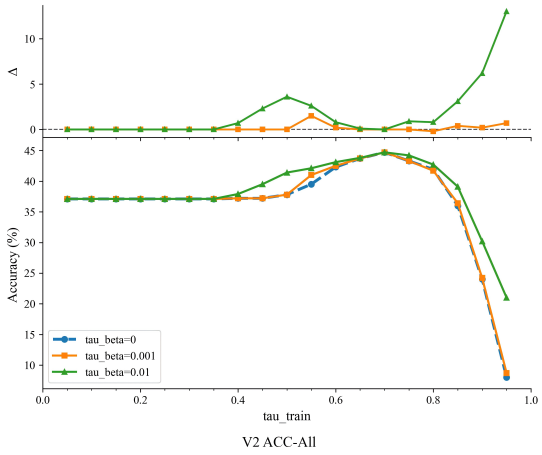

Figure 7. Impact of AutoTau on performance with different τ_{init} values. The improvement is more significant in the Strict Hungarian metric, reaching up to 5% for suboptimal τ_{init} values.

Maximizing the right-hand side over the ball $\{\|\Delta x\|_2 \leq \varepsilon\}$ yields the optimal step $\Delta x^\star = \varepsilon\, \nabla_x \mathcal{J}(x)/\|\nabla_x \mathcal{J}(x)\|_2$, with an improvement lower bound:

$$\mathcal{J}(x + \Delta x^\star) \geq \mathcal{J}(x) + \varepsilon \|\nabla_x \mathcal{J}(x)\|_2 - \tfrac{L}{2}\varepsilon^2.$$

This derivation validates the practicality of the one-step gradient rule, as it ensures a guaranteed increase in $\mathcal{J}(x)$ under reasonable smoothness assumptions.

D.2. Why MKEE Slightly Reduces Known-Class Accuracy?

Impact of Entropy Maximization on Known Classes. Consider a classification model that outputs a probability

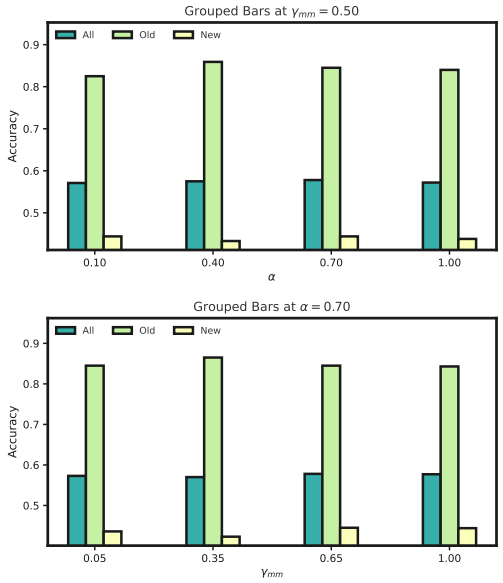

Figure 8. Parameter sensitivity analysis for α and γ_{mm}. The results show the performance variations with different values of these hyperparameters.

distribution $p(y|x)$. For known classes, the training objective minimizes the cross-entropy loss to encourage high-confidence predictions (i.e., entropy minimization):

$$\mathcal{L}_{\text{ce}} = -\mathbb{E}_{x \sim \mathcal{D}_{\text{known}}} \left[\sum_{c=1}^{K} y_c \log p_c(x) \right],$$

where y_c is the one-hot encoded label. This pushes the model toward Dirac-like distributions for known samples, resulting in low entropy. In contrast, MKEE generates pseudo-unknown samples that require high entropy (low

confidence). This conflict introduces an adversarial dynamic during optimization.

For a linear classifier with logits $\ell(x) = Wx+b$ and softmax probabilities $p_c(x) = \exp(\ell_c(x))/\sum_k \exp(\ell_k(x))$, the predictive entropy $H(p(y|x))$ is a function of $\ell(x)$. Minimizing $\mathcal{L}_{\text{ce}}$ drives $\ell(x_{\text{known}})$ to extreme values, reducing entropy, while entropy maximization for x_{pus} pulls $\ell(x_{\text{pus}})$ toward zero to equalize logits. When both objectives are combined, the total loss effectively includes an entropy regularization term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \beta \mathbb{E}_{x\sim\text{pseudo-unknown}} \left[H(p(y|x))\right].$$

This regularization softens decision boundaries, improving generalization to novel classes but reducing the sharpness of predictions for known classes, leading to a slight accuracy drop. Notably, MKEE achieves this indirectly through sample generation rather than direct loss modification, which avoids training instability associated with opposing loss terms.

Effect of Kernel Density Minimization. The kernel density term $\rho(x)$ in MKEE pushes samples away from dense regions of the known class distribution. While this exposes the model to out-of-distribution samples, it is beneficial for novel class discovery, and also biases decision boundaries away from known class centroids. Statistically, the density minimization acts as a penalty term:

$$\mathcal{L}_{\text{total}} \approx \mathcal{L}_{\text{ce}} + \beta_1 H(p(y|x_{\text{pus}})) - \beta_2 \rho(x_{\text{pus}}),$$

promoting robustness at the cost of marginal accuracy for known classes near decision boundaries. This explains the observed trade-off where known class accuracy decreases slightly while novel class accuracy improves substantially.

D.3. Correspondence with Experiments

Empirical results align with the theoretical analysis: MKEE consistently enhances novel class accuracy with a minor known class degradation (e.g., -1% on datasets like CIFAR-10). This trade-off is acceptable in OCD, where novel class discovery is prioritized. The adaptive threshold calibration in MKEE (Section 3.3 of the main text) mitigates this decline by dynamically balancing the novelty trigger. The parameters λ_1 and β allow tuning the trade-off, though future work could explore finer sample generation strategies to minimize known class impact while preserving novel gains.

E. Limitations Discussion

Balance Between New and Old Category Accuracy. In our experiments, we observed that the accuracy of the old categories may slightly degrade when using the MKEE approach. This is a trade-off we face in the pursuit of significantly improved new category discovery capabilities. While the proposed method excels at detecting new classes, this comes at the expense of some loss in performance on the old categories. In future work, a key challenge will be to develop methods that mitigate the forgetting of old category knowledge while maintaining the ability to effectively discover new categories. Techniques such as continual learning, knowledge distillation, or regularization strategies might be useful in addressing this limitation and ensuring more balanced performance across both known and novel categories.

Exploring the Integration of Textual Modality. Our current approach leverages the CLIP backbone, which primarily utilizes visual data. However, incorporating additional modalities, particularly textual information, into the OCD task presents an intriguing avenue for future exploration. The integration of text can potentially enhance the model's ability to generalize across categories and improve the discovery of novel classes, especially in scenarios where labeled data is sparse or ambiguous. Exploring how to effectively combine multimodal data, such as by aligning text and image representations, could further enrich the model's understanding of complex class relationships.