Prompt-Free Universal Region Proposal Network

Qihong Tang[1*], Changhan Liu[1*], Shaofeng Zhang[2], Wenbin Li[1], Qi Fan[1 ✉], Yang Gao[1]
[1]Nanjing University, [2]University of Science and Technology of China

Abstract

Identifying potential objects is critical for object recognition and analysis across various computer vision applications. Existing methods typically localize potential objects by relying on exemplar images, predefined categories, or textual descriptions. However, their reliance on image and text prompts often limits flexibility, restricting adaptability in real-world scenarios. In this paper, we introduce a novel Prompt-Free Universal Region Proposal Network (PF-RPN), which identifies potential objects without relying on external prompts. First, the Sparse Image-Aware Adapter (SIA) module performs initial localization of potential objects using a learnable query embedding dynamically updated with visual features. Next, the Cascade Self-Prompt (CSP) module identifies the remaining potential objects by leveraging the self-prompted learnable embedding, autonomously aggregating informative visual features in a cascading manner. Finally, the Centerness-Guided Query Selection (CG-QS) module facilitates the selection of high-quality query embeddings using a centerness scoring network. Our method can be optimized with limited data (e.g., 5% of MS COCO data) and applied directly to various object detection application domains for identifying potential objects without fine-tuning, such as underwater object detection, industrial defect detection, and remote sensing image object detection. Experimental results across 19 datasets validate the effectiveness of our method. Code is available at https://github.com/tangqh03/PF-RPN.

1. Introduction

Recent object detection methods with Region Proposal Networks (RPN) [12, 49] have achieved significant progress in various computer vision applications. The Region Proposal Network (RPN) generates a sparse set of proposal boxes for potential objects, which is a key component of object detection. However, existing RPN methods [34, 39, 64] often fail to identify potential target objects from unseen domains.

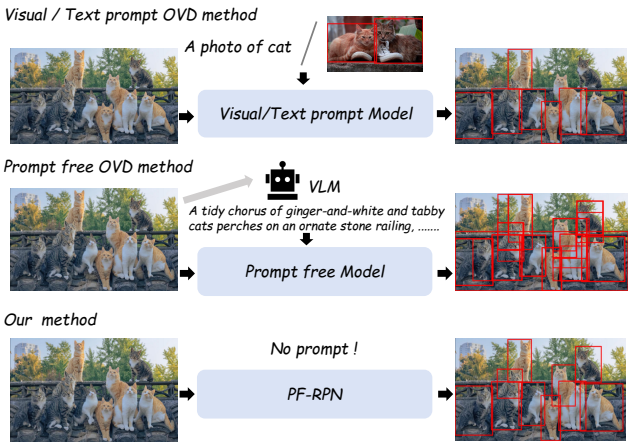

Figure 1. Existing visual/text prompt based OVD methods typically rely on predefined categories or exemplar images to propose potential objects for the target image. Recent prompt-free OVD methods often leverage VLMs to generate textual descriptions for the target image to localize potential objects which introduce considerable latency costs. In contrast, our PF-RPN doesn't require any external prompts and only utilizes visual features to generate high-quality proposals. Experimental results show the effectiveness of our PF-RPN in localizing potential objects.

This limitation significantly hinders object detection application in open-world scenarios.

Open-vocabulary object detection (OVD) models [6, 8, 11, 13, 17, 32, 42] have demonstrated impressive capabilities in localizing objects from unseen domains by leveraging category names or example images as prompts. Although OVD methods are well-suited as RPN detectors due to their strong generalization, their reliance on predefined categories and exemplar images limits flexibility in practical scenarios. For instance, in industrial defect detection and underwater object detection scenarios, the target categories and exemplar images are often unavailable, which substantially limits the application of these models. Although some prompt-free OVD models [22, 29, 45, 52] explore generative vision-language models (VLMs) to eliminate the need for manually provided prompts, they often introduce significant memory and latency costs. Therefore, it is necessary to propose an efficient region proposal network that can gen-

*Equal Contribution,
✉Correspondence to: Qi Fan <fanqi@nju.edu.cn>.

eralize across various domains without external prompts.

In this paper, we propose a novel *Prompt-Free Universal Region Proposal Network (PF-RPN)* for localizing potential objects, which can be applied to distinct unseen domains without the need for exemplar images or textual descriptions. Our model is optimized using limited data and can be directly applied to downstream tasks without requiring additional fine-tuning.

PF-RPN builds on the powerful OVD model by aggregating informative visual features through a learnable visual embedding, eliminating the need for manually provided prompts while retaining its strong generalization ability. Specifically, the learnable query embedding is initialized and updated by the proposed *Sparse Image-Aware Adapter (SIA)* module, which dynamically adjusts the embedding by selectively aggregating multi-level visual features. This adapter enables the model to capture salient visual details at various spatial resolutions, enhancing the localization of potential objects in complex visual scenes.

The SIA-adjusted learnable query embedding enables the model to identify salient objects with distinct visual appearances. However, the embedding may still struggle to capture challenging objects with unclear visual features, such as small or occluded objects. To mitigate this issue, we propose the *Cascade Self-Prompt (CSP)* module to identify the remaining challenging objects by iteratively refining the query embedding through a self-prompting mechanism. The query embedding is progressively updated by aggregating multi-scale, informative visual context, enabling the model to handle ambiguities associated with small or occluded objects more effectively.

Furthermore, we observe that query embeddings near the object center tend to generate more accurate proposals than those at the object edges. This observation motivates the design of the *Centerness-Guided Query Selection (CG-QS)* module, which selects queries based on the predicted centerness score, emphasizing the central region of objects during the query embedding selection process. Focusing on the centermost areas helps reduce false positives and improves the quality of the proposals generated by the model.

Compared with conventional and OVD-based RPN methods, our PF-RPN significantly improves proposal quality without requiring re-training or external prompts for unseen domains. Trained with limited data, PF-RPN demonstrates strong zero-shot generalization ability, achieving consistent improvements across 19 datasets spanning diverse domains and application scenarios. Specifically, PF-RPN achieves 6.0/7.5/6.6 AR improvement on CD-FSOD and 4.4/5.2/5.8 AR improvement on ODinW13 with 100/300/900 candidate boxes, respectively, substantially surpassing SOTA models. In summary, our advantages are as follows:

- We propose a novel Prompt-Free Universal Region Proposal Network (PF-RPN), a cutting-edge model which can accurately identify potential objects in practical open-world scenarios without any external prompts.
- We propose sparse image-aware adapter, cascade self-prompting and centerness-guided query selection, enabling our model to effectively retrieve potential objects by using only the visual features.
- Our PF-RPN achieves strong generalization performance with limited data (*e.g.*, 5% of COCO data) and can be directly applied to downstream tasks without additional fine-tuning. Experimental results on 19 cross-domain datasets demonstrate the effectiveness of our model.

2. Related Works

Open-Vocabulary Object Detection. Recent progress in open-vocabulary and grounded vision–language modeling [6, 8, 13, 18, 21, 27, 40, 46, 51, 55, 59, 61] has greatly improved detector generalization. GLIP [21] unifies detection and grounding for language-aware pre-training, and Grounding DINO [27] enhances open-set detection via vision–language fusion. DetCLIPv2 [51] further strengthens word–region alignment, while YOLO-World [6] and YOLOE [40] provide efficient vision–language fusion for accurate, real-time OVD. However, most methods still rely on text prompts or exemplar images for localization, limiting flexibility when external input is unavailable. Although YOLOE supports prompt-free detection, its zero-shot generalization is constrained by static text proxies. In contrast, our PF-RPN learns a visual embedding and refines it through self-prompting, removing the need for text prompts while preserving strong generalization.

Prompt-free Object Detection. Recent works [22, 29, 45, 52] explore prompt-free paradigms that generate object descriptions directly. GenerateU [22] formulates detection as a generative process that maps visual regions to free-form names, while CapDet [29] bridges detection and captioning by predicting category labels or region captions. DetCLIPv3 [52] integrates a caption head into an open-set detector and leverages auto-annotated data for pre-training. However, such models rely on large captioners, which are computationally expensive and often biased. Our PF-RPN uses a learnable embedding as a text proxy, achieving unbiased detection with low latency and memory cost.

Multimodal Large Language Models. Multimodal Large Language Models (MLLMs) extend LLMs with visual perception and reasoning. Early studies [9, 19, 37, 48, 50] focused on vision-language alignment for tasks such as captioning and VQA, while later works [1, 2, 5, 30, 41, 43, 47, 53, 54, 62] (*e.g.*, Qwen3-VL, DeepSeek-VL2) target fine-grained understanding for grounding and OCR. Despite their strong reasoning capability, MLLMs require massive computation and exhibit limited transfer to cross-domain detection. Our PF-RPN achieves comparable

zero-shot generalization without textual input or large-scale training, offering lower latency and deployment costs.

3. Method

3.1. Method Overview

Unlike existing prompt-free open-vocabulary object detection (PFOVD) methods [22, 45, 52] and OVD methods [6, 21, 27, 40], which rely on computationally expensive captioners to generate object names for image I or require manual user input of category names or exemplar images, our PF-RPN directly proposes potential objects across diverse domains without any text or visual prompts.

Fig. 2 illustrates the overall of PF-RPN. First, the image encoder (e.g., ResNet [14] or Swin Transformer [28]) extracts multi-level feature maps $F_i^I \in \mathbb{R}^{H_i \times W_i \times C}, i \in \{1, \cdots, 4\}$, where $H_i \times W_i$ denotes the spatial resolution of the i-th feature map and C is the channel dimension. Then, the *Sparse Image-Aware Adapter (SIA)* adaptively integrates the k most informative features with the learnable embedding $F^T \in \mathbb{R}^{1 \times C}$ via a routing mechanism and cross-attention. Subsequently, the *Cascade Self-Prompt (CSP)* module progressively refines F^T using feature maps from deep to shallow layers. Finally, the multi-level features F_i^I are flattened into $F^I \in \mathbb{R}^{H \times W \times C}$ and used as memory, following DETR-like frameworks [4, 27, 57, 63]. We replace the language-guided query selection in Grounding DINO [27] with our *Centerness-Guided Query Selection (CG-QS)* module to decode object proposals. The entire framework is jointly trained on classification datasets with pseudo bounding boxes and object detection datasets.

3.2. Sparse Image-Aware Adapter

Existing OVD methods [6, 21, 27, 40] mainly focus on aligning image and text features for scoring detected boxes, yet they often overlook the rich multi-level visual cues from the image encoder. Early works [25, 26] reveals that feature contributions vary across levels—shallow features are beneficial for small objects, while deeper ones capture large objects—indicating that naive fusion across all levels introduces redundancy and noise. To address this, we propose the *Sparse Image-Aware Adapter (SIA)*, a Mixture-of-Experts (MoE) module that adaptively selects and fuses the most informative feature levels with the learnable embedding F_T. Inspired by visual feature-based prompt tuning [60], SIA replaces text embeddings in pre-trained OVDs (*e.g.*, Grounding DINO [27]) with image-derived representations, bridging the modality gap.

Given the multi-level feature maps F_I^i, a global average pooling layer extracts compact features $\bar{F}_I^i \in \mathbb{R}^C$. An MoE router predicts their importance $w_i = \textit{Router}(\bar{F}_I^i)$, where *Router* is a lightweight MLP. We then select the top-k ($k \leq 4$) feature levels and normalize their weights via softmax. Finally, F_T acts as the query and the concatenated features $\left[\bar{F}_{\sigma(j)}^I, F_{\sigma(j)}^I\right]$ serve as key–value pairs in cross-attention [33, 38] to produce the updated embedding:

$$\tilde{F}^T = \sum_{j=1}^{k} \tilde{w}_{\sigma(j)} \cdot \text{Attn}\left(F_T, [\bar{F}_{\sigma(j)}^I, F_{\sigma(j)}^I]\right), \quad (1)$$

where $\sigma(j)$, $1 \leq j \leq k$ denotes the selected feature levels.

The proposed SIA module sparsely adapts multi-level visual features to the learnable embedding while maintaining consistency between object scales and feature levels. Moreover, by leveraging both global features $\bar{F}_{\sigma(j)}^I$ and local features $F_{\sigma(j)}^I$, the learnable embedding is enriched with both coarse- and fine-grained visual cues. As illustrated in Fig. 4, SIA significantly enhances the localization capability of the learnable embedding by emphasizing semantically relevant object regions and suppressing background noise. However, background activations are still observed, suggesting that a single-step adaptation is insufficient. To further refine the embedding and achieve more precise localization, we introduce the CSP module in the next section.

3.3. Cascade Self-Prompt

While the SIA module enriches the learnable embedding $\tilde{F}^T$ with scale-relevant cues and enhances its localization ability, we observe that some background regions may still be partially activated as shown in Fig. 4. This suggests that a single-step adaptation remains insufficient to fully suppress noisy responses. To further purify the embedding, we design a refinement mechanism that leverages the embedding's own visual activations.

Empirically, object-internal features exhibit stronger localization ability than the learnable embedding itself, and this finding is proved in our supplementary materials.

This motivates an iterative refinement scheme in which activated visual feature progressively guide $\tilde{F}^T$ toward more discriminative representations. Moreover, since deeper layers encode high-level semantics while shallower layers capture fine-grained structural details [24, 56], we perform the refinement in a deep-to-shallow cascade—first aggregating semantics, then integrating structure.

Based on these insights, we propose the *Cascade Self-Prompt (CSP)* module, which iteratively refines $\tilde{F}^T$ using multi-level features F_i^I. Starting from $\tilde{F}_0^T = \tilde{F}^T$, we generate a similarity mask at each level:

$$M_i = \mathbb{1}\left(\cos\left(\tilde{F}_{i-1}^T, F_i^I\right) > \delta\right), \quad (2)$$

where δ is a manually set threshold (set to 0.3), $\cos$ denotes the cosine similarity, and $\mathbb{1}$ is the indicator function. The embedding is then updated via masked average pooling:

$$\tilde{F}_i^T = \tilde{F}_{i-1}^T + \textit{MAP}(M_i, F_i^I), \quad (3)$$

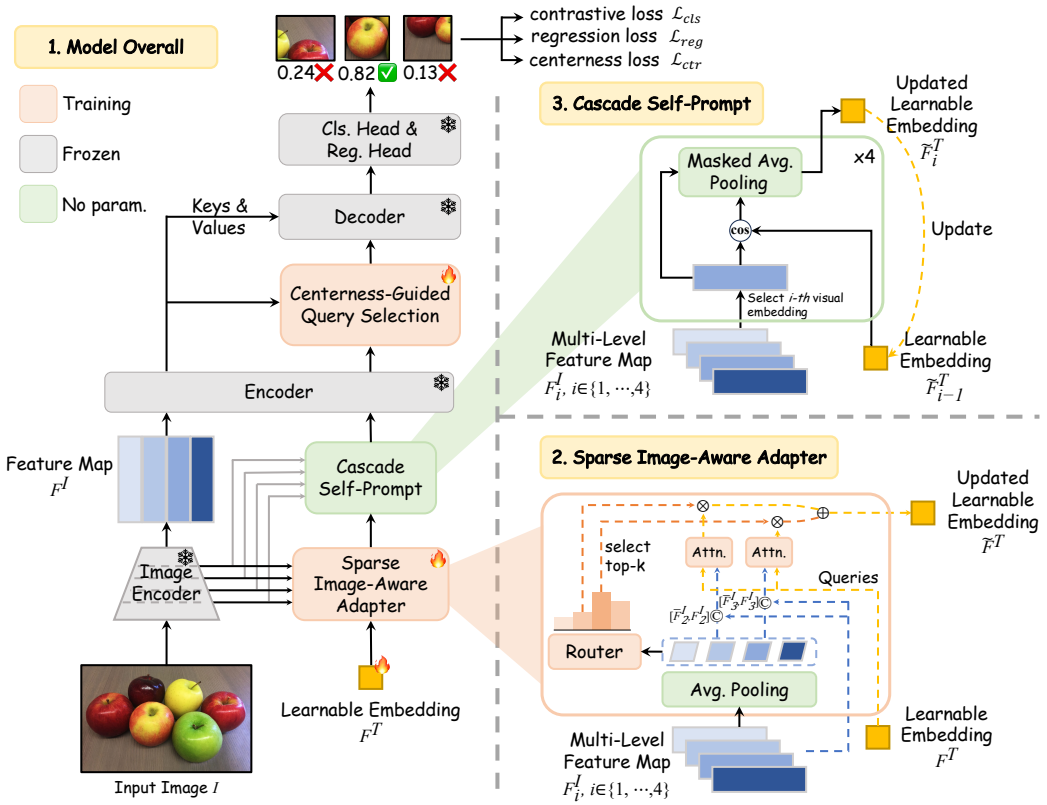

Figure 2. Overall architecture of our PF-RPN. It comprises three core components: (1) the Sparse Image-Aware Adapter (SIA) module, which adaptively integrates multi-level feature maps F_i^I with a learnable embedding F^T via a routing mechanism and cross-attention; (2) the Cascade Self-Prompt (CSP) module, which iteratively refines the embedding through masked average pooling across multiple visual levels; and (3) the Centerness-Guided Query Selection (CG-QS) module, which decodes the features into final predictions optimized by contrastive, regression, and centerness losses.

where *MAP* denotes the masked average pooling. By cascading this process from deep to shallow layers, CSP progressively expands object-consistent activations while suppressing background noise. Guided by the strong prior from SIA, the refinement jointly optimizes visual consistency and scoring reliability, yielding more precise and robust localization. Fig. 3 illustrates the effectiveness of this iterative process. To achieve an optimal balance between accuracy and efficiency, we set the number of iteration to 3.

3.4. Centerness-Guided Query Selection

After CSP module, we can localize potential object regions and obtain their corresponding queries. However, the importance of each query largely depends on its spatial location. As shown in Fig. 7, queries located near the object center tend to produce more accurate proposals than those near object boundaries. Therefore, we propose the *Centerness-Guided Query Selection (CG-QS)* module to estimate the likelihood that each query lies near the object center.

Specifically, a lightweight MLP is employed as a center scoring network to generate a center score g_i for each query f_i. Meanwhile, we compute the distances from the query to the left, right, top, and bottom edges of the corresponding ground-truth box to derive the center supervision c_i:

$$c_i = \sqrt{\frac{\min(l, r)}{\max(l, r)} \times \frac{\min(t, b)}{\max(t, b)}}. \tag{4}$$

When a query is closer to the ground-truth box center, the corresponding supervision c_i approaches 1, and the network is trained to make the predicted score g_i match c_i. The centerness loss is then defined as the L1 distance between the predicted center score g_i and its supervision c_i, $\mathcal{L}_{ctr} = \sum_{i=1}^{N} \|g_i - c_i\|_1$, where N denotes the total number of queries and $\|\cdot\|_1$ represents the L1 loss.

The proposed CG-QS module effectively prioritizes visual embeddings near object centers. During both training and inference, given classification scores computed by the dot product between the learnable embedding and the queries, we combine the center scores generated by the scoring network with these classification scores for query selection, and then use the resulting scores to determine the final candidate query set.

3.5. Objective Loss

Previous work [10] shows that the fine-tuning stage of detectors introduces bias into the image encoder, since detec-

Table 1. Comparison with OVD models, MLLMs, and RPNs. The best results are highlighted in **bold**. $AR_{100/300/900}$ denotes the average recall under 100/300/900 candidate boxes, respectively, and $AR_{s/m/l}$ denotes the average recall for small/medium/large objects. † indicates using the original class names from the corresponding dataset as text input, and ‡ indicates replacing the original class names with "object" to obtain a prompt-free setting. Our method achieves the best performance on both the CD-FSOD and ODinW13 benchmarks.

Datasets	Methods	Prompt Free	AR_{100}	AR_{300}	AR_{900}	AR_s	AR_m	AR_l
CD-FSOD	GDINO† [27]	✗	52.9	53.5	54.7	31.1	41.6	63.9
	GDINO‡ [27]	✓	54.7	57.8	61.6	34.1	49.3	67.0
	YOLOE-v8-L† [40]	✗	44.4	46.2	47.1	21.6	36.6	54.9
	YWorldv8-L† [6]	✗	49.6	51.1	51.6	25.1	42.7	60.6
	Qwen-VL† [2]	✗	20.1	20.1	20.1	1.0	3.0	26.5
	GLIP† [21]	✗	47.6	47.6	47.6	21.2	34.6	56.0
	GenerateU [22]	✓	47.7	54.1	55.7	28.1	48.3	69.4
	Open-Det [3]	✓	36.6	46.3	54.3	28.2	45.3	67.7
	RPN [34]	✓	32.0	39.0	45.7	29.9	43.0	54.3
	Cascade RPN [39]	✓	45.8	52.0	56.9	31.1	50.5	66.0
	Ours	✓	**60.7**	**65.3**	**68.2**	**38.5**	**61.9**	**80.3**
ODinW13	GDINO† [27]	✗	72.1	73.4	74.0	**45.6**	61.7	79.2
	GDINO‡ [27]	✓	69.1	70.9	72.4	40.8	64.6	78.4
	YOLOE-v8-L† [40]	✗	66.6	67.8	68.3	39.2	57.8	72.8
	YWorldv8-L† [6]	✗	69.1	70.3	71.5	37.5	62.2	75.4
	GLIP† [21]	✗	69.8	69.8	69.8	33.2	50.9	75.2
	GenerateU [22]	✓	67.3	71.5	72.2	32.8	63.1	80.0
	Open-Det [3]	✓	53.9	62.9	69.1	27.7	59.8	76.6
	RPN [34]	✓	49.0	52.4	55.7	35.3	54.0	59.8
	Cascade RPN [39]	✓	60.9	65.5	70.2	40.3	65.5	75.0
	Ours	✓	**76.5**	**78.6**	**79.8**	45.4	**71.9**	**85.8**

tion models are fine-tuned on detection datasets, whereas the image encoder is pretrained on classification datasets, *e.g.*, ImageNet [7]. To alleviate this bias, we jointly fine-tune our PF-RPN on 5% of the data from ImageNet with pseudo bounding boxes and COCO [23], thereby reducing the distribution gap between classification and detection data.

Following DETR-like frameworks [4, 21, 27, 57, 63], we employ the L1 loss and the GIoU loss [35] as the regression loss $\mathcal{L}_{reg}$, and use a contrastive loss between queries and the learnable embedding $\tilde{F}_4^T$ for classification scoring.

To prevent a few experts from being over-activated while others remain rarely used—resulting in load imbalance—we introduce an auxiliary loss $\mathcal{L}_{rt} = \text{std}(w_i), i \in \{1, \cdots, 4\}$ on the expert weights w_i to balance the load across experts and fully exploit the multi-level feature maps, where std denotes the empirical standard deviation. Minimizing $\mathcal{L}_{rt}$ encourages the expert weights w_i from the router to be more evenly distributed, improving load balance. Finally, the overall objective function is formulated as:

$$\mathcal{L} = \mathcal{L}_{reg} + \mathcal{L}_{cls} + \mathcal{L}_{rt} + \lambda \mathcal{L}_{ctr}, \tag{5}$$

where $\mathcal{L}_{reg}$ and $\mathcal{L}_{cls}$ follow the same configurations as in Grounding DINO [27].

4. Experiments

We adopt Grounding DINO [27] with a Swin-B backbone as our baseline. Our model is trained on 5% of the COCO [23] dataset (80 classes) and 5% of the ImageNet [7] dataset (1000 classes) and can be directly applied to downstream tasks without any further fine-tuning. Following previous work [21], we evaluate our model on the ODinW13 benchmark, which includes datasets from diverse domains such as wildlife photography, household objects, and aerial imagery. To further assess the generalization of our model, we also evaluate our model on the CD-FSOD benchmark, which consists of six cross-domain datasets with distinct domain shifts: ArTaxOr [31] (insect images), Clipart1k [16] (hand-drawn cartoon images), DIOR [20] (remote sensing images), DeepFish [36] (underwater fish images), NEU-DET [15] (industrial defect images), and UODD [44] (marine organism images). In our experiments, we use Average Recall (AR) as the evaluation metric to evaluate our PF-RPN's ability to propose potential objects. All experiments are conducted on four NVIDIA RTX 4090 GPUs.

4.1. Quantitative Results

Comparison with OVD Models, RPNs and MLLMs. As shown in Table 1, we compare our PF-RPN with typical open-vocabulary object detection (OVD) models. For OVD models, we feed the class names from the corresponding

Table 2. Results of module ablation studies. SIA denotes the Sparse Image-Aware Adapter, CSP denotes the Cascade Self-Prompt module and CG-QS denotes the Centerness-Guided Query Selection module. The best results are highlighted in **bold**.

SIA	CSP	CG-QS	AR_{100}	AR_{300}	AR_{900}	AR_s	AR_m	AR_l
×	×	×	52.9	53.5	54.7	31.1	41.6	63.9
✓	×	×	57.8	63.1	66.7	47.9	58.2	77.5
×	✓	×	58.1	62.9	65.8	37.7	56.7	77.0
×	×	✓	54.4	56.9	60.2	21.9	50.6	79.8
✓	×	✓	60.2	64.7	68.0	47.1	60.7	79.1
×	✓	✓	59.6	64.3	67.3	45.2	60.6	78.1
✓	✓	×	60.2	65.2	67.7	43.4	60.0	78.4
✓	✓	✓	**60.7**	**65.3**	**68.2**	**38.5**	**61.9**	**80.3**

dataset into the model to obtain detection boxes that serve as proposals. Meanwhile, to further investigate the impact of text prompts on model performance, we also evaluate its performance under the prompt-free setting by replacing the class names with "object" as the model text input. Our PF-RPN outperforms the baseline model Grounding DINO, achieving improvements of 7.8/11.8/13.5 AR on the CD-FSOD benchmark under 100/300/900 candidate boxes, respectively. On the ODinW13 benchmark, our PF-RPN further surpasses Grounding DINO by 4.4/5.2/5.8 AR under 100/300/900 candidate boxes. Compared with the OVD model YOLOE [40], our PF-RPN achieves performance gains of 16.3/19.1/21.1 AR. To further assess the generalization of our PF-RPN, we also compare it with MLLMs. Specifically, compared with Qwen2.5-VL-7B [2], our PF-RPN obtains improvements of 40.6/45.2/48.1 AR under 100/300/900 candidate boxes. In addition, compared with the Cascade RPN [39], our PF-RPN improves performance by 15.6/13.1/9.6 AR on the ODinW13 benchmark.

Module Ablation Studies. To evaluate the contribution of each module, we conduct the module ablation study on the CD-FSOD benchmark. As shown in Tab. 2, adding the SIA module raises the average performance to 57.8 AR_{100}, outperforming the baseline and indicating that visual features are more effective than text for localizing potential objects. Building on this, adding both the SIA and CSP modules further improves the performance to 60.2 AR_{100}, showing that the cascaded self-prompt strategy effectively reduces missed detections by iteratively updating the learnable embedding to retrieve more potential objects. Adding the SIA module and CG-QS modules improves performance to 59.6 AR_{100}, demonstrating that the center scoring network can accurately assess proposal quality and help the model select high-quality proposals. When combining all modules, our approach achieves the best performance of 60.7 AR_{100}, confirming the complementarity among these modules.

Data Ablation Studies. To investigate the influence of training data scale on our model, we conduct a data ablation experiment on the CD-FSOD benchmark. As shown in Tab. 3, increasing the proportion of detection data from COCO leads to consistent improvements in average recall (AR). Notably, the performance gain from using 1% to 5% of COCO is significantly larger than that from 5% to 10%, indicating diminishing returns when further expanding the data scale. Therefore, we adopt 5% of COCO as a trade-off between performance and efficiency. Furthermore, introducing classification data (ImageNet) yields an additional improvement in AR, demonstrating its effectiveness in alleviating the bias in the image encoder caused by detection-only training. This confirms that a small amount of classification data helps enhance cross-modal alignment and improves the generalization ability of our model.

Table 3. Results of data ablation studies. COCO denotes the COCO dataset and IN denotes the ImageNet dataset, The best results are highlighted in **bold**.

COCO	IN	AR_{100}	AR_{300}	AR_{900}	AR_s	AR_m	AR_l
1%	0	57.2	63.7	68.0	**42.2**	60.9	76.6
5%	0	58.8	64.1	67.8	40.2	59.0	78.6
10%	0	59.0	64.1	67.5	38.9	59.1	78.9
1%	5%	59.1	63.3	66.8	36.2	59.9	79.8
5%	5%	60.7	65.3	68.2	38.5	**61.9**	80.3
10%	5%	**60.9**	**65.5**	**68.5**	39.6	61.6	**80.5**

Table 4. Results with different backbones. The best results for each baseline are highlighted in **bold**. Our model effectively improves the performance of detectors with different backbone.

Method	AR_{100}	AR_{300}	AR_{900}	AR_s	AR_m	AR_l
GDINO (SwinB)	52.9	53.5	54.7	31.1	41.6	63.9
+ ours	**60.7**	**65.3**	**68.2**	**38.5**	**61.9**	**80.3**
GDINO (ResNet50)	40.4	43.4	44.6	21.7	32.5	54.9
+ ours	**45.6**	**51.4**	**58.0**	**28.1**	**50.6**	**73.7**

Comparison with Different Backbone. As shown in Tab. 4, the experimental results demonstrate that our model achieves strong performance across different backbone. Specifically, when integrating our model with a ResNet-50 backbone [14], the performance improves by 5.2 AR_{100}, while using a Swin-B backbone [28] leads to an improvement of 7.8 AR_{100}.

Ablation Study on MoE module in the SIA. To verify the effectiveness of the MoE module, we conduct ablation experiments as shown in Tab. 5. The removal of MoE leads to a consistent performance drop, further validating its importance. The experimental results demonstrate that while the attention mechanism can suppress irrelevant information, it operates *only within* individual feature levels and is *unable to select across levels*. In contrast, the MoE module *filters out irrelevant feature levels before* the attention stage. Since objects of different scales are best represented at distinct feature levels, relying solely on attention would inevitably introduce noise from non-informative levels.

Integrating into Well-Train Detectors. To evaluate the ex-

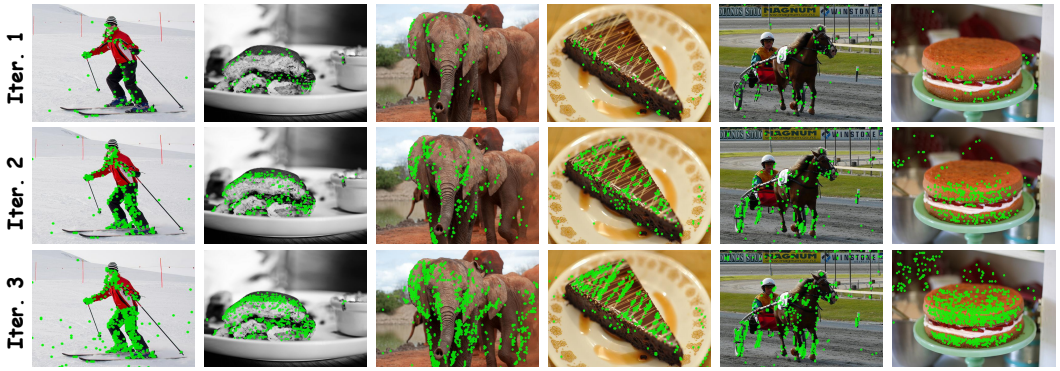

Figure 3. **Effect of iterations in the Cascade Self-Prompt module.** Visualization of region selection across different Cascade Self-Prompt iterations. Green points indicate the object regions selected by the model in the current iteration. As the number of iterations increases, the model progressively selects more object regions in the image, demonstrating the effectiveness of our cascade self-prompt mechanism.

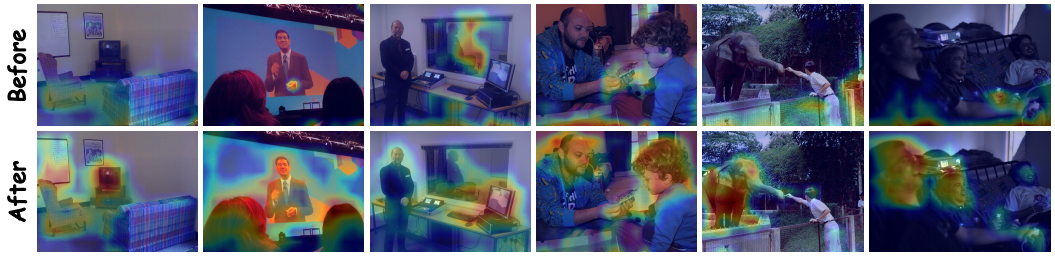

Figure 4. **Effect of the Sparse Image-Aware Adapter.** Visualization of similarity heatmaps between the learnable embedding and image features before and after the module update. Each pair of heatmaps (top: before, bottom: after) corresponds to the same image. After the update, the learnable embedding exhibits stronger responses in semantically relevant regions, indicating improved alignment between visual and learned representations and providing a stronger prior for the cascade self-prompt module.

Table 5. Ablation study on the effectiveness of the MoE module. The best results for each baseline are highlighted in **bold**.

Dataset	Method	AR_{100}	AR_{300}	AR_{900}	AR_s	AR_m	AR_l
CD-FSOD	w/o MoE	58.6	61.2	63.5	37.1	56.5	73.5
	Ours	**60.7**	**65.3**	**68.2**	**38.5**	**61.9**	**80.3**
ODinW13	w/o MoE	68.7	71.7	73.8	37.9	67.0	80.7
	Ours	**76.5**	**78.6**	**79.8**	**45.4**	**71.9**	**85.8**

tensibility of our model, we integrate our model into existing well-trained RPN-based detectors. As shown in Tab. 6, when we replace the original RPN in DE-ViT [49] with our proposed module, the detector achieves an improvement of 3.7 AP on the COCO dataset. Furthermore, to evaluate the generalization of our model in cross-domain scenarios, we integrate our model into the cross-domain detector CD-ViTO [12] and evaluate its performance on the CD-FSOD benchmark following the original setting. Experiment result shows that Integrating our model results in an improvement of 5.5 AP on the CD-FSOD benchmark.

4.2. Qualitative Visualizations

Cascade Self-Prompt Module. In the Cascade Self-Prompt (CSP) module, our key idea is to use visual features from some objects to iteratively retrieve the remaining potential objects. To validate the effectiveness of this

Table 6. Results of integrating our model into well-trained detectors. OD denotes the conventional object detection task evaluated on the COCO dataset and CDFSOD denotes the cross-domain few-shot object detection task evaluated on the CD-FSOD benchmark. The best results are highlighted in **bold**.

Method	Task	AP	AP_{50}	AR_{75}	AP_s	AP_m	AP_l
DE-VIT	OD	33.0	52.9	33.7	17.1	35.4	47.3
+ ours	OD	**36.7**	**56.2**	**38.8**	**23.5**	**40.1**	**50.7**
CD-VITO	CD-FSOD	29.6	49.1	30.0	5.7	**24.4**	37.6
+ ours	CD-FSOD	**35.1**	**59.6**	**35.0**	**9.0**	22.6	**41.9**

paradigm, we visualize the selected image regions at different iterations. As shown in Fig. 3, in the first iteration, the model can only select partial object regions. After updating the learnable embeddings with the selected features, it expands to cover more potential object regions. Through multiple iterations, the model is able to propose most potential object regions.

Sparse Image-Aware Adapter. In the Sparse Image-Aware Adapter (SIA), we dynamically adapt the learnable embedding using visual features, enabling it to propose objects from unseen categories. To evaluate the effectiveness of this module, we visualize the regions selected by the learnable embedding before and after the SIA update. As shown in Fig. 4, before the update, the learnable embedding assigns

Figure 5. **Effect of the Centerness-Guided Query Selection.** Visualization of query selection before and after applying the Centerness-Guided Query Selection (CG-QS) module. Each pair of heatmaps (top: before, bottom: after) corresponds to the same image. After applying the CG-QS module, the model tends to select queries near object centers, thereby generating more accurate proposals.

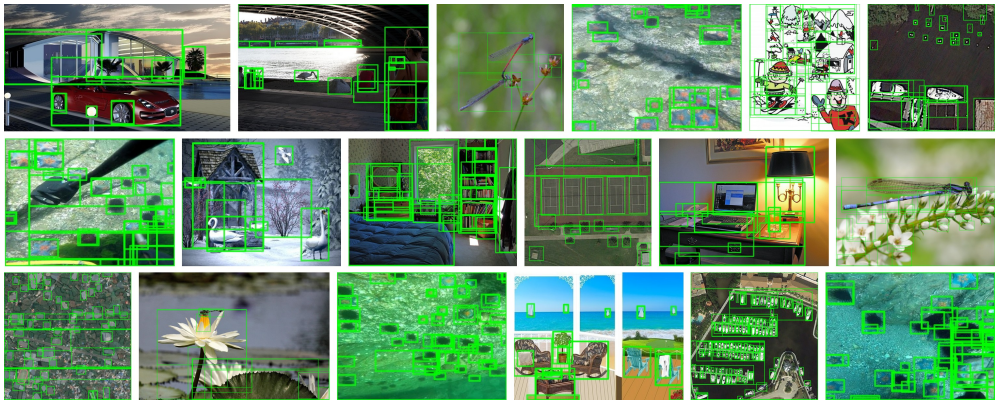

Figure 6. Qualitative results of **PF-RPN** on several object detection benchmarks. PF-RPN exhibits strong cross-domain generalization and localization ability, accurately proposing potential object regions without any domain-specific fine-tuning.

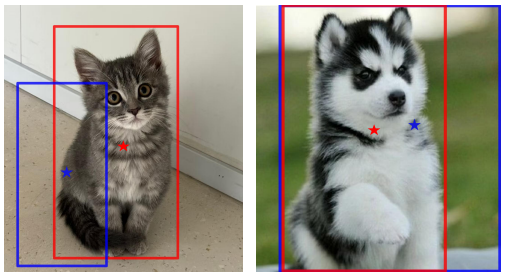

Figure 7. Prediction box comparison. The red box corresponds to the query indicated by the red star, while the blue box corresponds to the query indicated by the blue star. It can be observed that the query located near the object center produces a more accurate bounding box than queries near object boundaries.

high attention to background regions. If these regions are fed into the CSP module, the model tends to propose more background regions. In contrast, after being updated with visual features, the learnable embedding focuses on object regions and suppresses background distractions, underscoring the necessity of the SIA.

Centerness-Guided Query Selection. The core idea of CG-QS is that image queries located near an object's center tend to generate more accurate proposals than those near the object boundary. As shown in Fig. 7, when the model selects a center-area query (the red star), it produces a precise bounding box. In contrast, when the module selects queries from boundary regions, it typically results in notable localization errors. Motivated by this observation, we introduce a center loss to encourage the model to prioritize queries closer to object centers. To evaluate the effectiveness of the CG-QS strategy, we visualize the selected image queries, as shown in Fig. 5. After adding the CG-QS module, our model prefers to select queries at center locations, confirming the efficacy of the CG-QS strategy.

5. Conclusion

In this paper, we propose the Prompt-Free Universal Region Proposal Network (PF-RPN), which aims to address a critical limitation in computer vision: the task of proposing arbitrary potential objects typically depends on external prompts (*e.g.*, text descriptions or visual cues). To mitigate this limitation, PF-RPN introduces a learnable embedding serving as a proxy for text embeddings, enabling prompt-free arbitrary object proposal. We propose the Sparse Image-Aware Adapter and Cascade Self-Prompt modules, which enhance the model's localization capability as the similarity among visual embeddings is typically greater than that between the learnable embedding and visual embeddings. We further present the Centerness-Guided Query Selection module, which incorporates centerness and classification scores to select more appropriate queries for subsequent stages. Extensive experiments demonstrate our PF-RPN's superiority in zero-shot cross-domain object proposal, providing valuable insights for future research.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (62192783, 62276128, 62406140), Young Elite Scientists Sponsorship Program by China Association for Science and Technology (2023QNRC001), the Key Research and Development Program of Jiangsu Province under Grant (BE2023019) and Jiangsu Natural Science Foundation under Grant (BK20221441, BK20241200). The authors would like to thank Huawei Ascend Cloud Ecological Development Project for the support of Ascend 910 processors.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 5, 6

[3] Guiping Cao, Tao Wang, Wenjian Huang, Xiangyuan Lan, Jianguo Zhang, and Dongmei Jiang. Open-det: An efficient learning framework for open-ended detection. In *ICML*, 2025. 5, 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3, 5

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 2

[6] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *CVPR*, 2024. 1, 2, 3, 5

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[8] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 1, 2

[9] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *ACL*, 2022. 2

[10] Qi Fan, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with model calibration. In *ECCV*, 2022. 4

[11] Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *CVPR*, 2025. 1

[12] 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 *ECCV*, 2024. 1, 7, 2

[13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 1, 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 6

[15] Yibin Huang, Congying Qiu, and Kui Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 2020. 5

[16] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, 2018. 5

[17] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. In *ICML*, 2023. 1

[18] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. In *ICLR*, 2022. 2

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2

[20] 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*, 2020. 5

[21] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 2, 3, 5

[22] Chuang Lin, Yi Jiang, Lizhen Qu, Zehuan Yuan, and Jianfei Cai. Generative region-language pretraining for open-ended object detection. In *CVPR*, 2024. 1, 2, 3, 5

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3

[25] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3

[26] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018. 3

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang,

Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 2, 3, 5

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3, 6

[29] Yanxin Long, Youpeng Wen, Jianhua Han, Hang Xu, Pengzhen Ren, Wei Zhang, Shen Zhao, and Xiaodan Liang. Capdet: Unifying dense captioning and open-world detection pretraining. In *CVPR*, 2023. 1, 2

[30] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2

[31] Fatma MA Mazen. Arthropod taxonomy orders object detection in artaxor dataset using yolox. *Journal of Engineering and Applied Science*, 2023. 5

[32] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *NIPS*, 2023. 1

[33] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022. 3

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 2016. 1, 5, 2

[35] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 5

[36] Alzayat Saleh, Issam H Laradji, Dmitry A Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific reports*, 2020. 5

[37] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NIPS*, 2024. 2

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 3

[39] Thang Vu, Hyunjun Jang, Trung X Pham, and Chang Yoo. Cascade rpn: Delving into high-quality region proposal network with adaptive convolution. *NeurIPS*, 2019. 1, 5, 6, 2

[40] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything. In *ICCV*, 2025. 2, 3, 5, 6

[41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[42] Tao Wang. Learning to detect and segment for open vocabulary object detection. In *CVPR*, 2023. 1

[43] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2

[44] Yu Wei, Yi Wang, Baofeng Zhu, Chi Lin, Dan Wu, Xinwei Xue, and Ruili Wang. Underwater detection: A brief survey and a new multitask dataset. *International Journal of Network Dynamics and Intelligence*, 2024. 5

[45] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. In *ECCV*, 2024. 1, 2, 3

[46] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, 2023. 2

[47] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 2

[48] Chunyu Xie, Heng Cai, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Jianfei Song, Henrique Morimitsu, Lin Yao, Dexin Wang, Xiangzheng Zhang, et al. Ccmb: A large-scale chinese cross-modal benchmark. In *ACMMM*, 2023. 2

[49] Guanyu Xu, Zhiwei Hao, Yong Luo, Han Hu, Jianping An, and Shiwen Mao. Devit: Decomposing vision transformers for collaborative inference in edge devices. *CoRL*, 2024. 1, 7, 2

[50] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *ICCV*, 2025. 2

[51] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *CVPR*, 2023. 2

[52] Lewei Yao, Renjie Pi, Jianhua Han, Xiaodan Liang, Hang Xu, Wei Zhang, Zhenguo Li, and Dan Xu. Detclipv3: Towards versatile generative open-vocabulary object detection. In *CVPR*, 2024. 1, 2, 3

[53] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 2

[54] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025. 2

[55] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 2

[56] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 3

[57] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023. 3, 5

[58] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024. 2

[59] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 2

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 3

[61] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2

[62] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 2

[63] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 3, 5

[64] Wenbin Zou, Zhengyu Zhang, Yingqing Peng, Canqun Xiang, Shishun Tian, and Lu Zhang. Sc-rpn: A strong correlation learning framework for region proposal. *TIP*, 2021. 1