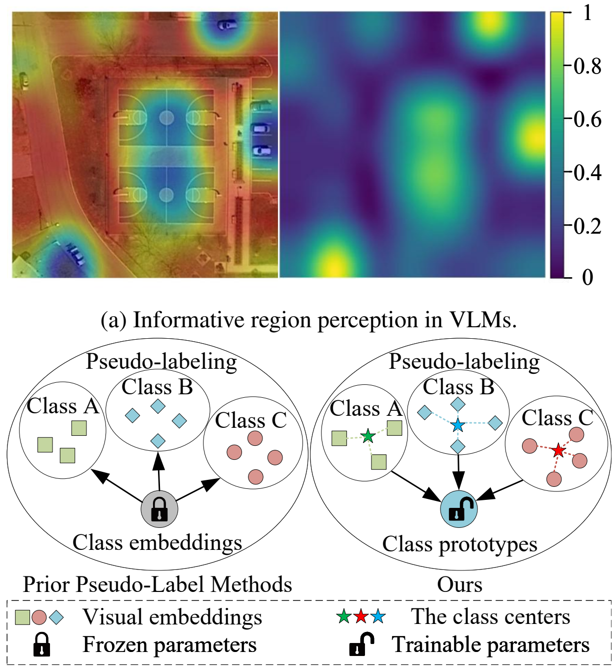
VK-Det: Visual Knowledge Guided Prototype Learning for Open-Vocabulary Aerial Object Detection

Jianhang Yao[1], **Yongbin Zheng**[1*], **Siqi Lu**[1], **Wanying Xu**[1], **Peng Sun**[1]
[1]National University of Defense Technology
yaojianhang0422@163.com, yaojianhang23@nudt.edu.cn, zybnudt@nudt.edu.cn,
lusiqi9803@163.com, wanying_xu@nudt.edu.cn, sunpeng@nudt.edu.cn

Abstract

To identify objects beyond predefined categories, open-vocabulary aerial object detection (OVAD) leverages the zero-shot capabilities of visual-language models (VLMs) to generalize from base to novel categories. Existing approaches typically utilize self-learning mechanisms with weak text supervision to generate region-level pseudo-labels to align detectors with VLMs semantic spaces. However, text dependence induces semantic bias, restricting open-vocabulary expansion to text-specified concepts. We propose **VK-Det**, a **V**isual **K**nowledge-guided open-vocabulary object **Det**ection framework *without* extra supervision. First, we discover and leverage vision encoder's inherent informative region perception to attain fine-grained localization and adaptive distillation. Second, we introduce a novel prototype-aware pseudo-labeling strategy. It models inter-class decision boundaries through feature clustering and maps detection regions to latent categories via prototype matching. This enhances attention to novel objects while compensating for missing supervision. Extensive experiments show state-of-the-art performance, achieving 30.1 mAP^N on DIOR and 23.3 mAP^N on DOTA, outperforming even extra supervised methods.



(a) Informative region perception in VLMs.

(b) Comparison of our method with existing methods.

Figure 1: (a) visualizes the attention heatmap from the visual encoder of VLMs for an aerial image. The heatmap is derived by averaging multi-layer attentions. (b) compares our pseudo-labeling approach with conventional methods.

Introduction

Aerial object detection (AOD), which involves precisely localizing and classifying objects within aerial images, is essential for earth observation tasks including security monitoring, disaster response, and urban management (Ding et al. 2022). While deep learning has substantially improved AOD performance on closed-set benchmarks (Redmon et al. 2015), existing methods remain limited to detecting only predefined object categories. To facilitate real-world deployment where countless unlabeled concepts exist, open-vocabulary aerial object detection (OVAD) is introduced to enable the recognition of novel objects (Li et al. 2023b).

Utilizing VLMs' zero-shot capabilities (e.g., RemoteCLIP (Liu et al. 2023)), OVAD replaces trainable classifier weights with frozen semantic embeddings from the text encoder of VLMs. Current research on OVAD primarily focuses on knowledge distillation and pseudo-labeling:

- Knowledge distillation transfers region-level semantic knowledge from VLMs to detectors, with ViLD (Gu et al. 2021) pioneering the alignment between detector region features and cropped image embeddings from VLMs.
- Pseudo-labeling employs self-learning or external data to generate pseudo-labeled data, thereby expanding category coverage through extra supervision. For instance, CastDet (Li et al. 2023b) employs a semi-supervised paradigm to produce high-quality pseudo-labels.

However, in AOD scenarios, knowledge distillation and pseudo-labeling underperform due to challenges in novel object localization, background interference, and textual noise, which ultimately lead to region-text misalignment. This compels us to confront an inherent issue: **whether we can automatically discover novel conceptual objects**

*Corresponding author.

from visual knowledge while simultaneously maximizing their category boundaries for knowledge transfer.

The visualization of attention heatmaps from VLMs' visual encoders reveals a notable phenomenon. This is evidenced by Fig. 1(a), where average attention maps across layers differentiate background from informative regions and assign higher weights to the latter without labels. Based on this observation, we design an adaptive selection mechanism and data augmentation method for small and elongated objects in aerial imagery. These components form the Adaptive Selection Knowledge Distillation (ASKD), enabling more efficient knowledge transfer. To better leverage informative regions, we propose a Prototype-Aware Pseudo-Label (PAPL) method. As shown in Fig. 1(b), unlike prior pseudo-label methods that classify class-agnostic visual embeddings using frozen class embeddings (Zhao et al. 2022), our method separates inter-class disparity via prototype learning and then generate trainable unknown class prototypes for classifier training. Finally, we propose the Synthetic Matching Inference (SMI) mechanism to evaluate the scores of novel classes through prototype matching and multi-level scoring. This approach integrates the distillation and prototype classifiers into a unified unknown category classifier, which works in conjunction with the localization network to assess the relevance of detected objects.

By combining the aforementioned modules, we introduce a novel framework termed **VK-Det**, a **Det**ector that relies only on **V**isual **K**nowledge for efficient knowledge distillation and pseudo-labeling optimization *without additional supervised signals or data*. We argue that relying solely on visual knowledge from VLMs enables the detector to learn its semantic knowledge efficiently and achieve performance comparable to methods that utilize extra supervision.

We evaluate VK-Det on two benchmark datasets, DIOR(Li et al. 2019) and DOTA(Ding et al. 2022). On DIOR, VK-Det achieves 30.1% mAP on novel categories, outperforming state-of-the-art methods. Notably, previous methods rely on extra supervised signals, whereas VK-Det attains superior performance without extra supervision. On DOTA, VK-Det achieves a performance of 23.3% mAP^N and 33.9% HM. Furthermore, our proposed PAPL method demonstrates superior performance and effectively mitigates the text illusion problem compared to approaches based on category-supervised pseudo-label generation.

The **main contributions** of this paper are as follows:

- We identify the inherent capability of **informative regions perception** for potential objects in VLMs. Leveraging this insight, we design an ASKD framework to extract informative, region-level embeddings for more effective and adaptive distillation.
- We introduce a PAPL approach that leverages prototype learning to generate high-quality pseudo-labels. Furthermore, we design prototype-based classifier and matching inference strategies to facilitate knowledge transfer.
- By integrating both components, we propose the VK-Det framework. Experiments on two standard benchmarks show that our method achieves state-of-the-art performance, surpassing even extra supervised methods.

Related Work

Open-Vocabulary Object Detection. The core idea of open-vocabulary object detection is to use joint visual-language modeling by leveraging pre-trained VLMs or image-text pairs as weakly supervised data to train a detector. Two dominant paradigms enable transferring image-level semantics from VLMs to region features: knowledge distillation and pseudo-labeling. Knowledge distillation extracts region-level semantic knowledge from pre-trained VLMs to empower detectors (Gu et al. 2021; Wang et al. 2023a; Ma et al. 2022; Li et al. 2023a). Pseudo-labeling enhances the annotation quality of novel objects through additional supervision (Zhao et al. 2022; Pham, Vu, and Nguyen 2023). Furthermore, current research on OVAD remains limited. DescReg (Zang et al. 2024) leverages triplet loss to preserve visual similarity structures in the classification space and enhance knowledge transfer. CastDet (Li et al. 2023b) employs a semi-supervised model with pseudo-labeled sequences to expand the class vocabulary. LAE-DINO (Pan et al. 2024) uses additional data to train the detector by combining dynamic vocabulary construction with visually guided text prompts.

Unlike existing OVAD methods, our approach optimizes knowledge distillation for aerial imagery without extra data. To eliminate the "extra supervision" bias from pseudo labels, which are novel class labels or textual signals that appear during training, we propose a prototype-aware method. This method generates category prototypes and matches them dynamically, effectively suppressing noisy supervision and ensuring reliable semantic updates.

Further details on related work are in **Appendix A**.

Methods

Preliminaries $\&$ Overview

Preliminaries. OVAD trains models on labeled data for base categories C_B, allowing them to localize and classify objects from novel categories C_N during inference, where $C_B \cap C_N = \emptyset$. Its core principle involves implicitly learning semantic features of unlabeled objects in training images, enabling the alignment between visual and textual features in a unified embedding space during inference.

Currently, to address the open-world localization problem, pre-trained Region Proposal Network (RPN) (Ren et al. 2015) or Object Localization Network (OLN) (Kim et al. 2021) models typically serve as fundamental components for OVAD. For an input image, they generate class-agnostic object proposals P, which consist of three mutually exclusive subsets: foreground proposals P_{fg} contain base-class C_B objects; targeted proposals P_{tg} contain unknown-class C_U instances; and untargeted proposals P_{ug} contain background-class C_U^{BG} instances such as aerial imagery of woods or houses, with $P = P_{fg} \cup P_{tg} \cup P_{ug}$. Notably, $C_U \supset C_N$ and $C_B \cap C_U = \emptyset$.

Knowledge distillation has emerged as an effective strategy for transferring semantic knowledge from VLMs to detectors. For each proposal $p \in P$, knowledge distillation crops the region, encodes it through the visual encoder of VLMs to generate visual embeddings v, uses the detector

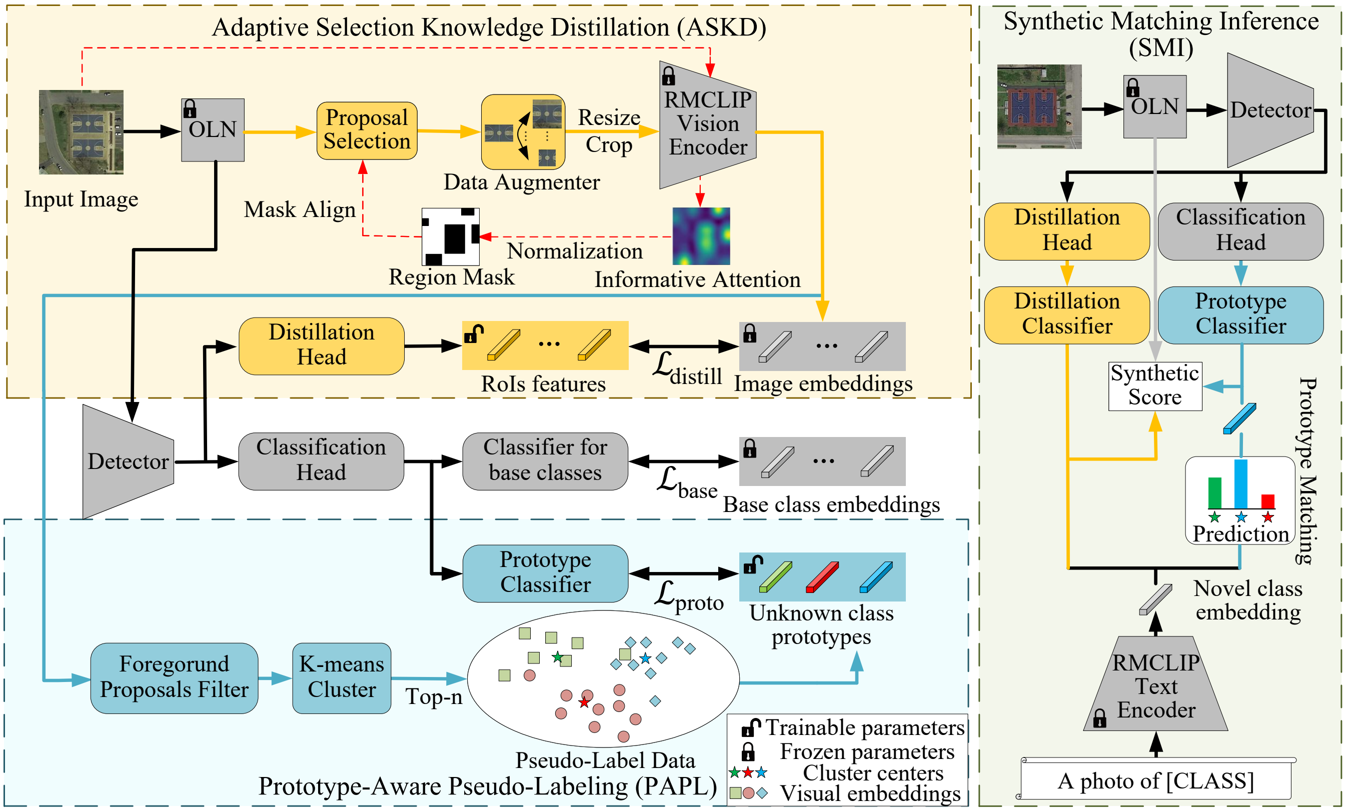


Figure 2: The overall architecture of our VK-Det. By utilizing ASKD and PAPL in training, the detector enables comprehensive learning of unlabeled objects. During inference, SMI systematically evaluate novel category objects. In the figure, RMCLIP denotes RemoteCLIP(Liu et al. 2023).

simultaneously with a Region of Interest (RoI) extractor to extract f_{roi}, and aligns these representations by minimizing their orthogonal similarity. In inference, by generating text embeddings for novel categories $\{t_c^N \mid c \in \mathcal{C}_N\}$ via VLMs' text encoders, the category of detection boxes is inferred by computing the similarity probability between f_{roi} of proposals and t_c, expressed as:

$$P(f_{roi}, t_c^N) = -\log \frac{\exp(\langle f_{roi}, t_c^N \rangle)}{\sum_{c \in \mathcal{C}_N} \exp(\langle f_{roi}, t_c^N \rangle)}, \quad (1)$$

where $\langle \cdot, \cdot \rangle$ presents their cosine similarity.

Pseudo-labeling enhances novel category detection by assigning labels C_N to proposals P via similarity matching between proposals and novel class text embeddings $\{t_c^N \mid c \in \mathcal{C}_N\}$. This process builds high-confidence pseudo-annotated data. With this data, the detector learns robust semantic knowledge by jointly optimizing bounding box regression and classification for novel categories.

Overview. In traditional distillation methods, distilling knowledge solely from P introduces noisy background features and hampers the learning of category correlations due to the inaccurate localization of unknown-class proposals. This significantly undermines the efficiency of knowledge distillation. To obtain more informative visual embeddings, we propose the ASKD module (Section 3.2). It addresses background interference and information destruction during knowledge extraction by exploiting VLMs ability to perceive informative regions in aerial images, enhancing feature alignment granularity. Moreover, current pseudo-labeling methods rely on additional supervised signals, inherently constraining open-vocabulary space expansion. Consequently, detection boundaries become limited by prior semantic knowledge, creating a category coverage bottleneck. To eliminate reliance on text embeddings, we propose an unsupervised pseudo-labeling method PAPL (Section 3.3) based on prototype learning. Finally, SMI (Section 3.4) is employed to integrate the outputs of the classifiers of ASKD and PAPL along with the localization network, thereby enabling a comprehensive determination of the existence probability of unknown category objects.

Adaptive Selection Knowledge Distillation

Informative Region Perception. VLMs exhibit strong zero-shot image-level classification capabilities, yet they struggle with fine-grained region localization. We observe that averaging attention maps across layers of the visual encoder assigns higher weights to informative regions. By leveraging spatial priors, our adaptive proposal selection module and data augmenter extract informative proposals $P_{inf} = P_{fg} \cup P_{tg}$ to facilitate efficient knowledge distillation. Unlike threshold-based and count-based filtering methods (Gu et al. 2021), our approach better preserves semantic correlations within VLMs' region embeddings.

Adaptive Proposal Selection. To address the spatial mismatch between low-resolution attention maps and high-resolution images, we propose an attention normalization method inspired by (He et al. 2017). Specifically, we apply a scaling factor λ and subsequently perform sigmoid activation to transform the original attention maps.

$$\tilde{Attn} = \sigma(Attn \cdot \lambda) \tag{2}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Subsequently, we integrate an adaptive shifting mechanism:

$$M = \tilde{Attn} + \max\Big(1 - \mathbb{E}[\tilde{Attn}], 0\Big) \tag{3}$$

This operation guarantees non-negative attention values while maintaining the distribution characteristics, thereby generating a normalized attention mask $M \in R^{H\times W}$. The corresponding region $R_i \subset M$ is extracted through bilinear interpolation for each proposal p_i. Subsequently, the regional average response is computed to select informative region proposals based on a predefined threshold criterion. Mathematically, this process can be formulated as:

$$\tilde{w_i} = \mathbb{I}\left[\frac{1}{\mid R_i \mid}\sum_{(x,y)\in R_i} M(x,y) \geq 1\right] \tag{4}$$

This thresholding mechanism effectively distinguishes between informative regions ($\tilde{w_i} = 1$) and non-informative regions ($\tilde{w_i} = 0$), forming a subset of proposals $P_{inf} = \{p_i \mid \tilde{w_i} = 1\}$ that is optimized for the detection task.

Data Augmenter Based on Max-Min Edge Jitter. In VLMs, the non-adaptive cropping mechanisms of visual encoders significantly hinder the feature extraction process. Specifically, center-cropping objects with extreme aspect ratios results in the loss of crucial informative features, which compromises the alignment of semantic spaces during knowledge distillation. Moreover, given that aerial objects often exhibit local similarities across categories, the appropriate contextual receptive fields are essential for achieving high detection performance (Li et al. 2024b).

To address this, we propose an aspect ratio adaptive data augmenter to generate an enhanced proposal set P_{aug}. Given a proposal $\{p_i = (x_1, y_1, x_2, y_2) \mid p_i \in P_{inf}\}$ with width $w = x_2 - x_1$ and height $h = y_2 - y_1$, its aspect ratio is defined as $r = \max(w/h, h/w)$, reflecting its geometric characteristics. Proposals with $\log(r) > \alpha$ are classified as having extreme aspect ratios, where α is a pre-defined threshold. For such proposals, we define the dimensions as follows:

$$l = \max(w, h), \quad s = \min(w, h) \tag{5}$$

We employ two distinct strategies to augment P_{inf}: Longer-side jittering perturbs l within δ while fixing the maximum size. It can be expressed as:

$$p_i' = \left(c_x - \frac{l_\delta}{2}, c_y - \frac{l_\delta}{2}, c_x + \frac{l_\delta}{2}, c_y + \frac{l_\delta}{2}\right), \tag{6}$$

$$l_\delta = l + \sigma \cdot s \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{7}$$

where $c_x = \frac{x_1+x_2}{2}, c_y = \frac{y_1+y_2}{2}$ denote centroid coordinates. Set the variance jitter coefficient to a fixed value σ.

Shorter-side jittering proportionally scales δ while fixing the minimum size.

$$p_i' = \left(c_x - \frac{s_\delta}{2}, c_y - \frac{s_\delta}{2}, c_x + \frac{s_\delta}{2}, c_y + \frac{s_\delta}{2}\right) \tag{8}$$

$$s_\delta = s + \sigma \cdot l \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{9}$$

These augmented region proposals P_{aug} enable the model to learn both local and global region views during training, thereby enhancing feature extraction for objects with extreme aspect ratios, such as aerial ships and bridges.

Loss Function. Once P_{aug} is obtained, informative region features can be extracted using the visual encoder of the VLMs. Distilling these features minimizes the distance of the feature space between the detector and the visual encoder of the VLMs.

Specifically, region features $f_{roi}(p_i')$ are generated using the detector's RoI extractor. Meanwhile, the corresponding cropped image features $v(p_i')$ are extracted using the visual encoder for the same proposal. The feature distillation is performed by minimizing the L_1 distance between two sets of features, thereby enforcing geometric consistency between region features and the VLMs' semantic space.

$$\mathcal{L}_{distill} = \frac{1}{|P_{aug}|}\sum_i \parallel f_{roi}(p_i') - v(p_i') \parallel_1 \tag{10}$$

For training base classes, we follow (Gu et al. 2021) by replacing the weights of the trainable classifiers with frozen text embeddings of base classes $\{t_c^B \mid c \in \mathcal{C}_B\}$, which are generated by the VLMs text encoder. Additionally, we introduce a learnable background embedding t_{bg}^B, such that $t^B = \{t_c^B, t_{bg}^B\}$. For each class-agnostic proposal p_i in the input image, the cross-entropy loss function is defined as:

$$\mathcal{L}_{base} = \mathbb{E}_{(f(p),t)}[P(f_{roi}(p_i), t^B)] \tag{11}$$

Prototype-Aware Pseudo-Labeling

Unsupervised Pseudo-Labeled Data Generation. Notably, P_{aug} contains numerous unknown category objects P_{tg}. To generate high-quality pseudo-labeled data of unknown categories, we propose a PAPL method. It learns class decision boundaries from informative visual embeddings using prototype learning.

Specifically, P_{aug} is filtered to remove proposals that contain base categories C_B, based on RandBox's anchor position conditions (Wang et al. 2023b), retaining only regions that potentially contain unknown categories. Their visual embeddings undergo K-means clustering (Vaze et al. 2022) to capture inter-class differences among potential unknown categories, producing k cluster centers $\{v_j\}_{j=1}^k$. To minimize intra-class noise, the n nearest neighbor embeddings to each center within the embedding space are selected. Their corresponding proposals form a clean pseudo-labeled dataset with labels ranging from unknown-1 to unknown-k (denoted as $\mathcal{C}_U = \{\text{unknown-1}, \dots, \text{unknown-}k\}$), corresponding to k cluster centers $\{v_j\}_{j=1}^k$.

However, blindly increasing k may scatter features from the same object category into different unknown class clusters due to feature variability introduced by local crop encoding. Training with only a subset of these datasets introduces bias in the detector's RoI feature representation and multi-scale feature selection for the affected categories. Therefore, the value of k should be appropriately chosen. This ensures that the detector learns comprehensive features and achieves consistent bounding box localization for unknown categories, accurately capturing their open semantic knowledge (Li et al. 2024a).

Trainable Class Prototype Setting. Given the clean pseudo-labeled data, we propose learnable class prototypes to replace the frozen text embeddings of novel classes. The lack of extra priors in unsupervised pseudo-labeled data hinders precise category learning. To enable effective learning from the pseudo-labeled data, we introduce k trainable class prototypes $\{u_c \mid c \in \mathcal{C}_U\}$, where each prototype corresponds to a specific unknown category ('unknown-j', $j = 1, 2, \ldots, k$).

These class prototypes incorporate adaptive proposal embeddings for novel categories and include an additional learnable background prototype u_{bg} to address the issue of background proposal misclassification within informative regions. Therefore, the complete set of class prototypes is defined as $u = \{u_c, u_{bg}\}$.

Although the detector does not have access to explicit semantic labels for the pseudo-labeled data, the training process encourages it to distinguish and utilize inter-class variations in visual features, which are encoded into the learnable class prototypes. For each proposal p_i in the pseudo-labeled dataset, the cross-entropy loss used to optimize an extra prototype classifier is defined as:

$$\mathcal{L}_{proto} = \mathbb{E}_{(f(p),u)} P(f_{roi}(p_i), u) \tag{12}$$

This loss promotes alignment between VLMs' feature space and the detector's class prototypes. Selecting the top-n proposal embeddings ensures tighter intra-class cohesion within the embedding space. As a result, the detector achieves improved discrimination of novel category features in informative regions. Further details on the prototype classifier are in **Appendix B**.

Synthetic Matching Inference

Building upon traditional object detection frameworks such as Faster R-CNN (Ren et al. 2015), the two aforementioned methods enable the detector to learn generalized object representations beyond the distribution of the training data.

To estimate confidence scores for novel objects, we introduce a SMI mechanism. Inspired by LP-OVOD (Pham, Vu, and Nguyen 2023), our framework aggregates distillation scores $Score_d$, prototype scores $Score_p$, and localization objectness $Score_l$ into a unified scoring scheme.

First, for each proposal p generated by OLN, we utilize similarity scores modulated by a temperature parameter τ as one of the confidence metrics in the distillation head:

$$Score_d = -\log \frac{\exp(\langle f_{roi}(p), t_c^N \rangle / \tau)}{\sum_{c \in \mathcal{C}_N} \exp(\langle f_{roi}(p), t_c^N \rangle / \tau)} \tag{13}$$

Additionally, regarding $Score_p$, matching text embeddings of novel categories to unknown class prototypes remains a challenging task.

To address it, our analysis indicates that a higher similarity between cluster centers and novel-category text embeddings corresponds to better alignment between object features and class prototypes. Therefore, for a given novel text embedding t_c^N, we identify its nearest neighbor cluster center v_i, referred to as 'unknown-i', among the cluster centers in orthogonal space, and directly select the corresponding prototype $\hat{u}_i$ for classification:

$$\hat{u}_i = \underset{j \in \{1,2,\ldots,k\}}{\arg\max} \left(\langle t_c^N, v_j \rangle \right), \tag{14}$$

where $\hat{u}_i$ denotes the i-th class prototype.

Based on this selection, $Score_p$ is defined as:

$$Score_p = -\log \frac{\exp(\langle f_{roi}(p), \hat{u}_i \rangle / \tau)}{\sum_{j \in \{1,2,\ldots,k\}} \exp(\langle f_{roi}(p), u_j \rangle / \tau)} \tag{15}$$

The classification scores associated with these prototypes are aggregated using a dynamic weighting mechanism, ultimately forming the confidence estimate for unknown category objects within PAPL. The category score for a proposal in an image can then be expressed as:

$$Score_{cls} = \sqrt{Score_d \cdot Score_p}, \tag{16}$$

Furthermore, OLN, which is trained on base category data, also generates objectness scores ($Score_l$) based on the localization quality of object regions. These scores evaluate object confidence from a positional accuracy perspective. The synthetic score is formulated as:

$$Score_s = \sqrt{Score_l \cdot Score_{cls}} \tag{17}$$

Experiments

Experimental Setups

Datasets and Metrics. To evaluate the effectiveness of VK-Det for OVAD, experiments are conducted on two established aerial benchmarks: DIOR and DOTA. Following established protocols (Zang et al. 2024), DIOR's categories are divided into 16 base categories and 4 novel categories, while DOTA's categories are split into 11 base categories and 4 novel categories. We conducted training on the training set and evaluated it on the validation set. The primary evaluation metric is mean Average Precision at an IoU threshold of 0.5. This includes base category performance (mAP^B), novel category performance (mAP^N), overall performance (mAP^A), and harmonic mean (HM), which balances detection capability between base and novel categories. Following Castdet (Li et al. 2023b), HM is calculated as:

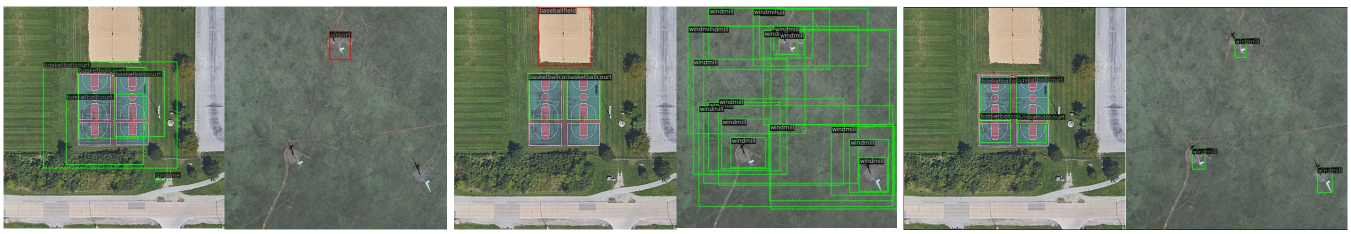
$$\mathrm{HM} = \frac{2 \cdot \mathrm{mAP}^B \cdot \mathrm{mAP}^N}{\mathrm{mAP}^B + \mathrm{mAP}^N} \tag{18}$$

Notably, mAP^N and HM are considered the primary evaluation metrics for OVAD on both datasets. Further details on dataset construction are in **Appendix D**.

Implementation details. Our method is implemented in the MMDetection toolbox (Chen et al. 2019). We use a Faster R-CNN (Ren et al. 2015) with a ResNet-50 (He et al. 2016) backbone as the detector. The pre-trained RemoteCLIP-ViT-B32 (Liu et al. 2023) serves as the pre-trained VLM. The training process consists of two stages: In

Method	Source	Backbone	$\mathcal{S}_u$	DIOR				DOTA			
				N	B	A	HM	N	B	A	HM
RRFS*	CVPR22	ResNet-101	×	2.8	41.9	38.1	5.2	2.2	47.1	38.1	4.2
ContrastZSD*	TPAMI22	ResNet-101	×	3.9	51.4	41.9	7.2	2.8	41.6	33.8	5.2
ViLD‡	ICLR22	Resnet-50	×	7.1	63.5	52.2	12.6	3.4	63.8	47.7	6.5
DescReg*	AAAI24	ResNet-101	✓	7.9	68.7	56.5	14.2	4.7	68.7	55.9	8.8
Castdet‡	ECCV24	ResNet-50	✓	<u>29.8</u>	75.5	66.5	**42.7**	<u>14.2</u>	64.3	50.9	<u>23.3</u>
Ours	-	ResNet-50	×	**30.1**	64.4	57.5	<u>41.0</u>	**23.3**	62.0	51.7	**33.9**

Table 1: Comparison with OVAD methods. where ‡ represents that the results of our own implementation, under the same experimental setup as ours. $*$ represents results quoted from the original paper. $\mathcal{S}_u$ represents the incorporation of extra supervised signals from unknown categories during training. N, B, and A present mAP^N, mAP^B, and mAP^A respectively.



(a) ViLD (b) Castdet (c) Ours

Figure 3: Visualization of open-vocabulary inference on DIOR dataset

the first stage, the distillation head is trained via ASKD for 20 epochs with a batch size of 32 on a single A800 GPU, using SGD optimization (lr = 1e-3, weight decay = 1e-4). In the second stage, we select the top 500 proposals most similar to 20 cluster centers as pseudo-labels and establish 20 category prototypes for the detector. The detector is then fine-tuned using pseudo-labeled data for 12 epochs with a batch size of 64, during which the backbone and neck are frozen, while other settings from the first stage are retained.

Comparison with State-of-the-Art Approaches

We evaluate VK-Det against OVAD methods (CastDet (Li et al. 2023b), DescReg (Zang et al. 2024)) and general open-vocabulary detectors (RRFS (Huang et al. 2022), ContrastZSD (Yan et al. 2021), ViLD (Gu et al. 2021)).

To ensure a fair comparison, we preprocessed both datasets by filtering out training set images that contained annotations of novel categories, maintaining consistency with our methodology. We train ViLD through RemoteCLIP-based image-region feature distillation without textual supervision, while CastDet leverages extra supervision from novel classes to construct dynamic pseudo-labeled sequences that guide semantic learning. Further details on relevant model training are in **Appendix E**.

As summarized in Table 1, our method demonstrates superior performance on both datasets compared to state-of-the-art open-vocabulary detectors. For instance, on DIOR, our method achieves 23.0% higher mAP^N than ViLD (no extra supervision) and 0.3% higher mAP^N than CastDet (extra supervision). This outperformance confirms our framework's effectiveness, establishing new SOTA.

To intuitively evaluate the effectiveness of VK-Det, Fig. 3 visualizes the detection results by comparing our detector with ViLD and CastDet methods. Correct novel class detections are shown in green, while erroneous detections appear in red. Our approach achieves precise novel class detection with minimal false positives. More results of the qualitative analysis are in **Appendix F**.

ASKD	PAPL	SMI	N	B	A	HM
✓	-	-	7.8	69.9	57.5	14.0
✓	✓	-	<u>20.4</u>	68.0	58.5	<u>31.4</u>
✓	-	✓	20.1	68.6	58.9	31.1
✓	✓	✓	**30.1**	64.4	57.5	**41.0**

Table 2: Ablation study of the VK-Det framework.

DOTA involves small objects and a large scale, which pose significant challenges. Our method achieves a 9.1% improvement over the state-of-the-art, demonstrating strong generalization across diverse aerial scenarios.

Ablation Study

We conducted additional ablation studies on the DIOR dataset, including component-wise method analysis, ASKD ablation, PAPL ablation, and SMI ablation experiments.

Ablation study of the VK-Det framework. Ablation studies evaluated core methods including ASKD, PAPL, and SMI in the VK-Det framework. Performance changes observed via stepwise integration in Table 2 demonstrate that ASKD boosted novel category detection, increasing mAP^N by 0.7% compared to ViLD and confirming the critical role of information region perception in ASKD; subsequent optimization of the combining localization network score (excluding prototype classifier score) further improved novel

Mask	Enhancer	N	B	A	HM
-	-	20.0	64.5	55.6	30.5
-	✓	23.2	64.0	55.8	34.1
✓	-	24.5	64.1	56.2	35.5
✓	✓	**30.1**	64.4	57.5	**41.0**

Table 3: Ablation study of the ASKD module.

-	N	B	A	HM
Extra supervision	28.1	64.0	56.8	39.0
Ours	**30.1**	64.4	57.5	**41.0**

Table 4: Ablation study of the PAPL module.

category mAP^N to 20.1%, reflecting OLN's open-world localization characteristics; balancing ASKD and PAPL classifier weights then increased novel category mAP^N to 20.4%, indicating successful learning of category boundaries in PAPL; finally, full method collaboration achieved an optimal novel category mAP^N of 30.1%, demonstrating that these two methods acquire distinct and mutually complementary open semantic knowledge.

Ablation study of the ASKD module. To further validate ASKD's effectiveness, Table 3 explores refined modules in ASKD. "Mask" presents masked proposal selection, while "Enhancer" denotes our data augmenter. The results demonstrate the superiority of our proposed feature selection: Independently enhancing a subset of proposals yields a 3.2% mAP^N gain over non-enhanced methods, while masked proposal selection improves performance by 4.5% mAP^N compared to unmasked knowledge distillation. Combining both approaches further boosts detector performance.

Ablation study of the PAPL module. Table 4 analyzes special pseudo-labeling cases by comparing our method with an extra supervised pseudo-labeling method for novel categories, where frozen text embeddings from fixed prompts generate quantitative pseudo-labels. Our approach outperforms the extra supervised pseudo-labeling method by 2.0% mAP^N, as textual supervision induces hallucinations and noise in image regions, causing category bounding box shifts. Conversely, PAPL maps feature embeddings into a latent category space, enabling unknown category matching and significantly enhancing pseudo-label quality.

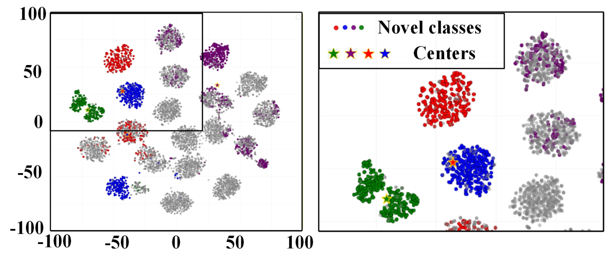
Ablation study of the SMI module. In the final inference stage, we conduct ablation studies to assess the impact of three scoring components: the distillation head score ($Score_d$), the classification head score ($Score_p$), and the localization network score ($Score_l$).

Table 5 demonstrates that relying exclusively on distillation or classification scores traps the model in local optima at 7.8% and 9.3% mAP^N, respectively. Merging two heads elevates mAP^N to 20.4% through coordinated category discrimination, while integrating all three scores attains peak performance of 30.1% mAP^N.

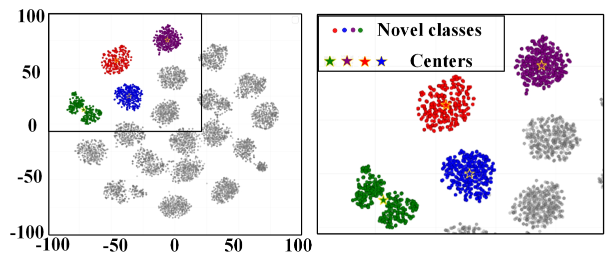
Furthermore, t-SNE visualizations (Linderman et al. 2017) of novel category features demonstrate the efficacy of PAPL in Fig. 4. The pseudo-labeled data generated by PAPL

Score_d	Score_p	Score_l	N	B	A	HM
✓	-	-	7.8	69.9	57.5	14.0
-	✓	-	9.3	69.6	57.5	16.4
✓	✓	-	20.4	68.0	58.5	31.4
-	✓	✓	12.6	64.4	54.0	21.1
✓	-	✓	24.8	64.1	56.2	35.7
✓	✓	✓	**30.1**	64.4	57.5	**41.0**

Table 5: Ablation study of the SMI module.



(a) Pseudo-labels for overlapping with ground truth labels.



(b) Pseudo-labels for text embedding selection.

Figure 4: Comparison of feature distributions: (a) Feature distribution in pseudo-labeled data with high IoU to the ground truth labels of novel classes; (b) Feature distribution in pseudo-labeled data exhibiting high similarity to the text embeddings of novel classes.

contains abundant novel category annotations; thus enabling the detector to learn distinguishable novel category features with minimal noise. This allows VK-Det to efficiently transfer novel category semantic knowledge from VLMs. More results of the ablation study are in **Appendix C**.

Conclusion

We propose to utilize informative region perception as guidance and construct a prototype learning-based classifier that efficiently and dynamically transfer knowledge from the visual encoder of VLMs. Unlike existing state-of-the-art methods that rely on extra supervision to generate pseudo-labels, we achieve efficient alignment of the detector feature space with the semantic space of VLMs through ASKD and PAPL, without additional data and supervision. Experiments show that when extra supervision is used to generate pseudo-labels, its robustness is lower than ours. This work could inspire further exploration of VLM visual knowledge spaces for dense prediction tasks. Future work will be devoted to developing efficient and lightweight methods for OVAD.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 62273353)

References

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; Zhang, Z.; Cheng, D.; Zhu, C.; Cheng, T.; Zhao, Q.; Li, B.; Lu, X.; Zhu, R.; Wu, Y.; Liu, K.; Dai, J.; Wang, J.; Shi, J.; Ouyang, W.; Loy, C. C.; and Lin, D. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *ArXiv*, abs/1906.07155.

Ding, J.; Xue, N.; Xia, G.-S.; Bai, X.; Yang, W.; Yang, M. Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2022. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 77787796.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *International Conference on Learning Representations*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017. Mask R-CNN.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Huang, P.; Han, J.; Cheng, D.; and Zhang, D. 2022. Robust Region Feature Synthesizer for Zero-Shot Object Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7612–7621.

Kim, D.; Lin, T.-Y.; Angelova, A.; Kweon, I.-S.; and Kuo, W. 2021. Learning Open-World Object Proposals without Learning to Classify. *IEEE Robotics and Automation Letters*, PP: 1–1.

Li, J.; Zhang, J.; Li, J.; Li, G.; Liu, S.; Lin, L.; and Li, G. 2024a. Learning Background Prompts to Discover Implicit Knowledge for Open Vocabulary Object Detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16678–16687.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2019. Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark. *ArXiv*, abs/1909.00133.

Li, L.; Miao, J.; Shi, D.; Tan, W.; Ren, Y.; Yang, Y.; and Pu, S. 2023a. Distilling DETR with Visual-Linguistic Knowledge for Open-Vocabulary Object Detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6478–6487.

Li, Y.; Guo, W.; Yang, X.; Liao, N.; He, D.; Zhou, J.; and Yu, W. 2023b. CastDet: Toward Open Vocabulary Aerial Object Detection with CLIP-Activated Student-Teacher Learning. In *European Conference on Computer Vision*.

Li, Y.; Li, X.; Dai, Y.; Hou, Q.; Liu, L.; Liu, Y.; Cheng, M.-M.; and Yang, J. 2024b. LSKNet: A Foundation Lightweight Backbone for Remote Sensing. *ArXiv*, abs/2403.11735.

Linderman, G. C.; Rachh, M.; Hoskins, J. G.; Steinerberger, S.; and Kluger, Y. 2017. Fast Interpolation-based t-SNE for Improved Visualization of Single-Cell RNA-Seq Data. *Nature methods*, 16: 243 – 245.

Liu, F.; Chen, D.; Guan, Z.-R.; Zhou, X.; Zhu, J.; and Zhou, J. 2023. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–16.

Ma, Z.; Luo, G.; Gao, J.; Li, L.; Chen, Y.; Wang, S.; Zhang, C.; and Hu, W. 2022. Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14054–14063.

Pan, J.; Liu, Y.; Fu, Y.; Ma, M.; Li, J.; Paudel, D. P.; van Gool, L.; and Huang, X. 2024. Locate Anything on Earth: Advancing Open-Vocabulary Object Detection for Remote Sensing Community. In *AAAI Conference on Artificial Intelligence*.

Pham, C.; Vu, T.; and Nguyen, K. 2023. LP-OVOD: Open-Vocabulary Object Detection by Linear Probing. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 768–777.

Redmon, J.; Divvala, S. K.; Girshick, R. B.; and Farhadi, A. 2015. You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39: 1137–1149.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Generalized Category Discovery. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7482–7491.

Wang, L.; Liu, Y.; Du, P.; Ding, Z.; Liao, Y.; Qi, Q.; Chen, B.; and Liu, S. 2023a. Object-Aware Distillation Pyramid for Open-Vocabulary Object Detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11186–11196.

Wang, Y.; Yue, Z.; Hua, X.; and Zhang, H. 2023b. Random Boxes Are Open-world Object Detectors. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6210–6220.

Yan, C.; Chang, X.; Luo, M.; Liu, H.; Zhang, X.; and Zheng, Q. 2021. Semantics-Guided Contrastive Network for Zero-Shot Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46: 1530–1544.

Zang, Z.; Lin, C.; Tang, C.; Wang, T.; and Lv, J. 2024. Zero-Shot Aerial Object Detection with Visual Description Regularization. *ArXiv*, abs/2402.18233.

Zhao, S.; Zhang, Z.; Schulter, S.; Zhao, L.; B.G, V. K.; Stathopoulos, A.; Chandraker, M.; and Metaxas, D. N. 2022. Exploiting Unlabeled Data with Vision and Language Models for Object Detection. *ArXiv*, abs/2207.08954.