BEYOND UNIFORM GRANULARITY: BENCHMARKING MIXED-GRANULARITY OPEN-SET RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most existing open-set recognition (OSR) benchmarks implicitly assume uniform category granularity across known class labels. In practice, however, known label sets often span multiple levels of category granularity due to differences in annotator expertise and other practical factors, resulting in a mismatch between such benchmarks and real-world recognition settings. To address this mismatch, we introduce Mixed-Granularity Open-Set Recognition (MG-OSR), where known classes may lie at different levels of an underlying category hierarchy while satisfying an antichain constraint[1], yet the hierarchy itself is not exposed to the model. We establish an MG-OSR benchmark and evaluate a broad range of conventional and vision-language-model-based OSR methods. We find that this mixed setting poses a threefold challenge to existing methods: (i) unseen subclasses of coarse-grained known classes are prone to being rejected as unknown; (ii) unknowns neighboring fine-grained known classes are prone to being accepted as known; and (iii) shared representation learning exhibits cross-granularity interference, reflected in reduced gradient coherence among sibling categories under the same taxonomic parent. We further show that widely used metrics, including AUROC and OSCR, can be unreliable indicators of granularity-dependent recognition quality in mixed-granularity settings, motivating a granularity-aware evaluation metric that we use to systematically reassess existing methods. Our study identifies mixed-granularity label spaces as an overlooked regime of OSR and provides a unified testbed for OSR beyond uniform granularity.

1 INTRODUCTION

Open-set recognition (OSR) addresses recognition in the presence of novel categories: a model must correctly classify samples from known classes while rejecting inputs from classes unseen during training Scheirer et al. (2012); Bendale & Boult (2016). Existing work has advanced OSR through open-space modeling, representation learning, synthetic unknowns, and calibrated decision boundaries Neal et al. (2018); Yoshihashi et al. (2019); Zhou et al. (2021); Chen et al. (2021). More recently, studies have revisited OSR benchmarking Vaze et al. (2022), while vision-language models (VLMs) have broadened the design space for open-set recognition Radford et al. (2021); Miller et al. (2024). Yet most existing OSR benchmarks share an implicit assumption: *known class labels are defined at a uniform level of category granularity*.

This assumption is convenient but can be mismatched with real-world labeling practice. Known label sets may span multiple levels of category granularity because annotators with different expertise, among other practical factors, can label concepts at different levels of specificity. For example, a recognition system may encounter the label space {*cat*, *golden retriever*, *husky*}, where *cat* is a coarse category while *golden retriever* and *husky* are substantially finer. Crucially, the model may receive such labels without any hierarchy indicating their relative granularity. This differs from hierarchical novelty recognition Lee et al. (2018); Pyakurel & Yu (2024); Hosoya et al. (2025), which assumes access to a taxonomy and exploits parent–child relations during learning or inference. It also differs from fine-grained OSR settings that study semantic proximity or globally controlled granularity Lang et al. (2024). Table 1 summarizes the differences between our proposed scenario

[1] An antichain constraint requires that no pair of known classes has an ancestor–descendant relation in the underlying category hierarchy; equivalently, no known class semantically subsumes another.

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.

Table 1: Comparison of related open-set recognition settings. MG-OSR considers mixed granularity within the known label set while keeping the underlying hierarchy hidden from the learner, such that unseen descendants of coarse known classes may still be treated as known.

Setting	Mixed granularity within known set	Hierarchy available to learner	Unseen descendants may remain known
Standard OSR	×	×	×
Hierarchical novelty detection	*Task-dependent*	✓	*Task-dependent*
Fine-grained OSR	×	×	×
MG-OSR (ours)	✓	×	✓

and existing settings. Our focus is instead a known label space in which *different levels of category granularity coexist while the underlying hierarchy remains hidden from the model*.

We formalize this regime as **Mixed-Granularity Open-Set Recognition (MG-OSR)**. Known classes may lie at different levels of an underlying category hierarchy while satisfying an *antichain constraint*: no pair of known classes has an ancestor–descendant relation, or equivalently, no known class semantically subsumes another. Thus, {*cat*, *golden retriever*, *husky*} is valid, whereas {*dog*, *golden retriever*, *husky*} is not. This seemingly simple change alters what it means for a category to be "unknown". An unseen subclass may still belong to the semantic scope of a coarse known label, whereas a closely related category may fall outside the scope of a fine-grained known label. **Thus, unseen does not necessarily mean unknown**. The hierarchy is used only for benchmark construction and evaluation and is never exposed to the model. This setting changes the risk structure of OSR. A coarse-grained known class may cover multiple subclasses while its training observations capture only part of this semantic extent, making unseen but valid subclasses prone to rejection as unknown. In contrast, a fine-grained known class requires tighter discrimination among closely related concepts, making neighboring unknowns prone to acceptance as known. Mixed granularity therefore does not merely make OSR harder; *it changes how OSR fails*.

To systematically characterize this regime, we establish an MG-OSR benchmark and evaluate a broad range of conventional and VLM-based OSR methods. Our experiments reveal a threefold challenge induceitd by mixed granularity: (i) unseen subclasses of coarse-grained known classes are prone to being rejected as unknown; (ii) unknowns neighboring fine-grained known classes are prone to being accepted as known; and (iii) shared representation learning exhibits cross-granularity interference, reflected in reduced gradient coherence among sibling categories under the same semantic parent. The first two failure modes consistently emerge across model families, while the third provides optimization-level evidence that mixed-granularity supervision can disrupt shared representation learning. Together, these findings show that mixed granularity affects both recognition behavior and representation learning.

These challenges further expose an evaluation mismatch in MG-OSR. AUROC measures the separability between known and unknown scores, while OSCR additionally incorporates correct classification of known samples Dhamija et al. (2018). Both, however, aggregate performance across categories and can obscure granularity-dependent failure modes. On our benchmark, models that severely reject valid subclasses of coarse-grained known classes or excessively accept unknowns near fine-grained known classes can still obtain competitive aggregate scores, leading to misleading or inconsistent method comparisons. We therefore introduce a **granularity-aware evaluation metric** that explicitly captures these asymmetric risks and use it to systematically reassess conventional and VLM-based OSR methods.

In summary, our contributions are three-fold:

- We introduce **Mixed-Granularity Open-Set Recognition (MG-OSR)** and establish a controlled benchmark in which known classes may occupy different levels of an underlying category hierarchy under an antichain constraint, while the hierarchy itself remains hidden from the model.
- We uncover **a threefold challenge** induced by mixed granularity: unseen subclasses of coarse-grained known classes are prone to false rejection, unknowns neighboring fine-

grained known classes are prone to false acceptance, and shared representation learning exhibits cross-granularity interference, reflected in reduced gradient coherence among sibling categories.

- We show that widely used OSR metrics such as AUROC and OSCR can be unreliable indicators of granularity-dependent recognition quality, and introduce a **granularity-aware evaluation metric** for systematically reassessing existing methods.

2 Mixed-Granularity Open-Set Recognition

2.1 Problem Formulation

Let $\mathcal{X}$ denote the input space and $\mathcal{H} = (\mathcal{V}, \preceq)$ an underlying category taxonomy, where $u \preceq v$ indicates that u is subsumed by v. Let $\mathcal{Y}^K = \{y_1, \dots, y_K\} \subset \mathcal{V}$ denote the known label set. We study **Mixed-Granularity Open-Set Recognition (MG-OSR)**, where known labels may occupy different levels of the taxonomy. To ensure that known labels are non-overlapping, we require $\mathcal{Y}^K$ to form an *antichain*:

$$\forall\, y_i \neq y_j, \qquad y_i \not\preceq y_j, \quad y_j \not\preceq y_i. \tag{1}$$

Thus, no known label subsumes another. The taxonomy $\mathcal{H}$ is used only for benchmark construction and evaluation and is hidden from the learner, which observes only the flat training set

$$\mathcal{D}_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^{N}, \qquad y_i \in \mathcal{Y}^K. \tag{2}$$

Let $\mathcal{A} \subseteq \mathcal{V}$ denote the set of atomic categories. We define the *taxonomic scope* of a known label y as

$$\mathcal{S}(y) = \{a \in \mathcal{A} : a \preceq y\}, \tag{3}$$

and let $\mathcal{S}_{\text{tr}}(y) \subseteq \mathcal{S}(y)$ denote the atomic categories observed during training. For a test input x with underlying atomic category $a(x)$, the target is

$$y^*(x) = \begin{cases} y, & a(x) \in \mathcal{S}(y) \text{ for some } y \in \mathcal{Y}^K, \\ \perp, & \text{otherwise}, \end{cases} \tag{4}$$

where $\perp$ denotes unknown. The antichain constraint ensures that the corresponding known label y is unique. A test sample is therefore either a *seen known* $a \in \mathcal{S}_{\text{tr}}(y)$, an *unseen known variant/subclass* $a \in \mathcal{S}(y) \setminus \mathcal{S}_{\text{tr}}(y)$, or a *genuine unknown* $a \notin \bigcup_{y \in \mathcal{Y}^K} \mathcal{S}(y)$. MG-OSR aims to learn

$$f_\theta : \mathcal{X} \to \mathcal{Y}^K \cup \{\perp\} \tag{5}$$

from flat supervision alone, accepting both seen and unseen variants/subclasses within each known label's taxonomic scope while rejecting samples outside all known scopes.

2.2 Benchmark Construction and Protocol

Datasets. We construct MG-OSR benchmarks on CIFAR-100, FGVC-Aircraft, ImageNet-1k, and iNaturalist 2021, using their established category taxonomies: CIFAR-100 superclasses, the Maker–Family–Variant hierarchy, WordNet, and biological taxonomy, respectively. The taxonomies are used only for benchmark construction and evaluation and are hidden from the learner.

Mixed-granularity splits. For each dataset, we sample disjoint taxonomic branches and assign them coarse- or fine-grained supervision under the antichain constraint in Section 2.1. A coarse known label corresponds to an internal node: a subset of its descendant leaves is observed during training, while the remaining descendants form *unseen known variants*. A fine known label corresponds to a leaf, with neighboring unseen categories reserved as *near unknowns*. Categories from branches not covered by any known label form *far unknowns*. The test set therefore comprises *seen knowns*, *unseen known variants*, *near unknowns*, and *far unknowns*; the first two should be accepted and the latter two rejected.

Protocol. We generate five splits (seeds 42$\sim$46), fixing the coarse/fine composition while resampling taxonomic branches. We use a fixed train/test partition for each dataset. Detailed dataset splits and implementation settings are provided in the supplementary material. All methods receive only images and flat mixed-granularity labels, without access to the taxonomy, category depth, or latent leaf identities. Table 2 summarizes the benchmark statistics.

Table 2: MG-OSR benchmark statistics across five splits. Known heads are reported as total (coarse / fine); coarse variants denote held-out descendants that remain known.

Dataset	Knowns (coarse / fine)	Unseen known variants	Near unknowns	Far unknowns
CIFAR-100	20 (5 / 15)	$\{10, 10, 10, 10, 10\}$	$\{10, 10, 10, 10, 10\}$	$\{50, 50, 50, 50, 50\}$
FGVC-Aircraft	20 (6 / 14)	$\{7, 7, 8, 9, 10\}$	$\{14, 14, 15, 16, 16\}$	$\{54, 54, 54, 54, 54\}$
ImageNet-1k	60 (18 / 42)	$\{35, 39, 40, 54, 77\}$	$\{49, 51, 54, 54, 55\}$	$\{640, 659, 676, 695, 696\}$
iNaturalist	200 (60 / 140)	$\{143, 155, 170, 174, 237\}$	$\{157, 161, 165, 207, 430\}$	$\{6773, 6850, 7108, 7195, 7430\}$

Table 3: Average Seen / Unseen accuracy (%) over 5 seeds on CIFAR100, FGVC-Aircraft, Imagenet, Inaturalist. $\Delta = \mathrm{UA} - \mathrm{SA}$.

	CIFAR100			FGVC-Aircraft			Imagenet			Inaturalist		
	S A	U A	Δ	S A	U A	Δ	S A	U A	Δ	S A	U A	Δ
Traditional												
ARPL+CS Chen et al. (2021)	90.5	59.7	-30.8	70.8	44.4	-26.4	88.9	55.9	-33.0	78.4	51.7	-26.7
MLS Vaze et al. (2022)	90.6	61.2	-29.4	63.2	38.0	-25.3	88.0	59.5	-28.4	81.7	55.6	-26.2
MixOE Zhang et al. (2023)	89.9	59.9	-30.0	59.1	36.2	-22.9	87.4	59.1	-28.3	79.3	53.5	-25.8
DEF Li et al. (2025)	90.5	60.9	-29.6	44.9	29.1	-15.8	58.0	47.5	-10.5	29.9	23.0	-6.9
VLM-based												
ZS CLIP Radford et al. (2021)	–	71.8	–	–	39.0	–	–	68.5	–	–	22.6	–
TANL Zhang et al. (2026)	–	64.1	–	–	41.4	–	–	64.6	–	–	20.2	–
LP Clip Radford et al. (2021)	92.5	73.3	-19.2	78.0	56.8	-21.2	93.3	69.2	-24.1	83.2	66.5	-16.7
CoOp Zhou et al. (2022b)	93.2	79.1	-14.1	76.0	57.6	-18.3	93.5	73.3	-20.2	79.2	63.5	-15.7
CoCoOp Zhou et al. (2022a)	93.1	76.6	-16.5	70.6	48.6	-22.0	93.3	74.1	-19.2	73.8	57.3	-16.5
A2PT Ren et al. (2023)	92.9	85.2	-7.7	73.2	59.8	-13.5	91.6	75.3	-16.2	66.6	52.2	-14.4
FSMoE Xie et al. (2026)	93.1	74.3	-18.8	77.6	55.0	-22.6	93.9	71.7	-22.2	81.2	62.1	-19.0
DOT-OOD Lu et al. (2026)	93.9	77.5	-16.4	83.8	60.1	-23.7	95.1	72.4	-22.7	86.5	67.7	-18.8

3 Understanding Mixed-Granularity OSR

To understand how existing recognizers behave under mixed granularity, we study twelve representative methods spanning conventional OSR and vision-language models (VLMs). The former include ARPL+CS Chen et al. (2021), MLS Vaze et al. (2022), MixOE Zhang et al. (2023), and DeF Li et al. (2025), all trained from scratch with VGG16-BN. The latter build on CLIP ViT-B/32 Radford et al. (2021) and cover a broad range of adaptation strategies: ZS-CLIP, TANL Zhang et al. (2026), LP-CLIP, CoOp Zhou et al. (2022b), CoCoOp Zhou et al. (2022a), A2PT Ren et al. (2023), FSMoE Xie et al. (2026), and DOT-OOD Lu et al. (2026). This collection ranges from training-free inference and linear probing to prompt learning and model fine-tuning, enabling us to examine whether mixed-granularity failures persist across different recognition paradigms. All methods use the same MG-OSR splits and retain their original open-set scores whenever available. Conventional methods follow their original training recipes on VGG16-BN. CLIP-based methods use ViT-B/32 and are trained for 50 epochs with a batch size of 128 (64 for FSMoE), with the visual encoder frozen except for DOT-OOD. Prompt-learning methods use 16 context tokens per class, while ZS-CLIP and DOT-OOD use the template "a photo of a [class]".

Unseen subclasses of coarse-grained known classes are prone to false rejection. We first examine whether a coarse-grained known class can generalize to valid subclasses that are not represented during training. A coarse label covers multiple subclasses within its taxonomic scope, only a subset of which may be observed during training. Thus, a test sample can be unseen at the subclass level while still belonging to a known class and should not be rejected as unknown.

To isolate this effect, we evaluate the same target subclasses under two training conditions: they are either included in (*seen*) or held out from (*unseen*) training, while their coarse parent remains known. Let SA and UA denote the corresponding Seen-Subclass and Unseen-Subclass Accuracies, and define

$$\Delta = \mathrm{UA} - \mathrm{SA}. \tag{6}$$

Since the test samples are identical, a negative Δ directly measures the degradation caused by missing subclass-level exposure.

As shown in Table 3, UA consistently drops across all datasets and both conventional and VLM-based recognizers, with reductions approaching 30 points in several settings. Even strong pretrained models remain affected; for example, A2PT drops by 7.7$\sim$16.2 points across the four benchmarks.



Figure 1: Known class and near unknown knownness score distributions at the TPR_{95} threshold (dashed line). Rows: ARPL+CS, MLS, CoOp, A2PT; columns: CIFAR-100, FGVC-Aircraft, ImageNet-1k, iNaturalist. The two masses overlap for every method and dataset — the overlap couples coarse false rejection with fine false acceptance and is not specific to one model family.

These results reveal a distinct coarse-side failure: samples within a known label's taxonomic scope can be rejected simply because their subclasses were unseen during training. We term this *unseen-subclass rejection*.

Unknowns neighboring fine-grained known classes are prone to false acceptance. We next examine the complementary failure around fine-grained known classes. Unlike coarse classes, which must generalize to unseen but valid subclasses, fine-grained classes require tight discrimination from closely related concepts. Consequently, neighboring categories may fall outside a known label's taxonomic scope while remaining highly similar to it.

To quantify this effect, we compare the knownness score distributions of known-class and *near-unknown* samples, where larger scores indicate stronger evidence of being known. We set the acceptance threshold at

$$\tau_{95} = Q_{0.05}(s(x) \mid x \in \mathcal{D}_{\text{Known}}), \tag{7}$$

which retains 95% of known class samples. The near-unknown mass above τ_{95} therefore measures their false-acceptance rate.

As shown in Figure 1, known and near-unknown score distributions overlap substantially across datasets and both conventional and VLM-based recognizers, causing many near unknowns to enter the acceptance region. We term this fine-side failure **near-unknown acceptance**. Together, the two failures expose a fundamental asymmetry in MG-OSR: coarse classes are difficult to cover broadly enough, whereas fine classes are difficult to bound tightly enough.

Mixed-granularity supervision induces cross-granularity interference. The asymmetric failures above suggest that coarse- and fine-grained supervision impose different demands on a shared representation: the former favors invariance across within-class variants, whereas the latter requires sensitivity to subtle inter-class differences. We next examine whether mixing these supervision granularities interferes with the optimization of the shared encoder.

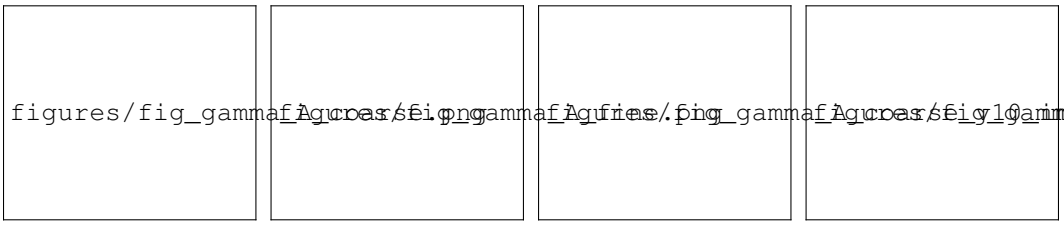

Figure 2: Cross-granularity interference in shared representation learning. Intra-branch gradient coherence Γ_A for branch A on CIFAR-100 and ImageNet when A is labelled coarse or fine. Blue and red denote fine- and coarse-grained supervision on the other branches, respectively. Matching granularity generally yields higher Γ_A, indicating reduced gradient coherence under granularity mismatch. Error bars show standard deviations over ten resampling rounds.

To quantify this effect, consider a taxonomic branch $A = \{a_1, \ldots, a_m\}$. For each subclass a_i, let

$$g_i = \nabla_\theta \mathcal{L}(\mathcal{B}_i, y_i), \qquad y_i = \begin{cases} A, & \text{if } A \text{ is coarse,} \\ a_i, & \text{if } A \text{ is fine,} \end{cases} \tag{8}$$

denote the encoder gradient induced by a subclass-specific mini-batch $\mathcal{B}_i$. We define the *intra-branch gradient coherence* as

$$\Gamma_A = \frac{2}{m(m-1)} \sum_{i<j} \cos(g_i, g_j), \tag{9}$$

where higher Γ_A indicates more aligned subclass-level updates within A.

We conduct controlled experiments on CIFAR-100 and ImageNet. Taking CIFAR-100 as an example, we fix A as aquatic mammals and progressively add other taxonomic branches ($\{B\}$, $\{B, C\}$, and $\{B, C, D\}$), which are labelled either all coarse or all fine, while independently setting A to coarse or fine. As shown in Figure 2, across both datasets, Γ_A is consistently higher when the other branches match the granularity of A and decreases when coarse- and fine-grained supervision are mixed. This consistent mismatch effect shows that the granularity assigned to other branches can alter the optimization signals within A, providing optimization-level evidence of cross-granularity interference.

4 Rethinking Evaluation under Mixed Granularity

4.1 Do Standard OSR Metrics Capture Granularity-Dependent Generalization?

The preceding analysis reveals two asymmetric challenges in MG-OSR: coarse-grained classes must generalize to unseen subclasses, whereas fine-grained classes must reject neighboring unknowns. This raises a natural question: *do standard OSR metrics capture these granularity-dependent capabilities?*

AUROC and OSCR summarize global known–unknown separability, with OSCR additionally accounting for correct known-class recognition. While useful for overall OSR performance, such aggregate metrics may obscure failures specific to label granularity. To examine this, we here focus on coarse-side generalization and define *Generalization Retention (GR)* using the SA and UA introduced above:

$$\mathrm{GR} = 100 + (\mathrm{UA} - \mathrm{SA}), \tag{10}$$

where $\mathrm{GR} = 100$ indicates no degradation from seen to unseen subclasses, and lower values indicate poorer retention. GR serves as a diagnostic probe rather than an alternative OSR metric.

Figure 3 compares GR with Global AUROC and OSCR across conventional and VLM-based methods on four datasets. **The two views are consistently misaligned**: methods with similar aggregate OSR scores can exhibit markedly different GR, while higher AUROC or OSCR does not consistently imply stronger generalization retention. Importantly, this mismatch arises both *within* each



Figure 3: Generalization retention (GR, Eq. equation 10) versus the Global open-set scores over the ten baselines (zero-shot methods without Seen Acc are excluded): top row, Global AUROC; bottom row, Global OSCR. Blue circles: VLM-based; red triangles: traditional. Ranks under Global AUROC and OSCR are inconsistent with the coarse-side retention (e.g., on FGVC-Aircraft the AUROC leader sits near the bottom in GR, and vice versa).

method family and *across* conventional and VLM-based recognizers, indicating that it is not specific to a particular recognition paradigm. The discrepancy is particularly pronounced on FGVC-Aircraft and iNaturalist.

These results reveal an **evaluation mismatch under mixed granularity**. AUROC and OSCR remain meaningful measures of global open-set performance, but do not necessarily reflect whether a model generalizes beyond observed subclasses to the full taxonomic scope of a coarse-grained known class. Thus, *aggregate OSR performance alone is insufficient to characterize granularity-dependent recognition quality*.

4.2 Granularity-Aware Evaluation and Re-evaluation of Existing OSR Methods

Granularity-aware evaluation. The evaluation mismatch above motivates a protocol that explicitly captures both sides of MG-OSR: accepting unseen variants within the taxonomic scope of coarse-grained known classes and rejecting near unknowns around fine-grained known classes.

Let $s(x)$ denote the knownness score, with larger values indicating stronger evidence of being known. We set the rejection threshold τ_{95} to the 5-th percentile of ID validation scores, corresponding to 95% TPR. At this operating point, we define

$$\text{CV-RR} = \frac{|\{x \in G : s(x) < \tau_{95}\}|}{|G|}, \qquad \text{UN-RR} = \frac{|\{x \in N : s(x) < \tau_{95}\}|}{|N|}, \tag{11}$$

where G denotes unseen known variants of coarse-grained classes and N denotes near unknowns around fine-grained classes. CV-RR measures erroneous rejection of valid coarse variants (lower is better), whereas UN-RR measures successful rejection of near unknowns (higher is better).

To capture whether a model succeeds on both sides, we define $\text{CV-AR} = 1 - \text{CV-RR}$ and take the harmonic mean of CV-AR and UN-RR:

$$\text{H-Score} = \frac{2 * \text{CV-AR} * \text{UN-RR}}{\text{CV-AR} + \text{UN-RR}}. \tag{12}$$

A high H-Score therefore requires both broad coarse-class coverage and strong fine-class rejection. We report all three metrics at TPR_{95}.

Re-evaluation of existing OSR methods. We re-evaluate conventional and VLM-based OSR methods under the proposed granularity-aware protocol. Table 4 yields three key observations.

First, existing methods struggle to balance the two sides of MG-OSR. Even the best H-Score reaches only 57.7% on CIFAR-100 and 53.3% on ImageNet-1k, and drops to 16.5% and 19.8% on

Table 4: Granularity-aware evaluation (%) at TPR_{95}, averaged over five splits. CV-RR measures erroneous rejection of unseen known variants (↓), UN-RR measures successful rejection of near unknowns (↑), and H-Score summarizes coarse-variant acceptance and near-unknown rejection (↑).

	CIFAR100			FGVC-Aircraft			Imagenet			Inaturalist		
	CR↓	UR↑	HS↑	CR↓	UR↑	HS↑	CR↓	UR↑	HS↑	CR↓	UR↑	HS↑
Traditional												
ARPL+CS Chen et al. (2021)	12.3	17.3	28.8	4.8	4.8	9.1	13.7	19.4	31.7	6.2	10.1	18.3
MLS Vaze et al. (2022)	13.5	21.0	33.7	6.4	5.3	10.0	15.8	22.6	35.5	7.1	10.1	18.2
MixOE Zhang et al. (2023)	10.5	18.4	30.5	6.0	4.9	9.3	16.1	23.8	37.0	7.1	11.1	19.8
DEF Li et al. (2025)	11.3	19.8	32.3	5.6	6.0	11.2	6.3	9.3	16.7	4.0	6.9	12.7
VLM-based												
ZS CLIP Radford et al. (2021)	5.8	13.3	22.8	4.9	5.4	10.2	8.4	10.0	18.0	4.9	4.9	9.2
TANL Zhang et al. (2026)	5.4	11.0	19.0	4.1	5.3	10.0	10.9	8.9	16.0	7.9	5.0	9.1
LP Clip Radford et al. (2021)	11.9	22.1	35.0	4.8	8.6	15.7	19.0	27.0	40.3	5.4	9.1	16.5
CoOp Zhou et al. (2022b)	11.8	26.7	40.6	4.5	6.6	12.3	21.7	29.3	42.5	5.4	7.2	13.2
CoCoOp Zhou et al. (2022a)	9.9	25.0	38.8	4.9	6.2	11.6	10.9	17.7	27.7	5.6	7.6	14.0
A2PT Ren et al. (2023)	3.9	16.3	27.8	3.0	6.8	12.6	12.1	26.6	40.8	3.0	7.8	14.2
FSMoE Xie et al. (2026)	12.4	31.1	45.6	3.6	8.4	15.5	17.4	23.3	36.1	6.0	7.5	13.8
DOT-OOD Lu et al. (2026)	18.0	44.7	57.7	6.5	9.1	16.5	28.6	42.7	53.3	7.5	11.0	19.6

FGVC-Aircraft and iNaturalist, respectively. These results expose a substantial gap in simultaneously maintaining coarse-class coverage and fine-class rejection.

Second, near-unknown rejection is a major bottleneck, particularly on fine-grained datasets. The best UN-RR reaches 44.7% on CIFAR-100 and 42.7% on ImageNet-1k, but only 9.1% on FGVC-Aircraft and 11.1% on iNaturalist. Thus, even the best-performing method on this criterion accepts roughly nine out of ten near unknowns on the latter two datasets, highlighting the difficulty of enforcing tight rejection boundaries around fine-grained known classes.

Third, improving near-unknown rejection can come at the expense of coarse-class coverage. For example, DOT-OOD achieves the highest UN-RR on CIFAR-100 and ImageNet-1k (44.7% and 42.7%), but also incurs the highest CV-RR on both datasets (18.0% and 28.6%). This trade-off illustrates why neither side alone is sufficient and motivates evaluating their balance through H-Score.

5 Related Work

Open-set recognition. Open-set recognition (OSR) studies classification in the presence of categories unseen during training Scheirer et al. (2012); Geng et al. (2020); Vaze et al. (2022); Park et al. (2024); Yang et al. (2024); Wang et al. (2025); Sun et al. (2026). Unlike closed-set recognition, OSR requires a model not only to discriminate among known classes but also to identify inputs that fall outside the known label space. This dual objective has motivated a broad range of approaches, including synthetic unknown generation Neal et al. (2018), representation learning Yoshihashi et al. (2019); Zhou et al. (2021), calibrated decision boundaries Chen et al. (2021), and studies of the connection between closed-set classification and open-set recognition Vaze et al. (2022). Together, these directions seek to improve the separation between known and unknown samples while preserving accurate recognition of known categories. More recently, vision-language models (VLMs) have expanded the OSR design space by providing semantically structured pretrained representations and enabling recognition through class names Radford et al. (2021); Miller et al. (2024). Building on these capabilities, subsequent methods adapt VLMs to open-set recognition through prompt learning, visual–textual adaptation, test-time negative labels, and fine-tuning Ren et al. (2023); Xie et al. (2026); Lu et al. (2026). Despite substantial methodological progress across both conventional and VLM-based approaches, existing OSR methods are predominantly developed and evaluated under known label spaces defined at a common level of semantic granularity. Consequently, their behavior remains less understood when the known label space itself contains categories at different levels of granularity. Our work studies this mixed-granularity regime and systematically evaluates how existing OSR methods behave when coarse- and fine-grained categories coexist within the same known label space.

Hierarchical and fine-grained novelty recognition. Prior work has extended conventional OSR by incorporating hierarchical structure or finer-grained semantic distinctions. Hierarchical novelty detection explicitly exploits taxonomic information to provide more informative predictions for un-

seen categories. Lee et al. Lee et al. (2018) associate novel samples with their closest known superclasses in a hierarchy, while Dengxiong and Kong Dengxiong & Kong (2023) formulate generalized OSR as ancestor search using hierarchical side information. Pyakurel and Yu Pyakurel & Yu (2024) further improve hierarchical novelty detection through fine-grained evidence allocation over non-leaf nodes, and Xu and Xiang Xu & Xiang (2025) leverage visual-semantic attributes and hierarchical relations for interpretable open-set recognition. These approaches use hierarchy as learner-side information to structure novelty predictions. A complementary line of work studies fine-grained novelty, where unknown categories are increasingly similar to the known set. Vaze et al. Vaze et al. (2022) consider semantic shift in OSR benchmarking, while Lang et al. Lang et al. (2024) systematically study label granularity, open–closed-set similarity, and hierarchical supervision from coarse- to fine-grained OSR. MG-OSR differs from both lines of work in that coarse- and fine-grained known labels coexist within the same flat recognition problem, while the underlying taxonomy is hidden from the learner.

6 Discussion and Limitation

Discussion. This work exposes a tension that is easy to overlook in conventional OSR: being unseen during training does not uniquely determine whether a sample should be treated as unknown. In MG-OSR, an unseen subclass of a coarse-grained known class should be accepted, whereas a neighboring category outside a fine-grained known class should be rejected. This observation leads to a more fundamental question:

What does "open" mean in open-set recognition?

Consider an unseen Samoyed. If *dog* is a known label, it should be accepted as a valid instance; if *Golden Retriever* is known instead, the same Samoyed should be rejected as unknown. The sample is unchanged; what changes is the granularity at which the known world is defined. Thus, **unseen does not necessarily mean unknown**: a category absent from training may still fall within the taxonomic scope of a known label.

This perspective suggests that unknownness is not an intrinsic property of a category, but is relative to the scope defined by the known labels. MG-OSR makes this dependence explicit: coarse labels define broad scopes that must accommodate unseen variants, whereas fine labels define narrow scopes that must exclude closely related categories. Mixed granularity therefore induces class-dependent known–unknown boundaries. From this perspective, the central question in OSR is not simply whether an input has been seen before, but whether it lies within the scope of what is known—and at what level of granularity.

Limitation. MG-OSR relies on existing taxonomies to define granularity and taxonomic scope, which may not fully capture context-dependent category boundaries in the real world. Our experiments are also limited to four visual recognition datasets and a fixed set of mixed-granularity constructions. Finally, this work focuses on exposing and evaluating mixed-granularity failures rather than solving them; developing methods that jointly preserve coarse-class coverage and fine-class rejection remains an important direction for future work.

7 Conclusion

We introduced Mixed-Granularity Open-Set Recognition (MG-OSR), where coarse- and fine-grained categories coexist within a flat known label space. Our benchmark reveals asymmetric failures in coarse-class coverage and fine-class rejection, together with cross-granularity interference in representation learning. We further show that aggregate OSR metrics can obscure these granularity-dependent failures and introduce a granularity-aware evaluation protocol that explicitly captures both sides. Together, our findings establish mixed granularity as an overlooked challenge for OSR and call for recognition systems that can accommodate heterogeneous label granularity. More broadly, they highlight a simple principle: **unseen does not necessarily mean unknown**—unknownness is relative to the scope of what is known.

AI USE, ETHICS, AND REPRODUCIBILITY STATEMENTS

AI Use Statement. Generative AI tools were used to assist with the presentation and writing of this paper, including refining the paper structure, drafting and editing text for clarity and concision. Generative AI was not used to generate experimental data, implement the evaluated methods, or execute experiments. All research decisions, experimental results, and conclusions were reviewed and validated by the authors. The authors also independently checked AI-assisted technical statements, equations, and references for correctness. We take full responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with AI assistance.

Ethics Statement. This work introduces a benchmark and evaluation framework for mixed-granularity open-set recognition using existing vision datasets and their established category taxonomies. We do not collect new personal data or conduct experiments involving human subjects. Our study focuses on benchmarking and analyzing recognition systems and does not target sensitive attributes or high-stakes decision-making applications. Nevertheless, recognition systems may inherit biases from their underlying data and should be carefully evaluated before deployment in real-world settings.

Reproducibility Statement. We take several steps to facilitate reproducibility. Sections 2 and 4 describe the benchmark construction and evaluation protocols, while Section 3 specifies the evaluated methods and experimental setup. Additional implementation details, benchmark splits, and supplementary experiments are provided in the Appendix. We also provide anonymized code for benchmark construction and evaluation, together with the split definitions used in our experiments, as supplementary material.

REFERENCES

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.

Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2021.

Xiwen Dengxiong and Yu Kong. Ancestor search: Generalized open set recognition via hyperbolic side information learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3992–4001. IEEE, 2023.

Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018.

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.

Yusuke Hosoya, Masanori Suganuma, and Takayuki Okatani. Rethinking open-set object detection: Issues, a new formulation, and taxonomy: Y. hosoya et al. *International Journal of Computer Vision*, 133(9):6145–6169, 2025.

Nico Lang, Vésteinn Snæbjarnarson, Elijah Cole, Oisin Mac Aodha, Christian Igel, and Serge Belongie. From coarse to fine-grained open-set recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17804–17814. IEEE, 2024.

Kibok Lee, Kimin Lee, Kyle Min, Yuting Zhang, Jinwoo Shin, and Honglak Lee. Hierarchical novelty detection for visual object recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1034–1042. IEEE, 2018.

Chaohua Li, Enhao Zhang, Chuanxing Geng, and Songcan Chen. Unlocking better closed-set alignment based on neural collapse for open-set recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18227–18235, 2025. doi: 10.1609/aaai.v39i17.34005.

Hengyang Lu, Xin Guo, Shuai Feng, Wenyu Jiang, Yuntao Du, Chang Xia, and Chenyou Fan. A novel fine-tuned clip-ood detection method with double loss constraint through optimal transport semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15448–15456, 2026.

Dimity Miller, Niko Sünderhauf, Alex Kenna, and Keita Mason. Open-set recognition in the age of vision-language models. In *European conference on computer vision*, pp. 1–18. Springer, 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.

Jaewoo Park, Hojin Park, Eunju Jeong, and Andrew Beng Jin Teoh. Understanding open-set recognition by jacobian norm and inter-class separation. *Pattern recognition*, 145:109942, 2024.

Spandan Pyakurel and Qi Yu. Hierarchical novelty detection via fine-grained evidence allocation. In *Forty-first International Conference on Machine Learning*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Hairui Ren, Fan Tang, Xingjia Pan, Juan Cao, Weiming Dong, Zhiwen Lin, Ke Yan, and Changsheng Xu. A^2pt: Anti-associative prompt tuning for open set visual recognition. *IEEE Transactions on Multimedia*, 26:8419–8431, 2023.

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.

Heyang Sun, Chuanxing Geng, and Songcan Chen. Embracing the power of known class bias in open set recognition from a reconstruction perspective. *IEEE Transactions on Image Processing*, 35:14–26, 2026.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International conference on learning representations*, 2022.

Yu Wang, Junxian Mu, Hongzhi Huang, Qilong Wang, Pengfei Zhu, and Qinghua Hu. Backmix: Regularizing open set recognition by removing underlying fore-background priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6172–6184, 2025.

Yifei Xie, Chuanxing Geng, Yahao Hu, Man Chen, Jun Chen, and Zhisong Pan. Full-spectrum prompt tuning with sparse moe for open-set recognition. *Neural Networks*, 196:108369, 2026. doi: 10.1016/j.neunet.2025.108369.

Zhuo Xu and Xiang Xiang. Learning visual-semantic hierarchical attribute space for interpretable open-set recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5697–5706, 2025.

Haifeng Yang, Chuanxing Geng, Pong C Yuen, and Songcan Chen. Dynamic against dynamic: an open-set self-learning framework. *arXiv preprint arXiv:2404.17830*, 2024.

Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4011–4020. IEEE, 2019.

Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure for out-of-distribution detection in fine-grained environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5531–5540, 2023.

Yabin Zhang, Maya Varma, Yunhe Gao, Jean-Benoit Delbrouck, Jiaming Liu, Chong Wang, and Curtis Langlotz. Activation matters: Test-time activated negative labels for ood detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4399–4408. IEEE, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.