HVGCD:Rethinking Generalized Category Discovery through Hypothesis–Verification

Baoqiang Zhang*[1], Kunze Huang*[1], Luyao Tang[2], and Xiaotong Tu†[1]

[1] Xiamen University,Xiamen 361102,China
[2] University of Hong Kong, Hong Kong, 999077, China
{zhangbq,kzhuang,xttu}@stu.xmu.edu.cn
lytang1999@gmail.com

Abstract. Generalized Category Discovery (GCD) aims to simultaneously recognize known categories and discover novel ones from partially labeled data without annotations. Existing methods rely on single-path global embeddings, implicitly performing only hypothesis formation and omitting an explicit verification mechanism, which leads to over-reliance on pretrained semantics and confident misclassification of unknown categories. We reformulate GCD as a Hypothesis–Verification representation framework and propose HVGCD, a structured representation framework that decomposes inference into a global hypothesis, hypothesis-conditioned verification evidence, and statistical grounding. The verification branch reconstructs discriminative evidence via content-adaptive prototypes, enabling the model to verify and refine its hypotheses under open-world uncertainty, while a maximum-entropy grounding term stabilizes inference in open-world settings. As an efficient module, HVGCD integrates seamlessly into existing GCD pipelines and consistently improves performance on both coarse- and fine-grained benchmarks.

Keywords: Generalized Category Discovery · Vision Transformers · Representation Learning

1 Introduction

Enabling visual models to recognize and generalize to unseen categories is a central objective in open-world vision systems [20]. Generalized Category Discovery(GCD) [54] seeks to unify old-class recognition and novel-class discovery under a partially labeled setting, where unlabeled data simultaneously contain both known and unknown categories. However, most existing approaches [38, 47, 63, 65, 66] formulate GCD as a static feature clustering problem, relying on pretrained models to produce global representations and performing category partitioning within this fixed semantic space.

We argue that recognizing unknown categories is inherently not a one-shot feature aggregation process, but rather a dynamic Hypothesis–Verification cognitive process [23,51]. Stable decisions do not arise from a single global judgment,

* Equal contribution. † Corresponding author.

instead, they result from the interaction between hypothesis formation and subsequent verification. The hypothesis provides semantic direction, while verification refines and constrains it. Together, they establish a structured reasoning process that balances flexibility with reliability [14].

In contrast, existing GCD methods [1, 6, 12, 29, 44] primarily rely on the global representations of pretrained ViTs [17] as instance embeddings. Under the hypothesis–verification perspective, this design effectively models only the hypothesis stage. The model is encouraged to produce consistent and confident global semantic assignments, yet it does not incorporate an explicit mechanism to reassess or refine these initial judgments.

In closed-world scenarios, where data distributions are aligned, a unidirectional hypothesis often aligns with the ground-truth category. However, in open-world settings [3, 40], the presence of unknown categories fundamentally invalidates this assumption. The global representation tends to gravitate toward the nearest known semantic prototype, leading to highly confident yet incorrect category assignments [4, 43]. Without a subsequent verification process, the model cannot effectively revise its initial bias using local evidence, leading to a structural failure of the reasoning loop.

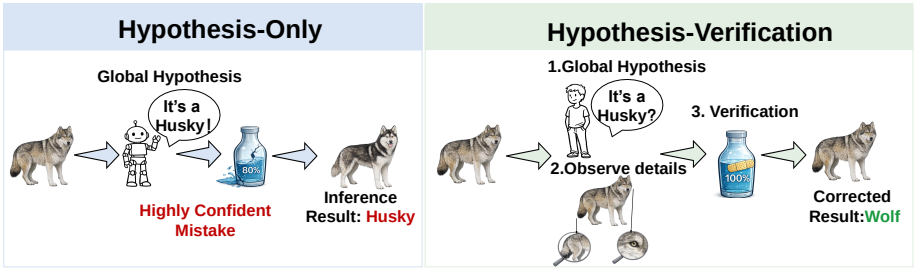

Fig. 1: Hypothesis-Only vs. Hypothesis-Verification in Generalized Category Discovery

As illustrated in Fig. 1, when the model encounters an unknown category such as a "wolf," its global representation often resembles that of a known class prototype (e.g., "husky"), leading the model to assign it to the nearest known category with high confidence. Due to the lack of attention to discriminative local details and the absence of an explicit verification step, this Hypothesis-Only inference paradigm, which operates in a unidirectional manner, is prone to confident misclassification.

In contrast, humans typically form an initial hypothesis based on the overall appearance of an object and then examine diagnostic visual cues to evaluate this interpretation [13, 62]. Rather than committing to an immediate decision, perception unfolds through a gradual iterative process in which an initial interpretation is evaluated and revised as further evidence is examined. This verification stage allows misinterpretations to be corrected and ultimately leads to more reliable semantic judgments. Therefore, the key challenge is not merely

to construct stronger hypotheses, but to restore the missing verification process, enabling representation learning to move beyond unidirectional inference toward hypothesis-guided refinement.

Importantly, incorporating patch-level tokens does not automatically constitute a reasoning mechanism [24, 28]. Reasoning is not defined by the quantity of information, but by whether supporting evidence is examined in a hypothesis-driven manner. Without an explicit objective, local features are merely aggregated in an unstructured way, which neither corrects bias nor stabilizes semantic decisions [7, 31, 52]. The essential requirement is thus to establish a genuinely hypothesis-oriented verification structure.

Based on the above analysis, we propose HVGCD, a Hypothesis–Verification framework for GCD, designed to restore a more complete Hypothesis–Verification reasoning process in GCD. We regard the global representation as the initial hypothesis and explicitly introduce a verification branch that performs structured querying and filtering over the patch sequence via content-adaptive semantic prototypes [7, 22]. This design enables the model to actively reconstruct hypothesis-relevant local evidence. Local information therefore serves not as passive supplementation, but as structured evidence that supports or challenges the initial hypothesis.

Furthermore, to account for the uncertainty inherent in open-world distributions [2], we introduce a statistical grounding component that provides a stable distributional anchor for the reasoning process. This component functions neither as hypothesis nor verification, instead, it imposes distribution-level constraints on both, preventing the verification mechanism from being dominated by local saliency biases.

Ultimately, HVGCD integrates hypothesis, verification, and statistical grounding into a unified framework, transforming representation from unidirectional semantic projection into a structured reasoning process. Under this formulation, unknown categories are no longer passively absorbed into the existing semantic space, but instead acquire more discriminative representations through structured hypothesis-guided verification.Our main contributions are as follows:

- We introduce a Hypothesis–Verification perspective for GCD, providing a unified interpretation of representation learning as a structured reasoning process rather than static feature clustering.
- We propose HVGCD, a framework that restores the hypothesis–verification reasoning loop by treating global representations as hypotheses and extracting hypothesis-guided local evidence.
- We conduct systematic evaluations on multiple GCD benchmarks, demonstrating the effectiveness of HVGCD.

2 Related Works

2.1 Hypothesis–Verification Mechanisms

The hypothesis–verification paradigm [13, 23, 51] originates from cognitive science, where an agent forms an initial hypothesis and iteratively refines it through

evidence evaluation. Recently, this principle has inspired verification mechanisms in machine learning [21, 39, 61] to enhance reliability and robustness.

In natural language generation, CoVe [16] proposes a Chain-of-Verification framework that reduces hallucinations through explicit verification steps, while TWEAK [45] improves faithfulness by validating candidate outputs during decoding. In vision, hypothesis–verification mechanisms have also been explored in deep learning architectures. For example, Slot Attention [37] iteratively refines object hypotheses through attention-based evidence aggregation, while DETR [30] validates object queries against image features via cross-attention, [36] measuring an energy-based confidence score, which helps identify unreliable predictions under distribution shifts. Unlike prior approaches that verify predictions at the output or decision level, we introduce an explicit hypothesis–verification structure at the representation level.

2.2 Generalized Category Discovery

In Generalized Category Discovery (GCD) [54], the objective is to leverage a small labeled dataset from known categories to partition unlabeled samples into both known and novel categories, better reflecting real-world image recognition scenarios. Subsequent research [5, 10, 27, 33, 59] has proposed various strategies to address these challenges.For example, SimGCD [63] replaces clustering with a classifier trained using pseudo-labels. SPTNet [60] employs spatial prompt tuning to encourage attention toward category-relevant regions. [9] improves fine-grained discrimination by modeling correspondences between global semantics and part-level features. ConGCD [50] focuses on enhancing the task through semantic visual primitives, whereas BIA [49] emphasizes preserving the overall feature manifold to prevent representation collapse. while [59] separates semantic and domain representations to mitigate domain shifts. However, most methods rely on a single global representation without verification, resulting in overconfident errors in open-world scenarios.

3 Methodology

3.1 Problem Definition of GCD

For each dataset, we consider a labeled subset denoted as $D^l = \{(x_i^l, y_i^l)\} \in \mathcal{X} \times \mathcal{Y}^l$. In addition, there is an unlabeled subset $D^u = \{(x_i^u, y_i^u)\} \in \mathcal{X} \times \mathcal{Y}^u$. The labeled set D^l contains only known classes, i.e., $\mathcal{Y}^l = C_{\text{known}}$. In contrast, the unlabeled set D^u contains both known and novel classes, i.e., $\mathcal{Y}^u = C_{\text{known}} \cup C_{\text{novel}}$. The task of the model is to cluster the samples in D^u into their corresponding known and novel categories. It is typically assumed that the total number of classes is known a priori [25, 44], though off-the-shelf methods [12, 26] can also be employed to estimate it.

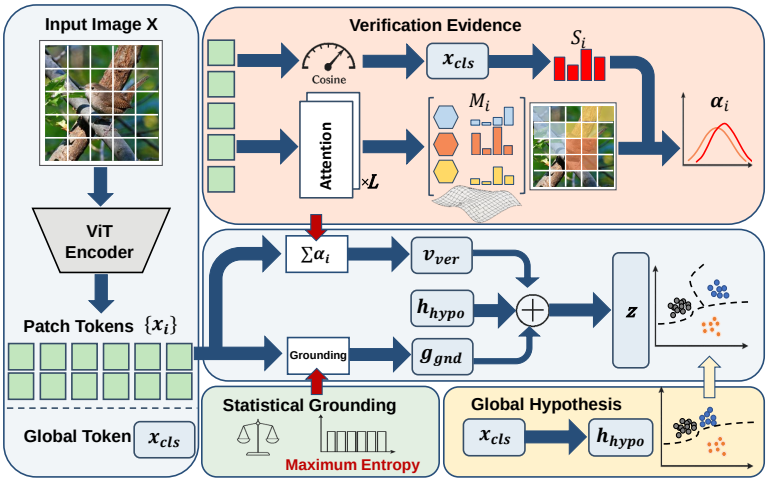

Fig. 2: Overview of the proposed HVGCD:(1) A global intuitive hypothesis is formed from holistic semantics.(2) Local discriminative cues are selectively reconstructed via prototype-guided verification.(3) Maximum-entropy statistical grounding stabilizes inference under open-world uncertainty.

3.2 Hypothesis, Verification, and Grounding

Existing GCD methods [56, 63, 64] typically formulate instance representation learning as a single embedding mapping, where a pretrained model projects high-dimensional visual input X into a global representation x_{cls}. Under this formulation, the inference structure of an instance can be expressed as:

$$z \approx h(x_{cls}). \tag{1}$$

where $h(\cdot)$ denotes a direct discriminative function operating on global features.

We argue that such a single-path inference paradigm effectively completes only the Hypothesis Formation stage. In the absence of an explicit Verification Mechanism, the model tends to over-rely on dominant semantics acquired during pretraining, while overlooking fine-grained residual cues that are crucial for distinguishing unknown categories. This results in Inference Truncation: when encountering novel classes, the model is unable to revise its initial hypothesis using local evidence, leading to highly confident yet erroneous predictions.

As shown in Fig. 2, to establish a more complete reasoning structure, the GCD representation process is reconceptualized as a dynamic Hypothesis–Verification system [51]. The objective extends beyond feature extraction; instead, a structured representation z is constructed to integrate global hypotheses and verification evidence. Specifically, we formulate z as consisting of three functional components:

- Intuitive Hypothesis (h_{hypo}): Derived from the global semantic vector x_{cls}, it captures the model's rapid, holistic judgment and provides the initial directional prior for inference.
- Verification Evidence (v_{ver}): Conditioned on the hypothesis, it actively searches for and reconstructs discriminative details from local visual tokens x_i, serving to confirm or refute the initial hypothesis and recover overlooked cues.
- Statistical Grounding (g_{gnd}): Under open-world distributional uncertainty, it provides an unbiased distribution-level anchor that constrains the reasoning process and prevents instability on atypical or ambiguous samples.

Formally, given the ViT output sequence $H = [x_{cls}, x_1, \ldots, x_N]$, we construct the following structured inference model:

$$z = \mathcal{T}(h_{hypo}, v_{ver}, g_{gnd}) = h_{hypo} + \lambda_v \cdot \psi(X \mid h_{hypo}) + \lambda_g \cdot \mathbb{E}[X], \tag{2}$$

where $\psi(\cdot \mid \cdot)$ denotes the hypothesis-conditioned verification operator, and $\mathbb{E}[\cdot]$ represents the statistical grounding term.

By introducing an explicit verification pathway together with a distribution-level stabilizing constraint, the instance representation evolves from a unidirectional semantic projection into a structured reasoning process.

3.3 Verification Evidence

The core of the verification mechanism is the construction of a hypothesis-conditioned operator $\psi(X \mid h_{hypo})$. Its objective is to extract, from noisy local signals, those discriminative residuals that can confirm or refute the initial hypothesis h_{hypo}. Direct aggregation of patch features, however, often introduces substantial hypothesis-irrelevant background noise. We therefore implement this conditional verification process through content-adaptive prototype induction coupled with a dual-gating mechanism.

To decouple structured semantics from high-dimensional signals, we induce a set of content-adaptive semantic prototypes [7,22] $C = \{c_k\}_{k=1}^{K}$ via a lightweight depthwise convolutional clustering module. These prototypes are dynamically generated cluster centroids conditioned on the current input features. Specifically, given the local patch sequence $X = \{x_i\}_{i=1}^{N}$, we first estimate soft cluster assignments:

$$z_{k,i} = \mathrm{Softmax}_i(\mathcal{G}(x_i)), \tag{3}$$

where $\mathcal{G}(\cdot)$ denotes a convolutional projection that produces cluster logits. The prototypes are then computed as weighted cluster centroids:

$$c_k = \sum_{i=1}^{N} z_{k,i} x_i. \tag{4}$$

These dynamically induced prototypes define a structured semantic manifold over the patch feature space. Rather than passively aggregating features, we

formulate verification as projection onto and reconstruction within this induced prototype manifold. Each patch token queries the induced prototypes through cross attention [52]:

$$\hat{x}_i = \text{CrossAttention}(Q(x_i), K(C), V(C)). \tag{5}$$

This operation performs a semantic manifold projection that preserves only prototype explainable components in the reconstructed signal $\hat{x}_i$, including potentially discriminative evidence such as characteristic textures or structural primitives. Since these signals recover fine grained information that may be suppressed during global compression, we refer to them as Discriminative Cues.

To support sufficiently deep reasoning while maintaining computational efficiency, we stack L layers of cluster guided attention blocks. Through this iterative refinement process, local discriminative cues become progressively aligned with the evolving global hypothesis.

Building upon this structured alignment, we instantiate the verification operator $\psi(X \mid h_{hypo})$ by explicitly measuring how strongly each reconstructed cue supports the current hypothesis. Concretely, we estimate a Posterior Verification Probability that quantifies the degree of evidence provided by the discriminative cues under the given hypothesis.

Let α_i denote the confidence that patch x_i constitutes valid verification evidence. We define validity as the joint occurrence of a Structural Informativeness Event (E_{inf}) and a Hypothesis Coherence Event (E_{coh}). Under the hypothesis–verification paradigm [13, 62], an effective cue must satisfy:

- It carries structurally informative semantic content that can participate in hypothesis confirmation or revision.
- It bears semantic relevance to the current global hypothesis.

The verification probability is thus expressed as $P(E_{inf}, E_{coh} \mid x_i, h_{hypo})$. Assuming conditional independence between E_{inf} and E_{coh} given the features, we decompose:

$$\alpha_i \approx P(\text{Valid} \mid x_i, h_{\text{hypo}}) \propto \underbrace{P(E_{\text{inf}} \mid \hat{x}_i)}_{\text{Intrinsic Saliency}} \cdot \underbrace{P(E_{\text{coh}} \mid x_i, h_{\text{hypo}})}_{\text{Hypothesis Conditioned}}. \tag{6}$$

This decomposition reveals the dual constraints underlying effective verification. First, the saliency prior E_{inf} measures the structural informativeness of the reconstructed cue. Effective verification requires evidence that contains meaningful semantic structure rather than incidental visual variations [58]. Consequently, E_{inf} functions as an informativeness prior that filters out signals lacking discriminative semantic content, ensuring that only structurally meaningful cues contribute to the verification process.

Since the reconstructed signal $\hat{x}_i$ is obtained via projection onto the dynamically induced prototype manifold, which filters out semantically unstructured variations and stabilizes the representation within the induced prototype manifold, it provides a more robust basis for saliency estimation. Accordingly, we

employ a lightweight prediction head $\phi(\cdot)$ to evaluate the saliency intensity of the reconstructed signal:

$$S_i = \sigma(\phi(\hat{x}_i)) \,. \tag{7}$$

Second, the coherence likelihood E_{coh} corresponds to the conditional constraint of the verification operator [15], namely suppressing semantic distractors that are irrelevant to the current hypothesis. We compute the consistency between the original local signal and the global hypothesis as:

$$M_i = \langle \text{Norm}(h_{\text{hypo}}), \text{Norm}(x_i) \rangle \,. \tag{8}$$

thereby preserving fine-grained semantic associations in the original feature space.Finally, to retain fine-grained texture details for precise verification, we define the verification evidence v_{ver} as a high-confidence weighted aggregation of the original local signals:

$$v_{\text{ver}} = \sum_{i=1}^{N} \alpha_i x_i, \quad \text{where } \alpha_i = \text{Softmax}\left(\frac{S_i \cdot M_i}{\tau}\right). \tag{9}$$

This mechanism enables refinement of the global hypothesis: when h_{hypo} is incomplete or partially biased, the aggregated verification evidence v_{ver} provides complementary or corrective semantic signals, enabling structured revision of the initial hypothesis, thereby completing a logical closed loop in feature space from unidirectional inference to reasoning.

3.4 Statistical Grounding

Although the verification mechanism restores discriminative cues, in open-world settings where samples originate from entirely unknown distributions, overly selective verification may introduce additional bias. To ensure the robustness of the overall reasoning system, we introduce a distribution-level statistical reference to stabilize the inference dynamics and prevent the verification process from drifting due to local saliency bias.

We derive the optimal form of this reference from the Principle of Maximum Entropy [32]. In the absence of prior knowledge, the least biased weighting scheme corresponds to the distribution with maximum entropy. We formulate the search for the optimal weight distribution w as the following constrained optimization problem:

$$w^* = \arg\max_{w} H(w) = -\sum_{i=1}^{N} w_i \log w_i, \quad \text{s.t.} \quad \sum_{i=1}^{N} w_i = 1. \tag{10}$$

This convex optimization problem admits a unique solution, namely the uniform distribution:

$$w_i^* = \frac{1}{N}. \tag{11}$$

Based on this solution, we define $\mathbb{E}[X]$ as the maximum-entropy expectation of local features, which we term the Statistical Grounding:

$$g_{\text{gnd}} = \sum_{i=1}^{N} w_i^* x_i = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{12}$$

The statistical grounding captures the first-order moment of the visual signal. It does not serve a discriminative role, rather, it functions as a distributional constraint that provides a stabilizing reference for both hypothesis formation and verification.

When the verification branch becomes unreliable under high uncertainty, the model representation reverts to this stable statistical grounding, thereby preventing representational collapse under unknown distributions and ensuring robust reasoning in open-world scenarios.

3.5 Integrated Inference and Alignment

The final instance representation z is defined as the unified inference result jointly derived from the global hypothesis, the verification evidence, and the statistical grounding:

$$z = h_{\text{hypo}} + \lambda_v v_{\text{ver}} + \lambda_g g_{\text{gnd}}, \tag{13}$$

where λ_v and λ_g are balancing coefficients. This multi-source integration strategy enables the model to maintain reliable classification on known categories, activate fine-grained verification for unknown instances, and revert to a stable grounding under extreme uncertainty.

To ensure that the verification process remains semantically consistent with the global hypothesis throughout training, we introduce a Semantic Alignment Regularization term L_{align}:

$$L_{\text{align}} = 1 - \langle \text{Norm}(v_{\text{ver}}), \text{Norm}(h_{\text{hypo}}) \rangle . \tag{14}$$

This constraint enforces directional consistency between the verification evidence and the global hypothesis on the semantic manifold, ensuring that v_{ver} serves as a refinement of h_{hypo} rather than a contradictory deviation.

The final optimization objective combines the primary loss of the GCD task with the alignment regularization term:

$$L_{\text{total}} = L_{\text{gcd}} + \gamma L_{\text{align}}. \tag{15}$$

Through this objective, HVGCD achieves a dynamic equilibrium among hypothesis formation, evidence verification, and statistical grounding within a unified framework, thereby enhancing generalization performance in complex open-world scenarios.Notably, HVGCD is a plug-and-play module. It operates solely in the embedding space without imposing architectural constraints on the backbone network, and can be seamlessly integrated into existing GCD frameworks.

Table 1: Main Results on Fine-Grained Image Classification Datasets

Backbone	Method	CUB-200			FGVC-Aircraft			Stanford-Cars			Herbarium19			Average		
		All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
DINOv1	GCD [54]	51.3	56.6	48.7	45.0	41.1	46.9	39.0	57.6	29.9	35.4	51.0	27.0	42.7	51.6	38.1
	GPC [65]	52.0	55.5	47.5	43.3	40.7	44.8	38.2	58.9	27.4	-	-	-	44.5	51.7	39.9
	XCon [18]	52.1	54.3	51.0	47.7	44.4	49.4	40.5	58.8	31.7	38.1	58.3	27.3	44.6	54.0	39.9
	PromptCAL [64]	62.9	64.4	62.1	52.2	52.2	52.3	50.2	70.1	40.6	37.0	52.0	28.9	50.7	60.0	46.0
	AMEND [1]	64.9	75.6	59.6	52.8	61.8	48.3	56.4	73.3	48.2	44.2	60.5	35.4	54.6	67.8	47.9
	μGCD [56]	65.7	68.0	64.6	53.8	55.4	53.0	56.5	68.1	50.9	-	-	-	58.7	63.8	56.2
	InfoSieve [46]	69.4	**77.9**	65.2	56.3	63.7	52.5	55.7	74.8	46.4	41.0	55.4	33.2	55.6	68.0	49.3
	SimGCD [63]	60.8	64.8	58.8	52.1	57.7	49.3	55.4	70.9	48.0	43.1	56.3	36.0	52.9	62.4	48.0
	+ Ours	66.5	70.5	64.6	54.1	59.8	51.2	**60.6**	75.7	**53.3**	44.6	57.8	37.5	56.4	65.9	51.6
	CMS [12]	67.2	77.0	62.3	52.5	61.0	48.3	51.7	71.2	42.2	35.9	55.0	25.5	51.8	66.1	44.6
	+ Ours	67.9	75.1	64.8	55.1	61.5	52.0	53.9	72.5	44.9	36.4	55.3	26.2	53.3	66.1	47.0
	LegoGCD [6]	61.2	71.4	56.1	50.9	59.1	46.8	57.1	**78.8**	46.6	44.7	58.3	37.3	53.5	66.9	46.7
	+ Ours	66.9	73.4	63.6	53.8	61.9	49.8	58.8	76.5	50.2	**45.8**	**60.1**	**38.2**	56.3	**68.0**	50.4
	SelEx [47]	77.0	74.4	78.3	55.2	63.6	51.0	54.5	75.0	44.7	36.0	45.7	30.8	55.7	64.7	51.2
	+ Ours	**78.5**	76.9	**79.3**	**65.2**	**68.0**	**63.8**	58.1	75.3	49.8	37.9	48.2	32.3	**59.9**	67.1	**56.3**
	Avg. Δ	+3.39	+2.02	+4.20	+4.36	2.45	+5.34	+3.14	+1.05	+4.15	+1.26	+1.52	+1.13	+3.04	+1.76	+3.71
DINOv2	SimGCD [63]	71.1	79.5	66.9	69.1	66.1	70.6	70.9	84.6	64.3	56.2	63.5	52.2	66.8	73.4	63.5
	+ Ours	75.3	82.1	71.9	70.9	72.0	70.3	75.0	85.5	69.9	**57.0**	65.3	**52.5**	69.5	76.2	66.2
	CMS [12]	80.1	80.9	79.7	68.9	76.2	64.3	82.8	92.6	78.1	46.4	65.4	36.3	69.6	78.8	64.6
	+ Ours	82.6	84.6	81.7	68.4	75.5	64.9	**84.5**	93.0	**80.4**	47.8	**65.7**	38.2	70.8	79.7	66.3
	SelEx [47]	89.8	87.0	91.2	80.5	78.1	81.6	83.1	93.7	77.0	45.2	55.3	39.7	74.6	78.5	72.4
	+ Ours	**90.8**	**87.3**	**92.5**	**81.8**	**82.1**	**81.7**	83.5	**93.9**	77.6	46.2	56.1	40.9	**75.6**	**79.8**	**73.1**
	Avg. Δ	+2.55	+2.20	+2.72	+0.87	+3.02	+0.12	+2.05	+0.47	+2.84	+1.06	+0.94	+1.12	+1.63	+1.66	+1.70

4 Experiments

4.1 Setup

Datasets. We evaluate the proposed method on seven widely used image recognition benchmarks, including four fine-grained datasets—CUB-200-2011 [57], Stanford Cars [34], FGVC-Aircraft [41], and Herbarium19 [48]—as well as three coarse-grained datasets, CIFAR10, CIFAR100 [35], and ImageNet100 [19].For CUB-200-2011, Stanford Cars, and FGVC-Aircraft, we adopt the class partitions defined in the Semantic Shift Benchmark (SSB) [53] to divide the categories into known and unknown subsets. For the remaining datasets, we follow the data splits introduced in [55]. Under the CIFAR100 setting, 80% of the categories are treated as known classes, while for all other benchmarks, 50% of the categories are designated as known. The labeled dataset D_l is constructed by randomly sampling 50% of the images from the known categories for each benchmark.

Evaluation metrics. We evaluate our method using clustering accuracy (ACC) with Hungarian matching, following prior work [63]. For the unlabeled dataset D^u, ACC is computed by comparing the ground-truth labels y_i^u with the predicted labels $\hat{y}_i^u$:

$$\text{ACC} = \frac{1}{|D^u|} \sum_{i=1}^{|D^u|} \mathbf{1}\left(y_i^u = h(\hat{y}_i^u)\right), \tag{16}$$

where $h(\cdot)$ denotes the optimal one-to-one mapping obtained by the Hungarian algorithm, which aligns predicted cluster assignments with the true class

Table 2: Main Results on Generic Image Classification Datasets(CMS not reported on CIFAR-10 due to instability under small category scale).

Backbone	Method	CIFAR-10			CIFAR-100			ImageNet-100			Average		
		All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
DINOv1	GCD [54]	91.5	**97.9**	88.2	73.0	76.2	66.5	74.1	89.8	66.3	79.5	88.0	73.7
	GPC [65]	90.6	97.6	87.0	75.4	84.6	60.1	75.3	93.4	66.7	80.4	91.9	71.3
	XCon [18]	96.0	97.3	95.4	74.2	81.2	60.3	77.6	93.5	69.7	82.6	90.7	75.1
	PromptCAL [64]	**97.9**	96.6	98.5	81.2	84.2	75.3	83.1	92.7	78.3	87.4	91.2	84.0
	PIM [11]	94.7	97.4	93.3	78.3	84.2	66.5	83.1	95.3	77.0	85.4	92.3	78.9
	DCCL [44]	96.3	96.5	96.9	75.3	76.8	70.2	80.5	90.5	76.2	84.0	87.9	81.1
	InfoSieve [46]	94.8	97.7	93.4	78.3	82.2	70.5	80.5	93.8	73.8	84.5	91.2	79.2
	SimGCD [63]	96.9	93.9	98.4	80.5	80.8	**79.9**	83.4	92.4	78.9	87.0	89.1	85.7
	+ Ours	96.8	92.5	**98.9**	81.9	82.9	**79.9**	85.2	93.0	81.2	88.0	89.5	**86.7**
	CMS [12]	-	-	-	78.0	84.6	64.9	81.8	95.5	75.0	89.9	90.1	70.0
	+ Ours	-	-	-	80.2	86.3	67.9	84.5	95.6	78.9	82.4	91.0	73.4
	LegoGCD [6]	97.3	95.8	98.0	79.4	81.9	74.5	84.0	94.9	78.5	86.9	90.9	83.7
	+ Ours	97.5	95.4	98.6	81.2	83.6	76.3	83.4	95.0	77.6	87.4	91.3	84.2
	SelEx [47]	96.3	**97.9**	95.5	80.9	85.1	72.6	84.0	94.5	78.7	87.1	92.5	82.3
	+ Ours	96.8	97.8	96.2	**82.7**	**86.9**	75.3	**86.4**	**95.7**	**81.8**	**88.6**	**93.5**	84.4
	Avg. Δ	+0.19	-0.63	+0.08	+1.78	+1.83	+1.90	+1.57	+0.49	+2.10	+1.08	+0.63	+1.36
DINOv2	SimGCD [63]	98.5	96.0	**99.7**	86.8	88.6	83.3	89.7	95.1	87.0	91.7	93.3	90.0
	+ Ours	**98.6**	96.2	**99.7**	88.3	88.6	**87.8**	92.4	95.7	90.7	**93.1**	93.5	**92.8**
	CMS [12]	-	-	-	90.2	92.5	85.5	**93.9**	**97.2**	**92.2**	92.0	94.9	88.9
	+ Ours	-	-	-	**90.6**	**92.9**	85.8	92.2	95.8	90.3	91.4	94.4	88.1
	SelEx [47]	98.3	**98.7**	98.1	87.9	91.1	81.6	90.3	96.5	87.1	92.1	**95.4**	88.9
	+ Ours	98.2	98.4	98.1	87.8	90.4	82.6	91.7	96.7	89.1	92.5	95.2	90.0
	Avg. Δ	+0.03	+0.05	-0.04	+0.57	-0.12	+1.95	+0.78	-0.23	+1.28	+0.38	-0.17	+1.00

labels.In addition, we report ACC separately for three category groups: All, Old, and New classes. This provides a more detailed evaluation of the model performance across different types of categories.

Implementation Details. We integrate HVGCD into several representative GCD frameworks, including SimGCD [63], SelEx [47], CMS [12], and LegoGCD [6], covering parametric, non-parametric, and knowledge-retention paradigms. For each baseline, we faithfully reproduce the original implementation and perform controlled comparisons under identical experimental settings to ensure fairness.

Models are initialized with ImageNet-1K-pretrained DINOv1 [8] weights, with $\gamma = 0.05$, $\lambda_v = 0.3$, $\lambda_g = 0.5$, $L = 3$, and $K = 100$. All remaining hyperparameters follow the respective original works. We further report results using ImageNet-22K-pretrained DINOv2 [42] features. All experiments are conducted on RTX-4090 GPUs with fixed random seeds to ensure reproducibility and prevent label leakage.

4.2 Results

In the fine-grained setting (Tab. 1), when combined with DINOv1, HVGCD improves the overall accuracy of four baseline methods by an average of **3.04%**. Notably, on the FGVC-Aircraft dataset, the performance on novel categories increases by **5.34%**. With DINOv2, HVGCD further improves the overall accuracy of three baselines by **1.63%**, with gains of **1.76%** on known categories and

1.70% on novel categories. In the coarse-grained setting (Tab. 2), HVGCD improves the overall accuracy by an average of **1.08%**, including **0.63%** improvement on known categories and **1.36%** on novel categories. When combined with stronger pretrained models, the overall accuracy further increases by **0.38%**, with an additional **1.00%** gain on novel categories, while the performance on known categories remains stable or slightly improved. These results demonstrate that our proposed method consistently enhances the performance of diverse GCD methods, serving as an efficient and generalizable plug-and-play framework.

4.3 Analysis

Component Effectiveness and Sensitivity. To verify the effectiveness of v_{ver} and g_{gnd}, we conduct ablation studies on CUB-200 and CIFAR-100. As shown in Fig. 3, introducing either Verification Evidence or Statistical Grounding significantly improves accuracy, with optimal performance achieved at $\lambda_v = 0.3$ and $\lambda_g = 0.5$, respectively. These values are therefore adopted as the default settings in subsequent experiments.

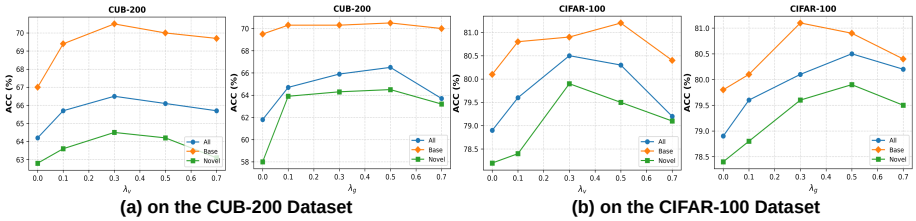

Fig. 3: Ablation study on HVGCD components

The results confirm the central role of v_{ver} in mining local discriminative cues, as well as the necessity of g_{gnd} in preventing the model from overfitting to local saliency biases. Their combination yields a synergistic improvement in both discriminative capability and reasoning robustness, ultimately establishing a more reliable inference framework.

Table 3: Sensitivity analysis of the semantic alignment regularization weight γ on different datasets.

	CUB-200			Stanford-Cars			CIFAR-100			Average		
L_{align}	All	Old	New	All	Old	New	All	Old	New	All	Old	New
0	65.8	**70.9**	63.3	59.5	72.2	52.6	81.2	82.6	78.4	68.8	75.2	64.8
0.05	**66.5**	70.5	**64.6**	**60.6**	75.7	**53.3**	**81.9**	**82.9**	**79.9**	**69.7**	**76.4**	**65.9**
0.1	65.7	70.2	62.7	60.1	**77.2**	51.9	80.4	81.7	79.6	68.8	**76.4**	64.6
0.15	61.8	69.5	58.0	58.4	73.3	48.4	79.9	81.2	79.3	66.7	74.7	61.9

Impact of Semantic Alignment Regularization. To further examine the effect of the semantic alignment loss L_{align}, we evaluate model performance under different regularization weights γ, as reported in Tab. 3. The results indicate that the model achieves optimal performance at $\gamma = 0.05$.

Introducing L_{align} enforces directional consistency between v_{ver} and the global hypothesis, ensuring that the verification branch refines semantic representations rather than introducing spurious noise, thereby enhancing novel category discovery. However, excessively strong regularization suppresses essential discriminative residuals within the verification branch, limiting the model's ability to capture subtle semantic variations and leading to performance degradation.

These observations suggest that maintaining a moderate alignment strength is critical for achieving a balanced and robust reasoning framework.

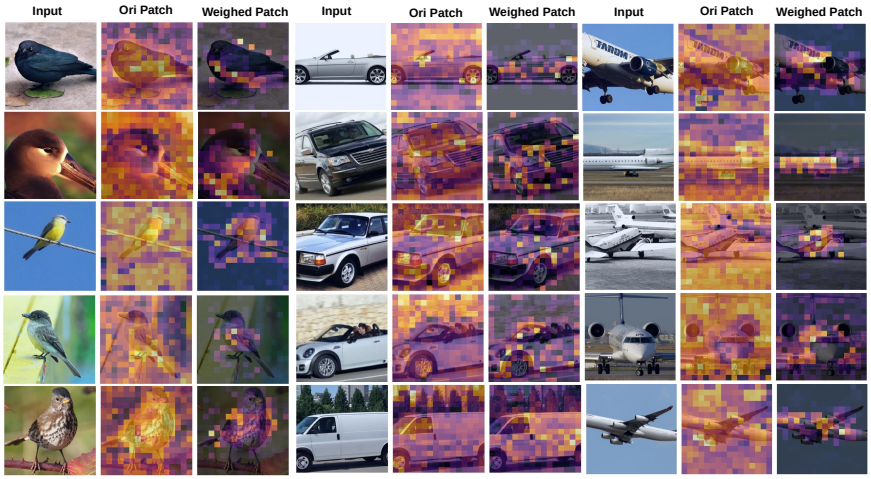

Fig. 4: Comparison of the original patch responses and the α_i-weighted patch responses. The former is primarily concentrated in texture-rich or high-contrast regions, whereas the latter focuses on structurally meaningful and discriminative regions

Analysis of Spatial Attention and Discriminative Cues. To illustrate the role of α_i, we visualize the original patch responses and the α_i-weighted patch responses, as shown in Fig. 4. The weighted patch responses exhibit a clearer semantic focus compared to the saliency-driven original responses.

Specifically, the model shifts its attention from texture- or contrast-dominated regions to structurally meaningful areas that provide discriminative cues for semantic verification. These regions often contain critical information that distinguishes visually similar categories or corrects the initial hypothesis. This result clearly indicates that our method effectively guides the model to contract its attention scope, transitioning from dispersed saliency-driven responses toward

more structured semantic verification regions, thereby achieving a more discriminative spatial focus.

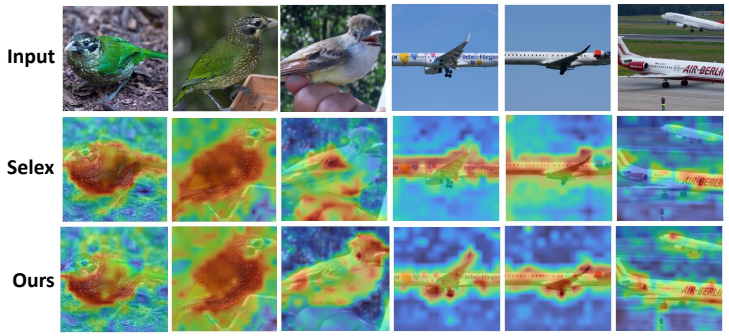

Fig. 5: Visualization on attention maps of SoTA and HVGCD

Attention Map Visualization and Comparison. As shown in Fig. 5, after introducing the verification enhancement mechanism, our method exhibits a more concentrated and structured semantic distribution compared to SelEx. The spatial responses contract from broadly covering the target object to focusing on key semantic components, resulting in a more compact heat distribution with stronger directional emphasis.

More importantly, this contraction does not merely reduce the spatial extent of activation, but reflects hypothesis-guided semantic verification. The model selectively emphasizes structurally informative regions that contribute to distinguishing similar categories or refining the initial global hypothesis. Such structured semantic focusing produces representations that are more compact and discriminative, better aligning with the semantic structural patterns required for reliable category discrimination in open-world scenarios.

5 Conclusion

In this work, we revisit Generalized Category Discovery from the perspective of structured inference and identify a fundamental limitation of existing methods: the absence of an explicit verification mechanism beyond global hypothesis formation. To address this issue, we propose HVGCD, a plug-and-play framework that decomposes instance representation into global hypothesis, hypothesis-conditioned verification evidence, and statistical grounding. By establishing a hypothesis-guided interaction between global semantics and local discriminative cues, our method enhances both novel category discovery and known-class recognition while maintaining robustness under open-world uncertainty. Extensive experiments on multiple coarse- and fine-grained benchmarks demonstrate the effectiveness and generality of the proposed framework.

Acknowledgements

This work was supported by the Xiaomi Foundation through the Xiaomi Young Scholar Program and the Fundamental Research Funds for the Central Universities under Grant No. 20720252020.

References

1. Banerjee, A., Kallooriyakath, L.S., Biswas, S.: Amend: Adaptive margin and expanded neighborhood for efficient generalized category discovery. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2101–2110 (2024)
2. Bendale, A., Boult, T.: Towards open world recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1893–1902 (2015)
3. Bendale, A., Boult, T.E.: Towards open set deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1563–1572 (2016)
4. Boult, T.E., Cruz, S., Dhamija, A.R., Gunther, M., Henrydoss, J., Scheirer, W.J.: Learning and the unknown: Surveying steps toward open world recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9801–9807 (2019)
5. Cao, K., Brbic, M., Leskovec, J.: Open-world semi-supervised learning. arXiv preprint arXiv:2102.03526 (2021)
6. Cao, X., Zheng, X., Wang, G., Yu, W., Shen, Y., Li, K., Lu, Y., Tian, Y.: Solving the catastrophic forgetting problem in generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16880–16889 (2024)
7. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
9. Cendra, F.J., Han, K.: Partco: Part-level correspondence priors enhance category discovery. arXiv preprint arXiv:2509.22769 (2025)
10. Cendra, F.J., Zhao, B., Han, K.: Promptccd: Learning gaussian mixture prompt pool for continual category discovery. In: European conference on computer vision. pp. 188–205. Springer (2024)
11. Chiaroni, F., Dolz, J., Masud, Z.I., Mitiche, A., Ben Ayed, I.: Parametric information maximization for generalized category discovery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1729–1739 (2023)
12. Choi, S., Kang, D., Cho, M.: Contrastive mean-shift learning for generalized category discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23094–23104 (2024)
13. Clark, A.: Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and brain sciences **36**(3), 181–204 (2013)
14. Clark, A.: Surfing uncertainty: Prediction, action, and the embodied mind. Oxford University Press (2015)
15. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)

16. Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., Weston, J.: Chain-of-verification reduces hallucination in large language models. In: Findings of the association for computational linguistics: ACL 2024. pp. 3563–3578 (2024)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
18. Fei, Y., Zhao, Z., Yang, S., Zhao, B.: Xcon: Learning with experts for fine-grained category discovery. arXiv preprint arXiv:2208.01898 (2022)
19. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: International conference on learning representations (2018)
20. Geng, C., Huang, S.j., Chen, S.: Recent advances in open set recognition: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3614–3631 (2020)
21. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
22. Grainger, R., Paniagua, T., Song, X., Cuntoor, N., Lee, M.W., Wu, T.: Paca-vit: learning patch-to-cluster attention in vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18568–18578 (2023)
23. Gregory, R.L.: Perceptions as hypotheses. Philosophical Transactions of the Royal Society of London. B, Biological Sciences **290**(1038), 181–197 (1980)
24. Guo, Y., Stutz, D., Schiele, B.: Robustifying token attention for vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17557–17568 (2023)
25. Han, K., Rebuffi, S.A., Ehrhardt, S., Vedaldi, A., Zisserman, A.: Autonovel: Automatically discovering and learning novel visual categories. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6767–6781 (2021)
26. Han, K., Vedaldi, A., Zisserman, A.: Learning to discover novel visual categories via deep transfer clustering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8401–8409 (2019)
27. Hao, S., Han, K., Wong, K.Y.K.: Cipr: An efficient framework with cross-instance positive relations for generalized category discovery. arXiv preprint arXiv:2304.06928 (2023)
28. Havtorn, J.D., Royer, A., Blankevoort, T., Bejnordi, B.E.: Msvit: Dynamic mixed-scale tokenization for vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 838–848 (2023)
29. He, Z., Liu, Y., Han, K.: Seal: Semantic-aware hierarchical learning for generalized category discovery. Advances in Neural Information Processing Systems **38**, 160765–160794 (2026)
30. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al.: Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795 (2021)
31. Jain, S., Wallace, B.C.: Attention is not explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 3543–3556 (2019)

32. Jaynes, E.T.: Information theory and statistical mechanics. Physical review **106**(4), 620 (1957)
33. Joseph, K., Paul, S., Aggarwal, G., Biswas, S., Rai, P., Han, K., Balasubramanian, V.N.: Novel class discovery without forgetting. In: European Conference on Computer Vision. pp. 570–586. Springer (2022)
34. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
35. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
36. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in neural information processing systems **33**, 21464–21475 (2020)
37. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in neural information processing systems **33**, 11525–11538 (2020)
38. Ma, S., Zhu, F., Zhang, X.Y., Liu, C.L.: Protogcd: Unified and unbiased prototype learning for generalized category discovery. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
39. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al.: Self-refine: Iterative refinement with self-feedback. Advances in neural information processing systems **36**, 46534–46594 (2023)
40. Mahdavi, A., Carvalho, M.: A survey on open set recognition. In: 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). pp. 37–44. IEEE (2021)
41. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
42. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
43. Parmar, J., Chouhan, S., Raychoudhury, V., Rathore, S.: Open-world machine learning: applications, challenges, and opportunities. ACM Computing Surveys **55**(10), 1–37 (2023)
44. Pu, N., Zhong, Z., Sebe, N.: Dynamic conceptional contrastive learning for generalized category discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7579–7588 (2023)
45. Qiu, Y., Embar, V., Cohen, S.B., Han, B.: Think while you write: Hypothesis verification promotes faithful knowledge-to-text generation. In: Findings of the Association for Computational Linguistics: NAACL 2024. pp. 1628–1644 (2024)
46. Rastegar, S., Doughty, H., Snoek, C.: Learn to categorize or categorize to learn? self-coding for generalized category discovery. Advances in Neural Information Processing Systems **36**, 72794–72818 (2023)
47. Rastegar, S., Salehi, M., Asano, Y.M., Doughty, H., Snoek, C.G.: Selex: Self-expertise in fine-grained generalized category discovery. In: European Conference on Computer Vision. pp. 440–458. Springer (2024)
48. Tan, K.C., Liu, Y., Ambrose, B., Tulig, M., Belongie, S.: The herbarium challenge 2019 dataset. arXiv preprint arXiv:1906.05372 (2019)
49. Tang, L., Huang, K., Chen, C., Chen, C.: Bures generalized category discovery. In: The Fourteenth International Conference on Learning Representations (2026)

50. Tang, L., Huang, K., Chen, C., Yuan, Y., Li, C., Tu, X., Ding, X., Huang, Y.: Dissecting generalized category discovery: Multiplex consensus under self-deconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 297–307 (2025)
51. Trevarthen, C.: Cognition and reality: principles and implications of cognitive psychology (1977)
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
53. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: A good closed-set classifier is all you need. In: International conference on learning representations (2021)
54. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7492–7501 (2022)
55. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7492–7501 (2022)
56. Vaze, S., Vedaldi, A., Zisserman, A.: No representation rules them all in category discovery. Advances in Neural Information Processing Systems **36**, 19962–19989 (2023)
57. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., et al.: The caltech-ucsd birds-200-2011 dataset. Tech. rep., Technical Report CNS-TR-2011-001, California Institute of Technology (2011)
58. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 24–25 (2020)
59. Wang, H., Vaze, S., Han, K.: Hilo: A learning framework for generalized category discovery robust to domain shifts. arXiv preprint arXiv:2408.04591 (2024)
60. Wang, H., Vaze, S., Han, K.: Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. arXiv preprint arXiv:2403.13684 (2024)
61. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
62. Wason, P.C.: On the failure to eliminate hypotheses in a conceptual task. Quarterly journal of experimental psychology **12**(3), 129–140 (1960)
63. Wen, X., Zhao, B., Qi, X.: Parametric classification for generalized category discovery: A baseline study. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16590–16600 (2023)
64. Zhang, S., Khan, S., Shen, Z., Naseer, M., Chen, G., Khan, F.S.: 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. pp. 3479–3488 (2023)
65. Zhao, B., Wen, X., Han, K.: Learning semi-supervised gaussian mixture models for generalized category discovery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16623–16633 (2023)
66. Zheng, H., Pu, N., Li, W., Sebe, N., Zhong, Z.: Generalized fine-grained category discovery with multi-granularity conceptual experts. arXiv preprint arXiv:2509.26227 (2025)