KNOWLEDGE LIES IN LAYERS: INTERNALIZING FOUNDATION VISUAL PRIORS FOR FEW-SHOT REMOTE SENSING OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot remote sensing object detection is limited by scarce annotations, hindering robust and generalizable representation learning. Existing methods increasingly leverage vision foundation models to enrich visual representations, but often rely on fixed-depth features or unified multi-layer integration, overlooking depth-dependent representation characteristics. Through a systematic layer-wise analysis of a vision foundation model, we find clear depth-dependent differences: low-to-mid layers exhibit stronger variation across spatial locations, while mid-to-high layers show stronger alignment between patch responses and object regions. Such a phenomenon implies that different network depths offer distinct advantages in spatial representation and object-aware semantics, and that exploiting these advantages may provide a lightweight solution to improve few-shot detection under limited supervision. Accordingly, we propose LayerFi-FSOD, a hierarchical foundation visual prior internalization framework that exploits foundation-model knowledge according to its depth-dependent representation characteristics. Specifically, the Adaptive Dense Prior Injection Module aggregates low-to-mid-layer dense priors to enhance multi-scale detection features, while the Dual-Grained Structural-Relational Guidance Module guides the student detector to absorb knowledge from different depths through local spatial structures and global semantic relations. Extensive experiments on the DIOR and NWPU VHR-10 datasets validate the effectiveness of LayerFi-FSOD. Notably, the vision foundation model is used only as a visual prior provider during training and is completely removed at inference, introducing no additional foundation-model inference overhead. The relevant code will be made publicly available upon acceptance.

1 INTRODUCTION

Few-shot remote sensing object detection aims to generalize detectors to novel categories using only a few annotated samples while maintaining accurate recognition and localization (Li et al., 2020; Zhou et al., 2025; Li et al., 2024; Guirguis et al., 2023). However, limited training samples cannot sufficiently cover the complex variations in object scale, appearance, and background, making detectors prone to overfitting and hindering the learning of robust and generalizable object representations (Ma et al., 2022). This challenge is further aggravated in remote sensing imagery by small objects, substantial scale variations, and complex backgrounds (Wang et al., 2019; Xiao & Marlet, 2020). Therefore, relying solely on limited task-specific annotations often fails to provide sufficient and reliable visual knowledge, making it crucial to introduce richer and more generalizable external visual priors to compensate for insufficient supervision (Li et al., 2025a). Vision foundation models (Caron et al., 2021; Radford et al., 2021) acquire rich and transferable visual representations through large-scale pretraining, offering a promising way to alleviate supervision scarcity in few-shot detection. Existing studies (Hu et al., 2026; Sun et al., 2025; Zhu et al., 2026) have explored transferring their semantic representations or dense visual features to downstream detection tasks to enhance feature learning under limited supervision. However, existing methods (Qin et al., 2026; Xu et al., 2026) typically rely on final-layer features or uniformly integrate multi-layer representations, without fully exploiting the visual information encoded at different network depths. Meanwhile, most approaches still retain the vision foundation model during testing, resulting in additional computational overhead and deployment burden (Cuttano et al., 2026; Fu et al., 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

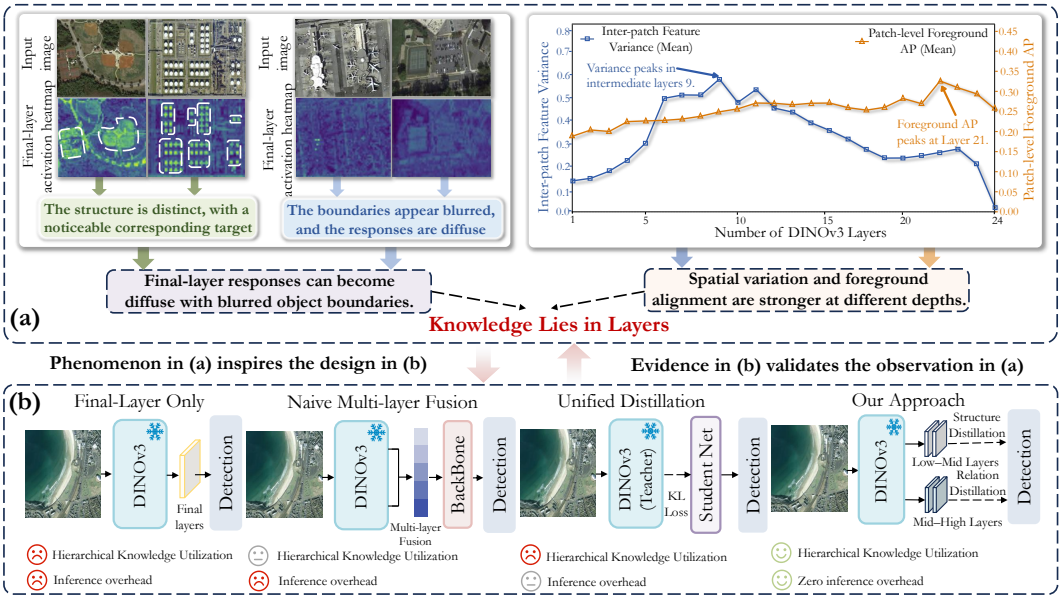

Figure 1: Motivation and key observations of LayerFi-FSOD. (a) The final layer of DINOv3 does not consistently preserve clear object structures, while the layer-wise analysis shows different depth-dependent trends in inter-patch feature variation and patch-level foreground AP. (b) Compared with final-layer-only utilization, naive multi-layer fusion, and unified distillation, LayerFi-FSOD exploits knowledge from different depths according to their characteristics and promotes prior absorption through local structural and global semantic relational guidance, without introducing additional inference overhead.

Among recent vision foundation models, DINOv3 (Siméoni et al., 2026) is a representative self-supervised visual encoder increasingly used for dense prediction and object detection (Cuttano et al., 2026; Faulkenberry & Prasad, 2026; Gu et al., 2026). Its dense hierarchical representations make it well suited for layer-wise analysis. As shown in Fig. 1(a), final-layer responses do not consistently preserve clear object structures, with some targets exhibiting diffuse activations and blurred boundaries. We therefore examine two layer-wise statistics: inter-patch feature variance, which measures spatial variation among patch representations, and patch-level foreground AP, which evaluates whether stronger patch responses align with object regions. For the latter, the ℓ_2 norm of each patch token is used as its response score, and patches overlapping ground-truth boxes are treated as foreground. The results reveal distinct depth-dependent trends: inter-patch variation is higher in intermediate layers, whereas foreground AP increases in deeper layers. Thus, low-to-mid layers exhibit stronger spatial variation, while mid-to-high layers show stronger responses on object regions, indicating that useful visual information is distributed differently across network depths.

Based on these observations, we propose LayerFi-FSOD. As shown in Fig. 1(b), it exploits representations from different depths according to their characteristics and absorbs the resulting visual priors into the detector during training, avoiding additional foundation-model inference overhead. Specifically, low-to-mid-layer dense features are adaptively aggregated to enhance local structural and scale-aware representations, while mid-to-high-layer features are used to model semantic relations among proposals. These priors are further absorbed through local spatial structure and global semantic relation guidance.

In summary, our contributions are threefold:

- **Insight.** We identify clear depth-dependent representation differences in a representative vision foundation model: low-to-mid layers exhibit stronger inter-patch variation, while mid-to-high layers show stronger alignment between patch responses and object regions.
- **Method.** We propose LayerFi-FSOD, which combines ADPIM and DSGM to exploit depth-dependent visual priors and promote their absorption through local spatial structures and global semantic relations.

- **Experiments.** Extensive experiments on multiple few-shot remote sensing object detection benchmarks demonstrate competitive performance across different few-shot settings, with student-only inference and no additional foundation-model inference overhead.

2 Related Work

Few-Shot Remote Sensing Object Detection. Few-Shot Object Detection (FSOD) aims to generalize detectors to novel categories using only a few annotated samples (zhou Li et al., 2023; Lin et al., 2025; Yan et al., 2025; Kang et al., 2019). Compared with natural images, remote sensing imagery typically exhibits complex backgrounds, substantial scale variations, and large intra-class appearance changes (Xia et al., 2018; Zhang et al., 2023c; Jia et al., 2025), making representation learning and cross-category generalization particularly challenging under few-shot supervision (Li et al., 2022a; Wang et al., 2025; Wolf et al., 2021). Existing few-shot remote sensing object detection methods mainly follow transfer-learning and meta-learning paradigms. Transfer-learning-based methods typically train detectors on well-annotated base categories and then adapt them to novel categories with only a few labeled samples (Sun et al., 2021; Jiang et al., 2023; Qiao et al., 2021). Representative methods include Retentive R-CNN (Fan et al., 2021), which mitigates catastrophic forgetting through bias balancing; RP-FSOD (Wu et al., 2024), which reduces intra-class variation and inter-class confusion through prototype refinement and personality filtering; and SCG-FSOD (Hu et al., 2026), which models semantic correlations between base and novel categories while incorporating contrastive learning. Meta-learning-based methods generally employ episodic training to model category-conditioned interactions between support samples and query images (Han et al., 2022; Zhang et al., 2023a). For example, P-CNN (Cheng et al., 2022) enhances category discrimination through prototype learning, while MLII-FSOD (Wang et al., 2024) alleviates feature misalignment through multi-level support–query interaction. Despite these advances, existing approaches still rely primarily on limited task-specific supervision and remain constrained by insufficient visual representations in extremely low-shot settings. To address this issue, we introduce DINOv3 (Siméoni et al., 2026), one of the most representative vision foundation models, as an external source of visual priors to provide richer structural and semantic information.

Vision Foundation Models for Object Detection. Vision foundation models acquire transferable and generalizable visual representations through large-scale pretraining, providing valuable external knowledge for alleviating supervision scarcity in few-shot learning. Self-supervised Vision Transformers such as DINOv2 (Oquab et al., 2024) and DINOv3 (Siméoni et al., 2026) learn dense representations containing rich semantic and spatial information and have increasingly been explored for object detection and related downstream tasks. DE-ViT (Zhang et al., 2025) constructs prototype-induced subspaces from frozen DINOv2 features for novel-object recognition and localization; FM-FSOD (Han & Lim, 2024) combines foundation-model visual representations with a large language model to model contextual relationships among proposals and categories; VFM-DETR (Liu et al., 2025) treats representations from different network depths as visual experts and selectively integrates them through routing mechanisms; and INSID3 (Cuttano et al., 2026) performs multi-grained in-context matching and localization using frozen DINOv3 dense features. Despite these advances, existing approaches either rely on fixed-depth representations or exploit multi-layer features without explicitly distinguishing their depth-dependent representation characteristics and functional roles. Moreover, many foundation-model-enhanced detectors retain pretrained visual models during inference (Faulkenberry & Prasad, 2026; Park et al., 2024; Gu et al., 2025), introducing additional computational and deployment costs. In contrast, our work first analyzes how visual representations vary across network depths and then exploits different depth ranges according to their characteristics, while promoting visual prior absorption into the student detector during training and eliminating dependence on the foundation model at inference.

3 Method

3.1 Problem Definition

Following the standard FSOD protocol (Li et al., 2022b; Liu et al., 2024; Xu et al., 2025), all object categories are divided into disjoint base and novel sets C_{base} and C_{novel}, with $C_{\text{base}} \cap C_{\text{novel}} = \emptyset$. The base dataset $D_{\text{base}} = \{(x_i, y_i) \mid y_i \in C_{\text{base}}\}$ contains abundant annotations, whereas D_{novel}

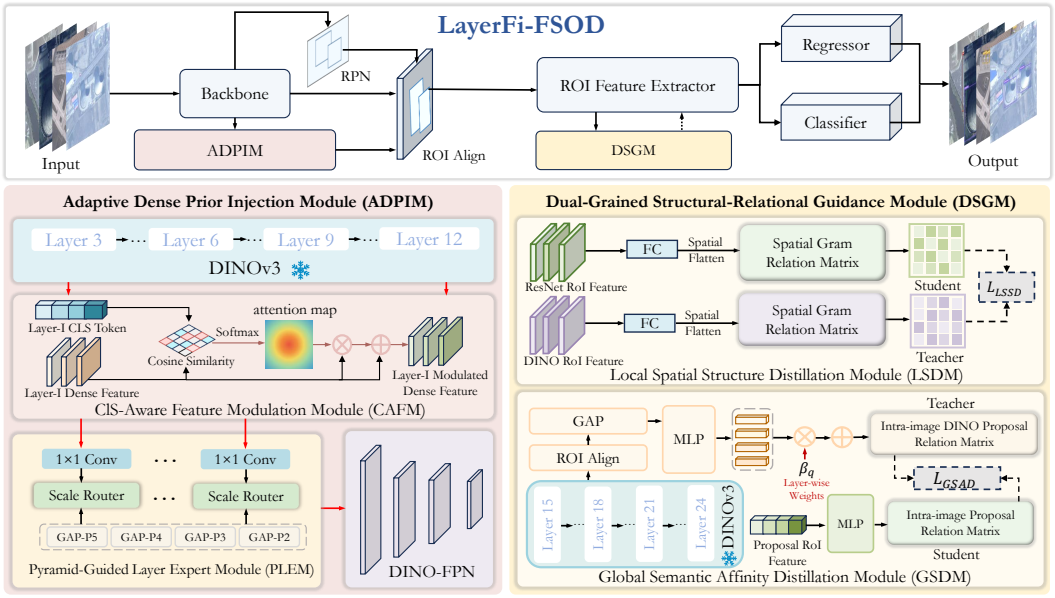

Figure 2: Overall framework of LayerFi-FSOD. The proposed framework employs a frozen DINOv3 only during training to provide hierarchical visual priors. Specifically, ADPIM primarily exploits dense spatial priors from low-to-mid layers, while DSGM further promotes the absorption of visual knowledge from different depths through local spatial structure guidance and global semantic affinity guidance.

provides only K annotated instances per novel category. The detector is first trained on D_{base} and then fine-tuned on a class-balanced set containing K instances per base and novel category. As shown in Fig. 2, LayerFi-FSOD introduces a frozen DINOv3 only during few-shot fine-tuning and removes all DINOv3-related branches at inference.

3.2 Adaptive Dense Prior Injection Module

According to the preceding analysis, low-to-mid DINOv3 layers exhibit stronger spatial variation, while representations from different depths are not equally suited to the scale requirements of different FPN levels. ADPIM therefore modulates and adaptively aggregates these dense representations to construct scale-adaptive visual priors.

CLS-Aware Feature Modulation Module (CAFM). CAFM uses the global semantic context encoded by the CLS token to modulate dense patch features. Given the CLS token $\mathbf{c}_k$ and the p-th patch feature $\mathbf{d}_{k,p}$ from DINOv3 layer k, we compute a patch-wise modulation weight based on their normalized semantic similarity as

$$a_{k,p} = N_k \frac{\exp\left(\langle \bar{\mathbf{c}}_k, \bar{\mathbf{d}}_{k,p} \rangle / \tau_c\right)}{\sum_{q=1}^{N_k} \exp\left(\langle \bar{\mathbf{c}}_k, \bar{\mathbf{d}}_{k,q} \rangle / \tau_c\right)}, \tag{1}$$

where $\bar{\mathbf{c}}_k$ and $\bar{\mathbf{d}}_{k,p}$ denote ℓ_2-normalized features, N_k is the number of spatial patches, and τ_c is the temperature parameter. The patch features are then modulated using a residual formulation as

$$\widetilde{\mathbf{d}}_{k,p} = \mathbf{d}_{k,p}\left(1 + \lambda_c a_{k,p}\right), \tag{2}$$

where λ_c controls the modulation strength. This residual formulation preserves the original dense features while emphasizing patches that are more semantically consistent with the global context.

Pyramid-Guided Layer Expert Module (PLEM). Different FPN levels (Lin et al., 2017) correspond to different object scales and thus require different combinations of low-to-mid DINOv3 representations. PLEM uses the current-level descriptor together with pyramid context to predict scale-specific routing weights. Let $\mathbf{P}_l$ denote the feature of the l-th FPN level, and let $\mathbf{g}_l = \text{GAP}(\mathbf{P}_l)$

denote its global descriptor obtained by global average pooling (GAP). We construct the pyramid context by concatenating the descriptors of all FPN levels, i.e., $\mathbf{g}_{\mathrm{ctx}} = [\mathbf{g}_2 \| \mathbf{g}_3 \| \mathbf{g}_4 \| \mathbf{g}_5]$, and use it together with the current-level descriptor to predict the routing weights as

$$\boldsymbol{\alpha}_l = \mathrm{Softmax}\left(\frac{\mathcal{R}_l\left([\mathbf{g}_l \| \mathbf{g}_{\mathrm{ctx}}]\right)}{\tau_r}\right), \qquad l \in \{2,3,4,5\}, \tag{3}$$

where $\|$ denotes feature concatenation, $\mathcal{R}_l$ is a scale-specific two-layer MLP with ReLU, and τ_r is the routing temperature. $\alpha_{l,k}$ denotes the contribution of DINOv3 layer k to pyramid level l.

After CLS-aware modulation, each selected DINOv3 feature is resized to the target FPN resolution, aligned by a layer-specific 1×1 convolution, and aggregated according to the routing weights as

$$\mathbf{Q}_l = \sum_{k \in \mathcal{K}_{\mathrm{lm}}} \alpha_{l,k}\, \phi_k\left(\mathrm{Resize}\left(\widetilde{\mathbf{D}}_k, H_l, W_l\right)\right), \tag{4}$$

where $\mathcal{K}_{\mathrm{lm}}$ denotes the selected low-to-mid DINOv3 layers, $\widetilde{\mathbf{D}}_k$ is the modulated dense feature map, and $\phi_k(\cdot)$ performs layer-specific channel alignment. $\mathbf{Q}_l$ is the resulting scale-adaptive dense prior at pyramid level l.

Finally, RoI features extracted from the detector FPN $\{\mathbf{P}_l\}_{l=2}^5$ and the constructed prior pyramid $\{\mathbf{Q}_l\}_{l=2}^5$ using the same proposals are fused through a gated residual connection as

$$\widehat{\mathbf{R}} = g\left(\{\mathbf{P}_l\}_{l=2}^5, \mathcal{B}\right) + \sigma(\gamma)\, \psi\left(g\left(\{\mathbf{Q}_l\}_{l=2}^5, \mathcal{B}\right)\right), \tag{5}$$

where $g(\mathbf{F}, b) = \mathrm{RoIAlign}(\mathbf{F}, b)$ denotes RoIAlign-based feature extraction from the feature pyramid $\mathbf{F}$ using proposal b, $b \in \mathcal{B}$, and $\mathcal{B}$ is the shared proposal set. $\psi(\cdot)$ consists of a 1×1 convolution followed by ReLU, γ is a learnable scalar, and $\sigma(\cdot)$ denotes the sigmoid function. The fused representation $\widehat{\mathbf{R}}$ is fed into the classification and regression branches.

3.3 Dual-Grained Structural-Relational Guidance Module

Although ADPIM enhances detection representations through visual-prior injection, feature injection alone does not explicitly promote prior absorption into the student detector. We therefore introduce DSGM, which transfers knowledge through local spatial structures and global semantic relations rather than directly matching feature values.

Local Spatial Structure Distillation Module (LSDM). Inspired by the Gram-anchoring objective in DINOv3 (Siméoni et al., 2026), LSDM transfers local spatial relations between corresponding positive RoIs. For each positive RoI m, the student feature and its corresponding DINO prior feature are projected into a shared embedding space and flattened into spatial token sequences $\mathbf{S}_m, \mathbf{T}_m \in \mathbb{R}^{N_r \times d}$. We construct token-wise self-similarity distributions to model pairwise spatial relations as

$$\mathbf{A}_m^S = \mathrm{Softmax}_{\mathrm{row}}\left(\frac{\overline{\mathbf{S}}_m \overline{\mathbf{S}}_m^\top}{\tau_g}\right), \qquad \mathbf{A}_m^T = \mathrm{Softmax}_{\mathrm{row}}\left(\frac{\overline{\mathbf{T}}_m \overline{\mathbf{T}}_m^\top}{\tau_g}\right), \tag{6}$$

where $\overline{\mathbf{S}}_m$ and $\overline{\mathbf{T}}_m$ denote feature-normalized spatial tokens, and τ_g is the temperature parameter. The Local Spatial Structure Distillation (LSSD) loss is defined as

$$\mathcal{L}_{\mathrm{LSSD}} = \frac{1}{M_+} \sum_{m=1}^{M_+} D_{\mathrm{KL}}\left(\mathrm{sg}\left[\mathbf{A}_m^T\right] \parallel \mathbf{A}_m^S\right), \tag{7}$$

where M_+ denotes the number of positive RoIs and $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation.

Global Semantic Affinity Distillation Module (GSDM). Since the mid-to-high layers of DINOv3 exhibit stronger object-aware representations, GSDM transfers their proposal-level semantic relations to the student detector. For each proposal b_m, features from the selected DINOv3 layers are regionally pooled, projected, and aggregated to obtain the teacher representation as

$$\mathbf{z}_m^T = \mathrm{Norm}\left(\sum_{q \in \mathcal{K}_{\mathrm{mh}}} \beta_q\, \phi_q^T\left(\mathrm{GAP}\left[g(\mathbf{D}_q, b_m)\right]\right)\right), \tag{8}$$

where $\mathcal{K}_{\mathrm{mh}}$ denotes the selected mid-to-high DINOv3 layers, $\mathbf{D}_q$ is the dense feature map from layer q, $\phi_q^T(\cdot)$ is the corresponding projection function, and $\mathrm{Norm}(\cdot)$ denotes ℓ_2 normalization. The layer weight is given by $\beta_q = \mathrm{Softmax}(\boldsymbol{\omega})_q$, where $\boldsymbol{\omega}$ denotes learnable layer logits. The student representation is obtained from the proposal classification feature as

$$\mathbf{z}_m^S = \mathrm{Norm}\left(\phi^S(\mathbf{f}_m^{\mathrm{cls}})\right), \tag{9}$$

where $\mathbf{f}_m^{\mathrm{cls}}$ denotes the classification feature of proposal b_m, and $\phi^S(\cdot)$ is the student projection function.

For each image i, we collect the teacher and student proposal embeddings as $\mathbf{Z}_i^T$ and $\mathbf{Z}_i^S$ and construct within-image semantic affinity distributions as

$$\mathbf{G}_i^T = \mathrm{Softmax}_{\mathrm{row}}\left(\frac{\mathbf{Z}_i^T(\mathbf{Z}_i^T)^\top}{\tau_t}\right), \qquad \mathbf{G}_i^S = \mathrm{Softmax}_{\mathrm{row}}\left(\frac{\mathbf{Z}_i^S(\mathbf{Z}_i^S)^\top}{\tau_s}\right), \tag{10}$$

where τ_t and τ_s denote the teacher and student temperatures, respectively. The Global Semantic Affinity Distillation (GSAD) loss is defined as

$$\mathcal{L}_{\mathrm{GSAD}} = \frac{1}{B'}\sum_{i=1}^{B'} D_{\mathrm{KL}}\left(\mathrm{sg}\left[\mathbf{G}_i^T\right] \parallel \mathbf{G}_i^S\right), \tag{11}$$

where B' denotes the number of images containing valid sampled proposals, and $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation.

3.4 Overall Loss

The base-training stage follows the original SAE-FSDet baseline without modification. During few-shot fine-tuning, we augment the baseline objective with the proposed local spatial and global semantic distillation objectives as

$$\mathcal{L} = \mathcal{L}_{\mathrm{det}} + \lambda_s\mathcal{L}_{\mathrm{LSSD}} + \lambda_g\mathcal{L}_{\mathrm{GSAD}}, \tag{12}$$

where $\mathcal{L}_{\mathrm{det}}$ denotes the original training objective of SAE-FSDet, while λ_s and λ_g control the contributions of the LSSD and GSAD objectives, respectively.

4 Experiments

In this section, we report the experimental results of LayerFi-FSOD on two remote sensing object detection datasets, DIOR (Li et al., 2020) and NWPU VHR-10 (Cheng et al., 2014). Following the standard protocols adopted in prior work (Liu et al., 2024; Xu et al., 2025), we report mAP_{50} on novel classes, i.e., the mean AP across novel classes at an IoU threshold of 0.5. Detailed dataset descriptions and experimental settings are provided in Appendix A.1.

4.1 Main Results

We report the results on DIOR and NWPU VHR-10 in Tables 1 and 2, respectively. LayerFi-FSOD achieves the highest novel-class mAP_{50} in 11 out of 12 evaluation settings. Compared with generic few-shot detectors such as Meta-RCNN (Yan et al., 2019) and FSCE (Sun et al., 2021), LayerFi-FSOD consistently delivers higher accuracy on both datasets. It also outperforms remote-sensing-specific methods, including G-FSDet (Zhang et al., 2023b) and the baseline SAE-FSDet (Liu et al., 2024), across all evaluated settings. On NWPU VHR-10, LayerFi-FSOD further surpasses RP-FSOD (Wu et al., 2024) and ST-FSOD (Sun et al., 2025) under every shot setting. Compared with SAE-FSDet, the gains reach 5.50 and 8.54 percentage points under the 3-shot and 5-shot settings, respectively, which are larger than the improvements of 2.30 and 0.98 points under the 10-shot and 20-shot settings. The larger gains in lower-shot regimes indicate that foundation-model visual priors provide useful knowledge beyond what the detector can learn from extremely limited task-specific annotations.

Among foundation-model-based approaches, the training-free No-Time-To-Train (Espinosa et al., 2026) and FSOD-VFM (Feng et al., 2026) show limited improvement as more samples become

Table 1: Comparison with state-of-the-art methods on DIOR. Results are reported as mAP_{50} on the novel classes under different few-shot settings. The best results are highlighted in bold, and the second-best results are underlined.

Split	Method	Publish	3-shot	5-shot	10-shot	20-shot
1	Meta-RCNN (Yan et al., 2019)	ICCV	12.02	13.09	14.07	14.45
	TFA w/cos (Wang et al., 2020)	ICML	16.07	15.36	16.45	18.93
	FsDetView (Xiao & Marlet, 2020)	ECCV	13.19	14.29	18.02	18.01
	FSOD (Fan et al., 2020)	CVPR	15.94	20.27	24.22	28.16
	FSCE (Sun et al., 2021)	CVPR	27.91	28.60	33.05	39.22
	MM-RCNN (Li et al., 2022a)	TGRS	19.80	23.90	28.80	30.80
	ICPE (Lu et al., 2023)	AAAI	11.68	12.34	12.95	14.33
	VFA (Han et al., 2023)	AAAI	21.94	21.29	23.32	24.28
	G-FSDet (Zhang et al., 2023b)	ISPRS	27.06	29.89	34.86	37.49
	SAE-FSDet (Liu et al., 2024)	TGRS	28.80	32.40	<u>37.09</u>	<u>42.46</u>
	No-Time-To-Train (Espinosa et al., 2026)	arXiv	22.24	20.49	23.08	22.54
	FSOD-VFM (Feng et al., 2026)	ICLR	**34.29**	<u>33.70</u>	35.68	36.07
	LayerFi-FSOD	–	<u>31.03</u>	**33.82**	**37.73**	**43.09**
2	Meta-RCNN (Yan et al., 2019)	ICCV	13.94	15.84	15.07	18.17
	TFA w/cos (Wang et al., 2020)	ICML	9.54	13.82	13.82	16.61
	FsDetView (Xiao & Marlet, 2020)	ECCV	14.28	15.95	15.37	16.96
	FSOD (Fan et al., 2020)	CVPR	11.84	12.98	17.17	18.46
	FSCE (Sun et al., 2021)	CVPR	<u>17.45</u>	20.42	22.22	24.96
	MM-RCNN (Li et al., 2022a)	TGRS	16.40	18.70	20.30	<u>27.10</u>
	ICPE (Lu et al., 2023)	AAAI	14.45	14.52	15.95	15.61
	VFA (Han et al., 2023)	AAAI	15.52	17.76	18.62	20.05
	G-FSDet (Zhang et al., 2023b)	ISPRS	13.80	15.70	18.56	19.94
	SAE-FSDet (Liu et al., 2024)	TGRS	17.27	<u>20.48</u>	<u>22.69</u>	26.75
	No-Time-To-Train (Espinosa et al., 2026)	arXiv	6.42	6.72	6.88	6.93
	FSOD-VFM (Feng et al., 2026)	ICLR	8.92	10.45	10.44	10.07
	LayerFi-FSOD	–	**19.16**	**21.54**	**25.01**	**27.71**

Table 2: Comparison with state-of-the-art methods on NWPU VHR-10. Results are reported as mAP_{50} on the novel classes under different few-shot settings. The best results are highlighted in bold, and the second-best results are underlined.

Method	Publish	3-shot	5-shot	10-shot	20-shot
Meta-RCNN (Yan et al., 2019)	ICCV	20.51	21.77	26.98	28.24
TFA w/cos (Wang et al., 2020)	ICML	8.80	9.49	9.26	10.83
FSCE (Sun et al., 2021)	CVPR	41.63	48.80	59.97	79.60
P-CNN (Cheng et al., 2022)	TGRS	41.80	49.17	63.29	66.83
MSOCL (Chen et al., 2022)	TGRS	45.00	49.04	63.27	78.43
VFA (Han et al., 2023)	AAAI	13.14	15.08	13.89	20.18
G-FSDet (Zhang et al., 2023b)	ISPRS	49.05	56.10	71.82	75.41
SAE-FSDet (Liu et al., 2024)	TGRS	57.96	59.40	71.02	<u>85.08</u>
RP-FSOD (Wu et al., 2024)	TCSVT	50.45	60.72	<u>72.91</u>	81.46
ST-FSOD (Sun et al., 2025)	TGRS	51.59	58.49	72.75	79.46
RIFR (Zhu et al., 2025)	TGRS	53.17	62.39	72.29	77.01
Domain-RAG (Li et al., 2025b)	NeurIPS	<u>59.99</u>	<u>65.78</u>	72.87	84.05
No-Time-To-Train (Espinosa et al., 2026)	arXiv	42.95	46.88	39.36	39.32
FSOD-VFM (Feng et al., 2026)	ICLR	42.54	41.86	41.66	41.83
LayerFi-FSOD	–	**63.46**	**67.94**	**73.32**	**86.06**

available. Although FSOD-VFM performs best under the 3-shot setting of DIOR Split 1, LayerFi-FSOD surpasses it with more training samples and exhibits consistent performance gains with increasing shots across all three evaluation splits. This suggests that incorporating foundation-model priors into the training process enables more effective task adaptation as supervision increases. Overall, LayerFi-FSOD maintains competitive accuracy among foundation-model-based methods while

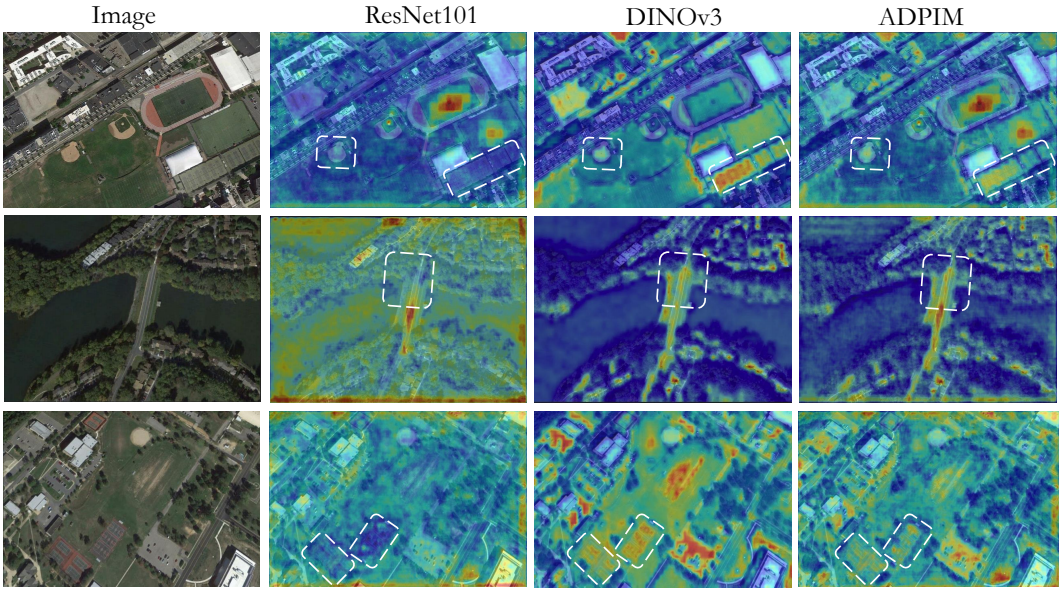

Figure 3: Visualization of feature responses after ADPIM injection. We compare the input images, feature responses from the ResNet-101 backbone and DINOv3, and the resulting LayerFi-FSOD representations after ADPIM injection.

requiring only the student detector at inference. Additional qualitative comparisons are provided in Appendix A.2.

4.2 Ablation Studies

Component Ablation. As shown in Table 3, both ADPIM and DSGM improve the student baseline individually, while their joint use achieves the best performance. Compared with the baseline, ADPIM improves mAP_{50} by 2.92 and 2.13 percentage points under the 5-shot and 10-shot settings, respectively, whereas DSGM brings gains of 7.54 and 1.48 points. Combining both modules raises the gains to 8.54 and 2.30 points, indicating that both visual-prior injection and subsequent prior absorption contribute to the final detector. Table 4 further examines the key design choices. For ADPIM, directly concatenating multi-layer features or replacing PLEM with fixed or globally learnable routing weights underperforms the full model, showing the importance of adapting DINOv3 representations to different FPN scales. Removing CAFM also degrades performance across all shot settings, particularly under 5-shot, suggesting that CLS-aware modulation helps emphasize useful object-related responses before multi-layer aggregation. For DSGM, removing either GSAD or LSSD results in lower performance. Removing LSSD causes decreases of 7.24 and 7.13 points under the 3-shot and 5-shot settings, respectively, indicating that local structural guidance is especially beneficial under limited supervision. GSAD provides additional gains by transferring proposal-level semantic relations, particularly under 5-shot and 10-shot. Overall, these results validate the key designs for internalizing vision foundation model priors.

Layer Selection. We further study the choice of DINOv3 layers in Table 5. Using only the final layer yields 60.86% and 62.55% mAP_{50} under the 3-shot and 5-shot settings, while using all candidate layers in both ADPIM and GSDM yields 60.58% and 65.88%. Both are inferior to assigning layers $\{3, 6, 9, 12\}$ to ADPIM and $\{15, 18, 21, 24\}$ to GSDM, which yields 63.46% and 67.94%, respectively. These results show that simply using deeper features or more layers is not sufficient. Instead, assigning different network depths according to their representation characteristics is more effective. This is consistent with our layer-wise analysis, where low-to-mid layers exhibit stronger spatial variation, while mid-to-high layers show stronger alignment between patch responses and object regions. The hyperparameter settings and corresponding analysis are provided in Appendix A.3.

Table 3: Ablation study of ADPIM and DSGM on NWPU VHR-10. Results are reported as mAP_{50} on the novel classes under the 5-shot and 10-shot settings. The best results are highlighted in bold.

	ADPIM	DSGM	5-shot	10-shot
I	×	×	59.40	71.02
II	✓	×	62.32	73.15
III	×	✓	66.94	72.50
IV	✓	✓	**67.94**	**73.32**

Table 4: Ablation study of different design choices in LayerFi-FSOD on NWPU VHR-10. Results are reported as mAP_{50} on the novel classes. The best results are highlighted in bold.

Component	Variant	3-shot	5-shot	10-shot	20-shot
ADPIM	Direct Concatenation	56.28	61.63	72.17	84.69
	Learnable Routing Weights	58.63	61.87	72.18	84.78
	Fixed Routing Weights	58.07	62.48	71.84	83.98
	w/o CAFM	63.11	62.56	72.97	85.13
DSGM	w/o GSAD	62.64	64.83	71.65	85.57
	w/o LSSD	56.22	60.81	73.10	84.29
LayerFi-FSOD	–	**63.46**	**67.94**	**73.32**	**86.06**

Table 5: Effect of different DINOv3 layer selections for ADPIM and GSDM. Results are reported as mAP_{50} on the novel classes of NWPU VHR-10. The best results are highlighted in bold.

DINOv3 Layers for ADPIM	DINOv3 Layers for GSDM	3-shot	5-shot
$\{24\}$	$\{24\}$	60.86	62.55
$\{9\}$	$\{21\}$	61.64	64.99
$\{9, 12\}$	$\{21, 24\}$	55.08	62.58
$\{3, 6, 9\}$	$\{12, 15, 18, 21, 24\}$	59.35	63.54
$\mathbf{\{3, 6, 9, 12\}}$	$\mathbf{\{15, 18, 21, 24\}}$	**63.46**	**67.94**
$\{3, 6, 9, 12, 15\}$	$\{18, 21, 24\}$	57.70	61.76
$\{3, 6, 9, 12, 15, 18, 21, 24\}$	$\{3, 6, 9, 12, 15, 18, 21, 24\}$	60.58	65.88

Table 6: Inference efficiency comparison on NWPU VHR-10. We compare different methods in terms of DINO usage and inference efficiency. ↑ indicates that higher is better, while ↓ indicates that lower is better.

Method	DINO usage		Inference Efficiency		
	Train	Test	Params (M) ↓	GFLOPs ↓	FPS ↑
SAE-FSDet	×	×	**60.89**	**181.13**	**28.06**
No-Time-To-Train	×	✓	528.80	5699.40	1.36
FSOD-VFM	×	✓	746.45	6078.86	1.08
LayerFi-FSOD w. DINOv3	✓	✓	368.31	3731.86	2.74
LayerFi-FSOD	✓	×	**60.89**	**181.13**	**28.06**

4.3 Comprehensive Analysis

Inference efficiency under different foundation-model utilization strategies is compared in Table 6. Methods that retain pretrained visual models during inference incur substantially higher computational costs. For example, No-Time-To-Train (Espinosa et al., 2026) and FSOD-VFM (Feng et al., 2026) require 5699.40 and 6078.86 GFLOPs, with throughputs of only 1.36 and 1.08 FPS, respectively. Similarly, retaining DINOv3 in LayerFi-FSOD increases the computation from 181.13 to 3731.86 GFLOPs and reduces throughput from 28.06 to 2.74 FPS. In contrast, after removing DINOv3, LayerFi-FSOD retains the same parameter count, GFLOPs, and FPS as SAE-FSDet while achieving stronger detection performance, demonstrating a favorable accuracy–efficiency trade-off. A more detailed performance–efficiency analysis is provided in Appendix A.4.

To examine the effect of ADPIM, Fig. 3 visualizes the feature responses of ResNet-101 (He et al., 2016), DINOv3, and LayerFi-FSOD. Compared with the relatively weak or diffuse responses of ResNet-101, DINOv3 provides stronger object-related activations, while ADPIM produces more concentrated target responses with higher contrast against the background. This indicates that ADPIM effectively exploits DINOv3 dense visual priors to enhance object-aware representations. Appendices A.5 and A.6 analyze localization quality and CLS-aware feature modulation, respectively.

5 Conclusion

This work addresses the limited supervision and insufficient utilization of foundation-model priors in few-shot remote sensing object detection by introducing LayerFi-FSOD. Through a layer-wise analysis of DINOv3, a representative vision foundation model, we identify clear depth-dependent differences: low-to-mid layers exhibit stronger spatial variation in dense representations, while mid-to-high layers show stronger alignment between patch responses and object regions. Based on these findings, LayerFi-FSOD exploits visual priors according to their representation characteristics across network depths and promotes their absorption into the student detector, alleviating the insufficient utilization of hierarchical foundation-model knowledge, while eliminating the need for the vision foundation model during inference. Extensive experiments validate the effectiveness of LayerFi-FSOD across different few-shot settings. In future work, we plan to investigate the applicability of this layer-aware prior utilization paradigm to broader vision foundation models and remote sensing tasks.

AI USE STATEMENT

In this work, we used generative AI tools for language refinement, including improving English expression, grammar, sentence fluency, and overall readability. We reviewed all AI-assisted revisions before incorporating them into the manuscript and take responsibility for the final content of this work.

ETHICS STATEMENT

This work follows standard research ethics. All compared methods and related works are appropriately cited. The datasets used in this study are publicly available and have been properly referenced. We have ensured that our data collection and processing procedures, as well as the release of our code, comply with applicable privacy and intellectual property laws.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we provide detailed descriptions of the experimental setup, methodology, and results, including the algorithms, model architecture, hyperparameter settings, and training procedures used in our experiments. Upon publication, we will release the source code together with all necessary scripts and configuration files required to reproduce the reported results.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

Jie Chen, Dengda Qin, Dongyang Hou, Jun Zhang, Min Deng, and Geng Sun. Multiscale object contrastive learning-derived few-shot object detection in vhr imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. doi: 10.1109/TGRS.2022.3229041.

Gong Cheng, Junwei Han, Peicheng Zhou, and Lei Guo. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98:119–132, 2014. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2014.10.002.

Gong Cheng, Bowei Yan, Peizhen Shi, Ke Li, Xiwen Yao, Lei Guo, and Junwei Han. Prototype-cnn for few-shot object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–10, 2022. doi: 10.1109/TGRS.2021.3078507.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. Insid3: Training-free in-context segmentation with dinov3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21638–21648, June 2026.

Miguel Espinosa, Chenhongyi Yang, Linus Ericsson, Steven McDonagh, and Elliot J. Crowley. No time to train! training-free reference-based instance segmentation, 2026.

Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4012–4021, 2020. doi: 10.1109/CVPR42600.2020.00407.

Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized few-shot object detection without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4527–4536, June 2021.

Ryan Faulkenberry and Saurabh Prasad. Dino soars: Dinov3 for open-vocabulary semantic segmentation of remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 7781–7791, June 2026.

Chen-Bin Feng, Youyang Sha, Longfei Liu, Yongjun YU, Chi Man VONG, Xuanlong Yu, and Xi SHEN. FSOD-VFM: Few-shot object detection with vision foundation models and graph diffusion. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qiu, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. In *European conference on computer vision*, pp. 247–264. Springer, 2024.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Univad: A training-free unified model for few-shot visual anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15194–15203, 2025. doi: 10.1109/CVPR52734.2025.01415.

Zuan Gu, Tianhan Gao, and Langxu Zhao. Zods-rs – zero-training oriented detection & segmentation for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 6322–6330, June 2026.

Karim Guirguis, Johannes Meier, George Eskandar, Matthias Kayser, Bin Yang, and Jürgen Beyerer. Niff: Alleviating forgetting in generalized few-shot object detection via neural instance feature forging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24193–24202, June 2023.

Guangxing Han and Ser-Nam Lim. Few-shot object detection with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28608–28618, June 2024.

Guangxing Han, Shiyuan Huang, Jiawei Ma, Yicheng He, and Shih-Fu Chang. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 780–789, 2022.

Jiaming Han, Yuqiang Ren, Jian Ding, Ke Yan, and Gui-Song Xia. Few-shot object detection via variational feature aggregation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i1.25153.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Hengchao Hu, Aobo Li, Jinjian Wu, Jie Feng, and Yaoqiang Jia. Scg-fsod: Semantic correlation-guided few-shot object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–15, 2026. doi: 10.1109/TGRS.2025.3645045.

Yuyu Jia, Chenchen Sun, Junyu Gao, and Qi Wang. Few-shot remote sensing scene classification via parameter-free attention and region matching. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227:265–275, 2025. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2025.05.026.

Xinyu Jiang, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Duoqian Miao. Few-shot object detection via improved classification features. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5375–5384, 2023. doi: 10.1109/WACV56688.2023.00535.

Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8419–8428, 2019. doi: 10.1109/ICCV.2019.00851.

Jianxiang Li, Yan Tian, Yiping Xu, Xinli Hu, Zili Zhang, Hu Wang, and Yiwen Xiao. Mm-rcnn: Toward few-shot object detection in remote sensing images with meta memory. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022a. doi: 10.1109/TGRS.2022.3228612.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2019.11.023.

Lingjun Li, Xiwen Yao, Gong Cheng, Mingliang Xu, Jungong Han, and Junwei Han. Solo-to-collaborative dual-attention network for one-shot object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022b. doi: 10.1109/TGRS.2021.3091003.

Wuzhou Li, Jiawei Zhou, Xiang Li, Yi Cao, Guang Jin, and Xuemin Zhang. Infrs: Incremental few-shot object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3475482.

Yan Li, Mingzhe Hao, Jiaman Ma, Amirkhan Temirbayev, Ying Li, Shijian Lu, Changjing Shang, and Qiang Shen. Hpmf: Hypergraph-guided prototype mining framework for few-shot object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025a. doi: 10.1109/TGRS.2025.3613849.

Yu Li, Xingyu Qiu, Yuqian Fu, Jie Chen, Tianwen Qian, Xu Zheng, Danda Pani Paudel, Yanwei Fu, Xuanjing Huang, Luc Van Gool, and Yu-Gang Jiang. Domain-RAG: Retrieval-guided compositional image generation for cross-domain few-shot object detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Hui Lin, Nan Li, Pengjuan Yao, Kexin Dong, Yuhan Guo, Danfeng Hong, Ying Zhang, and Congcong Wen. Generalization-enhanced few-shot object detection in remote sensing. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(6):5445–5460, 2025. doi: 10.1109/TCSVT.2025.3528262.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017. doi: 10.1109/CVPR.2017.106.

Changhan Liu, Xunzhi Xiang, Zixuan Duan, Wenbin Li, Qi Fan, and Yang Gao. DON't NEED RE-TRAINING: A mixture of DETR and vision foundation models for cross-domain few-shot object detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yanxing Liu, Zongxu Pan, Jianwei Yang, Bingchen Zhang, Guangyao Zhou, Yuxin Hu, and Qixiang Ye. Few-shot object detection in remote-sensing images via label-consistent classifier and gradual regression. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3369666.

Xiaonan Lu, Wenhui Diao, Yongqiang Mao, Junxi Li, Peijin Wang, Xian Sun, and Kun Fu. Breaking immutable: information-coupled prototype elaboration for few-shot object detection. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i2.25274.

Wenping Ma, Na Li, Hao Zhu, Licheng Jiao, Xu Tang, Yuwei Guo, and Biao Hou. Feature split–merge–enhancement network for remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022. doi: 10.1109/TGRS.2022.3140856.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification.

Keon-Hee Park, Kyungwoo Song, and Gyeong-Moon Park. Pre-trained vision and language transformers are few-shot incremental learners. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23881–23890, 2024. doi: 10.1109/CVPR52733.2024.02254.

Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8661–8670, 2021. doi: 10.1109/ICCV48922.2021.00856.

Yuan Qin, Chanling Pan, Jinyun Chen, Ruibo Chen, Jiaxing Chen, and Ruichao Qu. Mmdino: Dinov3-based universal framework for uni and multimodal remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–12, 2026. doi: 10.1109/TGRS.2026.3677346.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel HAZIZA, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. Featured Certification.

Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7352–7362, June 2021.

Chenchen Sun, Yuyu Jia, Han Han, Qiang Li, and Qi Wang. A semantic-guided framework for few-shot remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025. doi: 10.1109/TGRS.2025.3592484.

Jian Wang, Zeya Zhao, Jiang Shao, Xiaochun Zou, and Xinbo Zhao. Enhanced few-shot object detection for remote sensing images based on target characteristics. *Engineering Applications of Artificial Intelligence*, 148:110315, 2025. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2025.110315.

Lefan Wang, Shaohui Mei, Yi Wang, Jiawei Lian, Zonghao Han, and Xiaoning Chen. Few-shot object detection with multilevel information interaction for optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3410308.

Tao Wang, Xiaopeng Zhang, Li Yuan, and Jiashi Feng. Few-shot adaptive faster r-cnn. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7166–7175, 2019. doi: 10.1109/CVPR.2019.00734.

Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9919–9928. PMLR, 13–18 Jul 2020.

Stefan Wolf, Jonas Meier, Lars Sommer, and Jürgen Beyerer. Double head predictor based few-shot object detection for aerial imagery. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 721–731, 2021. doi: 10.1109/ICCVW54120.2021.00086.

Jiashan Wu, Chunbo Lang, Gong Cheng, Xingxing Xie, and Junwei Han. Retentive compensation and personality filtering for few-shot remote sensing object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):5805–5817, 2024. doi: 10.1109/TCSVT.2024.3367168.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 192–210, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58520-4.

Wei Xu, Gu Geng, Di Yuan, Rui Chen, and Qiao Liu. *DINOTrack: Leverage Differential Attention for Noise-Aware Visual Tracking with DINOv3*, pp. 1450–1458. Association for Computing Machinery, New York, NY, USA, 2026. ISBN 9798400726170.

Yang Xu, Jiahui Qin, Tianming Zhan, Huapeng Wu, Zhihui Wei, and Zebin Wu. Few-shot object detection in remote sensing images via dynamic adversarial contrastive-driven semantic-visual fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025. doi: 10.1109/TGRS.2025.3602969.

Bowei Yan, Gong Cheng, Chunbo Lang, Zhongling Huang, and Junwei Han. Global-integrated and drift-rectified imprinting for few-shot remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–11, 2025. doi: 10.1109/TGRS.2025.3546062.

Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9576–9585, 2019. doi: 10.1109/ICCV.2019.00967.

Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, Shijian Lu, and Eric P. Xing. Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12832–12843, 2023a. doi: 10.1109/TPAMI.2022.3195735.

Tianyang Zhang, Xiangrong Zhang, Peng Zhu, Xiuping Jia, Xu Tang, and Licheng Jiao. Generalized few-shot object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195:353–364, 2023b. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2022.12.004.

Xiangrong Zhang, Tianyang Zhang, Guanchun Wang, Peng Zhu, Xu Tang, Xiuping Jia, and Licheng Jiao. Remote sensing object detection meets deep learning: A metareview of challenges and advances. *IEEE Geoscience and Remote Sensing Magazine*, 11(4):8–44, 2023c. doi: 10.1109/MGRS.2023.3312347.

Xinyu Zhang, Yuhan Liu, Yuting Wang, and Abdeslam Boularias. Detect everything with few examples. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3986–4004. PMLR, 06–09 Nov 2025.

Jiawei Zhou, Wuzhou Li, Yi Cao, Hongtao Cai, Tianjin Huang, Gui-Song Xia, and Xiang Li. Few-shot oriented object detection in remote sensing images via memorable contrastive learning. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025. doi: 10.1109/TGRS.2025.3578632.

Wu zhou Li, Jia wei Zhou, Xiang Li, Yi Cao, and Guang Jin. Few-shot object detection on aerial imagery via deep metric learning and knowledge inheritance. *International Journal of Applied Earth Observation and Geoinformation*, 122:103397, 2023. ISSN 1569-8432. doi: https://doi.org/10.1016/j.jag.2023.103397.

Hao Zhu, Shuo Jin, Wenbin Liao, Jiayu Xiao, Yan Zhu, Siyue Yu, and Feng Dai. VIP: Visual-guided prompt evolution for efficient dense vision-language inference. In *Forty-third International Conference on Machine Learning*, 2026.

Zining Zhu, Peijin Wang, Wenhui Diao, Jinze Yang, Lingyu Kong, Hongqi Wang, and Xian Sun. Balancing attention to base and novel categories for few-shot object detection in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–19, 2025. doi: 10.1109/TGRS.2024.3512865.

Table 7: Base and novel class splits on DIOR and NWPU VHR-10.

Dataset	Novel Classes	Base Classes
DIOR Split 1	baseballfield, basketballcourt, bridge, chimney, ship	Remaining classes
DIOR Split 2	express service area, overpass, stadium, train station, windmill	Remaining classes
NWPU VHR-10	airplane, baseball-diamond, tennis-court	Remaining classes

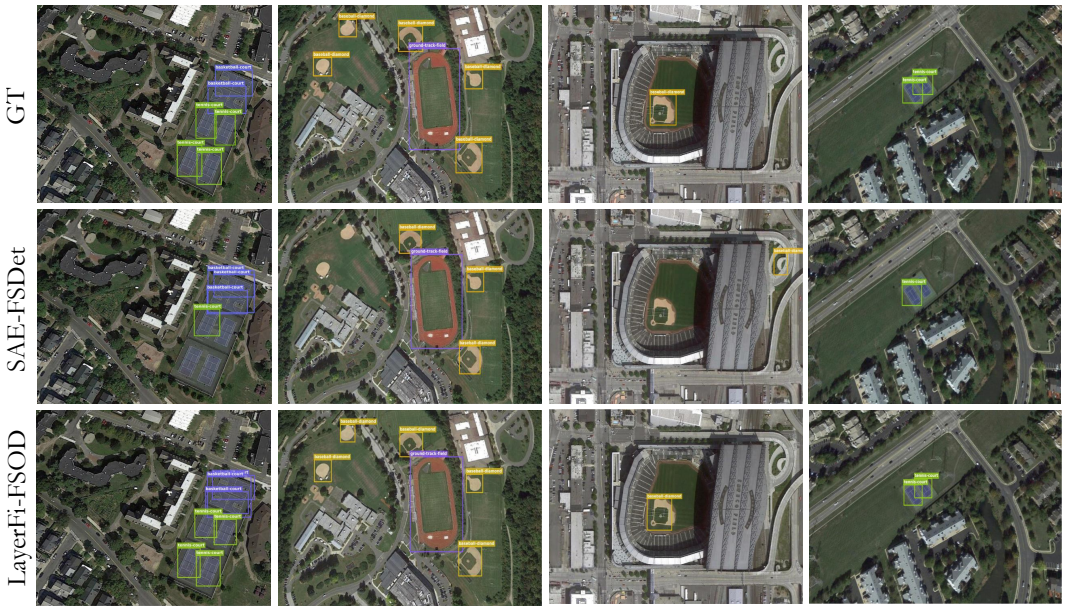

Figure 4: Qualitative comparison on NWPU VHR-10 under the 5-shot setting. Detection results of SAE-FSDet and LayerFi-FSOD are compared with the ground-truth annotations (GT).

A Appendix

A.1 Experimental Details

Datasets. We conduct experiments on two remote sensing object detection benchmarks, DIOR and NWPU VHR-10. DIOR is a large-scale remote sensing object detection dataset containing 23,463 images of size 800×800 pixels and 192,472 annotated object instances across 20 common object categories. It covers diverse object categories, complex backgrounds, and substantial scale variations, providing a challenging benchmark for evaluating detection performance in remote sensing scenes. NWPU VHR-10 is a very-high-resolution optical remote sensing dataset containing 800 images, including 650 positive images with at least one annotated object and 150 negative images without annotated objects. It covers 10 geospatial object categories, with images mainly collected from Google Earth and the ISPRS Vaihingen dataset. The two datasets differ in dataset scale, category composition, and scene complexity, allowing us to evaluate the effectiveness of LayerFi-FSOD under diverse remote sensing conditions. Following previous few-shot remote sensing object detection protocols, we adopt the standard base–novel class partitions, as summarized in Table 7.

Implementation Details. LayerFi-FSOD is built upon SAE-FSDet with a ResNet-101 backbone equipped with FPN. All experiments are conducted on a single NVIDIA A100-SXM4 GPU. The base-training stage follows the original SAE-FSDet settings, using SGD with a momentum of 0.9 and a weight decay of 1×10^{-4}. The model is trained for 18 epochs, with the learning rate set to 0.001 for the first 12 epochs and then decayed to 0.0001. During few-shot fine-tuning, the number of training iterations is adjusted according to the specific shot setting. A pretrained and frozen dinov3_vitl16_pretrain_sat493m model is introduced together with ADPIM and DSGM for visual prior utilization and absorption. DINOv3 remains frozen throughout fine-tuning and serves only as

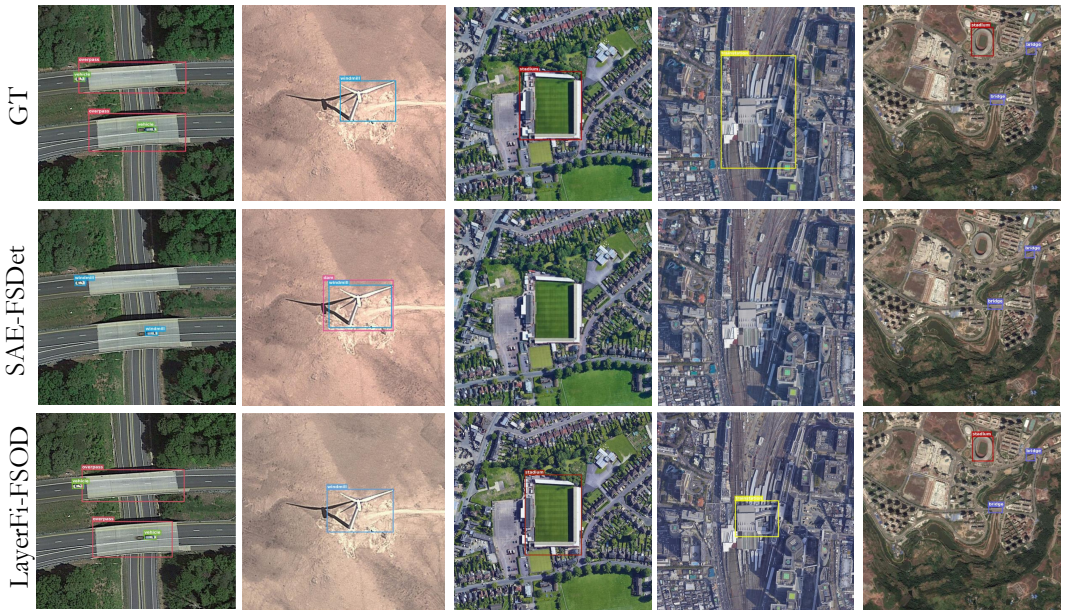

Figure 5: Qualitative comparison on DIOR Split 2 under the 3-shot setting. Detection results of SAE-FSDet and LayerFi-FSOD are compared with the ground-truth annotations (GT).

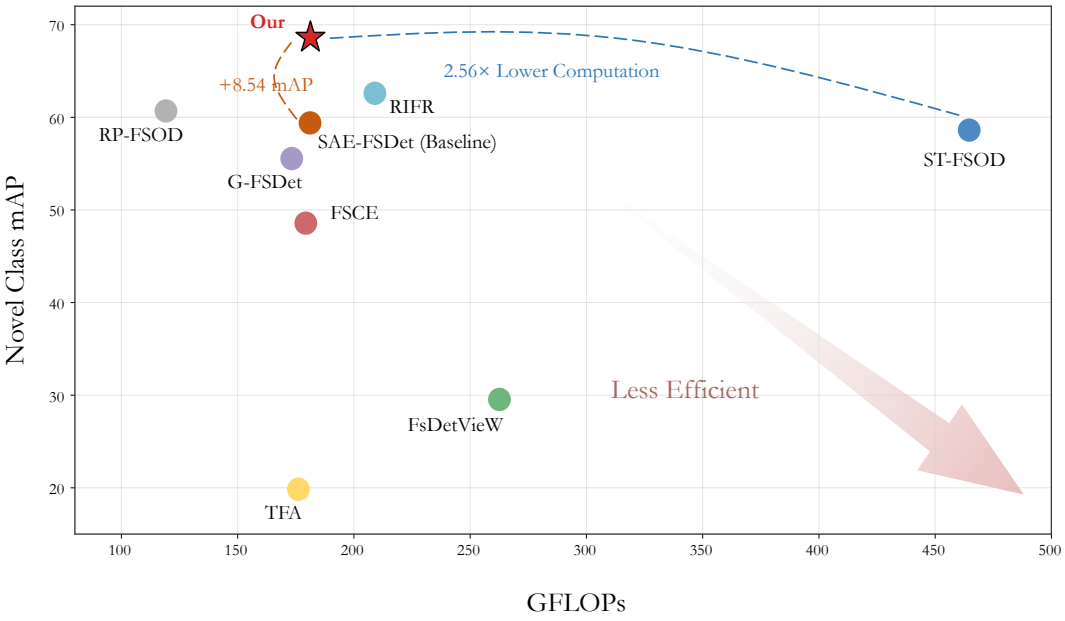

Figure 6: Performance–efficiency comparison with existing methods. We compare novel-class mAP_{50} under the 5-shot setting on NWPU VHR-10 and inference GFLOPs among different few-shot object detection methods. LayerFi-FSOD achieves a favorable trade-off between detection performance and computational efficiency.

a visual prior provider, while all DINOv3-related branches are removed after training. Consequently, the final detector retains the same student-only inference architecture as SAE-FSDet, enabling us to evaluate the benefit of foundation-model priors without introducing additional foundation-model inference cost. Evaluation follows the PASCAL VOC 2007 protocol, reporting mAP_{50} on the novel classes. The confidence and IoU thresholds for NMS are set to 0.05 and 0.5, respectively.

Table 8: Effect of the CAFM modulation coefficient λ_c. Results are reported as mAP_{50} on the novel classes of NWPU VHR-10, with the best results highlighted in bold.

λ_c	0.1	0.2	**0.3**	0.4
3-shot	62.57	63.36	**63.46**	61.47
5-shot	67.10	66.60	**67.94**	67.03

Table 9: Effect of the LSSD loss weight λ_s. Results are reported as mAP_{50} on the novel classes of NWPU VHR-10, with the best results highlighted in bold.

λ_s	0.05	**0.10**	0.15	0.20
3-shot	61.43	**63.46**	61.93	60.60
5-shot	67.11	**67.94**	66.45	65.90

A.2 Qualitative Comparisons

We provide additional qualitative comparisons between SAE-FSDet and LayerFi-FSOD on NWPU VHR-10 and DIOR. As shown in Figs. 4 and 5, LayerFi-FSOD produces more complete and accurate detection results across diverse remote sensing scenes. Compared with SAE-FSDet, our method reduces missed detections and improves object localization, particularly for challenging novel objects under limited supervision. These improvements indicate that the incorporated foundation visual priors not only strengthen novel-object recognition but also provide useful structural information for more reliable localization. The qualitative results therefore complement the quantitative comparisons in the main paper, showing that LayerFi-FSOD improves both the completeness and localization quality of novel-class detections in low-shot scenarios.

A.3 Hyperparameter Analysis

We analyze the sensitivity of LayerFi-FSOD to three key hyperparameters under the 3-shot and 5-shot settings on NWPU VHR-10. Specifically, λ_c controls the strength of CLS-aware feature modulation in CAFM, while λ_s and λ_g balance the contributions of the LSSD and GSAD objectives, respectively.

As shown in Table 8, $\lambda_c = 0.3$ achieves the best performance under both settings. Smaller values provide weaker modulation of the dense patch features, whereas increasing λ_c to 0.4 leads to a performance decrease, suggesting that excessively strong modulation may disturb the original dense representations. These results indicate that a moderate modulation strength provides a better balance between preserving the original features and emphasizing object-relevant responses.

Table 9 shows that $\lambda_s = 0.10$ performs best under both the 3-shot and 5-shot settings. Increasing the weight beyond 0.10 gradually reduces performance, indicating that local spatial structure distillation is most effective when used as an auxiliary constraint rather than dominating the detection objective. Similarly, Table 10 shows that a relatively small weight of $\lambda_g = 0.05$ is sufficient to obtain the best performance. In particular, the 5-shot performance consistently decreases as λ_g increases, suggesting that overly strong global relational constraints may interfere with task-specific adaptation under limited supervision.

Overall, the results show that LayerFi-FSOD benefits from moderate local structural guidance and relatively lightweight global relational guidance. We therefore set $\lambda_c = 0.3$, $\lambda_s = 0.10$, and $\lambda_g = 0.05$ in all experiments. In addition, the temperature coefficients are set to $\tau_c = 0.07$, $\tau_r = 2.0$, $\tau_g = 0.1$, $\tau_t = 0.07$, and $\tau_s = 0.10$.

A.4 Performance–Efficiency Comparison with Existing Methods

To evaluate the trade-off between detection performance and computational efficiency, we first compare LayerFi-FSOD with conventional few-shot remote sensing object detection methods that do not rely on vision foundation models. As shown in Fig. 6, LayerFi-FSOD is positioned in the upper-left region of the performance–efficiency space, achieving stronger novel-class detection performance

Table 10: Effect of the GSAD loss weight λ_g. Results are reported as mAP_{50} on the novel classes of NWPU VHR-10, with the best results highlighted in bold.

λ_g	**0.05**	0.10	0.15	0.20
3-shot	**63.46**	62.57	63.16	62.47
5-shot	**67.94**	66.41	65.83	64.02

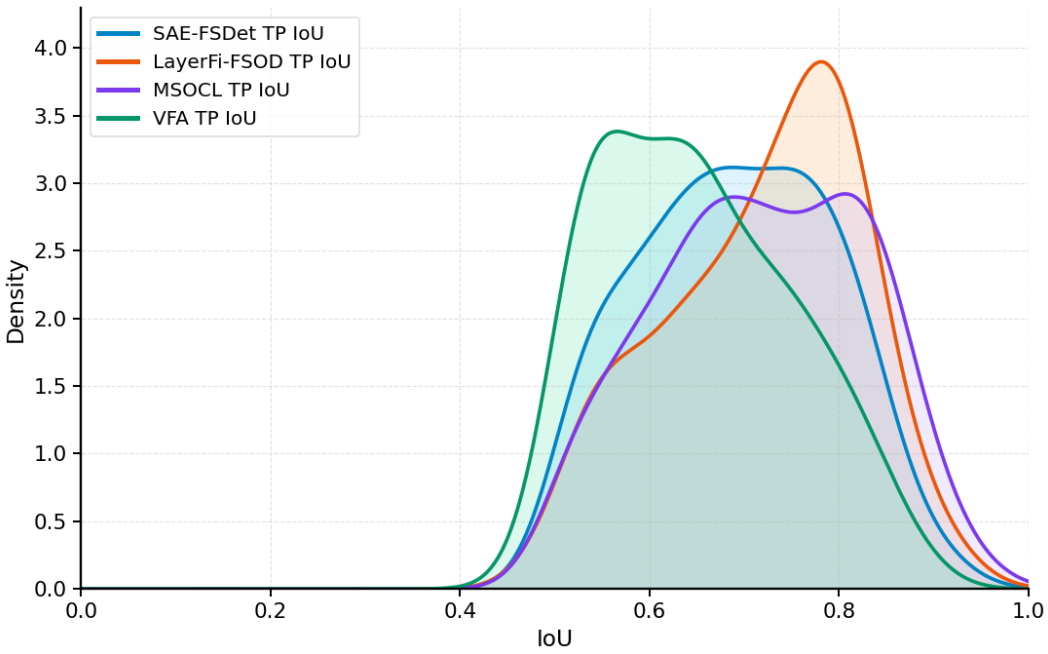

Figure 7: KDE visualization of TP IoU distributions on the novel classes of NWPU VHR-10. LayerFi-FSOD exhibits a distribution more concentrated toward the high-IoU region, indicating more accurate object localization.

with relatively low computational cost. Compared with methods of similar computational complexity, LayerFi-FSOD achieves higher detection accuracy, while methods with comparable performance generally require greater computation, indicating a favorable accuracy–efficiency trade-off.

We further compare LayerFi-FSOD with methods that leverage vision foundation models in Table 11, focusing on the computational impact of retaining pretrained visual models during inference. No-Time-To-Train (Espinosa et al., 2026) and FSOD-VFM (Feng et al., 2026) require 5699.40 and 6078.86 GFLOPs and achieve only 1.36 and 1.08 FPS, respectively. Similarly, retaining DINOv3 in LayerFi-FSOD increases the computation from 181.13 to 3731.86 GFLOPs and reduces throughput from 28.06 to 2.74 FPS, without yielding consistent accuracy gains: compared with the student-only model, performance decreases slightly under the 3-shot, 5-shot, and 10-shot settings, with only a marginal improvement under the 20-shot setting. After DINOv3 is removed, LayerFi-FSOD preserves the original inference cost of the student detector while improving mAP_{50} over SAE-FSDet by 5.50, 8.54, 2.30, and 0.98 percentage points under the 3-shot, 5-shot, 10-shot, and 20-shot settings, respectively.

Overall, LayerFi-FSOD internalizes useful visual priors into the student detector during few-shot fine-tuning and removes all foundation-model-related branches at test time. This design improves novel-class detection performance while preserving efficient student-only inference, thereby achieving a favorable balance between accuracy and computational efficiency.

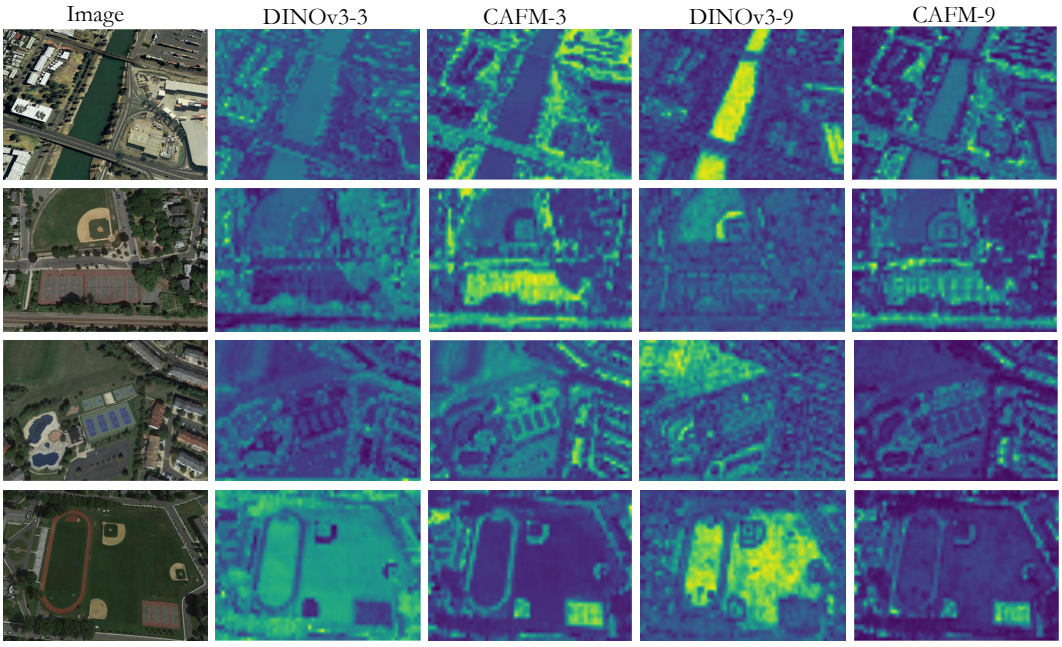

Figure 8: Visualization of CLS-aware feature modulation at different DINOv3 layers. We compare the original dense feature responses from the 3rd and 9th DINOv3 layers with their corresponding responses after CLS-aware modulation.

Table 11: **Accuracy–efficiency trade-off on NWPU VHR-10.** We compare different methods in terms of DINO usage, mAP_{50} (%), and inference efficiency. ↑ indicates that higher is better, while ↓ indicates that lower is better.

Method	DINO usage		mAP_{50} (%) ↑				Inference Efficiency		
	Train	Test	3-shot	5-shot	10-shot	20-shot	Params (M) ↓	GFLOPs ↓	FPS ↑
SAE-FSDet (Liu et al., 2024)	×	×	57.96	59.40	71.02	85.08	**60.89**	**181.13**	**28.06**
No-Time-To-Train (Espinosa et al., 2026)	×	✓	42.95	46.88	39.36	39.32	528.80	5699.40	1.36
FSOD-VFM (Feng et al., 2026)	×	✓	42.54	41.86	41.66	41.83	746.45	6078.86	1.08
LayerFi-FSOD w. DINOv3	✓	✓	63.01	67.16	72.88	**86.33**	368.31	3731.86	2.74
LayerFi-FSOD	✓	×	**63.46**	**67.94**	**73.32**	86.06	**60.89**	**181.13**	**28.06**

A.5 Localization Ability Analysis

To further investigate the localization ability of LayerFi-FSOD, we analyze the kernel density estimation (KDE) of IoU values for true-positive detections on the novel classes of NWPU VHR-10. As shown in Fig. 7, the IoU distribution of LayerFi-FSOD is more concentrated in the high-IoU region compared with SAE-FSDet (Liu et al., 2024), MSOCL (Chen et al., 2022), and VFA (Han et al., 2023), with a pronounced density around an IoU of 0.8. In contrast, the competing methods exhibit broader distributions or stronger responses in lower-IoU regions.

The shift toward higher IoU values indicates that, for correctly detected novel objects, the predicted bounding boxes of LayerFi-FSOD are more closely aligned with the ground-truth boxes. This suggests that the foundation visual priors not only benefit novel-object recognition but also contribute to more accurate spatial localization. Together with the detection results reported in the main paper, this analysis further demonstrates that LayerFi-FSOD improves both recognition and localization quality under limited novel-class supervision.

A.6 Visualization of CLS-Aware Feature Modulation

To better understand the effect of CAFM, we visualize the dense feature responses of DINOv3 before and after CLS-aware modulation at different network depths. As shown in Fig. 8, after CLS-

aware modulation, the feature responses become more concentrated around object-relevant regions, while their relative contrast against less relevant background regions is increased. This behavior is consistent with the design of CAFM, which uses the global semantic context encoded by the CLS token to reweight dense patch representations through a residual modulation mechanism. Rather than replacing the original dense features, CAFM selectively emphasizes patches that are more semantically consistent with the global context while retaining their underlying spatial information. The consistent response enhancement across different DINOv3 layers further supports the use of CLS-aware modulation before multi-layer prior aggregation in ADPIM.

A.7 Extended Related Work

Few-Shot Remote Sensing Object Detection. Few-Shot Object Detection (FSOD) aims to generalize detectors to novel categories using only a few annotated samples (zhou Li et al., 2023; Lin et al., 2025; Yan et al., 2025; Kang et al., 2019). Compared with natural images, remote sensing imagery typically contains complex backgrounds, substantial scale variations, and large intra-class appearance changes (Xia et al., 2018; Zhang et al., 2023c; Jia et al., 2025). These characteristics make it difficult to learn discriminative and transferable object representations from only a handful of labeled instances, thereby posing additional challenges to representation learning and cross-category generalization under few-shot supervision (Li et al., 2022a; Wang et al., 2025; Wolf et al., 2021).

Existing few-shot remote sensing object detection methods mainly follow transfer-learning and meta-learning paradigms. Transfer-learning-based approaches typically train a detector on well-annotated base categories and subsequently adapt it to novel categories with only a few labeled samples (Sun et al., 2021; Jiang et al., 2023; Qiao et al., 2021). This paradigm allows the detector to inherit representations learned from abundant base-category annotations, while the few-shot fine-tuning stage adapts the model to novel categories. However, because only limited novel-category supervision is available, effectively transferring previously learned knowledge while avoiding overfitting remains challenging. Representative methods address this issue from different perspectives. Retentive R-CNN (Fan et al., 2021) mitigates catastrophic forgetting through bias balancing, aiming to preserve previously learned knowledge during adaptation. RP-FSOD (Wu et al., 2024) reduces intra-class variation and inter-class confusion through prototype refinement and personality filtering, improving the discriminability of category representations. SCG-FSOD (Hu et al., 2026) explicitly models semantic correlations between base and novel categories and incorporates contrastive learning to facilitate knowledge transfer across categories. Meta-learning-based methods provide another major direction for few-shot detection. Instead of relying only on conventional fine-tuning, these methods generally adopt episodic training to simulate few-shot tasks during training and learn category-conditioned interactions between support samples and query images (Han et al., 2022; Zhang et al., 2023a). Such a formulation encourages the detector to extract transferable category information from a limited support set and use it to guide query-image detection. For instance, P-CNN (Cheng et al., 2022) enhances category discrimination through prototype learning, whereas MLII-FSOD (Wang et al., 2024) performs multi-level support–query interaction to alleviate feature misalignment between support and query representations.

Despite these advances, both transfer-learning- and meta-learning-based methods still rely primarily on task-specific annotations to learn or adapt their visual representations. Under extremely low-shot settings, the available samples may be insufficient to cover the substantial appearance, scale, and background variations commonly observed in remote sensing imagery. Consequently, the learned representations can remain limited by scarce supervision. To provide richer external visual knowledge, we introduce DINOv3 (Siméoni et al., 2026), one of the most representative vision foundation models, as an external source of visual priors and exploit its structural and semantic information during few-shot adaptation.

Vision Foundation Models for Object Detection. Vision foundation models acquire transferable and generalizable visual representations through large-scale pretraining, providing a valuable source of external knowledge for downstream tasks with limited supervision. In particular, self-supervised Vision Transformers such as DINOv2 (Oquab et al., 2024) and DINOv3 (Siméoni et al., 2026) learn dense visual representations containing rich semantic and spatial information. Their pretrained features can provide object-level semantics, local structural cues, and dense spatial responses that

are difficult to learn reliably from only a few task-specific annotations, making them increasingly attractive for object detection and related downstream tasks.

Recent studies have explored different ways of exploiting foundation-model representations for detection. DE-ViT (Zhang et al., 2025) constructs prototype-induced subspaces from frozen DINOv2 features for novel-object recognition and localization, demonstrating the value of pretrained dense representations for detecting previously unseen categories. FM-FSOD (Han & Lim, 2024) combines foundation-model visual representations with a large language model to model contextual relationships among proposals and categories, introducing additional high-level contextual knowledge into few-shot detection. VFM-DETR (Liu et al., 2025) treats representations from different network depths as visual experts and selectively integrates them through routing mechanisms, highlighting the potential value of multi-level foundation-model representations. INSID3 (Cuttano et al., 2026), in contrast, uses frozen DINOv3 dense features for multi-grained in-context matching and localization, directly leveraging pretrained visual representations for object-level correspondence. Although these methods demonstrate the effectiveness of vision foundation models for object detection, two limitations remain particularly relevant to our setting. First, existing approaches often rely on fixed-depth representations or integrate multi-layer features through a unified strategy, without explicitly accounting for the distinct representation characteristics encoded at different network depths. However, hierarchical Transformer representations evolve substantially from lower to deeper layers, suggesting that features from different depths may be better suited to different downstream objectives. Second, many foundation-model-enhanced detectors retain large pretrained visual encoders during inference (Faulkenberry & Prasad, 2026; Park et al., 2024; Gu et al., 2025). While this provides direct access to pretrained representations at test time, it also introduces additional computational and deployment costs.

Our work addresses these issues from a different perspective. Rather than treating the vision foundation model as a persistent inference component, we use it as a training-time visual prior provider. We explicitly exploit its depth-dependent representation characteristics: low-to-mid layers provide dense spatial priors, while mid-to-high layers guide proposal-level semantic relations. The resulting visual knowledge is progressively absorbed into the student detector during few-shot fine-tuning. Consequently, the vision foundation model and all prior-related branches can be removed at inference, allowing the final detector to retain a student-only inference architecture while benefiting from the visual priors acquired during training.