Self-Prompting Analogical Reasoning for UAV Object Detection

Nianxin Li[1], Mao Ye[1*], Lihua Zhou[1], Song Tang[2], Yan Gan[3], Zizhuo Liang[4], Xiatian Zhu[5]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, China
[2]Institute of Machine Intelligence (IMI), University of Shanghai for Science and Technology, China
[3]College of Computer Science, Chongqing University, China
[4]University of Sheffield
[5]University of Surrey
linianxin1220@gmail.com, cvlab.uestc@gmail.com

Abstract

Unmanned Aerial Vehicle Object Detection (UAVOD) presents unique challenges due to varying altitudes, dynamic backgrounds, and the small size of objects. Traditional detection methods often struggle with these challenges, as they typically rely on visual features only and fail to extract the semantic relations between the objects. To address these limitations, we propose a novel approach named Self-Prompting Analogical Reasoning (SPAR). Our method utilizes the vision-language model (CLIP) to generate context-aware prompts based on image features, providing rich semantic information that guides analogical reasoning. SPAR includes two main modules: self-prompting and analogical reasoning. Self-prompting module based on learnable description and CLIP-text encoder generates context-aware prompt by combining specific image feature; then an objectness prompt score map is produced by computing the similarity between pixel-level features and context-aware prompt. With this score map, multi-scale image features are enhanced and pixel-level features are chosen for graph construction. While for analogical reasoning module, graph nodes consist of category-level prompt nodes and pixel-level image feature nodes. Analogical inference is based on graph convolution. Under the guidance of category-level nodes, different-scale object features have been enhanced, which helps achieve more accurate detection of challenging objects. Extensive experiments illustrate that SPAR outperforms traditional methods, offering a more robust and accurate solution for UAVOD.

Introduction

With the rapid advancements in deep learning, significant progress has been made in the field of object detection. For example, single-stage models like YOLO (Redmon et al. 2016) and two-stage models like Faster-RCNN (Ren et al. 2015) have demonstrated impressive performance on popular datasets such as COCO (Lin et al. 2014) and PASCAL VOC (Everingham et al. 2015). However, the effectiveness of these technologies still falls short of expectations when applied to UAV (Unmanned Aerial Vehicle) imagery. UAVs often capture wide-area images from a high altitude, leading to objects appearing much smaller compared to those in ground-level images. This scale variability makes it difficult for detection algorithms to accurately identify and localize small objects.

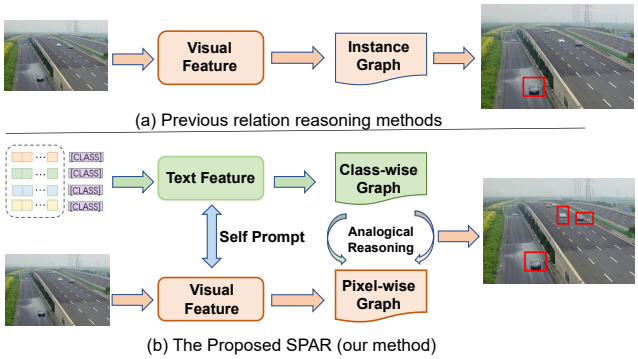

Figure 1: (a) The traditional relation reasoning methods only use the visual similarity to construct a relation graph. (b) Our method achieves self-prompting through the interaction of visual and text features, and simultaneously constructs relational graph. In this way, small objects can be detected by analogical reasoning.

Over the years, several methods have been proposed to tackle this challenge, these methods can generally be categorized into three routes. The first route segments the image into multiple regions and then selectively scales the regions containing dense objects by a certain factor to enhance their resolutions for object detection (Huang, Chen, and Huang 2022; Zhang et al. 2024). The second route introduces additional network modules to enhance the network's capability to extract meaningful features, such as attention (Woo et al. 2018), multi-scale feature fusion (Zhao et al. 2023), etc. The final route aims to enhance data diversity through image augmentation techniques such that a broader range of scenarios can be exploited during training (Yan et al. 2022; Garg, Mandal, and Narang 2021). All of the above methods do not account for the relationships between different objects and contextual information, this oversight means that while each object is detected individually, the interdependencies and interactions among objects and background are not considered.

*Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

It is a natural idea to incorporate relationships between objects and background to improve detection performance and get more meaningful interpretations of the scene. Recent advancements have explored incorporating relationship reasoning into object detection frameworks (Zhu et al. 2021a). These efforts primarily rely on visual feature similarity as the basis for establishing relationships between objects. While effective when visual features are clear and distinct, these methods face significant challenges when applied to UAV imagery. Objects in UAV images are often small in size, leading to a loss of critical visual features and making accurate relation reasoning difficult, as shown in Fig.1(a). Therefore, additional modal information is needed. Fortunately, vision-language multimodal models, e.g., CLIP (Bianchi et al. 2021), align text and images into a high-level semantic space, which provides a valuable auxiliary model for reasoning relationships between objects.

Based on the above analysis, we propose a novel method called Self-Prompt Analogical Reasoning (SPAR). Specifically, there are two primary modules. The first is self-prompting module, which outputs a context-aware prompt and multi-scale objectness prompt score maps. The context-aware prompt is generated by combining the embeddings of learnable class descriptions based CLIP-text encoder and specific image feature; while objectness score maps are computed by the similarity between the context-aware prompt and multi-scale pixel-level features. Another analogical reasoning module is based on multi-scale graphs. Graph nodes are composed of category-level and pixel-level nodes which are constructed by the class prompt and the filtered pixel-level features by score map, respectively. The edge weights are computed by learnable similarity measure. By applying graph convolution, based on language features embedded in the category-level node features, similar object features will been enhanced. Implementing our method based on YOLOv8 as the baseline achieves promising results on UAV image object detection datasets.

Our contributions can be summarized as follows: (1) A novel analogical reasoning framework for object detection is proposed based on vision-language model. Analogical reasoning has three steps: deduction, mapping and inference, which correspond to graph construction based on language feature, graph edge construction and graph reasoning respectively. In this way, the easier detection of objects can support the detection of smaller and challenging objects. (2) A self-prompting method is proposed which endows each image context-aware prompt and objectness prompt score map. The contextual information is implicitly extracted and feature representation is enhanced. (3) Implementing analogical reasoning through category-level and pixel-wise graph nodes enhances the features of objects that are difficult to be detected directly through visual features, enabling their successful detection through relational reasoning.

Related Works

Object detection on drone imagery. There are two main routes to address UAV object detection challenges. The first route emphasizes the importance of pointing and isolating areas within the image where objects are densely clustered (Feng et al. 2022; Han et al. 2021; Sun 2024), where the detection can be concentrated on. For example, ClusDet (Yang et al. 2019a) generates object cluster regions; Cascaded zoom-in(Meethal, Granger, and Pedersoli 2023) identifies density crops by adding "crop" as a new class, selecting high-quality density crops during inference. AdaZoom (Xu, Li, and Wang 2021) uses policy gradient to dynamically focus on small and dense object regions.

The second route involves enhancing detection performance by managing variability in object size and appearance in aerial imagery (Zhang et al. 2023; Lee et al. 2024; Wang et al. 2023; Li et al. 2024). This includes designing scale-robust models, using multi-scale feature extraction, incorporating data augmentation, and integrating attention mechanisms to better focus on relevant features and improve detection accuracy across different scales (Dadsetan et al. 2021). For example, TPH-YOLOv5 (Zhu et al. 2021b) enhances YOLOv5 by adding a prediction head for tiny objects; SIFDAL (Liu et al. 2024) improves detection accuracy by disentangling scale-invariant features and introducing a new multi-modal dataset with UAV-specific flight data.

Vision-language model for object detection. Since the introduction of CLIP, a substantial amount of works have applied vision-language models to downstream tasks (Liu et al. 2023). The applications on object detection can be roughly categorized into two types. The first one focuses on improving prompts by enhancing language representations to boost detection accuracy (Zang et al. 2024). For example, Coop (Zhou et al. 2022b,a) introduces a method that optimizes class prompts through contrastive learning to improve vision-language model performance on downstream tasks; EDA (Shi and Yang 2023) uses object-level supervision to learn dense-level alignment to maintain local fine-grained semantics; GLIP (Li et al. 2022) defines object detection as an association task by aligning each region/bounding box with a phrase in the text prompt.

While another one explores how to integrate language model into the existing detection networks. For example, ViLD (Gu et al. 2021) trains a student detector, whose region embeddings of detected boxes are aligned with the text and image embeddings inferred by the teacher; GridCLIP (Lin and Gong 2023) learns grid-level representations to adapt to the intrinsic principle of one-stage detection learning by expanding the conventional CLIP image-text holistic mapping to a more fine-grained grid-text alignment; ProposalCLIP (Shi et al. 2022) employs an unsupervised approach directly utilizes CLIP to label objects in image; RegionCLIP (Zhong et al. 2022) introduces a region-based vision-language pretraining method that learns to match image regions with their descriptions and associates region-text pairs; DenseCLIP (Rao et al. 2022) combines CLIP-based language features with visual features extends CLIP to dense detection. Although the binding of language and visual features can enhance the representation ability to a certain extent, the contextual information is still not fully utilized.

Relational reasoning for object detection. Relational reasoning in object detection involves understanding and

leveraging the relationships between objects within an image to improve detection accuracy and robustness (Hu et al. 2018). This approach goes beyond detecting objects in isolation by incorporating contextual and spatial relationships, providing valuable cues for object detection. Relational reasoning methods in object detection can be broadly categorized into three approaches. The first approach uses the contextual information surrounding objects to aid detection. For example, SMN (Chen and Gupta 2017) enhances object detection by efficiently modeling instance-level context that integrates object instances into a pseudo-image representation; GCRN (Acharya et al. 2022) detects out-of-context objects by capturing and analyzing contextual cues using a graph-based approach. The second approach optimizes predictions based on the expected spatial and semantic interactions between objects. For example, reasoning-RCNN (Xu et al. 2019) integrates global semantic knowledge and adaptive reasoning to refine both classification and bounding box regression; C-GCN (Fu et al. 2020) uses graph convolutional networks for small object detection, which encodes implicit pair-wise regional relationships and propagates semantic and spatial layout contextual information. Relational reasoning has been proven to be effective in object detection, but based solely on limited visual information, it cannot effectively achieve feature analogical reasoning. Therefore, we propose a joint reasoning approach that combines self-prompt and image features.

Methodology

Problem statement

Assuming UAV object detection training set $D_{train} = \{x_{tr}^i, y_{tr}^i\}_{i=1}^{N_{tr}}$, where $y_{tr}^i = (b_{tr}^i, c_{tr}^i)$ represents the boxes and classes of objects in the i-th image of training set, and it is further projected into pixel-wise one hot label $y_{tr,j}^i$ for the j-th pixel. N_{tr} is the cardinality of training set. The test set is $D_{test} = \{x_{te}^i, y_{te}^i\}_{i=1}^{N_{te}}$, where N_{te} is the cardinality of test set. $\mathcal{C} = \{c_i\}_{i=1}^{K}$ is a set of classes that need to be detected where c_i is the i-th class name and K is the number of classes. Our goal is to use the vision-language model CLIP to train a better object detector, improving detection performance in UAV images.

Overview. The proposed SPAR framework consists of two moduels: self-prompting and analogical reasoning, as illustrated in Fig. 2. Initially, the backbone $\mathcal{F}$ of YOLOv8 is utilized to extract multi-scale features $[C_3, C_4, C_5] = \mathcal{F}(x)$ for an images x. The self-prompting module is to output context-aware prompt and multi-scale objectness prompt score maps. A learnable textual context d_k and Multi-Head Cross Attention techinque (Cordonnier, Loukas, and Jaggi 2020) is used to modulate image feature to text embedding T_e for context-aware prompting; then, the resulted prompt T_{ei} is combined with the enhanced image feature map f_i to obtain score map s_i which denotes objectness in the corresponding position of feature maps for $i \in \{3, 4, 5\}$. The feature map f_i is also updated as f_i' in an attention-weighted manner according to the score map s_i.

Analogical reasoning module uses contextual information to assist in object detection based on graph reasoning. Multi-scale graphs are constructed corresponding to three scale backbone features f_i'. Each scale of graph nodes are composed of category-level nodes and pixel-level nodes. Category-level node is produced by each class text embedding $T_{ei}(k)$ for $k \in \{1, \cdots, K\}$, while pixel-level node is produced by the feature at the position whose score map value is greater than a threshold. The graph edge weight is computed based on similarity. Graph reasoning is performed based graph convolution such that object features from different locations can learn from each other. In the end, the updated node features are projected back to enhance the feature maps f_i' for obtaining better detection performance.

Self-Prompting Module

As we mentioned before, utilizing the contextual information around objects can help accurate object detection. However, UAV images always can not provide detailed object contextual descriptions. So we first leverage learnable description binding specific image feature for context-aware prompting to guide object detection; then based on this context-aware prompt, a prompt score map is obtained which denotes the possibility of various class objects appearing in the position of image feature maps.

Context-aware prompting. For incorporating contextual information, for each class c_i, its learnable description d_k is represented as

$$d_k = [p, c_k], 1 \leq k \leq K \tag{1}$$

where $p \in R^{N \times C}$ are the learnable textual contexts and $c_k \in R^C$ is the embedding for the k-th class name. C is feature dimension. Here the learnable prompt tokens p is used instead of fixed templates as the design of CoOp (Zhou et al. 2022b). Then, the description d_k will pass through the CLIP-text encoder $\mathcal{T}$ to obtain the text embedding $e_k \in R^C$ as $e_k = \mathcal{T}(d_k)$.

Next, we will associate a specific image feature with the text embedding. For an image x, the multi-scale features C_3, C_4 and C_5 are input into a max pooling layer to obtain 4×4 regions and flattened into a $16 \times C$ matrix respectively:

$$f_i = \text{Flatten}(\text{MaxPool}(C_i)), \quad \text{for } i \in \{3, 4, 5\}, \tag{2}$$

where $f_i \in R^{16 \times C}$. Then the multi-scale flattened features are concatenated as $f = \text{Concat}(f_3, f_4, f_5)$. Thus we obtain a visual feature $f \in \mathbb{R}^{48 \times C}$ that integrates multi-scale semantic information. This allows for the fusion of contextual information across different scales.

Multi-Head Cross-Attention Mechanism is used to fuse text and image features. The text feature $T_e = [e_1, e_2, \cdots, e_K]^T \in R^{K \times C}$ and the visual feature f are transformed into query vectors Q, key vectors K, and value vectors V respectively as

$$Q_h = TW_{Q_h}, K_h = fW_{K_h}, V_h = fW_{V_h} \tag{3}$$

where $W_{Q_h} \in \mathbb{R}^{C \times d_h}$, $W_{K_h} \in \mathbb{R}^{C \times d_h}$, and $W_{V_h} \in \mathbb{R}^{C \times d_h}$ are linear transformation matrices. h denotes the h-th attention head; d_h is the dimension of each head as $d_h = C/H$ where H is the number of attention heads. The attention $A_h \in R^{K \times 48}$ is computed as $A_h = \text{Softmax}(Q_h {K_h}^T / \sqrt{d_h})$.

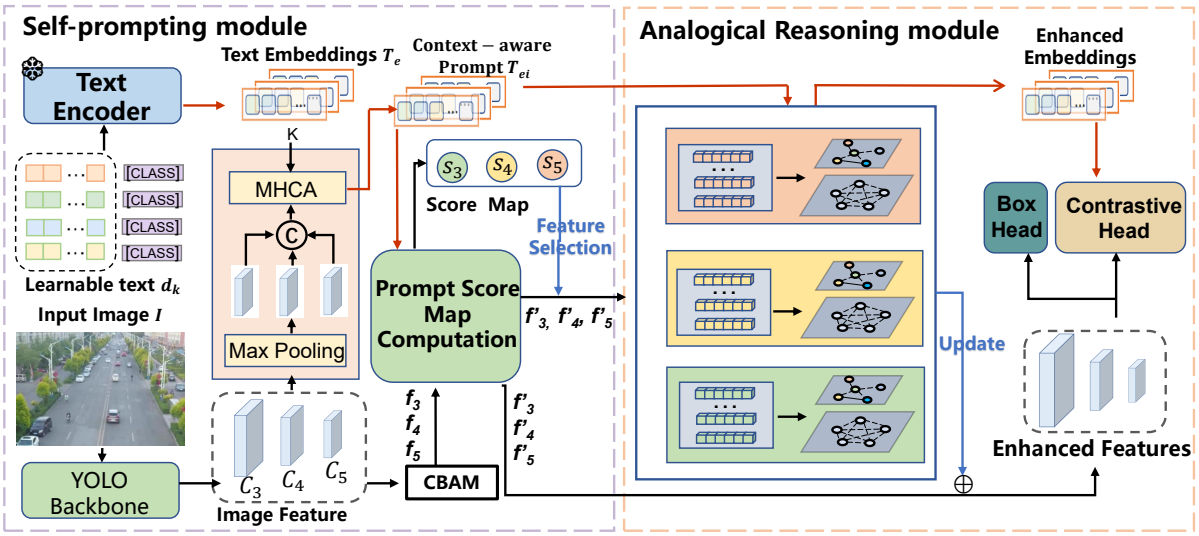

Figure 2: The proposed SPAR method consists of two main modules: self-prompting module and analogical reasoning module. Self-prompting module implicitly extracts contextual and semantic information from image, enhancing the text embeddings and utilizing them to generate three score maps that prompt the features focusing on important areas. Analogical reasoning is based on three graphs with respect to multi-scale features. Text feature based nodes and pixel-wise image feature based nodes are interacted and relation reasoning is performed, which helps achieve more accurate detection of challenging objects.

Attention is used to perform a weighted average of the value vectors to produce the output of each attention head, $O_h = A_h V_h \in R^{K\times d_h}$. Finally, the outputs from all attention heads are concatenated and processed through a linear transformation matrix W_O to generate the enhanced image feature tokens:

$$T_{sp} = \text{Concat}\left(O_1, O_2, \cdots, O_H\right) W_O \tag{4}$$

where $W_O \in R^{C\times C}$ and $T_{sp} \in R^{K\times C}$. By adding this context-enriched image feature to the initial text feature, we can obtain a new text feature that includes contextual semantic information,

$$T_{ei} = T_e + \alpha T_{sp}. \tag{5}$$

Remark. Through the above process, we implicitly integrate contextual information from the image into the text features. This results in more accurate and reliable text features. With these enhanced text features combined with the rich semantic information obtained through text-visual interaction, the classification accuracy will be improved.

Prompt score map computation. The strong zero-shot capability of CLIP model can highlight the feature regions that require significant attention while suppressing irrelevant areas, thereby further enhancing the representation capability of the feature map. This part will use the context-aware prompt to get the objectness score map.

We adjust the number of channels of features C_i $(i = 3, 4, 5)$ to C by modifying the number of output channels in the C2f module of YOLOv8. Inspired by the design of CBAM, we incorporate a channel attention on multi-scale feature $C_i \in \mathbb{R}^{H_i\times W_i\times C}$. First, we compute the global average pooling $f_{\text{avg}} \in R^C$ and global max pooling $f_{\text{max}} \in R^C$ of the feature C_i. Then the enhanced feature $f_i \in R^{H_i\times W_i\times C}$ is

$$f_i = \left(\sigma\left(\text{g}(f_{avg}) + \text{g}(f_{max})\right)\right) \odot C_i, \tag{6}$$

where σ denotes the Sigmoid activation function, g denotes the fully connected layer followed by ReLU activation and $\odot$ denotes element-wise multiplication along channel dimension.

The similarity map between the visual feature map $f_i \in R^{H_i\times W_i\times C}$ and context-aware prompt $T_{ei} \in R^{K\times C}$ is computed at each pixel as

$$p_i = \beta f_i \cdot T_{ei}^T + \gamma \tag{7}$$

where β and γ are learnable parameters, $p_i \in R^{H_i\times W_i\times K}$ represents the possibility of each class being present in every region in the feature map. The prompt score map $s_i \in R^{H_i\times W_i\times 1}$ is computed as $s_i = \max_k p_{x,y,k}$ by applying the max operation to p_i, where $p_{h,w,k}$ represents the value at position (h, w) for the k-th channel in the matrix p_i. This expression indicates that for each spatial position (h, w), we select the maximum value across the K channels in the matrix p_i. Furthermore, the visual feature map will be updated as

$$f'_i = f_i \odot \sigma(s_i)^T + f_i. \tag{8}$$

Analogical Reasoning

Analogical reasoning can provide semantic supplementation when processing unclear or incomplete data by leveraging information from similar situatons, which greatly aids in detecting small objects. Specifically, there are usually two types of relationships among objects in an image: category co-occurrence relationships, such as a bicycle often appearing with people or a car usually found on a road, and intra-class similarity relationships, where objects of the same

class typically share similar visual and distribution features. By using a graph to learn and memorize these relationships, we can infer hard-to-detect objects through those that are easily detected, based on their semantic associations or feature similarities. So our analogical reasoning comprise two parts: graph construction and graph reasoning.

Graph construction. Multi-scale graphs G_i are constructed with respect to the combinations of the text prompt T_{ei} and multi-scale image features f_i' for $i \in \{3, 4, 5\}$, respectively.

(1) **Graph node construction.** There are two kinds of graph nodes, i.e., category-level and pixel-level nodes. The text feature $T_{ei}(k)$ for class k, where $k \in \{1, \cdots, K\}$, corresponds to a graph node, totally resulting in a graph node matrix $H_c \in \mathbb{R}^{K \times C}$. While for pixel-level node, since single-stage object detection networks do not have a region proposal mechanism to obtain instance-level features, the feature $f_i'(h, w)$ corresponds to a node if $s_{i(h,w)} > \tau$, where τ is set empirically. Suppose there are M qualified nodes, then we get pixel-level node matrix $H_{i,p} \in R^{M \times C}$. Here, the positions are also recorded as $E = [(h_1, w_1), \cdots, (h_M, w_M)]$. It should be noted that the M is varying with different images. The node matrix of G_i is $H_i = H_c \cup H_{i,p}$.

(2) Graph edge construction. Following the work (Chen et al. 2019), for the graph G_i, we calculate the adjacency matrix $\tilde{A}^i = [\tilde{A}^i_{m,n}] \in \mathbb{R}^{N_v \times N_v}$, $N_v = K + M$, by performing a dot product between the feature vectors of the nodes m and n as follows:

$$\tilde{A^i}_{m,n} = \text{Conv}_{1\times1}(H_i)(m) \cdot \tilde{\Lambda}(H_i) \cdot \text{Conv}_{1\times1}(H_i)(n)^T$$

where $\text{Conv}_{1\times1}$ denotes a 1×1 convolutional layer followed by a ReLU activation function. The matrix $\tilde{\Lambda}(H_i) \in R^{C \times C}$ is a diagonal matrix designed to learn a more precise distance metric through the inner product, defined as:

$$\tilde{\Lambda}^i = \text{diag}(\text{Conv}_{1\times1}(\text{Avepool}(H_i))) \tag{9}$$

Here, $\text{Avepool}(\cdot)$ represents average pooling, and $\text{diag}(\cdot)$ reshapes a vector into a diagonal matrix. After this, we also perform normalization on the adjacency matrix,

$$\hat{A}^i = D^{-1/2}(\tilde{A}^i + I)D^{-1/2} \tag{10}$$

where $I \in R^{N_v \times N_v}$ is the identity matrix, and $D \in R^{N_v \times N_v}$ is the degree matrix, with the diagonal element $D(i, i)$ representing the degree of node i, i.e., $D_{ii} = \sum_j (\tilde{A}^i + I)_{ij}$. Normalizing the adjacency matrix helps alleviating feature amplification effects, ensures balanced feature contributions, improves gradient propagation stability, and accelerates model convergence. This makes the graph convolution network more stable and efficient during training and inference.

Graph reasoning. Analogical reasoning is typically divided into three steps: deduction, mapping, and inference. The construction of the graph and the establishment of relationships correspond to deduction and mapping respectively. Now, we use graph convolution to perform inference to update node features,

$$H_i{}^{(l+1)} = ReLU(L_i H_i{}^{(l)} W_i) \tag{11}$$

where the normalized graph Laplacian matrix $L_i \in R^{N_v \times N_v}$ is

$$L_i = I - \hat{A}^i. \tag{12}$$

$H_i^{(l)} \in R^{N_v \times C}$ is the node feature matrix at l-th iteration. $W_i \in R^{C \times C}$ is the learnable weight matrix which is independent of the number of graph nodes. It learns how to combine node features based on node similarity. The final $H_i^L \in R^{N_v \times C}$ is obtained after L iterations.

After splitting H_i^L into H_c^L and $H_{i,p}^L$, the updated node features are projected back to the multi-scale feature f_i' as

$$f_i^e(h_m, w_m) = f_i'(h_m, w_m) + H_{i,p}^L(h_m, w_m) \tag{13}$$

where (h_m, w_m) is the position recorded in E. While the text embeddings are update as T_{ei}^e based on H_c^L similarly. With the updated multi-scale features, we present a text contrastive head to obtain the classification score for each scale feature,

$$S_i(w, h) = Softmax((f_i^e(w, h)) \cdot (T_{ei}^e)^T) \tag{14}$$

where $S_i \in R^{H_i \times W_i \times K}$ is the classification result. The classification contrastive loss is defined as follows,

$$L_{\text{cls}}^i = -\frac{1}{N_{\text{pos}}^i} \sum_{j=1}^{H_i \times W_i} (y_j^i \log(p_j^i) + (1 - y_j^i) \log(1 - p_j^i))$$

where N_{pos}^i is the number of positive samples in the i scale feature map. p_j^i is the class prediction probability vector of the j-th pixel in S_i. y_j^i is the ground truth class label vector of the j-th pixel. In the end, combined with the Yolo loss, the overall loss is $L = \sum_i L_{cls}^i + L_{Yolo}$.

Remark. Analogical reasoning enhances visual feature by transmission between similar and related objects through graph convolution. For objects that cannot relate to other objects due to ambiguous features, indirect feature transmission is achieved using text nodes as intermediaries. This enriches feature representation and subsequently improves detection accuracy.

Experiment

Experimental Settings

Datasets. For the assement of our methods, we utilized two popular and challenging benchmark datasets in the field of aerial image detection: **VisDrone** (Du et al. 2019) dataset and **UAVDT** (Du et al. 2018) dataset. **VisDrone** dataset comprises 8599 images captured by drones, divided into 6471 for training, 548 for validation, and 1580 for testing, each with a resolution of approximately 2000 × 1500 pixels. The dataset includes ten categories of objects, with 540k annotated instances in the training set, mostly containing different categories of vehicles and pedestrians observed when the drone is flying through the streets. As the evaluation server is closed now, following the existing works, we used the validation set for evaluating the performance. **UAVDT** dataset is a comprehensive resource for drone-based tasks, including object detection, single-object tracking, and multi-object tracking. It comprises 24,143 training images and

Method	AP	AP50	AP75
FPN (NeurIPS, 2015)	16.9	30.7	17.2
Faster R-CNN (TPAMI, 2017)	12.1	23.5	10.8
CascadeRCNN (CVPR, 2018)	17.1	30.5	18.6
ClusDet (ICCV, 2019)	13.7	26.5	12.5
DREN (ICCVW, 2019)	15.1	-	-
AMRNet (Arxiv, 2019)	18.2	30.4	19.8
DMNet (CVPR Workshop, 2020)	14.7	24.6	16.3
TPH-YOLOv5 (ICCVW, 2021)	26.9	41.3	32.7
GLSAN (TIP, 2021)	17.0	28.1	18.8
AdaZoom (TMM, 2022)	20.1	34.5	21.5
Zoom&Reasoning Det (SPL, 2022)	21.8	34.9	24.8
UFPMP-Det (AAAI, 2022)	24.6	38.7	28.0
TPH-YOLOv5++ (MDPI, 2023)	30.1	43.5	34.3
EVORL (AAAI, 2024)	28.0	43.8	31.5
Proposed Method	**30.5**	**43.9**	**34.7**

Table 1: Comparison with the state-of-the-art methods on the UAVDT dataset in terms of MAP. The best results are highlight in bold.

16,592 testing images, each with an average resolution of 1024 × 540 pixels. This dataset features diverse and challenging scenarios. It is extensively used for the detection of various vehicle types, such as cars, trucks, and buses.

Evaluation metrics. We evaluate our method using standard object detection metrics(Lin et al. 2014), including mean Average Precision (mAP), mAP50, and mAP75. The mAP metric represents the average AP over IoU thresholds ranging from 0.50 to 0.95, with an interval of 0.05. The mAP50 and mAP75 metrics correspond to the AP values at IoU thresholds of 0.50 and 0.75, respectively. Higher values for these metrics indicate better performance.

Implementation details. The backbone detector used in our study is YOLOv8, and all of our models use an NVIDIA RTX3090 GPU for training and testing. In the training phase, we use part of pretrained model from YOLOv8, because SPAR and YOLOv8 share most part of backbone and some part of head, there are many weights can be transferred from YOLOv8x to our method, by using these weights we can save a lot of training time. The model on train set is trained for 100 epochs, and the first 2 epochs are used for warm-up. We use adam optimizer for training, and use 3e-4 as the initial learning rate.

Comparison with State-of-the-art Methods

Compared methods. We compare two categories of methods. The first category uses general object detection algorithms, such as Faster-RCNN and CascadeRCNN. The second category includes the methods specifically optimized for drone object detection scenarios, such as ClusDet (Yang et al. 2019b), UFPMP-DET, DMNet, ClusDet (Yang et al. 2019b), UFPMP-DET (Huang, Chen, and Huang 2022), DMNet (Fang and Li 2020), GLSAN (Deng et al. 2020), DREN (Zhang et al. 2019), AMRNet (Wei et al. 2020) (data augmentation) and NDFT (Wu et al. 2019).

Quantitative comparison. As shown in Table 1 and Table 2, the proposed method (SPAR) demonstrates superior performance in object detection tasks compared to various state-of-the-art methods on UAVDA and VisDrone datasets. Specifically, the results on UAVDA dataset show that SPAR achieves an AP of 30.5, surpassing many traditional methods, such as FPN and Faster R-CNN, and outperforming recent methods like EVORL (28.0). On VisDrone dataset, SPAR also achieves an AP of 42.8, surpassing most state-of-the-art methods. These results underscore SPAR's effectiveness in improving comprehensive performance, especially in detecting small and challenging objects in complex scenarios. Our method does not improve as much on the VisDrone dataset as on the UAVDA dataset because some categories in this dataset are semantically similar and visually confusing, such as pedestrian and person, tricycle and awning-tricycle. The binding of fine-grained text and visual features is a worthwhile research direction.

Method	AP	AP50	AP75
Faster R-CNN (TPAMI, 2017)	21.8	41.8	20.1
SAIC-FPN (Neurocomputing, 2019)	35.7	62.3	35.1
ClusDet (ICCV, 2019)	32.4	56.2	31.6
DMNet (CVPR Workshop, 2020)	29.4	49.3	30.6
GLSAN (TIP, 2021)	32.5	55.8	33.0
HRDNet (ICME, 2021)	35.5	62.0	35.1
TPH-YOLOv5(ICCVW, 2021)	42.1	63.1	**45.7**
Zoom&Reasoning Det (SPL, 2022)	39.0	66.5	39.7
UFPMP-Det (AAAI, 2022)	39.2	65.3	40.2
UFPMP-Det+MS (AAAI, 2022)	40.1	66.8	41.3
AdaZoom (TMM, 2022)	40.3	**66.9**	41.8
TPH-YOLOv5++ (ICCVW, 2021)	41.4	61.9	45.0
EVORL(AAAI, 2024)	42.2	66.0	44.5
Proposed Method	**42.8**	66.7	45.1

Table 2: Comparisons with state-of-the-art methods on the VisDrone dataset.

Method	AP^S	AP^M	AP^L
Faster R-CNN (TPAMI, 2017)	8.4	21.5	14.7
ClusDet (ICCV, 2019)	9.1	25.1	31.2
DMNet (CVPR Workshop, 2020)	9.3	26.2	35.2
AdaZoom (TMM, 2022)	14.2	29.2	28.4
Zoom&Reasoning Det (SPL, 2022)	15.3	32.7	30.8
EVORL(AAAI, 2024)	21.8	40.4	35.9
Proposed Method	**22.9**	**40.8**	**37.5**

Table 3: Performance comparison on different object sizes.

For a detailed evaluation of performance across various object sizes, Table 3 presents the average precision metrics of diffent scale objects: AP^S,AP^M and AP^L on the UAVDT dataset. These metrics measure detection accuracy for objects with areas smaller than 32 × 32 pixels, between 32 × 32 and 96 × 96 pixels, and larger than 96 × 96 pixels, respectively. The proposed method demonstrates substantial improvements over EVORL in all categories, achieving an AP^S of 22.9 compared to EVORL's 21.8, reflecting enhanced detection of small objects. For medium-sized objects, the proposed method's AP^M of 40.8 exceeds EVORL's 40.4, indicating superior performance. Additionally, the method achieves an AP^L of 37.5 for large objects, slightly better than EVORL's 35.9. These results confirm the

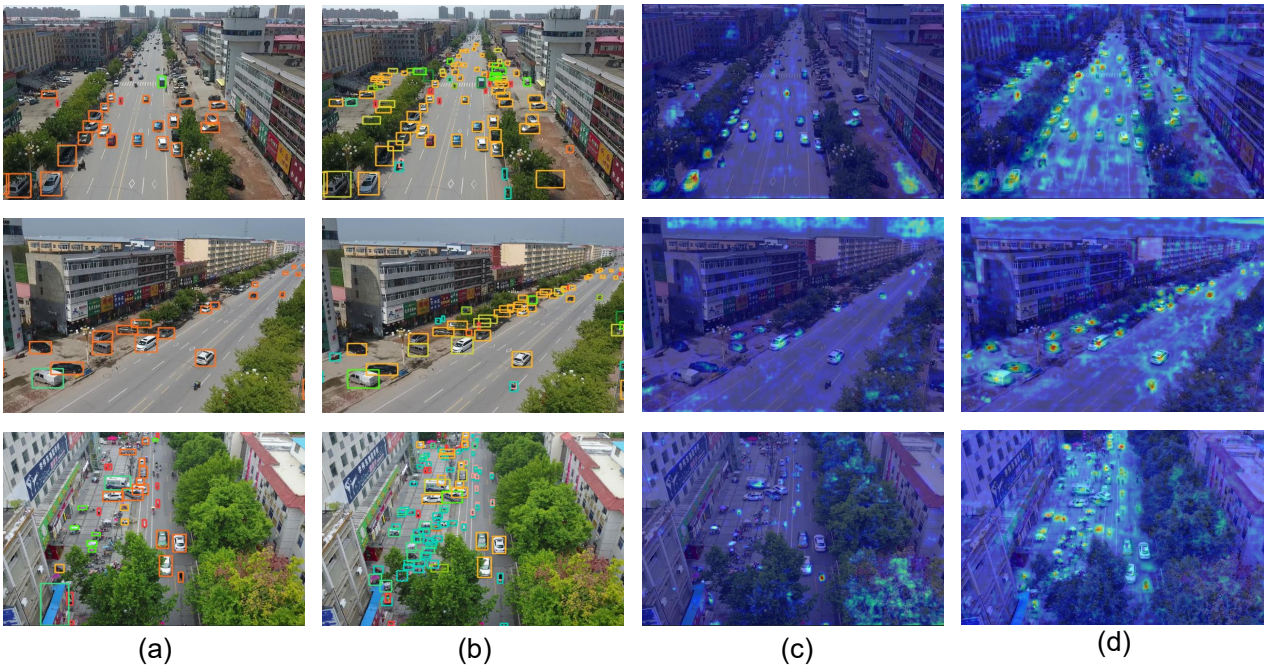

Figure 3: Visualization comparison. (a) The detection results of YOLOv8. (b) The detection results of SPAR. The small and dense objects are effectively detected as well. (c) Heatmap of YOLOv8, displaying less focused activation on small objects. (d) Heatmap with SPAR, showing more precise and concentrated activation on objects, particularly for small objects.

proposed method's overall effectiveness in improving detection accuracy across various object sizes.

Visualization comparison. Fig. 3 provides a comprehensive visualization of the performance of the SPAR method. The columns (a) and (b) show the detection results without and with SPAR respectively. The detection results in column (b) clearly demonstrate the superiority of the SPAR method, with small and dense objects being effectively detected. This improvement is further substantiated by the heatmaps shown in columns (c) and (d). Column (c), which represents the heatmap without SPAR, shows less focused activation, particularly for small objects. In contrast, column (d) illustrates the heatmap with SPAR, displaying more precise and concentrated activation on objects, with a marked improvement in the detection of small objects. These visual comparisons highlight the enhanced accuracy and robustness of the SPAR method in detecting small and densely packed objects.

Ablation Study

In order to show the effectiveness of each module, we conduct ablation experiments on VisDrone datasets, as shown in Table 4. Yolov8 is the baseline model; "+SP" is the baseline with only self-promting module; "+SP+AR(w/o text node)" denotes using self-prompting and analogical reasoning modules but the graph without category-level nodes. Based on Table 4, it can be concluded that all modules contribute to the final performance. Specifically, the combined using of category-level nodes and pixel-level nodes results in improved performance compared to using only pixel-level nodes; AP, AP50 and AP75 on VisDrone are increased by 0.3, 1.5 and 1.5 respectively. It proves the effectiveness of language guided analogical reasoning.

Method	AP	AP50	AP75
YOLOv8	41.3	63.3	40.9
+SP	42.1	64.0	41.7
+SP+AR(w/o text node)	42.5	65.2	43.6
SPAR	42.8	66.7	45.1

Table 4: Ablation study of SPAR on the VisDrone dataset.

Conclusion

We proposed a novel Self-Prompt Analogical Reasoning (SPAR) method to enhance object detection accuracy in UAV imagery. This approach integrates two key components. Self-prompting module generates context-aware prompts to enrich feature representation and also prompt objectness in feature maps based score maps. While analogical reasoning module employs graph-based reasoning to improve the detection of small and challenging objects. We formulated an analogical reasoning framework: deduction, mapping and inference. By constructing two kinds of graph nodes corresponding to text and visual features, knowledge deduction is performed; graph edge construction implements knowledge mapping; and graph convolution performs inference. Experiment results demonstrate superior performance compared to traditional methods.

Acknowledgments

This work is supported by the National Natural Science Foundation of China(62276048,62476169), Chengdu Science and Technology Projects (2023-YF06-00009-HZ) and Postdoctoral Fellowship Program of CPSF (GZC20233323).

References

Acharya, M.; Roy, A.; Koneripalli, K.; Jha, S.; Kanan, C.; and Divakaran, A. 2022. Detecting out-of-context objects using graph context reasoning network. In *IJCAI*.

Bianchi, F.; Attanasio, G.; Pisoni, R.; Terragni, S.; Sarti, G.; and Lakshmi, S. 2021. Contrastive Language-Image Pre-training for the Italian Language. arXiv:2108.08688.

Chen, X.; and Gupta, A. 2017. Spatial memory for context reasoning in object detection. In *Proceedings of the IEEE international conference on computer vision*, 4086–4096.

Chen, Y.; Rohrbach, M.; Yan, Z.; Shuicheng, Y.; Feng, J.; and Kalantidis, Y. 2019. Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 433–442.

Cordonnier, J.-B.; Loukas, A.; and Jaggi, M. 2020. Multi-Head Attention: Collaborate Instead of Concatenate. arXiv:2006.16362.

Dadsetan, S.; Rose, G.; Hovakimyan, N.; and Hobbs, J. 2021. Detection and prediction of nutrient deficiency stress using longitudinal aerial imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14729–14738.

Deng, S.; Li, S.; Xie, K.; Song, W.; Liao, X.; Hao, A.; and Qin, H. 2020. A global-local self-adaptive network for drone-view object detection. *IEEE Transactions on Image Processing*, 30: 1556–1569.

Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; and Tian, Q. 2018. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 370–386.

Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. 2019. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Everingham, M.; Eslami, S. M. A.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1): 98–136.

Fang, K.; and Li, W.-J. 2020. DMNet: difference minimization network for semi-supervised segmentation in medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 532–541. Springer.

Feng, X.; Yao, X.; Cheng, G.; and Han, J. 2022. Weakly supervised rotation-invariant aerial object detection network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14146–14155.

Fu, K.; Li, J.; Ma, L.; Mu, K.; and Tian, Y. 2020. Intrinsic Relationship Reasoning for Small Object Detection. arXiv:2009.00833.

Garg, P.; Mandal, M.; and Narang, P. 2021. Improving Aerial Instance Segmentation in the Dark with Self-Supervised Low Light Enhancement. arXiv:2102.05399.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation.

Han, J.; Ding, J.; Xue, N.; and Xia, G.-S. 2021. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2786–2795.

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2018. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3588–3597.

Huang, Y.; Chen, J.; and Huang, D. 2022. UFPMP-Det: Toward accurate and efficient object detection on drone imagery. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 1026–1033.

Lee, C.; Son, J.; Shon, H.; Jeon, Y.; and Kim, J. 2024. FRED: Towards a Full Rotation-Equivariance in Aerial Image Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2883–2891.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Li, S.; Ye, M.; Zhou, L.; Li, N.; Xiao, S.; Tang, S.; and Zhu, X. 2024. Cloud Object Detector Adaptation by Integrating Different Source Knowledge. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lin, J.; and Gong, S. 2023. GridCLIP: One-Stage Object Detection by Grid-Level CLIP Representation Learning. arXiv:2303.09252.

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 *European Conference on Computer Vision*, 740–755. Springer.

Liu, F.; Yao, L.; Zhang, C.; Wu, T.; Zhang, X.; Jiang, X.; and Zhou, J. 2024. Scale-Invariant Feature Disentanglement via Adversarial Learning for UAV-based Object Detection. arXiv:2405.15465.

Liu, S.; Zhang, H.; Qi, Y.; Wang, P.; Zhang, Y.; and Wu, Q. 2023. Aerialvln: Vision-and-language navigation for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15384–15394.

Meethal, A.; Granger, E.; and Pedersoli, M. 2023. Cascaded zoom-in detector for high resolution aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2046–2055.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided

dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18082–18091.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.

Shi, C.; and Yang, S. 2023. Edadet: Open-vocabulary object detection using early dense alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15724–15734.

Shi, H.; Hayat, M.; Wu, Y.; and Cai, J. 2022. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9611–9620.

Sun, H. 2024. Ultra-High Resolution Segmentation via Boundary-Enhanced Patch-Merging Transformer. arXiv:2412.10181.

Wang, Y.; Zou, H.; Yin, M.; and Zhang, X. 2023. Smff-yolo: A scale-adaptive yolo algorithm with multi-level feature fusion for object detection in uav scenes. *Remote Sensing*, 15(18): 4580.

Wei, Z.; Duan, C.; Song, X.; Tian, Y.; and Wang, H. 2020. AMRNet: Chips Augmentation in Aerial Images Object Detection. arXiv:2009.07168.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, Z.; Suresh, K.; Narayanan, P.; Xu, H.; Kwon, H.; and Wang, Z. 2019. Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1201–1210.

Xu, H.; Jiang, C.; Liang, X.; Lin, L.; and Li, Z. 2019. Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6419–6428.

Xu, J.; Li, Y.; and Wang, S. 2021. AdaZoom: Adaptive Zoom Network for Multi-Scale Object Detection in Large Scenes. arXiv:2106.10409.

Yan, Q.; Zheng, J.; Reding, S.; Li, S.; and Doytchinov, I. 2022. Crossloc: Scalable aerial localization assisted by multimodal synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17358–17368.

Yang, F.; Fan, H.; Chu, P.; Blasch, E.; and Ling, H. 2019a. Clustered object detection in aerial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8311–8320.

Yang, F.; Fan, H.; Chu, P.; Blasch, E.; and Ling, H. 2019b. Clustered object detection in aerial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8311–8320.

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.

Zhang, J.; Huang, J.; Chen, X.; and Zhang, D. 2019. How to fully exploit the abilities of aerial image detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Zhang, J.; Yang, X.; He, W.; Ren, J.; Zhang, Q.; Zhao, Y.; Bai, R.; He, X.; and Liu, J. 2024. Scale Optimization Using Evolutionary Reinforcement Learning for Object Detection on Drone Imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 410–418.

Zhang, Y.; Wu, C.; Guo, W.; Zhang, T.; and Li, W. 2023. CFANet: Efficient detection of UAV image based on cross-layer feature aggregation. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–11.

Zhao, Q.; Liu, B.; Lyu, S.; Wang, C.; and Zhang, H. 2023. Tph-yolov5++: Boosting object detection on drone-captured scenarios with cross-layer asymmetric transformer. *Remote Sensing*, 15(6): 1687.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16793–16803.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, C.; Chen, F.; Ahmed, U.; Shen, Z.; and Savvides, M. 2021a. Semantic relation reasoning for shot-stable few-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8782–8791.

Zhu, X.; Lyu, S.; Wang, X.; and Zhao, Q. 2021b. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2778–2788.