KNOWLEDGE RECOVERY FOR NOVEL CLASSES IN OPEN-VOCABULARY OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained open-vocabulary object detection (OVD) models can recognize novel classes without ground truth annotations in a target dataset. However, adapting these models with only base class supervision tends to compromise their generalization to novel classes. We find that treating unlabeled novel-class instances as background during adaptation suppresses two types of knowledge: class-agnostic foreground knowledge and class-discriminative semantic knowledge, leading to reduced detection performance for novel classes. To address this issue, we propose KR-OVD (Knowledge Recovery for Open-Vocabulary Detection), which provides two distinct recovery pathways. One pathway uses class-aware pseudo labels to directly supply the missing classification and localization supervision for novel classes. The other pathway introduces an Inter-Class Attention Path Aggregation Network (ICA-PAN) that models inter-class relationships to enable base-to-novel semantic transfer, indirectly recovering impaired novel-class knowledge. Experiments on OV-COCO and VisDrone2019-DET demonstrate that KR-OVD recovers novel-class recognition impaired during adaptation while preserving base-class performance.

1 INTRODUCTION

Traditional object detection operates under a closed-set assumption, where detectors are trained and evaluated on predefined category vocabularies, such as the 80 categories in MS-COCO and the 365 categories in Objects365 (Lin et al., 2014; Shao et al., 2019). Object detection architectures broadly fall into three paradigms: region-based detectors such as Faster R-CNN generate proposals before RoI-wise classification and box regression (Ren et al., 2017); dense one-stage detectors such as YOLO predict categories and boxes at spatial locations (Redmon et al., 2016); and query-based detectors such as DETR cast detection as Transformer-based set prediction (Carion et al., 2020). Despite their different designs, all remain constrained by the training vocabulary and cannot recognize categories outside it.

Open-vocabulary object detection (OVD) relaxes this constraint by exploiting the open-set semantics learned by large-scale vision–language models (VLMs) such as CLIP and ALIGN (Radford et al., 2021; Jia et al., 2021). By mapping category text and image or region representations into a shared semantic space, pretrained OVD detectors can recognize categories beyond the annotated vocabulary of a target dataset. Representative methods, from ViLD and RegionCLIP to DK-DETR and DeCo-DETR, are commonly evaluated in a generalized OVD setting in which detector training uses ground truth annotations for base categories, whereas inference covers both base and novel categories (Gu et al., 2022; Zhong et al., 2022; Li et al., 2023; Wang et al., 2026).

However, adapting pretrained detectors on base categories alone improves base-category detection while degrading novel-category recognition. NoOVD (Zhang et al., 2026) addresses background suppression through novel category discovery, self-distillation, and proposal confidence adjustment. These efforts mitigate novel class degradation, but a more detailed account of the information suppressed during adaptation remains lacking. This raises the question of what kind of pretrained knowledge is impaired by treating novel instances as background and how it can be recovered during adaptation.

Figure 1 illustrates the novel-category degradation caused by adaptation for both YOLO-World and YOLOE on OV-COCO and VisDrone2019-DET. On OV-COCO, novel AP_{50} drops from 68.0 to 25.7

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.

for YOLO-World and from 65.6 to 5.4 for YOLOE. On VisDrone2019-DET, the drops are from 15.1 to 13.6 and from 13.8 to 3.2, respectively. These results reveal a training–inference mismatch: novel instances are treated as background during adaptation but are expected to be recognized as foreground at inference. This background treatment suppresses class-agnostic foreground knowledge for identifying objects and class-discriminative semantic knowledge for recognizing their categories.

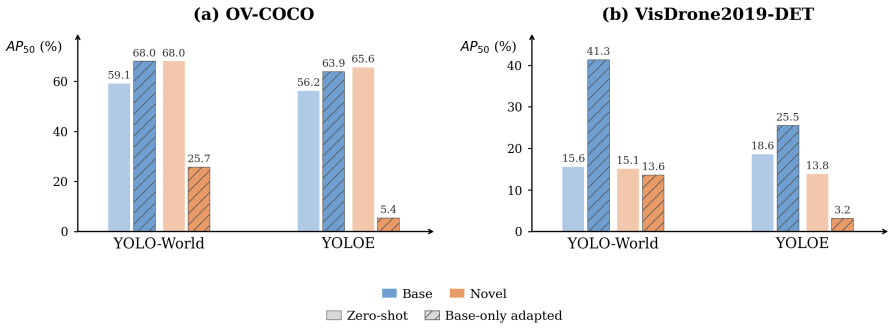


Figure 1: Effect of base-only adaptation on YOLO-World and YOLOE on (a) OV-COCO and (b) VisDrone2019-DET. Blue and orange bars report AP_{50}^{B} and AP_{50}^{N}, respectively. Unhatched bars denote pretrained zero-shot models, and hatched bars denote models after base-only adaptation. Across the two datasets, adaptation improves base class performance while reducing novel class performance for both detectors.

Motivated by this observation, we introduce KR-OVD to recover novel-class knowledge during base-only adaptation through two distinct pathways. The direct pathway uses class-specific pseudo labels from YOLO-World to supply missing classification and localization supervision for novel classes. The indirect pathway uses an Inter-Class Attention Path Aggregation Network (ICA-PAN) to model inter-class relationships and integrate category context into visual features, promoting semantic transfer from base to novel classes. Moreover, this framework focuses not on escalating methodological complexity, but on two distinct recovery mechanisms: explicit supervision supplementation and inter-class semantic transfer.

Our contributions are summarized as follows:

- We characterize the associated loss of class-agnostic foreground knowledge and class-discriminative semantic knowledge.
- We propose KR-OVD, which recovers novel-class knowledge impaired during adaptation through direct class-aware supervision supplementation and indirect semantic transfer based on inter-class relations.
- Experiments on OV-COCO and VisDrone2019-DET show that KR-OVD recovers novel class capability under base-only adaptation while preserving base class performance across distinct visual domains.

2 Related Work

2.1 Open-Vocabulary Object Detection

Open-vocabulary object detection (OVD) extends conventional detectors beyond a fixed category vocabulary by transferring semantic knowledge encoded in pretrained vision–language models. ViLD and RegionCLIP adapt this knowledge to region-level recognition through distillation or region–text alignment (Gu et al., 2022; Zhong et al., 2022), while Detic and PB-OVD broaden supervision through image-level labels or localized supervision derived from image–caption pairs (Zhou et al., 2022; Gao et al., 2022). OWL-ViT adapts an image–text-pretrained Vision Transformer to detection, whereas F-VLM retains a frozen VLM and trains only its detection heads (Minderer et al., 2022; Kuo et al., 2023). GLIP and Grounding DINO are pretrained with large-scale region–text grounding data (Li et al., 2022; Liu et al., 2023). More recently, YOLO-World introduced a lightweight architecture for real-time open-vocabulary detection (Cheng et al., 2024). We build KR-OVD on

YOLO-World rather than YOLOE (Wang et al., 2025) because its simple and extensible design facilitates the implementation and analysis of the two knowledge-recovery pathways.

NoOVD (Zhang et al., 2026) uses novel category discovery and self-distillation to mitigate the alignment of unlabeled novel features with background. At inference, it adjusts proposal confidence scores to improve novel proposal retention and recall. Our work instead investigates what knowledge is suppressed during adaptation and how it can be recovered through direct pseudo-supervision and indirect semantic transfer.

2.2 Pseudo-Labeling in OVD

In open-vocabulary object detection (OVD), ground truth annotations for novel classes are not used during training. To provide the missing supervision, models can generate pseudo labels which consist of predicted bounding boxes and category labels. The pseudo labels can provide supervision for unlabeled objects without using ground truth annotations for novel classes. Existing studies mainly focus on improving the quality and stability of pseudo-supervision. VL-PLM combines class-agnostic proposals with VLM semantics to generate category-aware pseudo labels (Zhao et al., 2022). SAS-Det stabilizes self-training through a split-and-fusion design with periodic teacher updates (Zhao et al., 2024), while LP-OVOD selects text-relevant proposals as pseudo annotations for novel class learning (Pham et al., 2024). MSPL further refines pseudo-label generation through multi-step localization, recognition, and background grounding (Choi et al., 2026), and CastDet extends this paradigm to aerial OVD using a RemoteCLIP teacher and a dynamic label queue (Li et al., 2025). Overall, these methods primarily improve the generation, selection, and optimization of pseudo labels, and thus recover novel class information through explicit pseudo-supervision.

In our work, pseudo-labeling provides a direct way to transfer pretrained vision–language knowledge into novel class supervision. Our analysis focuses on the types of knowledge recovered and the recovery pathways rather than the complexity of pseudo-label generation. Therefore, we adopt a simple strategy in which a frozen YOLO-World teacher directly generates pseudo labels. This setup allows us to compare class-agnostic and class-specific pseudo-supervision so that we can examine their respective contributions to knowledge recovery during adaptation.

3 Method

3.1 Overall Framework

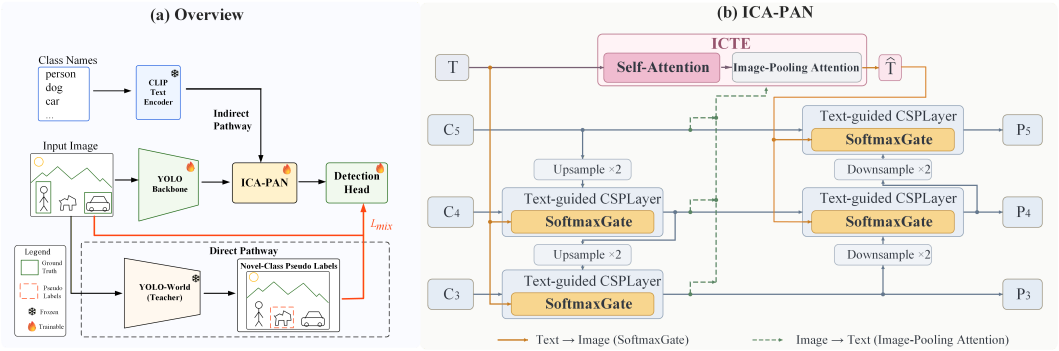


Figure 2: The proposed KR-OVD. **(a) Overall framework:** direct recovery through novel-class pseudo-supervision and indirect recovery through ICA-PAN. **(b) Structure of ICA-PAN:** ICTE enriches category text features with semantic and visual context, while SoftmaxGate injects category context during multi-scale visual fusion. Visual inputs C_3, C_4, and C_5 are processed through top-down and bottom-up paths to produce output features P_3, P_4, and P_5.

During adaptation of a pretrained OVD detector, only base classes $\mathcal{C}_{base}$ receive ground-truth box and category annotations, while novel classes $\mathcal{C}_{novel}$ remain unannotated, though their category names are available as text prompts. Treating unlabeled novel instances as background introduces

a training–inference mismatch, which suppresses class-agnostic foreground knowledge and class-discriminative semantic knowledge and further degrades novel-class detection performance.

To recover such suppressed novel-class knowledge under base-only adaptation, we present KR-OVD equipped with two complementary recovery pathways. One pathway directly introduces novel-class knowledge by leveraging novel-class pseudo labels to provide explicit supervision. The other pathway exploits the shared semantic knowledge potentially existing between base and novel classes to achieve semantic transfer from base to novel categories, which is instantiated by our proposed ICA-PAN. Specifically, ICA-PAN performs categorical semantic modeling and feature injection in two successive steps: it first enhances category text representations with inter-class relationships and image context via the Inter-Class Text Enhancer (ICTE), and then adaptively injects category semantics into multi-scale visual features through SoftmaxGate. Jointly, these two pathways promote semantic transfer from base to novel classes and restore the impaired novel-class capability. Figure 2 illustrates (a) the overall framework of KR-OVD and (b) the architecture of ICA-PAN.

3.2 Direct Recovery via Pseudo-Supervision

To supply the missing supervision for novel instances, the direct pathway uses class-aware pseudo labels generated by a frozen YOLO-World teacher. These labels provide classification and localization targets for novel instances, while ground truth annotations supervise base instances.

Mixed supervision construction For each training image I_i, the frozen teacher f_T is prompted with $\mathcal{C}_{novel}$ to produce box–category predictions and confidence scores:

$$f_T(I_i, \mathcal{C}_{novel}) = \{(\hat{B}_{ij}, \hat{t}_{ij}, s_{ij})\}_{j=1}^{M_i}. \tag{1}$$

Here, M_i is the number of novel class predictions before confidence filtering. For prediction j, $\hat{B}_{ij}$ denotes the bounding box, $\hat{t}_{ij} \in \mathcal{C}_{novel}$ the category text, and s_{ij} the confidence score. We retain predictions with confidence scores of at least 0.20 in the main experiments:

$$\widehat{\mathcal{Y}}_i^N = \{(\hat{B}_{ij}, \hat{t}_{ij}) \mid 1 \le j \le M_i,\ s_{ij} \ge 0.20\}. \tag{2}$$

The retained novel pseudo labels $\widehat{\mathcal{Y}}_i^N$ are combined with the base ground truth $\mathcal{Y}_i^B$ to construct the mixed supervision:

$$\mathcal{Y}_i^{mix} = \mathcal{Y}_i^B \cup \widehat{\mathcal{Y}}_i^N, \qquad \mathcal{D}_{mix} = \{(I_i, \mathcal{Y}_i^{mix})\}_i. \tag{3}$$

Here, i indexes all target training images, and $\mathcal{D}_{mix}$ is the resulting training dataset.

Detection objective We retain the original YOLO-World training objective. Base ground truth and novel pseudo labels are processed by the same Batch Task-Aligned Assigner and optimized using the same detection losses. The classification term $\mathcal{L}_{cls}$ applies sigmoid cross-entropy to the similarity-based category predictions and assigned targets. The localization terms consist of Complete-IoU regression loss $\mathcal{L}_{box}$ and distribution focal loss $\mathcal{L}_{dfl}$. The overall objective is

$$\mathcal{L}_{\text{mix}} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{box}\mathcal{L}_{box} + \lambda_{dfl}\mathcal{L}_{dfl}, \tag{4}$$

where the loss weights $(\lambda_{cls}, \lambda_{box}, \lambda_{dfl}) = (0.5, 7.5, 0.375)$ follow the YOLO-World configuration.

3.3 Indirect Recovery via ICA-PAN

We build KR-OVD upon YOLO-World, which comprises a visual backbone, a text encoder, RepVL-PAN and a detection head. RepVL-PAN fuses image and text embeddings via top-down and bottom-up flows, adopting max-sigmoid attention for text-driven visual modulation and Image-Pooling Attention for image-aware text refinement. Nevertheless, RepVL-PAN has two key drawbacks: it performs no explicit inter-class modeling, and category embeddings are not directly aggregated into visual features. Max-sigmoid attention merely scales raw visual features using single-channel spatial gates derived from image–text similarities—text semantics are never encoded into visual representations, and the max-response operation discards multi-category context. Meanwhile, Image-Pooling Attention enriches text embeddings with visual context but does not enable information exchange across categories. These deficits motivate explicit inter-class modeling and multi-category semantic injection for base-to-novel knowledge transfer.

To tackle these issues, we propose ICA-PAN for categorical semantic modeling and feature injection in two sequential stages, which enhances category text features through the Inter-Class Text Enhancer (ICTE) and then injects them into multi-scale visual features through SoftmaxGate. This facilitates base-to-novel semantic transfer and recovers novel-class knowledge suppressed under base-only adaptation.

ICTE: Inter-Class Text Enhancer The frozen CLIP text encoder maps the input category names into $T \in \mathbb{R}^{N\times d}$, where N is the number of input categories and $d = 512$ is the feature dimension. Although these embeddings capture semantic relationships, they lack explicit information exchange across the current category set. To enable this exchange, ICTE applies single-head self-attention (Vaswani et al., 2017), allowing each category to aggregate semantic context from other input categories. The resulting text features U are computed as

$$U = T + \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V. \tag{5}$$

Here, $Q, K, V \in \mathbb{R}^{N\times d}$ are the query, key, and value matrices obtained from T through three independent affine projections. Although U captures inter-class semantic context, it contains no visual information. To condition U on the current image, we apply the Image-Pooling Attention inherited from YOLO-World. Each top-down image feature map is projected by a 1×1 convolution and adaptively max-pooled to 3×3 spatial locations. The pooled features from three scales are flattened and concatenated into 27 image tokens, denoted by F. The enhanced text features $\hat{T}$ are obtained through cross-attention with a residual connection:

$$\hat{T} = U + \mathrm{MHA}(U, F, F). \tag{6}$$

Here, MHA denotes eight-head cross-attention with queries derived from U and keys and values derived from F. The enhanced features $\hat{T}$ guide visual fusion in the bottom-up SoftmaxGate layers.

SoftmaxGate In YOLO-World, max-sigmoid attention modulates visual features without aggregating semantic contributions across categories. To address this limitation, SoftmaxGate gathers category information into a semantic context at each spatial position and injects it into the visual features through a learned gate. The updated visual features are then scaled and shifted using a global summary of the input categories, combining spatial context injection with global channel modulation.

SoftmaxGate processes the output of the final bottleneck branch in each Text-guided CSPLayer. Its output is concatenated with the other CSP branches and fused by the output convolution. Let X and T_g denote the visual and text features entering SoftmaxGate, respectively. The text input T_g is the original text features T in the top-down path and the enhanced text features $\hat{T}$ in the bottom-up path. The computation proceeds in three steps:

(i) Soft category aggregation. We first weight category embeddings by their similarity to the visual features and aggregate them into semantic context at each spatial position. For one attention head, $E(X)$ denotes the visual embedding with spatial dimensions flattened, and $H_g = P_g(T_g)$ denotes the text embedding produced by an affine projection P_g. The projected text features serve as both keys and values. The category attention weights A and semantic context C are computed as

$$A = \mathrm{softmax}\left(\frac{E(X)H_g^\top}{\max(|\tau|, \epsilon)}\right), \qquad C = AH_g. \tag{7}$$

Softmax normalizes the attention weights across categories at each spatial position. When the visual and embedding channel dimensions differ, E uses a 1×1 convolution followed by normalization to align them. The learnable temperature τ is initialized to 0.1, with $\epsilon = 10^{-4}$ providing a numerical lower bound. The outputs of all heads are concatenated along channels and reshaped into the spatial context map C.

(ii) Gated residual injection. Given the aggregated context C, we next determine how strongly it contributes to the visual features at each position. A gate computed from both X and C controls the amount of aligned context $\phi(C)$ added through the residual connection:

$$G = \sigma(W_g * [X; C] + b_g), \qquad X' = X + G \odot \phi(C). \tag{8}$$

Here, G is the gate and X' is the updated visual feature map. The sigmoid function σ controls residual injection, whereas the preceding softmax determines the aggregation weights across categories. W_g and b_g are the weights and bias of the 1×1 gate convolution, and $*$ denotes convolution. $[X; C]$ denotes channel concatenation, and $\odot$ denotes element-wise multiplication. The channel-alignment map ϕ applies a 1×1 convolution to align the channel dimension of C with that of X when they differ. The gate bias is initialized to -2 to encourage small initial gating values.

(iii) Global channel modulation. After injecting spatial semantic context, we use the overall category semantics to modulate the channels of X' through feature-wise linear modulation (FiLM) (Perez et al., 2018). We average T_g over categories and pass the resulting global summary through two independent affine maps to obtain the scale increment γ and shift β. These parameters apply the same channel transformation across spatial positions:

$$Z = (1 + \gamma) \odot X' + \beta, \qquad Y = P_{\text{out}}(Z). \tag{9}$$

Here, γ and β are broadcast over spatial positions, Z denotes the modulated visual features, and Y denotes the projected output. Both affine maps have zero-initialized weights and biases, making FiLM initially an identity mapping. The output projection P_{out} consists of a 3×3 convolution followed by batch normalization.

4 Experiments

4.1 Experimental Setup

Datasets We conduct experiments on OV-COCO and VisDrone2019-DET. OV-COCO contains 48 base and 17 novel classes following the split of the COCO OVD benchmark (Bansal et al., 2018). To further evaluate generalization across visual scenarios, we use VisDrone2019-DET (Du et al., 2019) as an additional test domain. This dataset differs significantly from OV-COCO in aerial viewpoint, object scale, and scene distribution. Under the open-vocabulary detection setting, VisDrone2019-DET is divided into 7 base classes and 3 novel classes to evaluate novel class detection under cross-scenario conditions. The complete category splits are listed in Appendix A.1.

Evaluation Metrics We report Average Precision at an IoU threshold of 0.50 for base and novel categories, denoted as AP_{50}^{B} and AP_{50}^{N}, with all tabulated results expressed as percentages. We use AP_{50}^{N} to measure novel-capability preservation and recovery, while AP_{50}^{B} monitors base class adaptation performance. Text prompts for both base and novel categories are used simultaneously at inference under the standard OVD protocol.

Implementation Details We adopt YOLO-World-L v2.1 with a YOLOv8-L backbone and build KR-OVD upon it. All experiments are conducted on a single NVIDIA RTX 4090 for 80 epochs with a batch size of 8 and an initial learning rate of 1×10^{-5}. Except for the post-hoc threshold analysis in Appendix A.2, the confidence threshold for class-aware pseudo labels is fixed at 0.20 without dataset-specific tuning. In the training-data descriptions, O365 denotes Objects365v1 (Shao et al., 2019), and CC-Lite denotes CC-LiteV2 (Sharma et al., 2018).

4.2 Main Results

Table 1: Pretrained capability, base-only target adaptation, and knowledge recovery.

	OV-COCO		VisDrone2019-DET	
Model state	AP_{50}^{B}	AP_{50}^{N}	AP_{50}^{B}	AP_{50}^{N}
YOLO-World-L v2.1 (pretrained, zero-shot)	59.1	68.0	15.6	15.1
YOLO-World-L v2.1 (base-only adapted)	68.0	25.7	41.3	13.6
KR-OVD	66.5	50.7	40.2	17.9

Adaptation-induced degradation and recovery On OV-COCO, base-only adaptation raises AP_{50}^{B} from 59.1 to 68.0 while reducing AP_{50}^{N} from 68.0 to 25.7. KR-OVD restores novel AP to 50.7 and retains 66.5 base AP. On VisDrone2019-DET, the pretrained model achieves 15.6 base AP and 15.1 novel AP under the aerial-domain shift. After base-only adaptation, base AP increases to 41.3 and novel AP decreases to 13.6. KR-OVD achieves 40.2 base AP and 17.9 novel AP, with novel performance exceeding both the pretrained and base-only adapted models.

Comparison with existing OVD methods on OV-COCO Table 2 groups existing methods according to whether pseudo annotations are used for training. KR-OVD achieves the highest novel-class performance among the compared methods, reaching 50.7 AP_{50}^{N} and surpassing MSPL and NoOVD by 2.9 and 3.2 points, respectively. It also achieves 66.5 AP_{50}^{B}, exceeding both methods in base-class performance. Compared with the base-only adapted YOLO-World, KR-OVD improves novel AP by 25.0 points, with a 1.5-point decrease in base AP, demonstrating substantial recovery of novel-class capability while retaining competitive base-class performance. The compared methods differ in their backbones and extra training data. A dash in the Extra Training Data column denotes no additional training data.

Table 2: Comparison with existing methods on OV-COCO.

Method	Backbone	Extra Training Data	AP_{50}^{B}	AP_{50}^{N}
ViLD (Gu et al., 2022)	RN50	–	59.5	27.6
RegionCLIP (Zhong et al., 2022)	RN50	CC3M	57.1	31.4
Detic (Zhou et al., 2022)	RN50	COCO Caption	47.1	27.8
PB-OVD (Gao et al., 2022)	RN50	COCO Caption	46.1	30.8
F-VLM (Kuo et al., 2023)	RN50	LVIS-base	–	28.0
BARON (Wu et al., 2023a)	RN50	–	60.4	34.0
CORA (Wu et al., 2023b)	RN50	–	35.5	35.1
BIND (Zhang et al., 2024)	ViT-B/16	–	54.7	36.3
CLIPSelf (Wu et al., 2024)	ViT-B/16	–	–	37.6
LBP (Li et al., 2024)	RN50	–	58.7	37.8
DK-DETR (Li et al., 2023)	RN50	–	61.1	32.3
DeCo-DETR (Wang et al., 2026)	RN50	–	56.7	41.3
NoOVD (Zhang et al., 2026)	ViT-L/14	LVIS-base	65.8	47.5
YOLOE (Wang et al., 2025)	YOLOv8-L	O365+GoldG	63.9	5.4
YOLO-World	YOLOv8-L	O365+GoldG+CC-Lite	68.0	25.7
VL-PLM (Zhao et al., 2022)	RN50	Pseudo annotations	60.2	34.4
SAS-Det (Zhao et al., 2024)	RN50-C4	Pseudo annotations	58.5	37.4
LP-OVOD (Pham et al., 2024)	RN50	Pseudo annotations	60.5	40.5
MSPL (Choi et al., 2026)	RN50×4	Pseudo annotations	60.9	47.8
KR-OVD	YOLOv8-L	Pseudo annotations	66.5	50.7

Table 3: Comparison with existing methods on VisDrone2019-DET.

Method	Backbone	Extra Training Data	AP_{50}^{B}	AP_{50}^{N}
Grounding DINO (Liu et al., 2023)	Swin-T	O365+GoldG+Cap4M	59.3	6.7
YOLOE (Wang et al., 2025)	YOLOv8-L	O365+GoldG	25.5	3.2
CastDet (Li et al., 2025)	RN50-C4	Pseudo annotations	35.1	0.0
YOLO-World	YOLOv8-L	O365+GoldG+CC-Lite	41.3	13.6
KR-OVD	YOLOv8-L	Pseudo annotations	40.2	17.9

Comparison with existing OVD methods on VisDrone2019-DET Table 3 evaluates knowledge recovery under the challenging aerial-domain shift of VisDrone2019-DET. KR-OVD achieves the highest novel-class performance among the compared methods, obtaining 17.9 AP_{50}^{N}. It outperforms the strongest novel-class baseline, YOLO-World, by 4.3 points, while base AP decreases by 1.1 points to 40.2. Grounding DINO achieves higher base AP but substantially lower novel AP (6.7),

highlighting the distinction between base-class adaptation and novel-class generalization. These results demonstrate the effectiveness of KR-OVD in recovering novel-class capability under aerial-domain shift. For YOLOE and Grounding DINO, we use the same training settings as KR-OVD, with 80 epochs and an initial learning rate of 1×10^{-5}. CastDet is trained using its original settings.

4.3 Analysis of Pseudo-Supervision Types

Under base-only adaptation, we compare training without pseudo-supervision, class-agnostic pseudo labels that assign each pseudo box a generic *object* label, and class-specific pseudo labels that associate each box with an explicit category. Table 4 reports OV-COCO detection performance, while Figure 3 tracks novel class Recall@0.5 during adaptation.

The base-only adapted model without pseudo-supervision obtains 25.7 AP_{50}^{N}. Class-agnostic pseudo labels raise this value to 40.3 while changing AP_{50}^{B} only from 68.0 to 67.9, consistent with recovering missing foreground knowledge. Class-specific pseudo labels contain both boxes and category identities, further raising AP_{50}^{N} to 50.5 while retaining 67.5 AP_{50}^{B}. The direct pathway thus recovers both foreground and class-discriminative knowledge through explicit box–category supervision.

Table 4: Ablation of different pseudo-supervision types on OV-COCO.

Pseudo-supervision	AP_{50}^{B}	AP_{50}^{N}
w/o pseudo	68.0	25.7
+ class-agnostic pseudo	67.9	40.3
+ class-specific pseudo	67.5	50.5

Figure 3 shows how the three strategies affect novel class recall during adaptation. Recall is evaluated every five epochs, with the horizontal axis labeled at ten-epoch intervals. Without pseudo-supervision, Recall@0.5 decreases from approximately 72% at epoch 5 to about 15% at the end of training, consistent with progressive suppression of previously encoded novel capability. Class-agnostic pseudo labels slow this degradation, with recall decreasing from approximately 81% to 59%, while class-specific pseudo labels maintain recall near 81%. These dynamics support the distinction between recovering box-level foreground information and additionally restoring explicit category-discriminative supervision.

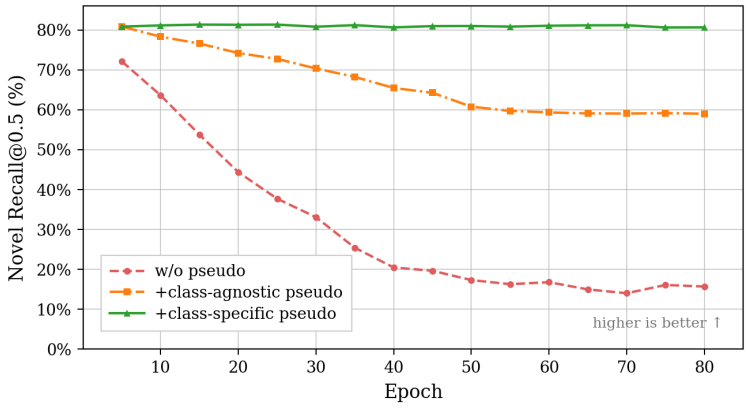


Figure 3: Novel class Recall@0.5 on OV-COCO throughout training under different pseudo-supervision strategies.

4.4 Ablation of Direct and Indirect Recovery Pathways

Table 5 presents an ablation of the direct class-specific pseudo-supervision pathway and the indirect ICA-PAN pathway on OV-COCO and VisDrone2019-DET.

OV-COCO. Both recovery pathways can operate individually on this benchmark. Class-specific pseudo-supervision improves AP_{50}^{N} from 25.7 to 50.5 ($+24.8$), while ICA-PAN alone raises it to 43.9 ($+18.2$) without explicit novel class pseudo-supervision. Combining them reaches 50.7. The non-additive gain indicates that the two mechanisms recover partially overlapping novel-class knowledge.

VisDrone2019-DET. Direct pseudo-supervision improves AP_{50}^{N} from 13.6 to 16.1, although AP_{50}^{B} decreases from 41.3 to 35.7. ICA-PAN alone yields only 4.7 novel AP. This result suggests that indirect semantic transfer alone is insufficient under the severe aerial-domain shift. When explicit novel class anchors are available through direct pseudo-supervision, adding ICA-PAN raises novel AP from 16.1 to 17.9 and base AP from 35.7 to 40.2, indicating that indirect transfer complements direct supervision when the domain shift is severe.

Taken together, the two recovery pathways exhibit scenario-dependent effects: both are effective on OV-COCO, whereas on VisDrone2019-DET the indirect pathway provides additional benefit mainly when combined with direct pseudo-supervision.

Table 5: Ablation of direct and indirect knowledge-recovery pathways on OV-COCO and VisDrone2019-DET.

		OV-COCO		VisDrone2019-DET	
Class-specific pseudo	ICA-PAN	AP_{50}^{B}	AP_{50}^{N}	AP_{50}^{B}	AP_{50}^{N}
		68.0	25.7	41.3	13.6
✓		67.5	50.5	35.7	16.1
	✓	68.2	43.9	40.1	4.7
✓	✓	66.5	50.7	40.2	17.9

4.5 ICA-PAN Component Analysis

Table 6 presents a hierarchical ablation on OV-COCO: the original YOLO-World, YOLO-World with SoftmaxGate, and the complete ICA-PAN with both SoftmaxGate and ICTE.

Table 6: Component analysis of ICA-PAN.

SoftmaxGate	ICTE	AP_{50}^{B}	AP_{50}^{N}
		68.0	25.7
✓		67.7	43.2
✓	✓	68.2	43.9

The original YOLO-World obtains 25.7 AP_{50}^{N} and 68.0 AP_{50}^{B}. SoftmaxGate raises novel AP from 25.7 to 43.2 with a slight base AP decrease from 68.0 to 67.7. Adding ICTE further improves novel and base AP by 0.7 and 0.5 points, respectively, supporting the benefit of inter-class relation modeling.

5 Conclusion

We present KR-OVD, a knowledge-recovery framework for open-vocabulary object detection that addresses adaptation-induced degradation of pretrained novel class capability. We recover the suppressed foreground and class-discriminative semantic knowledge through two pathways: class-aware pseudo-supervision and ICA-PAN-based inter-class semantic transfer. KR-OVD achieves 50.7 and 17.9 AP_{50}^{N} on OV-COCO and VisDrone2019-DET, respectively, while maintaining base class performance. These results establish knowledge recovery as an effective perspective for preserving novel class capability during base-only adaptation. We hope our findings encourage evaluation protocols that report both pretrained and adapted open-vocabulary performance.

AI USE STATEMENT

In this work, we used generative AI tools for polishing writing. We have not used generative AI tools for methodology development, experiment design, mathematical claims, implementation, and result interpretation. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 384–400, 2018.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pp. 213–229, 2020.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-time open-vocabulary object detection. *arXiv preprint arXiv:2401.17270*, 2024.

Hojun Choi, Youngsun Lim, Jaeyo Shin, and Hyunjung Shim. MSPL: Multi-step pseudo-labeling for open-vocabulary object detection, 2026. URL https://arxiv.org/abs/2510.14792v4.

Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 213–226, 2019. doi: 10.1109/ICCVW.2019.00030.

Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *Computer Vision–ECCV 2022*, volume 13670 of *Lecture Notes in Computer Science*, pp. 266–282. Springer, 2022. doi: 10.1007/978-3-031-20080-9_16.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021.

Weicheng Kuo, Yin Cui, Xiuye Gu, A. J. Piergiovanni, and Anelia Angelova. F-VLM: Open-vocabulary object detection upon frozen vision and language models. In *International Conference on Learning Representations*, 2023.

Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16678–16687, 2024.

Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, and Shiliang Pu. Distilling DETR with visual-linguistic knowledge for open-vocabulary object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6501–6510, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yan Li, Weiwei Guo, Xue Yang, Ning Liao, Dunyun He, Jiaqi Zhou, and Wenxian Yu. Toward open vocabulary aerial object detection with CLIP-activated student-teacher learning. In *Computer Vision–ECCV 2024*, volume 15144 of *Lecture Notes in Computer Science*, pp. 431–448. Springer, 2025. doi: 10.1007/978-3-031-73016-0_25.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pp. 740–755, 2014.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *Proceedings of the European Conference on Computer Vision*, 2022.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Chau Pham, Truong Vu, and Khoi Nguyen. LP-OVOD: Open-vocabulary object detection by linear probing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 779–788, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8429–8438, 2019.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2556–2565, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOE: Real-time seeing anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24591–24602, 2025.

Siheng Wang, Yanshu Li, Bohan Hu, Zhengdao Li, Haibo Zhan, Linshan Li, Weiming Liu, Ruizhi Qian, Guangxin Wu, Hao Zhang, Jifeng Shen, Piotr Koniusz, Zhengtao Yao, Junhao Dong, and Qiang Sun. DeCo-DETR: Decoupled cognition DETR for efficient open-vocabulary object detection. In *International Conference on Learning Representations*, 2026.

Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15254–15264, 2023a.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In *International Conference on Learning Representations*, 2024.

Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. CORA: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7031–7040, 2023b.

Heng Zhang, Qiuyu Zhao, Linyu Zheng, Hao Zeng, Zhiwei Ge, Tianhao Li, and Sulong Xu. Exploring region-word alignment in built-in detector for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16975–16984, 2024.

Yupeng Zhang, Ruize Han, Zhiwei Chen, Wei Feng, and Liang Wan. NoOVD: Novel category discovery and embedding for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6304–6313, June 2026.

Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, B. G. Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N. Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *Computer Vision–ECCV 2022*, volume 13669 of *Lecture Notes in Computer Science*, pp. 159–175. Springer, 2022. doi: 10.1007/978-3-031-20077-9_10.

Shiyu Zhao, Samuel Schulter, Long Zhao, Zhixing Zhang, B. G. Vijay Kumar, Yumin Suh, Manmohan Chandraker, and Dimitris N. Metaxas. Taming self-training for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13938–13947, 2024. doi: 10.1109/CVPR52733.2024.01322.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Proceedings of the European Conference on Computer Vision*, pp. 350–368, 2022.

A APPENDIX

A.1 CATEGORY SPLITS

A.1.1 OV-COCO CATEGORY SPLIT

OV-COCO uses 48 base classes: person, bicycle, car, motorcycle, train, truck, boat, bench, bird, horse, sheep, bear, zebra, giraffe, backpack, handbag, suitcase, frisbee, skis, kite, surfboard, bottle, fork, spoon, bowl, banana, apple, sandwich, orange, broccoli, carrot, pizza, donut, chair, bed, toilet, tv, laptop, mouse, remote, microwave, oven, toaster, refrigerator, book, clock, vase, toothbrush.

The 17 novel classes are: airplane, bus, cat, dog, cow, elephant, umbrella, tie, snowboard, skateboard, cup, knife, cake, couch, keyboard, sink, scissors.

A.1.2 VISDRONE2019-DET CATEGORY SPLIT

VisDrone2019-DET uses 7 base classes: pedestrian, people, bicycle, car, truck, tricycle, motor.

The 3 novel classes are: van, bus, awning-tricycle.

A.2 PSEUDO-LABEL CONFIDENCE THRESHOLD

We evaluate the trade-off between supervision noise and sample coverage under different confidence thresholds on VisDrone2019-DET.

Table 7: Results with different pseudo-label confidence thresholds.

Confidence threshold	Novel pseudo labels	AP_{50}^{B}	AP_{50}^{N}
0.15	31,524	40.2	17.2
0.20	21,181	40.2	17.9
0.25	14,965	40.3	18.7
0.30	10,924	40.7	18.8

We use a fixed confidence threshold of 0.20 in the main experiments without dataset-specific tuning. This post-hoc sensitivity analysis on VisDrone2019-DET shows that KR-OVD remains effective across the evaluated range, while higher-confidence pseudo labels yield modestly better base and novel AP despite substantially fewer pseudo annotations.