DroneFINE: Domain-Aware Parameter-Efficient Fine-Tuning of Vision-Language Detectors for Drone Images

Ke Wu[1,2,5,6], Yanan Zhang[3], Yingjie Gao[2], Wenhao Li[2], Chenyu Zhou[2], Xinzhu Ma[2], Jiaxin Chen[⋆2,4], and Di Huang[2]

[1] National College for Excellent Engineers, Beihang University, Beijing, China
[2] School of Computer Science and Engineering, Beihang University, Beijing, China
[3] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
[4] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[5] Innovation Center for Intelligent System Cognition and Decision-Making, Beijing, China
[6] Information Science Academy of China Electronics Technology Group Corporation, Beijing, China
{20373041,jiaxinchen}@buaa.edu.cn

Abstract. Object detection for Unmanned Aerial Vehicles (UAVs) working in open and dynamic environments is a highly challenging task. While Vision-Language Models (VLMs) have offered a powerful solution for universal object detection, adapting them to UAV scenarios remains non-trivial due to a substantial domain gap between VLM pre-training data and aerial imagery. The prevailing Parameter-Efficient Fine-Tuning (PEFT) methods prove ineffective in bridging this gap, as VLMs' "natural-scene, foreground-dominant" visual priors misalign with the "bird's-eye-view, background-dominant, small-object" characteristics of UAV data. To address this issue, we propose DroneFINE, a novel PEFT paradigm comprising two domain-aware complementary modules tailored for VLM-based drone image detectors. Specifically, a data-dependent, foreground-aware, and multi-path adaptation mechanism named HyperAdapter is designed, which overcomes the static structural constraints of PEFT. In addition, a background suppression algorithm named SemanticGate is developed, which is a text-conditioned guidance strategy that employs background vocabulary to actively guide the model in suppressing responses from irrelevant regions. Extensive experiments on VisDrone and UAVDT demonstrate that DroneFINE significantly outperforms existing PEFT methods and achieves performance comparable to full fine-tuning while substantially reducing the fine-tuned parameters.

Keywords: Object Detection · UAV · VLM · PEFT

1 Introduction

UAVs are playing an increasingly vital role in diverse applications such as intelligent inspection, emergency rescue, and logistics delivery, where object detec-

⋆ Corresponding author.

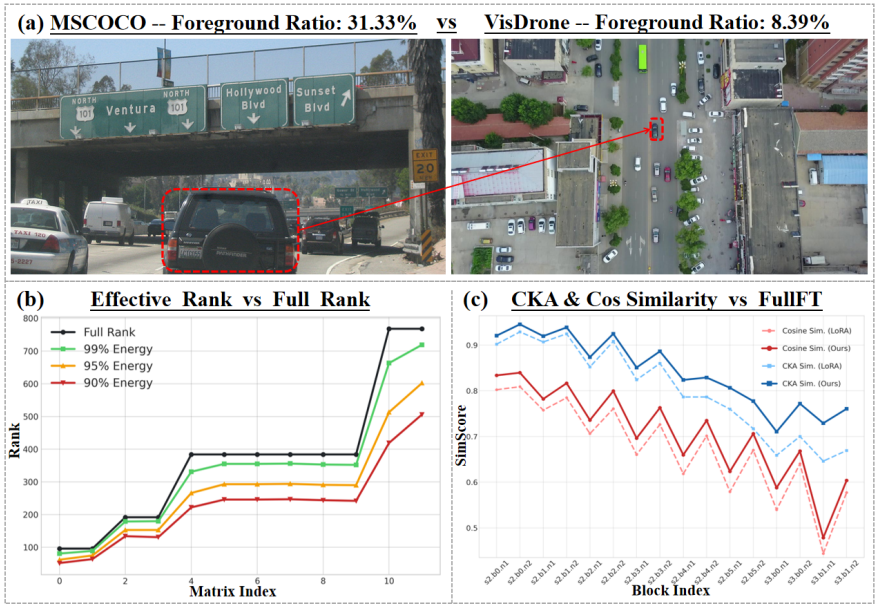

Fig. 1: Analysis of Domain Gap and Adaptation Bottlenecks for UAVs. (a) Comparison of image characteristics between common ground-level datasets [24] and aerial datasets [6]. **(b)** Rank-Energy($\|A\|_F^2$) plot of FFN layers' weights after Full Fine-Tuning (Full FT). **(c)** Feature similarity comparison between Full FT, standard LoRA [14](dashed line), and our method (solid line).

tion [3, 7, 8] serves as a critical perception task to enable these functions [30, 37, 39, 45, 48]. Recently, VLMs [3, 21, 29] have shown remarkable generalization and strong adaptability in universal object detection. Therefore, applying the broad knowledge of VLMs to drone detectors is expected to strengthen the robust perceptual capabilities of UAVs operating in complex and dynamic environments.

However, a fundamental domain gap between VLM pre-training data and aerial imagery prevents VLMs from leveraging their inherent strengths when directly applied to UAV scenarios. As shown in Fig. 1(a), the visual priors from foreground-dominant data [10, 21, 36, 38] generalize poorly to UAV imagery, which typically features bird's-eye-view and small objects in background-dominant scenes. While full fine-tuning or retraining [23, 33] might boost performance on UAV data, they often cause overfitting, representation collapse, or high computational costs, defeating our original purpose of using VLMs.

In this work, we investigate adapting VLMs to UAV scenarios via popular PEFT methods. Yet, our empirical study reveals that standard PEFT methods yield unsatisfactory transfer results. Furthermore, we analyze and summarize two critical flaws in these conventional methods: (i) Existing PEFT methods typically update weights in a low-rank space with limited adaptation capability. However, as shown in Fig. 1(b), fine-tuning a model for UAV imagery requires a high-rank space due to the complexity of both data and task, particularly in

deeper layers. (ii) The background-dominant nature of UAV imagery inevitably introduces significant noises from background, making it challenging to adapt the model to pay more attention to visual information in foreground areas.

To address the above issues, we propose a novel PEFT approach dubbed DroneFINE tailored for UAV imagery, by expanding the representation capability while effectively suppressing the interference from background. Specifically, we develop a dynamic adapter HyperAdapter, which strengthens the learning capability by utilizing queries to aggregate foreground features and generating multi-path depthwise separable convolutional kernels. Moreover, we design SemanticGate to unleash the generalization potential of the text branch in VLMs, which utilizes background text features to guide the learning of attention and employs a contrastive learning strategy to discriminate between foreground and background. Extensive experiments on UAV datasets show that DroneFINE outperforms existing PEFT methods by a large margin with minimal overhead, while achieving performance comparable to full fine-tuning. As shown in Fig. 1(c), the feature representations after fine-tuning via our method are significantly more aligned with full fine-tuning than LoRA [14]. This superior adaptation leads to a 1.7% mAP and 2.0% mAP50 improvement over LoRA, matching the performance of full fine-tuning on VisDrone [6]. Furthermore, our method achieves competitive performance with previous SOTA models on UAV imagery, while exhibiting stronger generalization and anti-forgetting capabilities.

The contributions of our work are summarized as below:

- We propose DroneFINE, a domain-aware PEFT paradigm for VLM-based drone image detectors that effectively addresses the substantial domain gap.
- We design HyperAdapter, which uses foreground-aware features to generate adaptive kernels, overcoming the representation constraints of PEFT.
- We develop a SemanticGate algorithm, which leverages background vocabulary to effectively mitigate background noise interference.
- Extensive experiments on UAV datasets demonstrate that DroneFINE significantly outperforms existing PEFT methods and achieves performance comparable to full fine-tuning while reducing the fine-tuned parameters.

2 Related Works

2.1 Object Detection on UAV imagery

Traditional research in UAV imagery object detection, distinct from ground-level, has focused on tackling domain-specific challenges, notably the prevalence of small objects and low foreground-to-background ratios. One line of works attempts to enhance small object discriminability through strategies like multi-scale refinement [41] or contextual feature enrichment [4, 20]. Another line of works focuses on mitigating background noise by explicitly separating foreground regions using techniques such as cluster-based proposals or density-aware sampling [15, 16, 42]. Despite progress on specific datasets, these methods share a fundamental flaw: an insular, closed-set design. They are tightly coupled to their

training distribution, not only in terms of predefined object categories but also specific scene characteristics. Consequently, adapting them to new environments or objects requires a cumbersome and resource-intensive modification process. This rigidity makes it highly inefficient to tailor UAV models for varying task-specific requirements. In sharp contrast, our work leverages VLMs as a powerful, feature-rich foundation to overcome this architectural rigidity. We demonstrate that rather than engineering complex, domain-specific modules, the robust representational power of VLMs can be directed towards targeted UAV tasks through remarkably simple adaptation mechanisms. This approach charts a path for an efficient and elegant paradigm: utilizing a strong foundation model alongside a widely applicable adaptation strategy to solve specific UAV detection challenges.

2.2 VLM Adaptation

VLMs with their strong image-text alignment and generalization capabilities, offer a viable path to address the traditional models' bottlenecks. VLM detectors identify objects via vision-language matching, enabling the detection of a vast range of objects. GLIP [21] pioneered VLM detectors through large-scale pre-training by fundamentally unifying object detection and phrase grounding. The GroundingDino [29] series introduced cross-modal attention to DINO [44], becoming a popular base model. YOLO-World [3] integrated CLIP [34] with YOLOv8 [19], gaining significant traction due to YOLO's widespread adoption. To apply these powerful, large-scale models to downstream tasks, PEFT methods [13, 14, 26–28, 47] have emerged as the key technology. By fine-tuning only a small subset of parameters, PEFT has achieved considerable success in adapting VLMs for general, ground-level vision tasks, effectively mitigating the high overhead and catastrophic forgetting associated with full fine-tuning.

However, adapting VLMs to the unique UAV domain remains an underexplored and challenging area. Several works have attempted to adapt VLMs to the aerial domain via full retraining, such as SPAR [23] and LAE-DINO [33]. Although these methods demonstrate strong performance, their reliance on massive data and computation renders them impractical for rapid iteration. Furthermore, this heavyweight approach is prone to overfitting on limited domain-specific data and risks catastrophic forgetting of the VLM's invaluable pre-trained knowledge. Given the drawbacks of full training, a natural alternative is to apply general PEFT methods directly. However, our preliminary experiments confirm that standard methods like LoRA [14] and Adapter [13] perform poorly on UAV imagery. This suggests a critical limitation: the simple representational structures of general PEFT methods are insufficient to bridge the fundamental domain gap between ground-level and aerial data. Based on this, we propose DroneFINE, a novel PEFT paradigm specifically designed to efficiently and effectively adapt VLMs to the challenging UAV domain by enhancing representational capacity and actively suppressing interference.

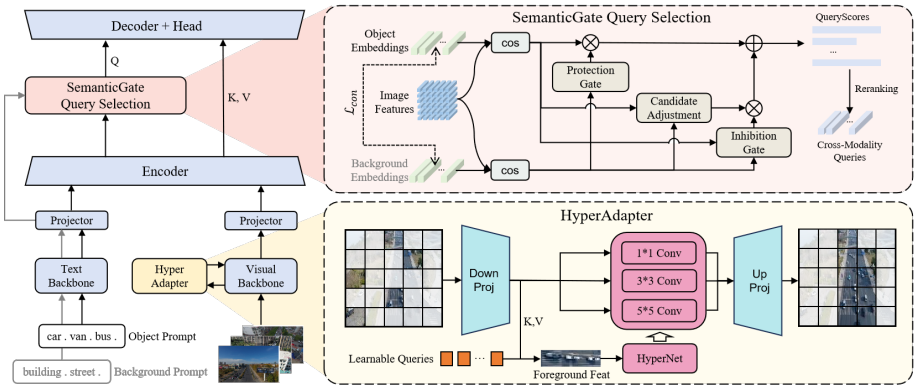

Fig. 2: Framework of DroneFINE. The model based on GroundingDINO comprises five parts: (i) frozen visual and text backbones for feature extraction, (ii) projectors for cross-modal alignment, (iii) a detection decoder and head, (iv) **HyperAdapter** as a serial adapter for the visual backbone, and (v) **SemanticGate** for enhanced GroundingDINO's language-guided query selection.

3 Methodology

3.1 Framework Overview

As illustrated in Fig. 2, our method builds upon the pre-trained vision-language object detector, GroundingDINO [29]. To improve the model's performance on UAV imagery, we introduce two novel modules dedicated to effective domain adaptation. Specifically, we propose the HyperAdapter to bolster the visual backbone's representational capacity via **foreground-aware** dynamic multi-path feature extraction. This is complemented by the SemanticGate, which refines the query selection mechanism through suppressing **background interference** and re-ranking queries, thus focusing the model's attention on the objects.

3.2 Preliminary

Full fine-tuning is the most intuitive approach for adapting a pre-trained vision-language detector to UAV imagery, as it involves updating all of the model's parameters. Formally, given a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, the optimization objective is defined as:

$$\theta \leftarrow \arg\min_{\theta} \mathcal{L}(D, \theta), \tag{1}$$

where θ denotes all the parameters, and $\mathcal{L}$ denotes the training loss of the model.

Despite its simplicity and effectiveness, full fine-tuning remains computationally expensive and prone to overfitting. To address these limitations, we adopt the parameter-efficient paradigm of adapter-tuning [13]. Adapter-tuning freezes

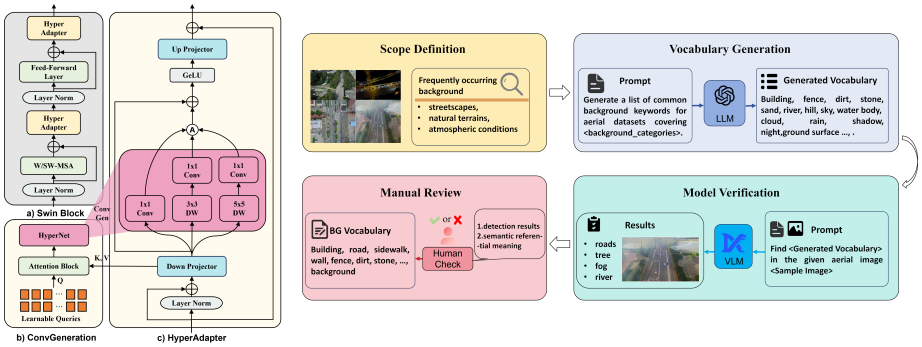

Fig. 3: (L)Architecture of HyperAdapter: Similar to a conventional visual adapter, the HyperAdapter is positioned after the attention and MLP layers. It introduces a foreground-aware multi-path convolution generation mechanism (highlighted in red in the figure) into the adapter. This mechanism utilizes queries to aggregate foreground features, which are then used to generate convolutions. These convolutions, in turn, leverage inductive bias for enhanced feature extraction.**(R)Process of Background Vocabulary Generation**:A background vocabulary is generated across streetscapes, natural terrains, and atmospheric conditions based on background distribution analysis of sampled images. It is validated via DINO-X [35] and manual check to ensure detectability and semantic clarity, resulting in 20 finalized vocabulary items.

the pre-trained backbone and updates the lightweight, inserted adapter modules, its optimization objective is:

$$\omega \leftarrow \arg\min_{\omega} \mathcal{L}(D, \theta_F, \omega), \tag{2}$$

where θ_F denotes the fixed backbone parameters, and ω represents the trainable parameters, including the adapters and the remaining parts of the detector.

3.3 HyperAdapter

Motivation. The profound domain gap between VLM pre-training data and UAV imagery is fundamentally characterized by dramatic shifts in object appearance. Beyond this profound domain gap, UAV imagery exhibits high intra-domain variability, where the appearance, scale, and spatial distribution of objects fluctuate markedly across samples. Specifically, under bird's-eye-view, the extremely limited scale of densely distributed targets (*e.g.*, pedestrians, bicycles) reduces them from textured objects into flattened geometric contours, rendering them as mere sparse pixel clusters. Furthermore, dominant environmental backgrounds overwhelm these already fragile features, making them nearly indistinguishable from noise. As observed in our experiments, this combination of fragile visual features and massive intra-domain variability causes VLMs to fail predominantly on small and densely distributed objects.

Conventional Architecture Limitations. Effective adaptation methods must possess sufficient capacity to reconstruct robust visual features against this representational collapse. As our Singular Value Decomposition (SVD) analysis illustrated in Fig. 1(b), when decomposing the weight matrices of the FFNs via SVD,

a remarkably high rank is required to maintain a low approximation error under UAV imagery. This indicates that simple low-rank linear approximations cannot adequately capture the complex spatial details of aerial objects. While incorporating static convolutional operations [18, 43] moderately alleviates this issue by introducing spatial inductive biases, they still exhibit critical representational bottlenecks in the UAV domain. Static structures applying a globally fixed set of filters lack the flexibility to accommodate extreme intra-domain variability and severe background noise. Crucially, as evidenced by our experiments on UAVDT, these static kernels lack the ability to dynamically guide the model toward discriminative regions. Instead of selectively emphasizing object-relevant semantics, the fixed parameter space forces a uniform feature extraction paradigm across the entire image. This lack of dynamic routing dilutes the model's focus, causing it to expend valuable representational capacity on redundant background noise.

HyperAdapter's Architecture. To tackle these issues, we introduce HyperAdapter, which is characterized by its foreground-aware architecture and multi-path convolution generation. Specifically, as shown in Fig. 3, HyperAdapter employs learnable foreground queries to extract essential foreground information. Because these queries are optimized across the entire training set, they encapsulate dataset-level inter-image priors, while their cross-attention with individual feature maps extracts specific intra-image contexts. Subsequently, these foreground features guide the generation of multi-path convolutions tailored for multi-scale object handling. This dynamic process substantially boosts the representation capacity of traditional low-rank adapters. As demonstrated by the feature visualization in Fig. 4, our mechanism successfully guides the model to focus precisely on the foreground regions, effectively verifying its capability to capture objects amid complex backgrounds. The complete computation flow of HyperAdapter is defined as:

$$\mathbf{W}_{dyn} = H(\mathbf{z}), \tag{3}$$

$$\mathbf{x}^{\text{flat}} = \text{Reshape}\left(\mathbf{W}_{down}\tilde{\mathbf{x}}\right), \tag{4}$$

$$\mathbf{p}_{out} = \text{Reshape}\left(\text{DynConv}\left(\mathbf{x}^{\text{flat}}; \mathbf{W}_{dyn}\right)\right), \tag{5}$$

$$\mathbf{x_{out}} = \mathbf{x} + \mathbf{W}_{up}\left(\text{GELU}\left(\mathbf{p}_{out}\right)\right), \tag{6}$$

where $\tilde{\mathbf{x}}$ denotes the scaled input features as in [43], which are subsequently projected into the low-rank space by $\mathbf{W}_{down}$. The hypernetwork H [11] maps the foreground feature $\mathbf{z}$ to data-dependent weights $\mathbf{W}_{dyn}$, which parameterize the DynConv module comprising parallel 1×1, 3×3, and 5×5 branches. Finally, the refined features are projected back by $\mathbf{W}_{up}$ and added to the input $\mathbf{x}$.

Foreground-Aware Module. To improve object perception, we propose leveraging aggregated foreground information to generate feature extractors. This ensures that the generated features can perform thorough extraction on regions that are semantically similar to the foreground cues. Given the background dominance in UAV imagery, we utilize M learnable queries $\mathbf{q}$ to aggregate foreground features via Cross-Attention instead of simple pooling. First, the attention weights $\mathbf{A}$ are computed between the queries and the flattened input

features $\mathbf{x}^{\text{flat}}$:

$$\mathbf{A} = \text{softmax}\left(\mathbf{q}(\mathbf{x}^{\text{flat}})^T\right) \in \mathbb{R}^{B\times M\times HW}. \tag{7}$$

Then, we perform a weighted aggregation of spatial features based on $\mathbf{A}$ and average the resulting M vectors to obtain the global foreground features $\mathbf{z}$. For the i-th sample within the same batch, this process is formulated as:

$$\mathbf{z_i} = \frac{1}{M}\sum_{j=1}^{M}\left(\sum_{k=1}^{HW}\mathbf{A}_{i,j,k}\cdot\mathbf{x}_{i,k}^{\text{flat}}\right). \tag{8}$$

This query-based pooling captures global foreground semantics implicitly, avoiding auxiliary computational overhead.

Parameter-Efficient Multi-Path Convolution Generation. With the global foreground features $\mathbf{z}$ extracted, the hypernetwork H transforms them into dynamic weights. Dynamic parameter generation often presents challenges in training stability and parameter efficiency, as validated in recent studies [2, 22, 31]. Drawing inspiration from existing studies, we first employ a single shared hypernetwork across all adapter layers to generate convolution kernels. This strategy not only economizes parameters but also ensures effective training by aggregating gradient feedback from multiple depth levels. Building on this robust foundation, we tailor the architecture to the specifics of UAV imagery. We introduce a multi-path design that serves a dual purpose: it enhances the perception of objects across varying scales and enriches gradient feedback to the generator through diverse spatial processing branches. To minimize the generated parameter count, we also compress the H's hidden dimensions, utilize depthwise separable convolutions, and employ a group-shared parameter mechanism. By dynamically generating a small number of convolution parameters, our method achieves a balance between performance and efficiency.

3.4 SemanticGate

Motivation. The open-vocabulary capability stands as a distinct advantage of VLMs over traditional models. While standard detection protocols narrow their focus strictly to objects of interest, they often overlook the VLMs' inherent capacity to recognize non-target background contexts. In UAV imagery, we observe a critical asymmetric feature alignment: while foreground features struggle to align with text prompts due to severe domain gaps, dominant background elements remain highly detectable.

Background Vocabulary Generation. Recognizing that the model allocates finite attention resources across the image, we propose a counter-intuitive strategy: instead of passively ignoring the background, we explicitly identify and actively suppress these background responses. This operation explicitly redirects the model's finite attention toward the more challenging foreground objects. As shown in Fig. 3, we establish this systematic background prior through a dedicated four-stage pipeline. To ensure comprehensive semantic coverage, we first sample representative UAV images to analyze common background distributions, identifying three dominant contexts: streetscapes, natural terrains, and

Algorithm 1 SemanticGate Query Selection: Enhanced GroundingDino's Language-guided Query Selection with Background Suppression

Input: $\mathbf{F}_{img}$, $\mathbf{F}_{text}$, $\mathbf{F}_{back}$, K
Output: *topk_ idx*
$\mathbf{S}_{orig} \leftarrow$ `torch.einsum("bic,btc->bit",` $\mathbf{F}_{img}, \mathbf{F}_{text}$`)`
$\mathbf{s}_o \leftarrow \mathbf{S}_{orig}$`.max(dim=-1)[0]`
$\mathbf{S}_{back} \leftarrow$ `torch.einsum("bic,btc->bit",` $\mathbf{F}_{img}, \mathbf{F}_{back}$`)`
$\mathbf{s}_{bg} \leftarrow \mathbf{S}_{back}$`.max(dim=-1)[0]`
$\mathbf{S}_{adj} \leftarrow \text{SemanticGateModule}(\mathbf{S}_{orig}, \mathbf{s}_o, \mathbf{s}_{bg})$
$\mathbf{s}_{adj} \leftarrow \mathbf{S}_{adj}$`.max(dim=-1)[0]`
topk_ idx $\leftarrow$ `torch.topk(`$\mathbf{s}_{adj}, K$`, dim=1)`[1]
return *topk_ idx*

atmospheric conditions. Based on these empirical observations, we prompt a LLM(GPT-4) to generate a diverse set of candidate keywords. Since not all textual concepts are visually salient from a bird's-eye view, we subsequently evaluate their empirical detectability by grounding these candidates in real UAV images using DINO-X [35]. Finally, a strict manual review is conducted to filter out terms with unstable detection results or ambiguous referential meanings. This rigorous process yields 20 finalized vocabulary items (*e.g.*, "road", "building"), whose embeddings are pre-computed to minimize runtime computational costs. Details of this generation process are provided in the Supplementary Material.

Gating Query Selection. Original query selection in GroundingDino is a mechanism similar to anchor box generation. It influences the position of the queries, which effectively dictates the model's regions of attention within the image. To redirect the model's attention to foreground regions, we propose SemanticGate Query Selection. We compute the similarity between background text features $\mathbf{F}_{back}$ and image features $\mathbf{F}_{img}$, resulting in a max background similarity score, $\mathbf{s}_{bg}$. This is compared against the max original (foreground) score, $\mathbf{s}_o$. As shown in Algo. 1:

$$\mathbf{s}_o = \max(\mathbf{S}_{orig}) \quad \text{and} \quad \mathbf{s}_{bg} = \max(\mathbf{S}_{back}), \tag{9}$$

where $\mathbf{S}_{orig}$ and $\mathbf{S}_{back}$ are the similarity matrices in Algo. 1. Specifically, $\mathbf{S}_{orig} \in \mathbb{R}^{N_{img} \times N_{text}}$ and $\mathbf{S}_{back} \in \mathbb{R}^{N_{img} \times N_{back}}$, where N_{img}, N_{text}, and N_{back} denote the number of image, text, and background tokens respectively.

To utilize these similarities, we adopt a gating mechanism inspired by traditional LSTM [12]'s forget, input, and output gates. It is designed to suppress query candidates where the specific background score s_{bg_can} is high and the foreground score s_{o_can} is low, indicating a probable background patch.

To resolve the competitive tension between $\mathbf{s}_{bg}$ and $\mathbf{s}_o$, we formulate a dynamic gating mechanism comprising three synergistic components: a Protection Gate (PG), an Inhibition Gate (IG), and a Candidate Adjustment (CA) module. PG evaluates the foreground confidence to determine the retention ratio of the original score $\mathbf{S}_{orig}$; for candidates where s_{o_can} is high, $\mathbf{g}_{pg}$ approaches 1, safely "protecting" potential target queries. Conversely, IG controls the activation of

the suppression mechanism; when a candidate exhibits distinct background dominance (high s_{bg_can} and low s_{o_can}), $\mathbf{g}_{ig}$ approaches 1 to enable suppression. Finally, CA quantifies the magnitude of the potential score adjustment, which is bounded by a tanh function and scaled by a hyperparameter α. Formally, these components are jointly defined as:

$$\begin{aligned} \text{(PG)} \quad & \mathbf{g}_{pg} = \sigma(w_{pg}^{bg} \cdot \mathbf{s}_{bg} + w_{pg}^{o} \cdot \mathbf{s}_{o} + b_{pg}), \\ \text{(IG)} \quad & \mathbf{g}_{ig} = \sigma(w_{ig}^{bg} \cdot \mathbf{s}_{bg} + w_{ig}^{o} \cdot \mathbf{s}_{o} + b_{ig}), \\ \text{(CA)} \quad & \mathbf{a}_{sig} = w_{ca}^{bg} \cdot \mathbf{s}_{bg} + w_{ca}^{o} \cdot \mathbf{s}_{o} + b_{ca}, \quad \mathbf{a}_{adj} = \tanh(\mathbf{a}_{sig}) \times \alpha. \end{aligned} \tag{10}$$

These components construct the final adjusted score matrix $\mathbf{S}_{adj}$, where suppression is applied via $\mathbf{g}_{ig}$ and the original score is preserved via $\mathbf{g}_{pg}$:

$$\mathbf{S}_{adj} = \mathbf{S}_{orig} \odot \mathbf{g}_{pg} + \mathbf{a}_{adj} \odot \mathbf{g}_{ig}. \tag{11}$$

The gating mechanism above critically depends on the assumption that $\mathbf{s}_{bg}$ and $\mathbf{s}_{o}$ are meaningful and distinct. To enforce this separability, we introduce a feature contrast solution. We apply an InfoNCE [32]-based contrastive loss, $\mathcal{L}_{con}$, on the text embedding features $\mathbf{F}_{text}$ and $\mathbf{F}_{back}$ used for vision-language matching.

$$\begin{aligned} \mathbf{F}_{comb} &= \text{Concat}([\mathbf{F}_{back}, \mathbf{F}_{text}]), \\ \mathcal{L}_{con} &= \text{TokenContrastiveLoss}(\mathbf{F}_{comb}). \end{aligned} \tag{12}$$

This loss pushes text tokens of different semantics (*e.g.* "road" vs "car") apart in the embedding space, ensuring that our background suppression mechanism operates on high-quality, separable features. Detailed statistical discussion of SemanticGate is relegated to the Supplementary Material.

4 Experiments

Datasets: To validate the performance of our model on aerial imagery, we select two datasets for training and validation: VisDrone [6] and UAVDT [5]. Overall, VisDrone challenges models with densely distributed small objects across diverse scenes, while UAVDT poses a high risk of overfitting due to the severe inter-frame similarity in its video sequences.
Evaluation Metrics: We evaluate model performance using the object detection metrics mAP_{50} and mAP (IoU=0.50:0.95) on the corresponding datasets.
Implementation Details: We adopt GroundingDino [29], implemented with MMDetection [1, 46], as our base model. For all fine-tuning strategies, we employ a low-rank space with $dim = 64$ to ensure a fair comparison. All adaptation methods are implemented by freezing both the text backbone and the visual backbone, while fine-tuning the decoder and detection head, restricting adjustments only to modality-specific modules.

4.1 Quantitative Evaluation

We conduct a comparative analysis, comparing our method with several representative and recent PEFT methods in Tab. 1. Based on these comparisons, we draw the following conclusions:

Table 1: Comparison of accuracy and parameter efficiency with state-of-the-art fine-tuning approaches on VisDrone and UAVDT.

Method	Δ Params.	Δ Trainable.	VisDrone (%)		UAVDT (%)	
			mAP	mAP_{50}	mAP	mAP_{50}
Zero-shot	0	–	16.5	26.9	9.5	18.5
Baseline	0	0	37.1	59.6	24.7	40.9
Full Fine-tuning	0	27.5M	**38.4**	**61.3**	22.4	38.6
Visual-Side Fine-tuning						
Adapter [13]	1.14M	1.14M	37.4	60.1	22.5	39.3
LoRA [14]	1.13M	1.13M	36.7	59.3	<u>25.0</u>	40.8
NormTuning [9]	0	0.025M	37.1	59.6	22.7	40.1
Mona [43]	1.4M	1.4M	<u>37.9</u>	60.5	22.2	37.6
VPT [17]	0.08M	0.08M	37.1	59.6	22.3	37.8
Text-Side Fine-tuning						
CoOp [47]	0.012M	0.012M	37.3	60.0	24.9	41.4
Shine [28]	0.02M	0.02M	37.0	60.1	24.3	<u>42.4</u>
DroneFINE-T (Ours)	1.54M	1.54M	**38.4**	<u>61.2</u>	**26.5**	**43.6**

Table 2: Performance comparison with UAV-based object detectors on VisDrone (left) and UAVDT (right). Other VLMs are pre-trained on the LAE-1M dataset and fine-tuned on the corresponding dataset, unlike our proposed DroneFINE.

Method	Backbone	mAP (%)	mAP_{50} (%)
Traditional UAV Models			
ClusDet [42] (ICCV'19)	ResNet-50	26.7	50.6
UFPMP-Det [16] (AAAI'22)	ResNet-50	36.6	62.4
CEASC [4] (CVPR'23)	ResNet-50	28.7	50.7
QueryDet [41] (CVPR'22)	ResNet-50	28.3	48.1
Recent SOTA UAV Models			
RemDet-X [20] (AAAI'25)	YOLOv8X	<u>40.0</u>	61.9
ESOD [25] (TIP'24)	YOLOv5	37.9	62.3
Dome-DETR [15] (ACM MM'25)	HGNetv2L	39.0	61.1
VLMs			
YOLO-World [3] (CVPR'24)	YOLOv8L	–	55.3
RT-OVAD [40] (arXiv'25)	ResNet-50	–	<u>64.6</u>
LAE-DINO [33] (AAAI'25)	Swin-T	–	56.4
DroneFINE-T (Ours)	Swin-T	38.4	61.2
DroneFINE-B (Ours)	Swin-B	38.8	61.9
DroneFINE-L (Ours)	Swin-L	**42.3**	**65.8**

Method	mAP (%)	mAP_{50} (%)
Traditional UAV Models		
ClusDet [42] (ICCV'19)	13.7	26.5
UFPMP-Det [16] (AAAI'22)	<u>24.6</u>	38.7
CEASC [4] (CVPR'23)	17.1	30.9
Recent SOTA UAV Models		
RemDet-L [20] (AAAI'25)	20.6	34.5
ESOD [25] (TIP'24)	23.6	**47.6**
VLMs		
YOLO-World [3] (CVPR'24)	–	35.8
RT-OVAD [40] (arXiv'25)	–	40.5
LAE-DINO [33] (AAAI'25)	–	36.5
DroneFINE-T (Ours)	**26.5**	<u>43.6</u>

General VLMs cannot achieve good performance when directly applied to aerial imagery. Although VLMs are expected to show strong generalization, the huge domain gap prevents them from achieving good zero-shot performance on UAV images. We observe that categories such as "people", and "awning-tricycle" yield very low mAP_{50} (approx. 2%-20%). This highlights the significant discrepancy between aerial images and the VLM's pre-training data.

Text-side adaptation prevents overfitting via semantic alignment but struggles to capture dense, small objects. Text-side methods [47] [28] use minimal parameter updates for high-dimensional semantic alignment. This lightweight approach provides strong regularization, effectively preventing overfitting on highly redundant UAVDT video sequences. However, on VisDrone, such high-level semantic guidance cannot compensate for the lack of low-level visual features required to detect densely distributed small objects. To overcome this limitation, DroneFINE jointly optimizes high-level semantics and low-level visual features, achieving stable performance improvements.

Visual-side adaptation is constrained by static representational structures, failing to adapt to the inherent characteristics of extreme data distributions. Visual-side methods [14] [13] directly augment spatial features

but are restricted by their linear architectures. Even when introducing convs like Mona [43], the static designs lack the representational flexibility to align with specific data characteristics. Consequently, their representational inadequacy leads to limited performance gains on the spatially complex VisDrone dataset, while triggering severe overfitting on UAVDT, where static kernels fail to adequately capture spatial context.

Our method improves performance by addressing both foreground and background. DroneFINE utilizes a foreground-aware dynamic multi-path convolution and a background suppression module. It achieves 38.4% mAP and 61.2% mAP_{50} on VisDrone. For the first time, it matches the performance of the 27.5M-parameter full fine-tuning method by training only 1.54M parameters, which constitutes 5.6% of the parameters required for full fine-tuning, thereby demonstrating the method's distinct advantage. On UAVDT, DroneFINE achieves gains of 1.8% in mAP and 2.7% in mAP_{50} compared to the baseline, which fine-tunes the decoder and detection head while freezing the remaining parameters. Furthermore, it outperforms all existing PEFT methods and even surpasses full fine-tuning on UAVDT.

To more objectively evaluate the effectiveness of our method, we conduct performance comparisons with current mainstream detectors for UAV imagery on VisDrone and UAVDT in Tab. 2. On VisDrone, our model DroneFINE-L surpasses the previous SOTA, achieving an improvement of 2.3% in mAP and 3.4% in mAP_{50}. Compared to other VLM detectors trained on LAE-1M [33], our model shows an improvement of approximately 1–10% in mAP_{50}. On UAVDT, our model DroneFINE-T achieves SOTA mAP, surpassing the recent SOTA ESOD [25] by 2.9%. Although our mAP_{50} is slightly lower, this is because ESOD re-crops the foreground regions, prioritizing foreground detection over precise localization at higher IoU thresholds. Moreover, despite lacking large-scale aerial pre-training, DroneFine matches RT-OVAD [40] and LAE-DINO [33], confirming the viability of adaptation approaches. Further details regarding the comparison are provided in the Supplementary Material.

4.2 Qualitative Evaluation

Fig. 4 illustrates a qualitative comparison between our method and the baseline on the VisDrone dataset. In the first row, the baseline misclassifies dumpsters as cars, incorrectly associating their aerial appearance with the pre-learned "car" concept, highlighting its poor cross-domain alignment. Furthermore, the second and third rows demonstrate the superior robustness of our model on challenging samples. Specifically, our method accurately detects the distant truck in the second row and the bicycle in the third row, effectively handling both small objects and rare categories where the baseline fails. In Fig. 4, warmer colors indicate stronger attention, demonstrating that our foreground-aware mechanism effectively focuses on salient foreground objects such as vehicles and pedestrians.

Fig. 4: Detection Results and Attention Map on VisDrone. (Left) Visualization of detection results on the VisDrone. **(Right)** Visualization of foreground-aware attention maps. Original images and corresponding attention maps generated by HyperAdapter's foreground-aware module.

Table 3: Ablation study of different components on the VisDrone and UAVDT datasets. "Full" indicates the module with all internal mechanisms.

Method	VisDrone (%)		UAVDT (%)	
	mAP	**mAP$_{50}$**	**mAP**	**mAP$_{50}$**
Baseline	37.1	59.6	24.7	40.9
HyperAdapter (Full)	38.1	61.1	26.1	41.9
w/o Foreground Awareness	37.8	60.2	24.9	41.0
SemanticGate (Full)	37.9	60.8	26.0	42.0
w/o Contrastive Loss	37.5	60.4	25.2	42.1
Mona + CoOp	37.5	60.3	24.5	41.7
DroneFINE-T (Full Model)	**38.4**	**61.2**	**26.5**	**43.6**

4.3 Ablation Study

Ablation on the main components. To analyze the contribution of each component, we conduct ablation studies on the modules and their internal mechanisms across the VisDrone and UAVDT datasets. As shown in Tab. 3, the proposed components generally improve the model's fine-tuning performance. Compared to the Baseline, each module consistently increases mAP$_{50}$ by 1.2-1.5% on VisDrone and 1.0-1.1% on UAVDT. Building on this, we delve deeper into the internal mechanisms of the two modules. The experiments show that dynamically generating convolutions improves performance, but its effect is limited without foreground awareness. The foreground-aware module is crucial, providing a 0.9% mAP$_{50}$ improvement on VisDrone and UAVDT. Additionally, although removing the contrastive loss reduces overall mAP and mAP$_{50}$ on VisDrone, the standalone SemanticGate still outperforms the baseline. Besides, we experiment with a simple combination of text-side (CoOp) and visual-side (Mona) fine-tuning. However, the performance is significantly inferior to DroneFINE, highlighting the complementarity and effectiveness of our method.

Table 4: Module ablation on VisDrone and additional performance evaluations. In (d), "Full" contains street, terrain, and atmosphere vocabularies, while "Random" denotes target-irrelevant words.

(a) Foreground-aware module ablation.

Method	mAP	mAP_{50}
Baseline	37.1	59.6
Adapter + ObjSeeker	37.3	60.0
HyperAda.	**38.1**	**61.1**
HyperAda. + ObjSeeker	38.0	**61.1**

(b) Rank ablation.

Rank	mAP	mAP_{50}
32	37.3	60.0
64	**38.4**	**61.2**
96	37.9	60.9

(c) Rank-scaled methods.

Method	Δ Params.	mAP	mAP_{50}
Baseline	0	37.1	59.6
LoRA ($r = 128$)	2.26M	37.1	59.6
Mona ($r = 128$)	2.97M	37.9	60.6
HyperAda. (Ours)	**1.54M**	**38.1**	**61.1**

(d) Background vocabulary ablation.

Vocabulary	#Words	mAP	mAP_{50}
Full vocabulary	20	**37.9**	**60.8**
Streetscapes	5	37.6	60.2
Natural terrains	8	37.8	60.3
Atmospheric conditions	7	37.2	59.6
Random-5	5	37.0	59.6
Random-10	10	36.6	59.2
Random-20	20	36.9	59.4

(e) Anti-forgetting & speed.

Method	mAP	FPS
Full FT	42.5	**3.22**
Mona	41.9	2.99
Ours	**44.9**	2.79

Ablation on the Foreground-Aware Module in HyperAdapter. We also introduce another foreground-aware module, ObjSeeker from ESOD [25], into the Adapter, which identifies the foreground by predicting a density map. While this module (Adapter + ObjSeeker) slightly improve performance over the baseline, its overall performance is weaker than our proposed HyperAdapter, as shown in Tab. 4a. We also attempt to replace the attention scores of HyperAdapter with density map prediction scores. However, this yield no performance improvement. This indicates that effective global aggregated features for highlighting foreground parts can be learned implicitly. Moreover, adding an auxiliary task head to each Adapter significantly increases computational cost. In contrast, our model achieves a better balance between performance and computational overhead. Finally, we investigate the impact of the rank on model performance. As shown in Tab. 4b, we first conduct an ablation study to identify the optimal rank configuration for HyperAdapter. Furthermore, we compare our approach with other methods across varying rank scales (Tab. 4c), which further demonstrates the effectiveness and parameter efficiency of our proposed method.

Analysis of Background Semantics and Query Re-ranking in SemanticGate. Tab. 4d shows that SemanticGate relies on meaningful background semantics rather than arbitrary text tokens. Quantitatively, the full vocabulary achieves 37.9%/60.8% mAP/mAP_{50}, outperforming the best single-category vocabulary by 0.1%/0.5% and the size-matched Random-20 vocabulary by 1.0%/1.4% points. Moreover, increasing the random vocabulary from 5 to 20 words does not

yield consistent gains, indicating that the improvement arises from complementary background semantics rather than vocabulary size. We also quantify the changes in queries before and after SemanticGate. On VisDrone, for query selection's top-20 queries, GT-box IoU coverage improves/degrades from pre- to post-SemanticGate on 19.3%/6.0% of images; Recall@0.5 improves/degrades on 6.8%/2.0%. Notably, foreground similarity slightly decreases on 23.9% of images after gating, indicating that the gains are not from simply boosting foreground scores. Instead, SemanticGate uses background scores to re-rank queries and improve GT coverage. A single-gate variant drops 0.7% mAP_{50} on VisDrone, supporting the multi-gate design.

4.4 Efficiency and Anti-Forgetting Evaluation

To evaluate the practical utility of DroneFINE, we assess its inference efficiency and anti-forgetting capability on COCO. As shown in Tab. 4e, our method achieves a COCO mAP of 44.9%, significantly outperforming Full FT (42.5%) and Mona (41.9%). This demonstrates that DroneFINE effectively preserves the inherent generalization of the VLM. Compared to the original model, inference speed decreases by only 0.43 FPS on a single 2080 Ti. Considering DroneFINE's superior performance, this trade-off is entirely reasonable.

5 Conclusion

Existing PEFT methods fail to bridge the substantial domain gap between ground-level and aerial imagery, limiting the deployment of VLMs in UAV-based object detection. In this paper, we analyze the underlying causes of this failure, attributing it to two main factors: the insufficient representation capacity of original PEFT structures, and the severe interference from complex background noise. To tackle these issues, we propose a novel domain-aware PEFT framework dubbed DroneFINE. Architecturally, DroneFINE employs a foreground-aware dynamic multi-path convolution to significantly enhance foreground representations. Mechanistically, it introduces a background vocabulary coupled with a gating mechanism to effectively suppress background interference. Experimental results demonstrate that DroneFINE significantly improves performance, paving a new path for the application of VLMs in the UAV domain.

Acknowledgement

This work was supported in part by the Beijing Natural Science Foundation under Grant 4242044, the National Key R&D Program of China under Grant 2021YFA1000401, the National Natural Science Foundation of China under Grants U23B2012, U22B2049, and 62506111, the China Postdoctoral Science Foundation under Grant 2025M781468, the Postdoctoral Fellowship Program of the CPSF under Grant GZC20251098, the Research Program of the State Key Laboratory of Virtual Reality Technology and Systems, and the Fundamental Research Funds for the Central Universities.

References

1. 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., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
2. Chen, Z., Zeng, Y., Chen, Z., Gao, H., Chen, L., Liu, J., Zhao, F.: Vfm-adapter: Adapting visual foundation models for dense prediction with dynamic hybrid operation mapping. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2385–2393 (2025)
3. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16901–16911 (2024)
4. Du, B., Huang, Y., Chen, J., Huang, D.: Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13435–13444 (2023)
5. Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q.: The unmanned aerial vehicle benchmark: Object detection and tracking. In: Proceedings of the European conference on computer vision (ECCV). pp. 370–386 (2018)
6. Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., Peng, T., Zheng, J., Wang, X., Zhang, Y., 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. 0–0 (2019)
7. Gao, Y., Zhang, Y., Cai, Z., Huang, D.: Test-time adaptive object detection with foundation model. Advances in Neural Information Processing Systems **38**, 95228–95251 (2026)
8. Gao, Y., Zhang, Y., Huang, Z., Liu, N., Huang, D.: Ps-ttl: Prototype-based soft-labels and test-time learning for few-shot object detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8691–8700 (2024)
9. Giannou, A., Rajput, S., Papailiopoulos, D.: The expressive power of tuning only the normalization layers. arXiv preprint arXiv:2302.07937 (2023)
10. Gupta, T., Marten, R., Kembhavi, A., Hoiem, D.: Grit: General robust image task benchmark. arXiv preprint arXiv:2204.13653 (2022)
11. Ha, D., Dai, A., Le, Q.V.: Hypernetworks. arXiv preprint arXiv:1609.09106 (2016)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
13. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799. PMLR (2019)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
15. Hu, Z., Wu, P., Chen, J., Zhu, H., Wang, Y., Peng, Y., Li, H., Sun, X.: Domedetr: Detr with density-oriented feature-query manipulation for efficient tiny object detection. arXiv preprint arXiv:2505.05741 (2025)
16. Huang, Y., Chen, J., Huang, D.: Ufpmp-det: Toward accurate and efficient object detection on drone imagery. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 1026–1033 (2022)

17. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
18. Jie, S., Deng, Z.H.: Convolutional bypasses are better vision transformer adapters. arXiv preprint arXiv:2207.07039 (2022)
19. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolov8 (2023), `https://github.com/ultralytics/ultralytics`
20. Li, C., Zhao, R., Wang, Z., Xu, H., Zhu, X.: Remdet: Rethinking efficient model design for uav object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4643–4651 (2025)
21. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10975 (2022)
22. Li, M., Chen, J., Feng, W., Li, B., Dai, F., Zhao, S., He, Q.: Hyperlora: Parameter-efficient adaptive generation for portrait synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13114–13123 (2025)
23. Li, N., Ye, M., Zhou, L., Tang, S., Gan, Y., Liang, Z., Zhu, X.: Self-prompting analogical reasoning for uav object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 18412–18420 (2025)
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755. Springer (2014)
25. Liu, K., Fu, Z., Jin, S., Chen, Z., Zhou, F., Jiang, R., Chen, Y., Ye, J.: Esod: efficient small object detection on high-resolution images. IEEE Transactions on Image Processing (2024)
26. Liu, L., Wang, N., Chen, C., Liu, D., Yang, X., Gao, X., Liu, T.: Frequency-based comprehensive prompt learning for vision-language models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
27. Liu, L., Wang, N., Yang, X., Gao, X., Liu, T.: Surrogate prompt learning: Towards efficient and diverse prompt learning for vision-language models. In: Forty-second International Conference on Machine Learning (2025)
28. Liu, M., Hayes, T.L., Ricci, E., Csurka, G., Volpi, R.: Shine: Semantic hierarchy nexus for open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16634–16644 (2024)
29. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
30. Liu, S., Zhang, H., Qi, Y., Wang, P., Zhang, Y., Wu, Q.: Aerialvln: Vision-and-language navigation for uavs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15384–15394 (2023)
31. Lv, C., Li, L., Zhang, S., Chen, G., Qi, F., Zhang, N., Zheng, H.T.: Hyperlora: Efficient cross-task generalization via constrained low-rank adapters generation. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 16376–16393 (2024)
32. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
33. Pan, J., Liu, Y., Fu, Y., Ma, M., Li, J., Paudel, D.P., Van Gool, L., Huang, X.: Locate anything on earth: Advancing open-vocabulary object detection for remote

sensing community. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6281–6289 (2025)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
35. Ren, T., Chen, Y., Jiang, Q., Zeng, Z., Xiong, Y., Liu, W., Ma, Z., Shen, J., Gao, Y., Jiang, X., et al.: Dino-x: A unified vision model for open-world object detection and understanding. arXiv preprint arXiv:2411.14347 (2024)
36. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019)
37. Tian, Y., Lin, F., Li, Y., Zhang, T., Zhang, Q., Fu, X., Huang, J., Dai, X., Wang, Y., Tian, C., et al.: Uavs meet llms: Overviews and perspectives towards agentic low-altitude mobility. Information Fusion **122**, 103158 (2025)
38. Wang, J., Zhang, P., Chu, T., Cao, Y., Zhou, Y., Wu, T., Wang, B., He, C., Lin, D.: V3det: Vast vocabulary visual detection dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19844–19854 (2023)
39. Wang, X., Yang, D., Wang, Z., Kwan, H., Chen, J., Wu, W., Li, H., Liao, Y., Liu, S.: Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. arXiv preprint arXiv:2410.07087 (2024)
40. Wei, G., Yuan, X., Liu, Y., Shang, Z., Xue, X., Wang, P., Yao, K., Zhao, C., Zhang, H., Xiao, R.: Rt-ovad: Real-time open-vocabulary aerial object detection via image-text collaboration. arXiv preprint arXiv:2408.12246 (2024)
41. Yang, C., Huang, Z., Wang, N.: Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 13668–13677 (2022)
42. Yang, F., Fan, H., Chu, P., Blasch, E., Ling, H.: Clustered object detection in aerial images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8311–8320 (2019)
43. Yin, D., Hu, L., Li, B., Zhang, Y., Yang, X.: 5%> 100%: Breaking performance shackles of full fine-tuning on visual recognition tasks. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20071–20081 (2025)
44. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: The Eleventh International Conference on Learning Representations
45. Zhang, W., Gao, C., Yu, S., Peng, R., Zhao, B., Zhang, Q., Cui, J., Chen, X., Li, Y.: Citynavagent: Aerial vision-and-language navigation with hierarchical semantic planning and global memory. arXiv preprint arXiv:2505.05622 (2025)
46. Zhao, X., Chen, Y., Xu, S., Li, X., Wang, X., Li, Y., Huang, H.: An open and comprehensive pipeline for unified object grounding and detection. arXiv preprint arXiv:2401.02361 (2024)
47. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
48. Zhou, Y., Li, J., Ou, C., Yan, D., Zhang, H., Xue, X.: Open-vocabulary object detection in uav imagery: A review and future perspectives. Drones **9**(8), 557 (2025)