SOAR: Semi-Supervised Open-Vocabulary Aerial Object Detection via Dual-Aware Enhanced Prior Denoising

Xu Liu[1,2], **Yihong Huang**[1,2], **Dan Zhang**[1,2], **Lingling Li**[1,2*], **Long Sun**[1,2], **Licheng Jiao**[1,2]

[1] School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi, China
[2] The Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xi'an, Shaanxi, China

Abstract

Open-Vocabulary Object Detection (OVOD) shows promise in remote sensing (RS), but due to its unique value, there are challenges such as the predominance of background regions, sparse labels, limited semantic information, and difficulties in semi-supervised training. To tackle these challenges, we propose the Semi-Supervised Open-Vocabulary Aerial Object Detection with Dual-Perception Prior Denoising (SOAR), which explicitly models the background embeddings of each scene to indirectly construct foreground priors, thereby capitalizing on the abundant background information present in RS imagery. We further introduce a query enhancement module that integrates language and foreground prior information to enhance the effectiveness of query selection and feature augmentation. During the decoding stage of semi-supervised training, we perform denoising and reconstruction of the foreground priors to generate pseudo-labels that support the training process. Additionally, we address the sparsity of label information through expansion and aggregation techniques, further improving model performance. Experimental evaluations reveal that, in the open-vocabulary object detection task on the DIOR dataset, our method achieves a mean Average Precision (mAP) of 68.5% and Harmonic Mean (HM) of 55.9%, outperforming the previous state-of-the-art model's mAP of 61.6% and HM of 53.6%. Our approach offers a novel solution to the open-vocabulary challenge in aerial object detection.

Code — https://github.com/Man-PaperRejected/SOAR

1 Introduction

Aerial object detection is crucial for geospatial analysis, but traditional deep learning models (Ahmad et al. 2025; Chen et al. 2025; Li et al. 2023b; Wang et al. 2023a; Wang, Gao, and Zhang 2025; Liang and Luo 2024; Xue et al. 2024; Zhang et al. 2024; Liu et al. 2025), being inherently closed-set, struggle with novel categories. Their reliance on fixed annotations is problematic, as fine-grained aerial labeling is expertise-intensive. Consequently, aerial datasets (Xu et al. 2022; Li et al. 2020; Ding et al. 2022; Vis 2023; Lam et al. 2018) lag significantly behind general vision datasets (e.g., COCO (Lin et al. 2014), LVIS (Gupta, Dollar, and Girshick 2019)) in scale and diversity, limiting real-world adaptability.

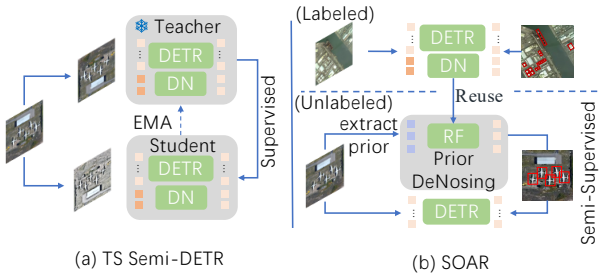

Figure 1: SOAR vs. Teacher-Student semi-supervised training architectures. (a) Teacher-Student paradigm. (b) SOAR approach: Instead of a separate teacher, SOAR reuses the denoising (DN) module as a refiner (RF) to generate supervised signals for semi-supervised training.

Vision-language models (VLMs) (Wang et al. 2023b; Li et al. 2023a; Zhang et al. 2023b; Sun et al. 2024) offer a solution by building scalable visual-semantic associations, enabling open-vocabulary object detection (OVOD) (Cheng et al. 2024; Lin et al. 2024; Yao et al. 2024; Liu et al. 2023; Zareian et al. 2020; Jia et al. 2021; Wang et al. 2024). However, applying OVOD to aerial imagery faces unique challenges compared to natural images: (1) Multi-scale objects with weak features hinder background separation (He et al. 2017; Du et al. 2022; Gu et al. 2021). (2) Complex, dominant backgrounds reduce training efficiency. (3) Prohibitive annotation costs severely restrict model performance.

Recent aerial OVOD works show promise. DescReg (Zang et al. 2024) enhances textual semantics by introducing a description regularization method. CastDet (Li et al. 2024b) employs a student-teacher framework, increasing the model's detection capability for unseen categories. LAE (Pan et al. 2025) addressed the domain gap with a large dataset (LAE-1M) and model adaptations (DVC, VisGT). Despite these efforts, critical challenges persist: (1) Effectively leveraging complex background features. (2) Optimizing the trade-off between architectural complexity and performance.

In this paper, we propose a novel semi-supervised model

*Corresponding author.

architecture, termed SOAR. As shown in Figure 1, this architecture simplifies the complex teacher-student framework (e.g., as seen in (Li et al. 2024b; Caron et al. 2021)) traditionally employed for pseudo-label generation by leveraging prior denoising to achieve this goal. Our approach is inspired by the contrastive denoising training (CDN) proposed in DINO(Zhang et al. 2022). For labeled data, the CDN is trained by constructing noisy ground truth (GT) and using the original GT as the supervision signal. For unlabeled data, we propose to dynamically and explicitly model background information during the image feature extraction process, thereby obtaining implicit foreground priors. We treat these priors as noisy GT and reuse the CDN module to denoise them, yielding pseudo-labels. We further enhance label semantics through our multi-granular label system (MGL), which optimizes the use of limited labeled data by efficiently leveraging its inherent structure and information. This approach, designed to maximize the value of scarce annotations, has been validated in the OVOD model(Liu et al. 2024) for natural images, where it improves model performance. Additionally, we introduce the dual-aware query enhancement (DAQE) module. This module first utilizes a prior-aware query selector (PAQS) to select proposal queries exhibiting a stronger correlation with the foreground. Subsequently, it employs a language-aware query enhancer (LAQE) to infuse these selected queries with language information, thereby enabling more efficient completion of the open-vocabulary object detection (OVOD) task. The contributions of our work can be summarized as follows:

1. A novel semi-supervised approach is proposed for open-vocabulary object detection in aerial imagery, named SOAR. It innovatively uses the model's internal mechanisms and extracted priors for self-consistent pseudo-label generation.
2. To inject positional priors and linguistic information into the queries, we propose the dual-aware query enhancement (DAQE) Module.
3. We introduce an offline multi-granularity label expansion module (MGL) to address the limited semantic information in aerial image category labels.

2 Related Work

2.1 Open-Vocabulary Object Detection

Traditional detectors (e.g., Faster R-CNN (Ren et al. 2015), YOLO (Bochkovskiy, Wang, and Liao 2020; Wang, Yeh, and Liao 2024; Alif and Hussain 2025), RetinaNet (Lin et al. 2017)) excel in closed-set scenarios but struggle with novel objects unseen during training. While early attempts like Zero-Shot Detection (ZSD) (Bansal et al. 2018; Rahman, Khan, and Porikli 2018; Huang et al. 2022; Demirel, Cinbis, and Ikizler-Cinbis 2018) used semantic embeddings, their performance is often constrained.

The rise of vision-language models (VLMs) like CLIP (Radford et al. 2021), BLIP-V2 (Li et al. 2023a), RAM (Zhang et al. 2023b), and BEIT-3 (Wang et al. 2023b) significantly advanced open-vocabulary learning through effective vision-language alignment. This spurred open-vocabulary object detection (OVOD), enabling detection from free-form text. Initial OVOD works (e.g., OV-RCNN (Zareian et al. 2020), ViLD (Gu et al. 2021)) primarily distilled knowledge from VLMs. Subsequent methods refined localization and generalization via techniques like learnable prompts, PromptDet (Feng et al. 2022), and improved region-text alignment (DetPro (Du et al. 2022)). Recent advancements include real-time OVOD (YOLO-World (Cheng et al. 2024)) and generative methods (Lin et al. 2024).

In aerial imagery, OVOD methods like DescReg (Zang et al. 2024) enhanced semantics, CastDet (Li et al. 2024b) employed a multi-teacher framework, and LAE (Pan et al. 2025) tackled the domain gap with a large dataset and model adaptations. Our work differs by proposing a semi-supervised framework specifically for aerial OVOD with limited labeled data. Instead of complex distillation (Li et al. 2024b) or direct VLM adaptation (Pan et al. 2025).

2.2 Semi-Supervised Object Detection

Semi-supervised object detection (SSOD) enhances performance using limited labeled and abundant unlabeled data. Pseudo-labeling is a dominant strategy, with techniques refining predictions via fixed thresholds (STAC (Sohn et al. 2020)), reliability estimation (Soft Teacher (Xu et al. 2021)), or addressing class imbalance (Unbiased Teacher (Liu et al. 2021)). SSOD has been applied to both two-stage (Tang et al. 2021; feng Zhou et al. 2021), which often uses sophisticated teachers and strong augmentation for better pseudo-labels, and one-stage (Sohn et al. 2020; Xu et al. 2023) detectors, as well as DETR architectures (Wang et al. 2022; Zhang et al. 2023a).

3 Method

3.1 Dynamic Prior Representation

We propose dynamically constructing scene foreground priors to address this. While inspired by recent work in background utilization(e.g. (Du et al. 2022; Li et al. 2024a)), it argue that directly applying global self-attention for a unified background embedding is suboptimal due to the spatial heterogeneity and semantic diversity of remote sensing backgrounds. Unlike natural images, remote sensing images have larger background regions, making per-image background modeling feasible.

Dynamic Background Representation: To accomplish dynamic implicit background modeling, we need to address the following problems step by step. The first question: **how to model the variable and complex background?** A previous method(Du et al. 2022) attempted the same, but concluded that directly and rigidly modeling the background, and representing the variable background with fixed feature representations, is difficult to converge, especially for aerial imagery. Therefore, we propose to model the background implicitly, by leveraging the existing features. Given that the background predominantly occupies remote sensing images, this approach appears feasible. Then, we must solve another problem: **how to model background information from existing features?** A straightforward approach would be to

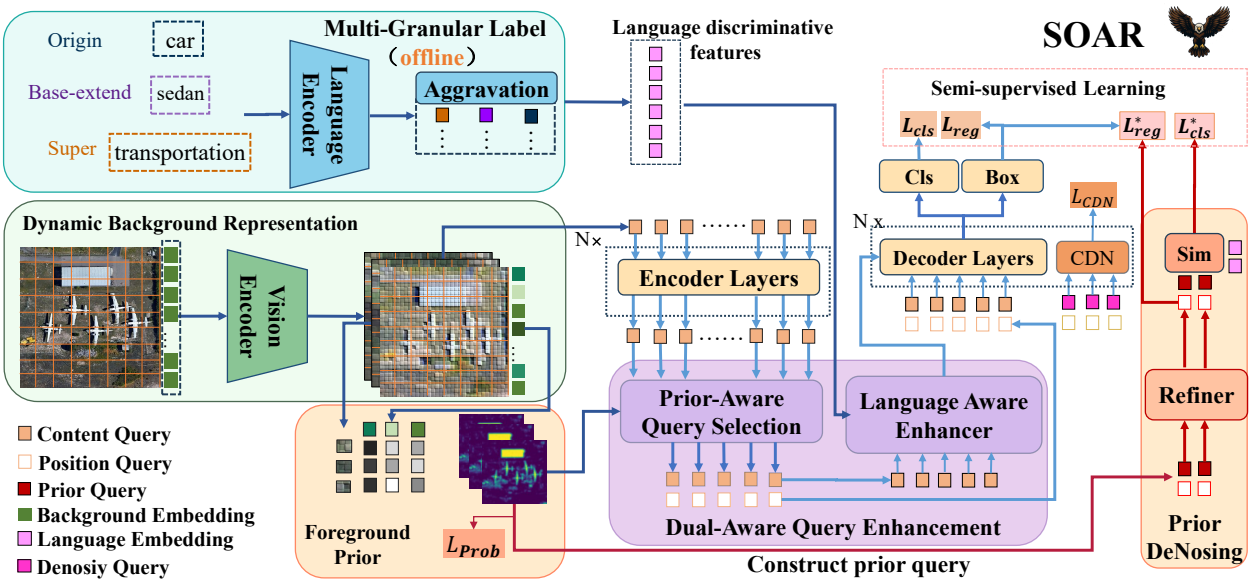

Figure 2: Overall architecture of SOAR. It consists of four main components: (1) Multi-Granular Label: Multi-granularity label generation and aggregation, expanding semantic information. (2) Dynamic Background Representation: Image feature extraction and background feature construction. (3) Dual-Aware Query Enhancement: Proposal query filtering and feature enhancement using prior location and language information. (4) Foreground prior construction and denoising: Indirect foreground prior construction via background features, prior box extraction, and refiner-based prior denoising to generate pseudo-labels. Red arrows indicate the semi-supervised training flow. Losses with a superscript * are semi-supervised training losses.

directly use the mean of global tokens as background features. After experimentation, we found that using this feature to calculate a similarity matrix with global tokens and performing histogram statistics revealed a small overall variance. This implies that it cannot effectively distinguish between background and foreground. Hence, we propose to introduce learnable tokens. While extracting image features, we allow the model to autonomously summarize and refine background information. Finally, we are left with one last question: **how many background tokens should we set, and how should we initialize them?** To initialize our nine background tokens, we compared three methods: random, global average, and regional average. The latter involves dividing the image into a 3x3 grid and initializing each token with the mean features of its corresponding region. Experiments showed this approach delivered optimal results for both foreground prior map construction. We also designed an attention mask to ensure each background token attends exclusively to its region, preventing feature leakage. Algorithm 1 illustrates its process.

Foreground Prior Maps: After performing image feature extraction and background modeling, we compute the similarity between the multi-scale image features and the background features. Subsequently, an inverse operation is applied to this similarity to obtain multi-scale foreground priors. For the i-th scale Foreground Prior:

$$\text{Prior}^i_{fg} = Merge(Prior^i_w), \forall w \in W \quad (1)$$

$$\text{Prior}^i_w = -SIM(b^i_w, Img^i(x,y)_w), \forall x,y \in Area_w, \quad (2)$$

where $Prior^i_{fg}$ denotes the *i-th* scale foreground prior,$Prior^i_w$ is the prior for the *w-th* region (scale *i*), b^i_w represents the background features for region *w* (scale *i*), SIM is the cosine similarity, and Merge aggregates the region priors $Prior^i_w$ to a complete prior map.

In our background modeling approach, adhering to implicit learning principles, we eschew direct background supervision. Instead, we utilize bounding box (Bbox) labels to generate foreground-centric supervision signals. These signals are created by constructing Gaussian responses within the box extents.

Gaussian Response: The Prior^i_{gt} is constructed by superimposing Gaussian responses:

$$\text{Prior}^i_{gt} = \sum_{\text{box}} \exp\left(-\left(\frac{(x-c_x)^2}{2\sigma_x^2} + \frac{(y-c_y)^2}{2\sigma_y^2}\right)\right), \quad (3)$$

where c_x and c_y are the center points, and σ_x and σ_y are the standard deviations defined by the width and height. We optimize the prior map using an L_2 loss, defined as:

$$\mathcal{L}_{\text{pm}} = \frac{1}{S}\sum_{S}^{i} \left\|\text{Prior}^i_{fg} - \text{Prior}^i_{gt}\right\|_2^2. \quad (4)$$

3.2 Dual-Aware Query Enhancement

As shown in Figure 3, a tailored query selection and enhancement module is proposed, drawing inspiration from previous works e.g., (Li et al. 2022), (Wang et al. 2024), (Zhang et al. 2022). Leveraging foreground prior information and linguistic information from text, we introduce a

Algorithm 1: Dynamic Background Modeling

Input: Image embeddings $\mathbf{E}_{\text{img}} \in \mathbb{R}^{B\times N\times D}$ (N=H×W)
Output: Foreground tokens $\mathbf{T}_{\text{fg}}$, Background tokens $\mathbf{T}_{\text{bg\_out}}$

Split into 3×3 Regions:
$\mathbf{E}_{\text{regions}}, \mathbf{I}_{\text{regions}} \leftarrow \text{SplitIntoRegions}(\mathbf{E}_{\text{img}})$
Compute Background Tokens:
$\mathbf{T}_{\text{bg}} \leftarrow \text{GlobalAvgPool}(\mathbf{E}_{\text{regions}})$
Concatenate All Tokens:
$\mathbf{T}_{\text{all}} \leftarrow \text{Concat}(\mathbf{E}_{\text{img}}, \mathbf{T}_{\text{bg}})$
: Create Attention Mask Base on region indices:
$\mathbf{M}_{\text{attn}} \leftarrow \text{CreateMask}(\mathbf{I}_{\text{regions}})$
Forward Pass through Vision Model:
$\mathbf{T}_{\text{out}} \leftarrow \text{Model}(\mathbf{T}_{\text{all}}, \mathbf{M}_{\text{attn}})$
Separate Output Tokens:
$\mathbf{T}_{\text{fg}}, \mathbf{T}_{\text{bg\_out}} \leftarrow \text{SplitOutputs}(\mathbf{T}_{\text{out}})$
return $\mathbf{T}_{\text{fg}}, \mathbf{T}_{\text{bg\_out}}$ =0

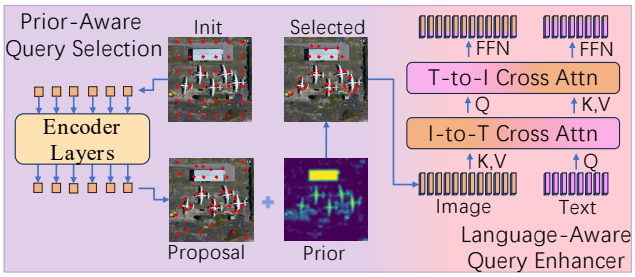

Figure 3: Architecture of the dual-aware query enhancement (DAQE). PAQS selects proposal queries closer to the foreground by centering its operation on the peak points of the foreground prior. Subsequently, LAQE employs bidirectional cross-attention to infuse these selected queries with language information.

dual-aware query enhancement module, which comprises two core components: a prior-aware query selector and a language-aware enhancer. Through these components, it aims to improve query quality, enabling a more precise capture of foreground targets while integrating linguistic context.

Prior-Aware Query Selector: To identify queries better aligned with foreground targets from a large pool of proposal queries, a prior-informed filtering mechanism is designed. Specifically, a foreground prior map is utilized to guide the query selection process. The steps are delineated as follows:

1. **Selection of Initial Cluster Centers**: On the foreground prior map $P \in \mathbb{R}^{H\times W}$, we select N_p initial cluster centers $C = \{c_1, c_2, \ldots, c_{N_p}\}$ by identifying local peaks.

$$c_i = \underset{\substack{(x,y)\in\mathcal{N}_i \\ P(x,y)>\text{threshold}}}{\text{argmax}} P(x,y), \tag{5}$$

   where $\mathcal{N}_i$ represents the *i-th* local neighborhood.

2. **Clustering Process**: After obtaining the initial cluster centers, we perform clustering on all reference points $R \in \mathbb{R}^{N\times 2}$ output by the encoder, where N is the total number of reference points. We simply adopt the L_2 distance from reference points to prior centers as a selection criterion. Specifically, for each reference point r_i, we compute its L_2 distances to all prior centers and use the minimum distance $Dist_i$ as its score.

$$Dist_i = Min(\|r_i - c_k\|_2^2), \forall c_k \in C.ss \tag{6}$$

   Then, the reference points are sorted by score in ascending order, and the queries for the top *K* points are selected. After clustering, yield K selected query pairs $Q_{select} = \{Q_{pos}^{enc}, Q_{content}^{enc}\}$.

$$Q_{select} = Q_{proposal}[TopK(-Dist)]. \tag{7}$$

Language-Aware Enhancer: After filtering potential foreground content queries, a bidirectional attention mechanism is employed to infuse linguistic information, further enhancing their semantic representation. Specifically, two cross-attention modules are designed to facilitate bidirectional information exchange between images and text. These are detailed as follows:

Image-to-Text Attention (I2T-Attn): In this module, image queries serve as the Key and Value vectors, while language features act as the Query. Let the selected content queries be $Q_i \in \mathbb{R}^{K\times D}$ and language features be $F_t \in \mathbb{R}^{L\times D}$ (where L is the text sequence length). The attention computation is:

$$\text{I2T-Attn}(Q_i, F_t) = \text{softmax}\left(\frac{F_t(Q_i)^T}{\sqrt{D}}\right)Q_i. \tag{8}$$

yielding enhanced language features F_t' fused with vision information.

Text-to-Image Attention (T2I-Attn): In this module, vision features serve as query vectors, while language features act as keys and values. The attention computation is:

$$\text{T2I-Attn}(F_t, Q_i) = \text{softmax}\left(\frac{Q_i(F_t')^T}{\sqrt{D}}\right)F_t. \tag{9}$$

producing enhanced vision features Q_i' integrated with language information.

3.3 Prior Refinement Based Denoising

Unlike traditional approaches that predominantly rely on student-teacher models to generate pseudo-labels for semi-supervised training, we innovatively leverage a denoising process to produce pseudo-labels, naming this Refiner.

Extracting Targets from the Foreground Prior Map: We follow the approach in Algorithm 2 to extract candidate bounding boxes from the foreground prior map. and initialize content queries by using text features obtained from the language encoder, specifically by performing global average pooling (GAP) on the text features $T \in \mathbb{R}^{C*D}$ across C classes with D dimensions and repeating K times.

$$Q_{\text{content}}^{rf} = \text{GAP}(T).repeat(K) \in \mathbb{R}^{K\times D}. \tag{10}$$

Algorithm 2: Foreground Target Extraction from Prior Map

Input: Foreground prior map $Prior_{fg}$ Output: Refiner positional queries $Q_{pos}^{rf} \in \mathbb{R}^{K\times 4}$

Noise Suppression via Gaussian Filtering:

$Prior_{smooth} \leftarrow Prior_{fg} * G_\sigma$

Contrast Enhancement using Gamma Correction:

$Prior_{enhanced} \leftarrow (Prior_{smooth})^\gamma$

Foreground Segmentation via Otsu's Thresholding:

$T_{otsu} \leftarrow$ OtsuThreshold$(Prior_{enhanced})$

$$Prior(x,y) \leftarrow \begin{cases} 1 & \text{if } Prior_{enhanced}(x,y) > T_{otsu} \\ 0 & \text{otherwise} \end{cases}$$

$\mathcal{B} \leftarrow$ ExtractBoundingBoxes$(Prior_{binary})$

Prior Bounding Box Noise Injection:

for each bounding box $b = (x_c, y_c, w, h) \in \mathcal{B}$ do:

$\Delta x, \Delta y \sim \mathcal{U}(-1, 1)$

$x_c' \leftarrow x_c + \Delta x \cdot w/2$; $y_c' \leftarrow y_c + \Delta y \cdot h/2$

$\Delta w, \Delta h \sim \mathcal{U}(0.5, 1.5)$

$w' \leftarrow w \cdot \Delta w$; $h' \leftarrow h \cdot \Delta h$

Update b with (x_c', y_c', w', h')

Generate Positional Queries:

$Q_{pos}^{rf} \leftarrow$ Stack$([b_i, \dots])$

return Q_{pos}^{rf} =0

Refiner for Pseudo-labels Generation: The refiner is created by repurposing the CDN module as Refiner to generate pseudo-labels through prior target denoising. Position and content queries serve as Query for the Refiner decoder, while DAQE-enhanced proposals content queries act as Key and Value. This refines position queries, denoises content queries ($Q_{\text{RF}} = [Q_{pos}^{rf}, Q_{content}^{rf}]$), and ultimately yields corresponding pseudo-labels $B \in \mathbb{R}^{K\times 4}$ and $C \in \mathbb{R}^{K\times D}$.

$$[B, C] = Refiner(Q = Q_{\text{RF}}, K = Q_{\text{DAQE}}, V = Q_{\text{DAQE}}). \tag{11}$$

We employ the following optimization objective for semi-supervised training, where $\hat{c}$ and $\hat{b}$ represent the outputs from the DINO Transformer encoder:

$$\mathcal{L}_u = \mu\mathcal{L}_{\text{cls}}(\hat{c}, C) + \lambda\mathcal{L}_{\text{reg}}(\hat{b}, B). \tag{12}$$

Hybrid Training: The model is trained in two stages. Initially, the model lacks foreground prior knowledge, so the first stage uses only labeled data.

Stage1: Using only labeled data, we optimize for detection box regression($\mathcal{L}_{reg}$), classification($\mathcal{L}_{cls}$), and CDN loss($\mathcal{L}_{cdn}$). To prepare for efficient unlabeled data training later, we probabilistically replace CDN with Refiner settings, using pseudo-labels from the prior map and GT labels for optimization with loss function $\mathcal{L}_{rf}$, where B_s and C_s are the ground truth of labeled data.

$$\mathcal{L}_s = \mu\mathcal{L}_{\text{cls}}(\hat{c}, C_s) + \lambda\mathcal{L}_{\text{reg}}(\hat{b}, B_s) + \mu\mathcal{L}_{\text{cdn}}. \tag{13}$$

$$\mathcal{L}_{rf} = \mathcal{L}_{\text{cls}}(C, C_s) + \mathcal{L}_{\text{reg}}(B, B_s). \tag{14}$$

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_{rf} + \mathcal{L}_{pm}. \tag{15}$$

Stage2: Foreground priors are simultaneously extracted for both data types, then sequentially processed by an encoder and DAQE for feature extraction/enhancement. The labeled data's noisy GT undergoes CDN denoising training. Unlabeled data uses prior boxes for extraction and a Refiner for pseudo-labels, completing the semi-supervised training.

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_u + \mathcal{L}_{pm}. \tag{16}$$

3.4 Multi-Granular Label Generation

To enrich the semantic information conveyed by category labels in remote sensing datasets, a multi-granularity labeling approach is proposed. For each category, we construct label groups that incorporate hierarchical semantic information: superordinate labels (representing broader, more abstract concepts), the original label (the category itself), and alternative labels (denoting synonyms of categories). This structured representation is designed to facilitate more effective feature extraction and model training.

Label Construction Process: We leverage a large language model (LLM) to generate multi-granularity labels through a designed prompt. The prompt instructs the LLM to produce *N* superordinate labels and *M* alternative labels based on the category list, ensuring the label system balances breadth and depth. The prompt is defined in appendix.

Feature Extraction and Aggregation: After label generation, all operations are performed offline.

Feature Extraction: Each label group is converted into semantic feature embedding via a pre-trained language encoder (e.g., Remote-CLIP).

Feature Aggregation: The feature vectors are fed into an aggregator to produce an enhanced label representation, integrating core semantics with hierarchical information. We experimented with three aggregation methods: Global averaging, Global maximum, and weighted aggregation.

4 Experiment

4.1 Implement Details

Method is evaluated on two challenging remote sensing image object detection datasets: DOTA(Ding et al. 2022) and DIOR(Li et al. 2020).

Model Setting and Evaluation: We employ Swin-T as the image encoder, paired with Remote-CLIP as the text encoder to ensure image-text semantic consistency. The model's performance is comprehensively evaluated using Mean Average Precision (mAP) and the Harmonic Mean (HM). The mAP is calculated with an Intersection over Union (IoU) threshold of 0.5 to assess detection accuracy for both base and novel categories, while the HM is used to balance performance between them, defined as:

$$\text{HM} = \frac{2 \cdot \text{mAP}_{\text{base}} \cdot \text{mAP}_{\text{novel}}}{\text{mAP}_{\text{base}} + \text{mAP}_{\text{novel}}}. \tag{17}$$

Dataset	Method	Source	Detector	GZSD				ZSD			
				mAP	mAP-base	mAP-novel	HM	APO	BC	GTF	WM
DIOR	ViLD(Gu et al. 2021)	ICLR21	Faster R-CNN	45.7	53.8	14.2	22.4	9.1	11.8	26.7	9.1
	OV-DETR(Zang et al. 2022)	ECCV22	Deformable DETR	54.0	62.6	19.9	30.2	29.1	19.7	25.6	5.2
	Detic(Zhou et al. 2022)	ECCV22	CenterNet2	36.9	45.3	3.5	6.5	0.7	2.2	11.2	0.0
	GroundingDINO(Liu et al. 2023)	ECCV24	DINO	57.3	70.8	3.2	6.2	0.3	3.3	9.1	0.2
	GLIP(Lin et al. 2024)	CVPR22	GLIP	56.0	69.1	3.6	6.9	0.0	2.4	9.1	3.0
	BARON(Wu et al. 2023)	CVPR23	Faster R-CNN	50.6	59.8	15.3	24.4	14.1	16.6	30.4	0.1
	YOLO-World(Cheng et al. 2024)	CVPR24	YOLOv8-M	57.7	70.2	8.0	14.4	0.0	19.8	12.2	0.0
	CastDet(Li et al. 2024b)	ECCV24	Faster R-CNN	61.6	65.7	**45.2**	53.6	45.4	**78.2**	41.3	15.9
	SOAR	-	DINO	**68.5**	**74.3**	44.9	**55.9**	**49.8**	70.2	**42.1**	**17.5**

Table 1: Comparison with the state-of-the-art detectors on DIOR dataset. The short names in ZSD are "APO": Airport, "BC": Basketball-Court, "GTF": Ground-Track-Filed, "WM": Wind Mill.

Figure 4: Detection results, which seen and unseen (APO, BC, GTF, WM) categories are included.

Dataset	Method	Source	GZSD			ZSD
			mAP			mAP
			Base	Novel	HM	
DOTA	RRFS *(Huang et al. 2022)*	CVPR22	47.1	2.2	4.2	2.9
	ContrastZSD *(Yan et al. 2024)*	TPAMI22	41.6	2.8	5.2	6.0
	DescReg *(Zang et al. 2024)*	AAAI24	68.7	4.7	8.8	8.5
	GroundingDINO *(Liu et al. 2023)*	ECCV24	66.0	5.8	10.6	7.9
	CastDet *(Li et al. 2024b)*	ECCV24	60.3	29.2	39.3	29.2
	SOAR *(Ours)*	-	**74.5**	**29.5**	**42.3**	**30.1**

Table 2: Comparison with the SOTA detectors on DOTA dataset.

4.2 Main Result

Generalized Zero-Shot Detection (GZSD) Performance on DIOR and DOTA: Overall Performance: As shown in Table 1, SOAR excels in GZSD, achieving 68.5% mAP and 55.9% HM, leading all compared methods. SOAR outperforms CastDet by 6.9% mAP and 2.3% HM, and YOLO-World (57.7% mAP) by 10.8% mAP. This demonstrates SOAR's superior object identification and balance of precision and recall. However, we find that its detection accuracy for novel classes is slightly lower than that of CastDet. After observing the DOTA and DIOR datasets, we speculate that this might be because the distinction between objects and the background in the DIOR dataset is lower than that

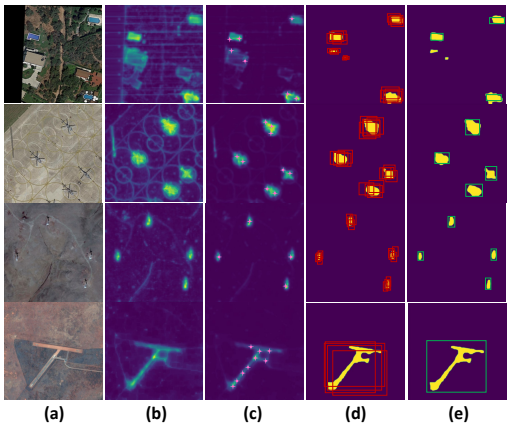

Figure 5: Prior box extraction and pseudo-label generation. (a) original image. (b) weighted prior map. (c) preprocessed prior map, where the pink star markers indicate regional peak coordinates used for query filtering in PAQS. (d) prior box extraction and diverse noise box generation. (e) pseudo-labels after Refiner processing.

in the DOTA dataset. Some detection results are shown in Figure 4. As shown in Table 2, SOAR excels in both base (74.5%) and novel (29.5%) class mAP, surpassing all compared methods in novel class mAP. Its overall HM (42.3%) is also the highest, demonstrating a strong balance between base and novel class detection accuracy.

Pseudo-label Quality Assessment: As shown in Table 3, to evaluate the effectiveness of prior box extraction and the denoising mechanism, we conducted experiments based on Stage2. The mAR of 63.4% indicates that the foreground prior map and box extraction can effectively locate foreground objects. The refined pseudo - labels by the Refiner achieve an mAP of 49.5%, but with a relatively low mAR, as a high threshold was set to obtain high - quality pseudo - labels. These data demonstrate the effectiveness of the denoising mechanism.

4.3 Ablation Study

Dual-Aware Query Enhancement: As shown in Table 6, our ablation studies on DAQE demonstrated that LAQE has

Target	First iter		Last iter	
	mAP	mAR	mAP	mAR
Label_{corase}	-	59.2%	-	63.4%
Label_{rf}	38.7%	21.4%	49.5%	32.1%
SOAR	11.4%	31.3%	44.9%	51.7%

Table 3: Performance comparison of different label types and the SOAR model across iterations on stage2.

Number	mAP	Time
900 (default)	68.5	3.21 images/second
450 (half)	64.2	5.45 images/second
450 (w/o PAQS)	41.7	6.11 images/second

Table 4: Ablation study for learnable queries

a substantial positive impact on performance, especially for novel categories, as expected since it injects language information. We found PAQS to have a less pronounced effect. Further experiments indicated that the large number of predefined queries (900) compared to foreground prior queries meant that most foreground prior queries were selected even without PAQS. Accordingly, we reduced the number of selected queries to 450, which resulted in a minor decrease in model performance but a notable increase in inference speed, show in Table 4.

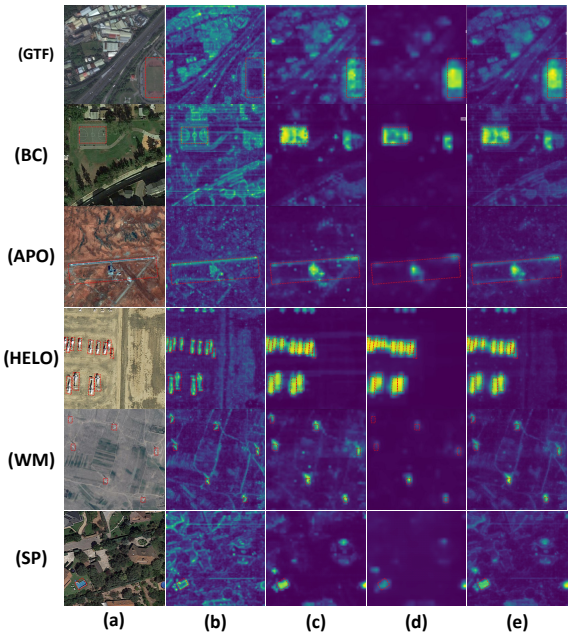

Figure 6: Unseen category foreground prior maps. (a) Original image. (b) High-level semantic prior map. (c) Mid-level semantic prior map. (d) Low-level semantic prior map. (e) Weighted prior map.

Initialization of Dynamic Background Tokens: As shown in Table 7, we present experimental results comparing different initialization strategies for background tokens. We found that initializing background tokens using Global

Label_{ori}	Label_{ext}	mAP	mAP_{base}	mAP_{novel}
1.0	0	68.1	74.3	42.3
0.5	0.5	68.5	74.3	44.9
max		68.3	74.3	43.8
-	rand	67.3	74.1	43.6

Table 5: Ablation experiments on the multi-granular label module. Here, "max": global max pooling; "rand": randomly selecting from the extended labels.

PAQS	LAQE	mAP	mAP_{base}	mAP_{novel}
✓	✓	**68.5**	**74.3**	**44.9**
✓	-	48.4	69.1	29.6
-	✓	67.7	73.2	42.8
-	-	45.7	67.9	28.2

Table 6: Ablation experiments of the dual-aware module. PAQS stands for position-aware query selector; LAQE stands for language-aware query enhancer.

Average Pooling (GAP) significantly outperforms random initialization. Furthermore, applying GAP on a regional basis yields a more fine-grained background representation, leading to improved model performance. Also, we observed that incorporating region-specific positional encodings for background tokens, which are then applied to their corresponding image tokens, further enhances performance.

Different Setting of Multi-Granular Label: As presented in Table 5, 1.0 represents full utilization of the feature embedding, while 0.5 signifies averaging. We observed that MGL provided minimal performance improvement, as the test data did not include categories beyond the original labels. To overcome this, we employed GPT to generate approximately 3 new labels per category, which were then mixed with training-time extended labels and randomly sampled during testing.

BG token Init	mAP	mAP_{base}	mAP_{novel}
Random	67.1	74.2	42.2
GAP	67.9	74.3	43.1
GAP_{area}	68.5	74.3	43.9
GAP_{area}+Pos	68.5	**74.3**	**44.9**

Table 7: Experiments on background token initialization methods. GAP refers to global average pooling; "area": GAP within different regions;"Pos": position embedding.

5 Conclusion

This paper proposes the SOAR, a semi-supervised detector leveraging prior denoising. By adapting a denoising process for pseudo-label generation, SOAR leads to a streamlined architecture. The DAQE module further optimizes query representation through prior-based filtering and cross-modal interaction. To boost novel class detection, multi-granularity aggregated labels are constructed for richer semantic context. SOAR achieves SOTA performance on the DIOR dataset with 68.5% mAP and 55.9% HM.

Acknowledgments

This work was supported in part by the Joint Funds of the National Natural Science Foundation of China (U22B2054), the National Natural Science Foundation of China (62076192, 62276199, 62431020 and 62276201), the 111 Project, the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence (HMHAI-202404and HMHAI-202405).

References

2023. VisDrone 2023 Challenage Zero-shot Object Detection.

Ahmad, U.; Liang, J.; Ma, T.; Yu, K.; Mehmood, F.; and Banoori, F. 2025. Small Aerial Object Detection through GAN-Integrated Feature Pyramid Networks. *Applied Soft Computing*.

Alif, M. A. R.; and Hussain, M. 2025. YOLOv12: A Breakdown of the Key Architectural Features. *ArXiv*, abs/2502.14740.

Bansal, A.; Sikka, K.; Sharma, G.; Chellappa, R.; and Divakaran, A. 2018. Zero-Shot Object Detection. In *European Conference on Computer Vision*.

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *ArXiv*, abs/2004.10934.

Caron, M.; Touvron, H.; Misra, I.; J'egou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9630–9640.

Chen, Y.; Ye, Z.; Sun, H.; Gong, T.; Xiong, S.; and Lu, X. 2025. Global–Local Fusion With Semantic Information Guidance for Accurate Small Object Detection in UAV Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–15.

Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16901–16911.

Demirel, B.; Cinbis, R. G.; and Ikizler-Cinbis, N. 2018. Zero-Shot Object Detection by Hybrid Region Embedding. In *British Machine Vision Conference*.

Ding, J.; Xue, N.; Xia, G.-S.; Bai, X.; Yang, W.; Yang, M. Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2022. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7778–7796.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to Prompt for Open-Vocabulary Object Detection with Vision-Language Model. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14064–14073.

Feng, C.; Zhong, Y.; Jie, Z.; Chu, X.; Ren, H.; Wei, X.; Xie, W.; and Ma, L. 2022. PromptDet: Towards Open-Vocabulary Detection Using Uncurated Images. In *European Conference on Computer Vision*.

feng Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; and Li, H. 2021. Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4079–4088.

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

Gupta, A.; Dollar, P.; and Girshick, R. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

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

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning*.

Lam, D.; Kuzma, R.; McGee, K.; Dooley, S.; Laielli, M.; Klaric, M. K.; Bulatov, Y.; and McCord, B. 2018. xView: Objects in Context in Overhead Imagery. *ArXiv*, abs/1802.07856.

Li, F.; Zhang, H.; guang Liu, S.; Guo, J.; shuan Ni, L. M.; and Zhang, L. 2022. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13609–13617.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*.

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

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159: 296–307.

Li, L.; Yao, X.; Wang, X. B.; Hong, D.; Cheng, G.; and Han, J. 2023b. Robust Few-Shot Aerial Image Object Detection via Unbiased Proposals Filtration. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–11.

Li, Y.; Guo, W.; Yang, X.; Liao, N.; He, D.; Zhou, J.; and Yu, W. 2024b. Toward open vocabulary aerial object detection with clip-activated student-teacher learning. In *European Conference on Computer Vision*, 431–448. Springer.

Liang, B.; and Luo, H. 2024. MEANet: An effective and lightweight solution for salient object detection in optical remote sensing images. *Expert Systems with Applications*, 238: 121778.

Lin, C.; Jiang, Y.-X.; Qu, L.; Yuan, Z.; and Cai, J. 2024. Generative Region-Language Pretraining for Open-Ended Object Detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13958–13968.

Lin, T.-Y.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Liu, M.; Hayes, T. L.; Ricci, E.; Csurka, G.; and Volpi, R. 2024. SHiNe: Semantic Hierarchy Nexus for Open-Vocabulary Object Detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16634–16644.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; yue Li, C.; Yang, J.; Su, H.; Zhu, J.-J.; and Zhang, L. 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *European Conference on Computer Vision*.

Liu, W.; Zhang, Y.; Wang, X.; and Zhang, L. 2025. Deep Multi-Level Contrastive Clustering for Multi-Modal Remote Sensing Images. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 1239–1247. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720352.

Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2021. Unbiased Teacher for Semi-Supervised Object Detection. *ArXiv*, abs/2102.09480.

Pan, J.; Liu, Y.; Fu, Y.; Ma, M.; Li, J.; Paudel, D. P.; Gool, L. V.; and Huang, X. 2025. Locate Anything on Earth: Advancing Open-Vocabulary Object Detection for Remote Sensing Community. arXiv:2408.09110.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Rahman, S.; Khan, S. H.; and Porikli, F. M. 2018. Zero-Shot Object Detection: Learning to Simultaneously Recognize and Localize Novel Concepts. *ArXiv*, abs/1803.06049.

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

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020. A Simple Semi-Supervised Learning Framework for Object Detection. *ArXiv*, abs/2005.04757.

Sun, X.; Yang, Z.; Xie, R.; Lian, F.; Kang, Z.; and Xu, C. 2024. LightVLP: A Lightweight Vision-Language Pre-training via Gated Interactive Masked AutoEncoders. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 10499–10510. Torino, Italia: ELRA and ICCL.

Tang, Y.; Chen, W.; Luo, Y.; and Zhang, Y. 2021. Humble Teachers Teach Better Students for Semi-Supervised Object Detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3131–3140.

Wang, C.-Y.; Yeh, I.-H.; and Liao, H. 2024. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *ArXiv*, abs/2402.13616.

Wang, G.; Chen, Y.; An, P.; Hong, H.; Hu, J.; and Huang, T. 2023a. UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios. *Sensors (Basel, Switzerland)*, 23.

Wang, H.; Ren, P.; Jie, Z.; Dong, X.; Feng, C.; Qian, Y.; Ma, L.; Jiang, D.; Wang, Y.; Lan, X.; and Liang, X. 2024. OV-DINO: Unified Open-Vocabulary Detection with Language-Aware Selective Fusion. *ArXiv*, abs/2407.07844.

Wang, J.; Gao, J.; and Zhang, B. 2025. A small object detection model in aerial images based on CPDD-YOLOv8. *Scientific Reports*, 15.

Wang, P.; Cai, Z.; Yang, H.; Swaminathan, G.; Vasconcelos, N.; Schiele, B.; and Soatto, S. 2022. Omni-DETR: Omni-Supervised Object Detection with Transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9357–9366.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; et al. 2023b. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19175–19186.

Wu, S.; Zhang, W.; Jin, S.; Liu, W.; and Loy, C. C. 2023. Aligning Bag of Regions for Open-Vocabulary Object Detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15254–15264.

Xu, B.; Chen, M.; Guan, W.; and Hu, L. 2023. Efficient Teacher: Semi-Supervised Object Detection for YOLOv5. *ArXiv*, abs/2302.07577.

Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; and Xia, G.-S. 2022. Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190: 79–93.

Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-End Semi-Supervised Object Detection with Soft Teacher. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3040–3049.

Xue, C.; Xia, Y.; Wu, M.; Chen, Z.; Cheng, F.; and Yun, L. 2024. EL-YOLO: An efficient and lightweight low-altitude aerial objects detector for onboard applications. *Expert Systems with Applications*, 256: 124848.

Yan, C.; Chang, X.; Luo, M.; Liu, H.; Zhang, X.; and Zheng, Q. 2024. Semantics-Guided Contrastive Network for Zero-Shot Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3): 1530–1544.

Yao, L.; Pi, R.; Han, J.; Liang, X.; Xu, H.; Zhang, W.; Li, Z.; and Xu, D. 2024. DetCLIPv3: Towards Versatile Generative Open-Vocabulary Object Detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5610–5619.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. *Open-Vocabulary DETR with Conditional Matching*, 106–122. Springer Nature Switzerland. ISBN 9783031200779.

Zang, Z.; Lin, C.; Tang, C.; Wang, T.; and Lv, J. 2024. Zero-shot aerial object detection with visual description regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6926–6934.

Zareian, A.; Rosa, K. D.; Hu, D. H.; and Chang, S.-F. 2020. Open-Vocabulary Object Detection Using Captions. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14388–14397.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. arXiv:2203.03605.

Zhang, J.; Lin, X.; Zhang, W.; Wang, K.; Tan, X.; Han, J.; Ding, E.; Wang, J.; and Li, G. 2023a. Semi-DETR: Semi-Supervised Object Detection with Detection Transformers. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23809–23818.

Zhang, Y.; Huang, X.; Ma, J.; Li, Z.; Luo, Z.; Xie, Y.; Qin, Y.; Luo, T.; Li, Y.; Liu, S.; Guo, Y.; and Zhang, L. 2023b. Recognize Anything: A Strong Image Tagging Model. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1724–1732.

Zhang, Y.; Yan, S.; Zhang, L.; and Du, B. 2024. Fast Projected Fuzzy Clustering With Anchor Guidance for Multimodal Remote Sensing Imagery. *IEEE Transactions on Image Processing*, 33: 4640–4653.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting Twenty-Thousand Classes Using Image-Level Supervision. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 350–368. Cham: Springer Nature Switzerland. ISBN 978-3-031-20077-9.