VL-Rotate: Vision Model Modulated by Language Model for Few-Shot Rotated Object Detection

Weihan Yin, Qinying Gu, Yaoyun Zhang, Lin Zhu, Qiao Sun, Liujia Yang, Xinbing Wang, and Nanyang Ye

***Abstract*—Rotated object detection (ROD) demands precise localization and angle prediction in dense scenes, yet the full potential of integrating natural language for improvement remains largely unexplored, especially in few-shot learning for out-of-distribution (OoD) scenarios. In this study, we introduce VL-Rotate, an effective vision model that integrates text-based prior knowledge from CLIP's text encoder to improve object representations in embedding space, and selectively deactivate classification features by a gradient-guided regularization method. We incorporate two innovative modules: CLIP-guided Fine-Tuning (CFT) and Masked Feature Heuristics Dropout (MFHD), guiding the model's fine-tuning throughout the training phase. Aimed at elevating detection accuracy and bolstering few-shot OoD inference capabilities, we conducted experiments in two areas of OoD research: domain adaptation and domain generalization. Compared to prior works, VL-Rotate achieves state-of-the-art results across all experiments, reaching an improvement up to 45.09% and 5.24% respectively on these two tasks, demonstrating the benefits of natural language guidance and text-image alignment. The experimental results validate the model's effectiveness and potential in advancing ROD.**

***Index Terms*—Rotated object detection, vision language learning, out-of-distribution generalization, few-shot learning.**

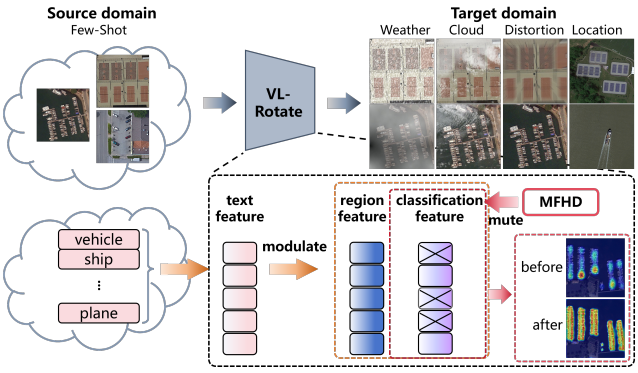

Fig. 1. An overview of our work. VL-Rotate aims to learn from a k-shot source domain and generalize to the target domain with unseen data. Our approach integrates text-based prior knowledge to modulate object features and mutes classification features with Masked Feature Heuristics Dropout (MFHD) to broaden feature participation, stabilize predictions, and improve generalization.

I. Introduction

ROTATED Object Detection (ROD) is a rapidly advancing area in computer vision, with recent innovations [1]–[4] driving significant progress in applications like object detection in remote-sensing images. Given that objects in aerial images are often densely packed, elongated, and arbitrarily oriented, oriented bounding boxes (OBB) have become the preferred method over traditional horizontal boxes for object localization, with many well-designed detectors showing promising results on challenging datasets.

Current research predominantly emphasizes the refinement of network architectures, feature extraction techniques, and loss functions under the assumption of independent and identically distributed (i.i.d.) data to elevate detection accuracy. However, ROD faces challenges when dealing with out-of-distribution (OoD) data in aerial images. The complexity of remote sensing environments—affected by dynamic weather, cloud cover, varying illumination, and seasonal changes—introduces uncertainty and incomplete information. Besides, technical disparities across data sources create inconsistencies in image resolution, noise, and color spaces, further complicating cross-domain generalization. The diversity in object states across geographic locations and movement patterns exacerbates these difficulties. Therefore, it is crucial for ROD models to address OoD conditions to maintain robust performance.

Remote sensing images often feature thousands of densely packed objects, such as cars or buildings, from a top-down perspective, which, coupled with complex OoD conditions, largely increases labeling costs. Privacy and national security concerns further limit the availability of public training data. Class imbalance adds to the challenge, particularly in detecting rare targets. Few-shot setting (FS) could emerge as a viable solution by refining features from limited samples, allowing models to quickly adapt to new classes and improving resource efficiency under OoD cases.

We observe that remote sensing images provide essential top-down visual information, while text offers semantic and abstract context, making cross-modal learning—especially the integration of vision and language—promising for advancing ROD. The top-down perspective of remote sensing images provides models with fixed top-view information of objects, while semantic information offers invariant anchor points in high-dimensional space. This creates the possibility for aligning visual features with semantic features in high-dimensional space.

Weihan Yin, Yaoyun Zhang, Lin Zhu, Liujia Yang, Xinbing Wang, and Nanyang Ye are with Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: yinweihan@sjtu.edu.cn; dancloud@sjtu.edu.cn; zhulin_sjtu@sjtu.edu.cn; yangliujia1008@sjtu.edu.cn; xwang8@sjtu.edu.cn; ynylincoln@sjtu.edu.cn).

Qinying Gu is with Shanghai Artifical Inteligence Laboratory, Shanghai 200232, China (e-mail: guqinying@pjlab.org.cn).

Qiao Sun is with Fudan University, Shanghai 200433, China (e-mail: qiaosun22@m.fudan.edu.cn). Yin and Gu contribute equally to this paper. Ye is the corresponding author.

Natural language descriptions of object attributes, shapes, or contexts are crucial for understanding categories and locations, improving model generalization under OoD and few-shot scenarios, while also reducing the demand for training samples to address few-shot learning challenges. While large-scale image-text pairs have been used for robust feature representation in pre-trained models, the unique challenges of ROD, such as complex backgrounds and rotated objects limit the effective use of textual information for detection. To date, no proposed method has fully harnessed the potential of language to improve ROD performance.

To address these limitations, we propose a novel approach named Vision Model Modulated by Language Knowledge for Few-Shot Rotated Object Detection (VL-Rotate). As shown in Fig. 1, our method leverages language representations within a few-shot setting to enhance the prediction of rotated objects in OoD scenarios. Our main contributions are as follows:

1) We propose a unique approach that integrates text-based prior knowledge to modulate object feature representations during fine-tuning, empowering the detector to achieve adequate generalization capabilities under unseen and complex data conditions.
2) We propose a novel dropout method that leverages gradients and GSNR to mute classification features, encouraging broader feature participation to achieve more stable predictions and enhance generalization on unseen data.
3) We conducted extensive experiments under few-shot settings on domain adaptation & generalization tasks, where VL-Rotate outperformed the baseline with up to 6.43% and 2.21% mAP gains on unseen data. To our knowledge, VL-Rotate is the pioneering work to integrate vision-language models for few-shot OoD ROD, and it is versatile, enhancing both classification and regression across single-stage, refine-stage, and two-stage detectors.

II. Related Works

A. Rotated Object Detection

Rotated object detection (ROD) is a challenging task involving dense object prediction and rotated bounding box prediction. In real-world applications, rotated object detection holds broad prospects and can be applied to scenarios such as remote sensing imagery and text detection. Traditional object detection employs horizontal bounding boxes (HBB) for annotation, which leads to overlapping bounding boxes among densely packed objects in crowded scenes. Moreover, when objects exhibit elongated shapes that are not aligned with the horizontal or vertical axes, such HBB annotations inevitably include irrelevant background regions. This not only interferes with the model's ability to localize targets accurately but also significantly degrades detection performance. In contrast, rotated object detection adopts oriented bounding boxes to achieve precise localization. This approach effectively eliminates interference between adjacent objects in dense scenarios and resolves the aforementioned issues, making it the predominant annotation method for rotated object detection tasks.

Novel methods have been proposed to address this problem, falling into three main categories according to their processing of candidate regions: two-stage detector [5]–[8], refine-stage detector [4], [9]–[14] and single-stage detector [1]–[3], [15], [16].

Two-stage detectors, building on the RCNN framework [17]–[19], excel in accuracy by employing deep convolutional networks, a Region Proposal Network (RPN) [20] for proposing regions, and task-specific heads. In recent studies, RoI Transformer [5] learns the transformation from Horizontal RoIs to Rotated RoIs and extracts rotation-invariant features from the Rotated RoIs, alleviating the misalignment between objects and RoIs. Redet [6] designs a RiRoI Align to extract rotated-invariant features and encodes both rotated equivariance and rotation invariance. Oriented R-CNN [7] introduces an oriented region proposal network for generating high-quality oriented proposals, coupled with an oriented detection head for region refinement and recognition. This framework has demonstrated superior performance on two benchmark rotated object detection datasets.

Single-stage detectors, consisting of a backbone, a feature pyramid network (FPN) [21], and task-specific heads, are valued for their speed and parameter efficiency, as they eliminate the need for extensive pre-defined region proposals. Recently, there has been a growing trend of exploring single-stage detectors. RTMDet [1] offers an efficient real-time detection solution through innovative architectural design. The framework is constructed with fundamental blocks composed of large-kernel depth-wise convolutional layers, which significantly enhance the model's capability to capture global contextual information. To further improve detection accuracy, RTMDet introduces a dynamic label assignment strategy incorporating soft labels, enabling more precise and high-quality matching between anchors and ground truth objects. This comprehensive approach simultaneously addresses both computational efficiency and detection performance in real-time scenarios. PSC [2] introduces the phase-shifting coder to cope with various periodic fuzzy problems in ROD and provides a unified framework by mapping the rotational periodicity of different cycles into the phase of different frequencies. H2RBox [3] presents a novel perspective by addressing the practical challenge of annotation costs. Recognizing that OBB annotation incurs significantly higher costs than HBB annotation during dataset preparation, the method introduces a plugin based on weakly supervised and self-supervised learning by training rotated object detection models using only HBB annotations. Specifically, H2RBox leverages instance segmentation results to compute minimum area rectangles, thereby deriving rotated prediction boxes.

Another type of detector that has garnered attention is the refine-stage detector, which builds upon one-stage detectors by eliminating anchor-based feature representations in favor of an anchor-free approach. This method learns anchor point representations and offset representations, improving accuracy without significantly sacrificing inference speed. RepPoints [11] is a classic refine-stage detector that

uses a set of points to represent objects, and learns the offset representation for semantically significant local areas during the refinement stage. SASM [10] introduces a shape-adaptive sample selection mechanism comprising two novel strategies: the Shape-Adaptive Selection Strategy (SASS) and Shape-Adaptive Measurement Strategy (SAMS). The SASS dynamically selects training samples based on both object geometry and feature distribution characteristics, while the SAMS quantitatively evaluates the quality of selected positive samples by assessing their localization potential. KFIoU [4] introduces a novel regression loss that addresses boundary point issues through the Gaussian Wasserstein distance. It comprises two terms: the first is a scale-insensitive center point loss designed to rapidly minimize the distance between the centers of two bounding boxes; the second employs the product of Gaussian distributions to inherently simulate the mechanism of SkewIoU. Oriented RepPoints [11] introduced an adaptive points representation to capture the geometric information of objects and proposed a corresponding quality assessment for adaptive points learning.

B. Out-of-Distribution Generalization

In recent years, various OoD generalization methods have been proposed to address distribution shifts. These methods can be categorized as follows [22]:

(1) **Domain generalization-based method** The goal of these methods is to train models on one or multiple related but distinct source domains while enabling them to achieve desirable generalization performance on unseen target domains. Methods based on domain generalization aim to learn transferable and robust feature representations across different domains. Common approaches include domain adversarial learning [23], [24], gradient-based methods [25], [26], meta learning [27], [28] and transfer learning [29], [30]. Domain adversarial learning encourages learning domain-invariant features through adversarial losses or adversarial training strategies. Gradien-based methods achieve out-of-distribution generalization by developing novel network update algorithms or gradient-based image augmentation techniques. Meta learning allows a model to be trained on multiple tasks to learn meta-features that can generalize across different tasks. Transfer learning leverages pre-training and fine-tuning to extract more generalizable features applicable across different domains.

(2) **Invariant representation learning** Exemplified by Invariant Risk Minimization (IRM) [31], this approach builds upon the theory of causally invariant features, transforming the optimization problem into an empirical risk minimization term and an invariant risk minimization term, ultimately deriving the training loss function. IRM explores causal relationships in data across different environments based on causal invariant features, aiming to learn cross-domain invariant features while preventing the model from overly relying on spurious features associated with specific environments. Invariant risk minimization games, proposed by [32], treat IRM as finding the Nash equilibrium of an ensemble game among multiple environments. Pareto Invariant Risk Minimization [33] introduces a multi-objective optimization perspective to understand the OoD optimization process and proposes a new optimization scheme to improve the robustness of OoD objectives. Sparse Invariant Risk Minimization [34] proposes an effective paradigm to address the contradiction between the generalization ability of IRM and overfitting.

(3) **Stable learning** This method combines causal inference with machine learning to tackle the OoD generalization problem from a different perspective. Stable learning methods include data augmentation [35] and Bayesian methods [36], etc. Current methods address OoD generalization uniquely, yet a specific focus on the OoD generalization problem in ROD is lacking. The variability in remote-sensing images, influenced by factors like weather and lighting, presents a valuable opportunity to study distribution shifts in ROD.

C. Vision-Language Pre-trained Models

Recent advancements in large-scale vision-language pre-training have notably enhanced downstream task performance. Prior works, such as VL-BERT [37], LX-MERT [38], and UNITER [39], have focused on learning joint representations from image-text pairs. Among these, Contrastive Language-Image Pretraining (CLIP) [40] stands out by effectively learning vision-language representations through maximizing image-text feature similarity using over 400 million pairs. CLIP's framework has inspired developments in few-shot learning, with models such as CoOp [41] and CoCoOp [42] innovating in prompt-tuning for better few-shot generalization, and CLIP-Adapter [43] introducing modifications for fine-tuning visual backbones in downstream applications. Furthermore, CLIP has been employed in various downstream fine-tuning tasks. DetCLIP [44] effectively integrates the CLIP paradigm with object detection tasks, DenseCLIP [45] applies CLIP to address pixel-text matching problems, and CLIP-ReID [46] leverages CLIP's cross-modal descriptive capabilities to accomplish image re-identification tasks. In recent years, cross-modal learning which integrates vision and language information, also known as vision-language alignment learning, has gradually emerged as a mainstream trend and has been widely applied in image recognition, particularly in tasks such as image classification and object detection [40], [42], [43], [45], [47], [48].

Recent advancements in large-scale vision-language pre-training have notably enhanced downstream task performance. Prior works [37]–[39] have focused on learning joint representations from image-text pairs. VL-BERT [37] is a general-purpose vision-language representation model that employs an extended Transformer [49] as its backbone. It takes both vision and language embedding as input, undergoes pre-training on large-scale conceptual caption datasets and pure text corpora, demonstrating the effectiveness of vision-language fusion in downstream tasks. LXMERT [38] follows an encoder-only paradigm and proposes a large-scale Transformer-based language model, which consists of an object-relation encoder, a language encoder, and a cross-modal encoder. UNITER [39] is a cross-modal pre-training model built upon Faster R-CNN [19] and BERT [50]. By adopting conditional masking in its pre-training tasks, it learns a universal image-text representation for joint vision-language tasks.

Among these, Contrastive Language-Image Pretraining (CLIP) [40] stands out by effectively learning vision-language representations through maximizing image-text feature similarity using over 400 million pairs. CLIP's framework has inspired developments in few-shot learning, with models such as CoOp [41] and CoCoOp [42] innovating in prompt-tuning for better few-shot generalization. CoOp [41] builds upon prompt-tuning by transforming the fixed prompts in CLIP into learnable ones, which are further fine-tuned on downstream tasks to better align text features with visual features in a high-dimensional joint latent space. On the basis of CoOp, CoCoOp [42] was proposed to address the overfitting issue of CoOp on base classes. By introducing a lightweight network that converts image features into instance-conditional vectors and combines them with word embeddings, CoCoOp enhances generalization performance on unseen classes. CLIP-Adapter [43] introduces a novel feature adapter structure into the visual backbones. By fine-tuning this structure on downstream applications, the model can adapt more efficiently while mitigating potential overfitting issues in few-shot learning.

Even in more complex tasks such as object detection and instance segmentation, CLIP-based research has emerged, including DenseCLIP [45] and RegionCLIP [48]. Since many CLIP-based vision-language models in classification tasks are trained at the image level, their direct transfer to region-level object detection often yields suboptimal performance. To address this, DenseCLIP proposes a novel dense prediction framework that reformulates CLIP's image-text alignment problem as a pixel-text matching task, enabling pixel-level dense prediction. RegionCLIP shares a similar motivation with DenseCLIP, aiming to bridge the gap between image-level classification and region-level detection. However, instead of pixel-text matching, RegionCLIP constructs region-text pairs and trains the model to align region-level visual features with textual features in a high-dimensional space, thereby learning fine-grained region-aware representations.

D. Out-of-Distribution in Rotated Object Detection

Remote sensing images often suffer from distribution shifts between training and test domains due to uneven sampling location distributions and susceptibility to unstable factors like weather and illumination conditions, presenting significant research potential. To date, scholars have begun exploring out-of-distribution generalization for rotated object detection through both Domain Generalization (DG) and Domain Adaptation (DA) approaches.

In DG, GOOD [51] addresses the limitation that current rotated object detection methods are predominantly studied under the I.I.D assumption which training anda test data are assumed to follow the same distribution. This significantly deviates from real-world scenarios. A domain-generalized rotated object detection task was proposed to investigate the detector's generalizability to arbitrary unseen targets. Given the diverse stylistic variations across domains, both object feature representation and orientation prediction become more challenging. GOOD employs CLIP to generate style-augmented data hallucinations, which are then learned through two key components: 1) rotation-aware content consistency learning enables the detector to acquire stable orientation representations from style-augmented data, while 2) style consistency learning further stabilizes the generalization capability of content representations across varying image styles. Through comprehensive generalization-oriented experimental setups, their proposed model achieves state-of-the-art performance on the DOTA dataset [52].

In DA, SFOD [53] addresses the scarcity of aerial imagery and its substantial variations across diverse geographical conditions, temporal changes, and weather patterns. SFOD proposes a source-free object detection framework to tackle domain shift challenges. The approach first employs a self-training paradigm to enhance the baseline model's performance. Subsequently, CLIP is integrated into the self-training framework, leveraging its semantic alignment capabilities to guide pseudo-label generation. Specifically, label noise is mitigated in self-training by aggregating CLIP's output scores with the classification confidence from the teacher model, thereby improving adaptation robustness.

III. Proposed Method

Traditional ROD methods rely on pre-trained weights and require substantial labeled data for downstream fine-tuning. In scenarios with limited samples, models risk overfitting, failing to capture the diversity of features and only memorizing specific instances without generalizing to new orientations. Moreover, significant distribution shifts between the few-shot training and test sets can lead to biased predictions due to spurious correlations in unseen domains.

To address these challenges, we propose VL-Rotate, which leverages language-guided text representations to modulate object-invariant features and iteratively deactivate features, encouraging all features to participate in making more stable predictions. Our approach also enables the efficient guidance of classification and regression features, allowing for rapid, plug-and-play deployment across various single-stage, two-stage, and refine-stage detectors. We employed the widely-used single-stage detector RetinaNet [54] as an example framework to illustrate how we build our method on top of it. The RetinaNet pipeline, depicted in Fig. 2, consists of a backbone network, a Feature Pyramid Network (FPN) [21], and task-specific heads for classification and regression.

A. CLIP-Guided Fine-Tuning

The large-scale vision-language model CLIP was designed to describe objects using semantic and abstract text concepts, enhancing object understanding. However, adapting CLIP from upstream classification to downstream ROD presents challenges, as ROD requires not only classification but also precise region and angle predictions, complicating the fusion of visual and textual information.

To address this issue, we proposed a CLIP-guided Fine-Tuning (CFT) method that leverages text information of CLIP to modulate the feature representations, enhancing the generalization ability under unseen data conditions in ROD. Given a k-shot image set $X_{tr} = \{x_i\} \in D_s, i \in [1, k]$ from

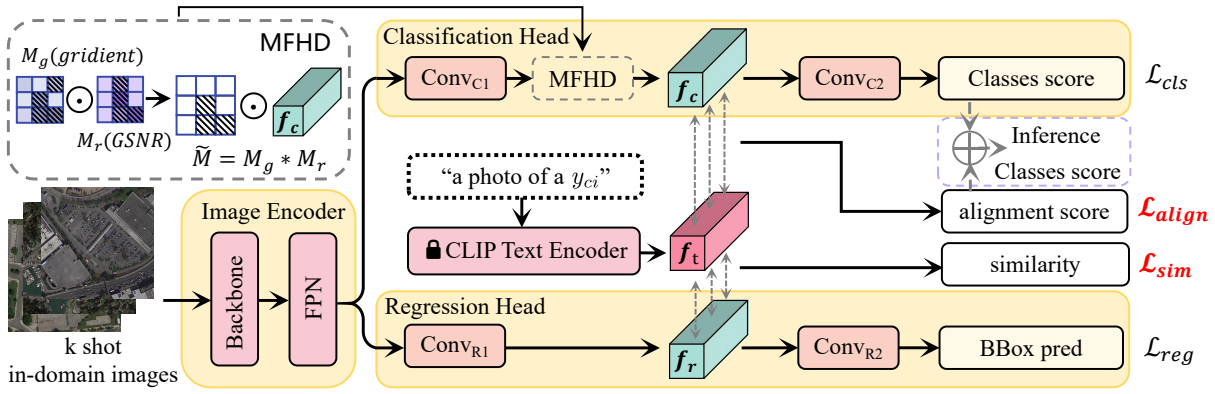

Fig. 2. The overall framework of the proposed VL-Rotate. RetinaNet is shown as the baseline, encompassing an Image Encoder and task-specific heads. VL-Rotate includes Masked Feature Heuristics Dropout (MFHD) and CLIP-Guided Fine-Tuning (CFT). During training, MFHD utilizes gradient and GSNR to mute the feature representations in f_c, encouraging the network to make predictions through more alternative features. CFT leverages text features f_t of CLIP to modulate f_c and f_r with text-classification heuristic alignment score and the best matching text-region fine-grained similarity, guiding the model to learn category-related textual descriptions. Final category scores are calculated by aggregating the alignment scores and classification scores in inference.

source domain D_s, as the training set, and a category set $Y_c = y_{ci}, i \in [1, m]$ containing category text, our goal is to fine-tune the model for effective generalization in the unseen target domain D_t.

1) Text-Classification Heuristic Alignment: We first introduce a Text-Category Heuristic Alignment (TCHA) technique that uses classical text tokens to guide the model in learning from imprecise textual descriptions. As shown in Fig. 2, the single-stage detector extracts image features using a backbone $I(\cdot)$ and a FPN, producing multi-scale output features $f_{fpn} = FPN(I(x))$. The classification head then applies a series of convolutional layers $Conv_{C1}(\cdot)$ to derive classification features $f_c \in \mathbb{R}^{b\times c\times h\times w}$ from f_{fpn}, where b, c, h, and w represent the batch size, channels, height, and width of the feature map. These features are further processed through $Conv_{C2}(\cdot)$ to output the classification results for each anchor or point.

Following CLIP's framework, we design a text description P_c as "a photo of a y_{ci}" and feed it into the CLIP text encoder $T(\cdot)$ to generate text features $f_t \in \mathbb{R}^{m\times c_t}$, where c_t is the dimension. We modify $Conv_{C1}(\cdot)$ to match the output channel dimension to c_t, enabling f_c to facilitate alignment learning and classification. By leveraging pre-trained knowledge from CLIP's text encoder, f_c is heuristically fine-tuned with text guidance, enhancing robustness in OoD inference.

During training, considering that f_t and f_c reside in different embedding spaces, we freeze the text encoder and fine-tune the detector. To guide alignment learning, we introduce an alignment loss, $\mathcal{L}_{align}$ which is computed by taking the inner product between f_t and f_c, yielding alignment scores $s_{align} = f_c \cdot f_t^T$ for classification, where f_c is reshaped to $\mathbb{R}^{b\times(h\times w)\times c_t}$. The original classification head and the alignment learning component are fine-tuned independently to avoid interference. $\mathcal{L}_{align}$ shares the same form as the classification loss $\mathcal{L}_{cls}$ used in RetinaNet and is minimized to align text features and classification features in high-dimensional space. $\mathcal{L}_{align}$ is defined as:

$$\mathcal{L}_{align} = -\alpha'(1 - s_{align})^{\gamma'} \log(s_{align}) \tag{1}$$

where α' is the balanced weight, set to 0.25 by default, and γ' is the focusing parameter, set to 2 by default.

During inference, the prediction results s_{cls} of each categories from $Conv_{C2}(\cdot)$ and the alignment scores s_{align} are combined to form the final classification result s:

$$s = \lambda s_{cls} + (1 - \lambda) s_{align} \tag{2}$$

where $\lambda = 0.5$ balances the two components, merging the model's intrinsic classification ability with text-based prior knowledge for more stable predictions. The motivation for combining s_{cls} and s_{align} stems from the distinct roles they play. Specifically, s_{cls} represents the result of learning the in-domain data distribution through the classification branch, whereas s_{align} reflects the similarity results guided by textual priors to align classification features with textual descriptions. By integrating these two scores, VL-Rotate can achieve accurate predictions for in-domain objects while leveraging textual descriptions to enhance generalization performance on unseen data distributions, thereby yielding more robust predictions and reducing the likelihood of generating erroneous labels from either source.

2) Text-Region Fine-grained Similarity: Despite the intuitive notion that textual information is agnostic to regions, it encapsulates descriptive features relevant to various categories, aiding the model in distinguishing between foreground and background. Motivated by this insight, we introduce a novel Text-Region Fine-grained Similarity (TRFS) technique in the CFT framework.

TRFS promotes the learning of fine-grained text-region correspondences, reinforcing each other during training, and improving the model's ability to understand the nuanced relationships between textual descriptions and visual regions.

In the regression head, the initial convolutional layer, $Conv_{R1}(\cdot)$, extracts region features $f_r \in \mathbb{R}^{b\times c\times h\times w}$ from f_{fpn}. Subsequently, $Conv_{R2}(\cdot)$ processes f_r to generate the final regression predictions. To facilitate this transition, we modify the output channel dimension of $Conv_{R1}(\cdot)$ to c_t, reshaping the features to $f_r \in \mathbb{R}^{b\times(h\times w)\times c_t}$, where $n = h \times w$ denotes the number of regions. Parallel to the classification

branch, we employ a text prompt P_r = "a photo of a y_{ci}" to extract region-related text features f_t from the CLIP text encoder.

The text-region similarity between the text feature f_{t_i} for the i-th category and all region features f_r is denoted as:

$$\Omega(f_r, f_{t_i})_i = \frac{1}{N}\sum_{j=1}^{N} f_{r_j} f_{t_i}^T \tag{3}$$

The total text-region similarity $\Omega(f_r, f_t)$ is calculated by summing these individual similarities in (3):

$$\Omega(f_r, f_{t_i}) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} f_{r_j} f_{t_i}^T \tag{4}$$

This measure reflects the similarity between the image x and the category set Y_C. However, since it includes all region features, it may incorporate background regions unrelated to the text, especially in remote-sensing images where objects are typically small, introducing noise into the similarity measure. To mitigate it, we select the region feature $\hat{f}_{r_i}$ in f_r that maximizes $\hat{f}_{r_i} f_{t_i}^T$ for the text feature f_{t_i}. This leads to the optimal-matching text-region similarity $\bar{\Omega}(f_r, f_t)$:

$$\bar{\Omega}(f_r, f_t) = \frac{1}{M}\sum_{i=1}^{M} \hat{f}_{r_j} f_{t_i}^T \tag{5}$$

Clearly, the total text-region similarity $\Omega(f_r, f_t)$ is maximized when considering only the most compatible region feature, such that $\Omega(f_r, f_t) \leq \bar{\Omega}(f_r, f_t)$.

However, this optimal-matching approach assumes a one-to-one correspondence between text and region features. In aerial images, where objects are densely packed, a one-to-many relationship often exists, with multiple objects of the same category appearing in the image. Thus, the optimal-matching similarity may not fully capture the text-region relationship, particularly in ROD where balancing the desired similarity with this one-to-many relationship is crucial. To address this, we introduce a softmax-weighted sum method to encode the probability distribution of text features across all region features. For the text features f_{t_i} of the i-th category and region features f_{r_j} of the j-th region, the softmax probability for selecting $f_{r_j} f_{t_i}^T$ is given by:

$$\text{softmax}(f_{r_j}, f_{t_i}^T) = \frac{\exp(f_{r_j} f_{t_i}^T/\gamma)}{\sum_r \exp(f_r f_{t_i}^T/\gamma)} \tag{6}$$

where γ is the hyperparameter controlling the sharpness of the softmax probability distribution.

The softmax probability is then incorporated into (4) to derive the final matching text-region similarity:

$$\bar{\Omega}(f_r, f_t) = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N} \text{softmax}(f_{r_j}, f_{t_i}^T) f_{r_j} f_{t_i}^T \tag{7}$$

This refined similarity accounts for all region features, appropriately weighting each and emphasizing those most aligned with the text. The corresponding text-region similarity loss is expressed as $\mathcal{L}_{sim}$:

$$\mathcal{L}_{sim} = -\frac{1}{B}\log\frac{\exp(\bar{\Omega}(f_r, f_{t_i})/\gamma)}{\sum_r \exp(\bar{\Omega}(f_r, f_{t_i})/\gamma)} \tag{8}$$

Here, B represents the batch size in a single iteration. This loss aids in training the model to learn a more refined text-region similarity to improve robustness to distributional shifts in OoD ROD. During training, the regression head and the TRFS branch are fine-tuned independently, while the TRFS branch is discarded during inference. The primary goal is to leverage text priors during training to enhance the region features' ability to distinguish foreground from background, aligning with the regression head's focus on localization without assuming classification responsibilities.

3) Overall Training Loss: Following RetinaNet, the total loss is calculated as:

$$\mathcal{L} = \omega_1\mathcal{L}_{cls} + \omega_2\mathcal{L}_{reg} + \omega_3\mathcal{L}_{align} + \omega_4\mathcal{L}_{sim} \tag{9}$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$, $\mathcal{L}_{align}$, $\mathcal{L}_{sim}$ represent the classification loss, regression loss, alignment loss, and refined text-region similarity matching loss. We use focal loss [54] for $\mathcal{L}_{cls}$, $\mathcal{L}_{align}$, and $\mathcal{L}_{sim}$, and GIoU loss [55] for $\mathcal{L}_{reg}$. The weights ω_1, ω_2, ω_3 and ω_4 are empirically set to 1:1:2:2.

B. Masked Feature Heuristics Dropout

Generalizing to unseen target domains poses a significant challenge, especially in few-shot cases where the model's performance can suffer due to its tendency to memorize specific features from limited data. Traditional regularization techniques like Dropout [61], which work by randomly deactivating network parameters, are often employed to address this issue. However, in few-shot settings, this random approach can inadvertently mute important features, limiting the model's ability to learn effectively.

To address this, inspired by [25], [62], we develop an advanced regularization method that strategically deactivates features based on gradient information rather than randomness. This approach, called Masked Feature Heuristics Dropout (MFHD), uses high gradients (i.e. gradients of parameters w.r.t the loss function) which denote the contribution of feature elements and high Gradient Signal-to-Noise Ratio (GSNR) [63] which measures the reliability of these contributions in OoD to create a mask that prevents the model from over-relying on "local optimal predictions" tied to the source domain, thereby enhancing generalization on unseen data. This approach can be likened to decision-making in a group: while individuals tend to rely on a leader's past correct decisions, unforeseen situations may increase the leader's likelihood of error. In such cases, collective input from all members enhances the group's resilience.

Unlike standard dropout methods that require extensive tuning and increased computational load, MFHD is applied specifically on the classification features f_c (see Fig. 2). This is because classification tasks are particularly vulnerable to memorizing specific instances instead of learning generalized features, while regression tasks require high precision, where even small errors can severely impact performance. This targeted approach helps maintain stability in the regression branch and ensures accurate predictions.

TABLE I

RESULT COMPARISON BETWEEN THE PROPOSED VL-ROTATE AND CD-FSOD DETECTORS (CF), DA & DG DETECTORS (DADG) AND TYPICAL ROD DETECTORS IN DOMAIN ADAPTATION TASK. THE CORRUPTIONS IN DIOR-C CAN BE CATEGORIZED INTO FOUR GROUPS: NOISE (**GAUSSIAN**, **SHOT**, **IMPULSE**, **SPECKLE**), BLUR (**DEFOCUS**, **GLASS**, **MOTION**, **ZOOM**, **GAUSSIAN**), WEATHER (**SNOW**, **FROST**, **FOG**, **BRIGHTNESS**, **SPATTER**), AND DIGITAL (**CONTRAST**, **ELASTIC TRANSFORM**, **PIXELATE**, **JPEG COMPRESSION**, **SATURATE**). FOR OOD EVALUATION, MODELS ARE FINE-TUNED ON 64-SHOT SAMPLES FROM THE SOURCE DOMAIN DIOR AND THEN DIRECTLY TESTED ON DIOR-CLOUDY AND DIOR-C. WE REPORT THE AVERAGE MAP (OOD MAP, %) ON BOTH DATASETS. ID EVALUATION (ID MAP, %) USES THE SAME TRAINING PROTOCOL BUT TEST ON DIOR. FRCNN DENOTES FASTER RCNN AND ORP DENOTES ORIENTED REPPOINTS.

	Method	DIOR-Cloudy	DIOR-C																			OoD	ID
		Cloudy	Ga	Sh	Im	Sp	De	Gl	Mo	Zo	Ga	Sn	Fr	Fo	Br	Sp	Co	El	Pi	JP	Sa	mAP	mAP
CF	CD-ViTO [56]	21.15	19.26	18.39	19.68	19.47	20.07	16.17	18.97	6.07	21.06	18.53	13.62	21.19	24.74	21.62	21.03	20.88	23.62	23.79	25.79	19.76	26.08
	Distill-FSOD [57]	28.13	18.23	18.85	20.07	22.18	27.68	17.68	23.79	10.73	29.56	17.50	18.63	31.82	35.41	26.48	29.28	30.39	30.74	28.31	37.57	25.15	38.52
DADG	IRG-SFDA [58]	15.30	4.12	3.82	5.36	7.34	14.58	12.39	13.01	6.54	16.29	7.52	8.72	17.67	19.89	13.77	15.87	18.48	17.85	18.29	20.76	12.88	21.77
	SFOD [53]	21.09	12.64	11.86	12.45	15.41	19.50	15.64	19.08	12.17	21.57	13.06	15.41	25.32	26.73	16.18	24.25	24.68	22.39	23.13	27.34	19.00	27.76
	OA-DG [59]	28.26	21.88	**21.23**	**21.60**	23.72	25.51	15.57	23.24	12.38	27.24	17.00	20.02	30.12	33.60	23.63	27.03	29.73	28.50	30.59	35.81	24.83	36.56
Typical ROD	*two-stage*:																						
	Faster RCNN OBB [20]	28.67	12.86	12.49	13.08	15.13	22.17	18.69	22.24	12.99	23.94	15.58	18.12	26.76	35.17	22.50	22.68	31.83	30.41	32.21	36.97	22.72	38.79
	Oriented RCNN [7]	31.09	16.23	15.56	15.50	18.68	22.92	19.28	23.98	13.65	24.14	15.70	19.83	28.21	38.60	24.26	25.19	34.80	33.42	36.13	41.06	24.91	42.83
	RoI Transformer [5]	33.34	15.36	14.16	15.32	17.86	23.17	21.61	25.67	16.12	24.48	16.68	20.62	29.28	40.08	24.44	24.61	37.99	35.98	38.01	43.28	25.90	44.94
	ReDet [6]	36.73	**22.21**	20.77	20.99	23.52	28.51	25.37	28.81	16.01	30.92	20.60	24.97	35.85	42.04	30.99	34.31	37.45	37.37	39.31	45.04	30.09	47.12
	FRCNN OBB+VL-Rotate	31.45	14.67	13.76	15.13	18.19	22.70	19.78	23.73	14.75	24.68	16.86	19.17	27.86	36.09	23.75	23.57	33.44	32.38	35.34	39.38	24.33	41.85
		+2.78	+1.81	+1.27	+2.05	+3.06	+0.53	+1.09	+1.49	+1.76	+0.74	+1.28	+1.05	+1.10	+0.92	+1.25	+0.89	+1.61	+1.97	+3.13	+2.41	+1.61	+3.06
	ReDet+VL-Rotate	38.94	21.58	19.17	20.89	22.97	**31.21**	25.77	**30.06**	17.39	**32.50**	**24.22**	**26.15**	**38.93**	42.54	**32.65**	**36.95**	38.66	38.07	40.19	46.25	**31.25**	47.78
		+2.21	-0.63	-1.60	-0.10	-0.55	+2.70	+0.40	+1.25	+1.38	+1.58	+3.62	+1.18	+3.08	+0.50	+1.66	+2.64	+1.21	+0.70	+0.88	+1.21	+1.17	+0.66
	single-stage:																						
	RetinaNet OBB [54]	17.02	9.00	8.99	9.17	10.00	12.50	12.75	12.66	8.02	14.17	11.72	12.67	15.11	21.48	15.83	13.57	20.15	18.25	19.76	22.32	14.26	23.22
	H2RBox [3]	18.07	7.67	6.26	8.92	9.42	14.83	14.14	16.06	9.54	16.17	10.60	9.61	15.91	23.07	13.39	14.00	21.10	19.52	20.66	23.41	14.62	25.17
	RTMDet-l [1]	27.13	16.59	15.84	16.61	19.23	20.36	19.67	22.27	11.58	20.85	16.84	18.36	22.31	30.57	25.38	22.84	29.60	31.01	32.79	36.04	22.79	37.09
	FCOS OBB-PSC [2]	30.49	14.51	13.43	15.20	16.90	22.56	20.30	22.64	12.18	24.49	17.33	19.48	26.64	35.70	24.18	23.38	32.09	32.30	34.26	38.72	23.84	39.97
	FCOS OBB [15]	31.79	13.88	14.06	14.81	16.77	23.86	19.24	23.88	13.66	25.78	18.56	20.92	30.68	35.99	24.37	27.50	33.06	32.30	34.70	39.70	24.78	41.71
	Rotated ATSS [16]	32.48	16.33	14.83	16.79	19.49	25.42	21.00	24.98	13.65	27.21	17.56	19.71	28.99	38.38	24.60	26.28	33.67	35.57	36.72	41.64	25.77	43.52
	RetinaNet OBB+VL-Rotate	26.44	12.68	11.85	11.34	13.72	19.58	16.84	20.21	11.29	20.57	13.01	17.34	24.17	32.45	19.85	20.77	29.51	28.02	29.99	34.07	20.69	35.34
		+9.42	+3.68	+2.86	+2.17	+3.72	+7.08	+4.09	+7.55	+3.27	+6.40	+1.29	+4.67	+9.06	+10.97	+4.02	+7.20	+9.36	+9.77	+10.23	+11.75	+6.43	+12.12
	RTMDet-l+VL-Rotate	34.37	19.12	18.17	18.94	22.38	21.29	17.68	24.17	10.30	22.73	20.98	22.94	33.01	39.83	29.58	29.12	29.97	31.77	33.72	42.42	26.12	44.64
		+7.24	+2.53	+2.33	+2.33	+3.15	+0.93	-1.99	+1.90	-1.28	+1.88	+4.14	+4.58	+10.7	+9.26	+4.20	+6.28	+0.37	+0.76	+0.93	+6.38	+3.33	+7.55
	refine-stage:																						
	S^2A-Net [60]	27.11	14.31	12.81	14.04	16.01	19.45	16.00	20.11	11.67	20.13	13.44	16.53	24.73	32.08	22.52	19.57	29.02	27.32	30.33	34.36	21.08	36.28
	R^3Det [12]	29.97	16.89	15.02	16.16	17.97	21.61	19.15	22.11	13.93	23.34	16.66	20.14	27.28	34.37	24.52	22.93	32.94	31.88	34.44	36.55	23.89	38.05
	RepPoints OBB [13]	30.91	11.70	11.67	11.39	14.66	21.48	21.92	23.48	14.26	23.34	17.91	19.76	26.83	34.36	27.19	24.25	32.61	31.74	33.90	36.74	23.51	38.22
	SASM [10]	36.19	13.03	11.63	12.19	15.21	24.25	24.50	26.99	16.96	26.29	20.82	23.51	32.53	42.09	29.96	27.21	39.68	36.64	41.51	45.62	27.34	47.86
	CFA [14]	37.77	18.39	17.45	18.30	21.28	26.52	23.20	27.51	16.88	27.87	22.18	23.16	34.98	43.95	30.54	32.95	39.83	38.72	43.07	47.38	29.60	49.07
	Oriented RepPoints [11]	37.71	20.31	19.36	19.89	23.59	28.01	25.66	27.19	15.79	29.95	19.87	24.16	34.60	43.15	31.22	29.89	40.46	39.18	43.05	47.71	30.04	49.38
	RepPoints OBB+VL-Rotate	31.43	12.40	11.66	11.86	15.47	21.86	21.65	23.31	14.76	24.05	18.44	20.30	26.67	35.64	26.86	25.02	34.84	32.90	35.22	39.52	24.19	40.53
		+0.52	+0.70	-0.01	+0.47	+0.81	+0.38	-0.27	-0.17	+0.50	+0.71	+0.53	+0.54	-0.16	+1.28	-0.33	+0.77	+2.23	+1.16	+1.32	+2.78	+0.69	+2.31
	SASM+VL-Rotate	38.67	12.38	11.35	11.37	15.27	28.06	**25.88**	29.28	**17.68**	30.21	20.98	24.09	34.90	44.03	30.97	31.73	**43.33**	**40.06**	**44.31**	48.14	29.13	50.81
		+2.48	-0.65	-0.28	-0.82	+0.06	+3.81	+1.38	+2.29	+0.72	+3.92	+0.16	+0.58	+2.37	+1.94	+1.01	+4.52	+3.65	+3.42	+2.80	+2.52	+1.79	+2.95
	ORP+VL-Rotate	**39.26**	21.01	19.81	20.61	**24.54**	25.45	23.18	25.32	16.14	27.52	21.25	25.10	36.23	**44.96**	30.93	30.63	40.41	39.62	44.06	**49.46**	30.27	**51.37**
		+1.55	+0.70	+0.45	+0.72	+0.95	-2.56	-2.48	-1.87	+0.35	-2.43	+1.38	+0.94	+1.63	+1.81	-0.29	+0.74	-0.05	+0.44	+1.01	+1.75	+1.99	+0.24

MFHD mutes the channels in f_c to obtain $\tilde{f}_c = \tilde{M} \odot f_c$, where "$\odot$" denotes element-wise product. $\tilde{M}$ is the mask to determine which feature in f_c should be muted, given by:

$$\tilde{M} = M_g \odot M_r \tag{10}$$

Given the gradients $g_c = \frac{\partial \mathcal{L}_{cls}(f_c, y_c)}{\partial \theta_c}$ of the classification loss $\mathcal{L}_{cls}$ with respect to the parameters θ_c of the top layers of $Conv_{C1}(\cdot)$, where y_c is the classification label, a first mask $M_g = \{m_g(i)\}$ by zeroing out the top p % of the most significant elements in g_c is calculated for the i-th element: $m_g(i)$ set to 0 if $g_c(i) \geq \mathcal{G}_p$ otherwise to 1, where $\mathcal{G}_p$ represents the threshold for the top p %. Next, MFHD computes GSNR for the parameters θ_c, defined as

$$r_c = \frac{\mathrm{E}^2_{(x,y_c)\sim\mathbb{D}}(g_c)}{\mathrm{Var}_{(x,y_c)\sim\mathbb{D}}(g_c)} \tag{11}$$

A second mask $M_r = \{m_r(i)\}$ is generated based on r_c, using a threshold $\mathcal{R}_p$ of the top p %. For the i-th element, $m_r(i)$ set to 0 if $r_c(i) \geq \mathcal{R}_p$ otherwise to 1. Empirically, we set p to 30%.

Additionally, a well-designed dropout schedule is critical. Applying MFHD throughout the entire training phase could interfere with the model's ability to learn generalizable features. Therefore, MFHD is activated after the first half of the training epochs, allowing the model to focus on learning general features early and on generalization capabilities later to avoid overfitting.

TABLE II

RESULT COMPARISON BETWEEN THE PROPOSED VL-ROTATE AND COMPETITORS ON DOMAIN GENERALIZATION TASK. WE REPORT THE ID MAP ON DOTA VALIDATION SET AND THE OOD MAP ON DIOR TEST SET.

	Method	ID mAP	OoD mAP
CF	CD-ViTO [56]	13.64	13.00
	Distill-FSOD [57]	31.96	28.29
DADG	SFOD [53]	-	-
	IRG-SFDA [58]	23.20	19.43
	OA-DG [59]	46.40	32.31
Typical ROD	*two-stage*:		
	Faster RCNN OBB [20]	48.21	44.64
	Oriented RCNN [7]	51.69	45.36
	RoI Transformer [5]	54.08	47.83
	ReDet [6]	54.23	48.39
	refine-stage:		
	Reppoints OBB [13]	50.08	46.05
	R^3Det [12]	50.80	45.19
	S^2A-Net [60]	51.07	47.12
	CFA [14]	51.82	47.01
	SASM [10]	53.89	50.02
	Oriented Reppoints [11]	54.96	49.00
	single-stage:		
	H2RBox [3]	36.99	37.34
	RetinaNet OBB [54]	44.10	42.19
	FCOS OBB-PSC [2]	50.11	44.40
	FCOS OBB [15]	50.84	47.33
	Rotated ATSS [16]	51.70	46.39
	RTMDet-l [1]	54.15	50.87
	RetinaNet OBB+VL-Rotate	46.17	44.40
		+2.07	+2.21
	RTMDet-l+VL-Rotate	**56.24**	**51.89**
		+2.09	+1.02

IV. EXPERIMENT

In this section, we employ two classical tasks to characterize the OoD dilemma: domain adaptation (DA) task and domain generalization (DG) task. Additionally, we conduct a series of ablation experiments to evaluate the effectiveness of VL-Rotate and rule out potential confounding factors.

A. Datasets

1) DOTA: DOTA [52] is a public large aerial image dataset for rotated object detection, which includes over 2,800 aerial images with 188,000 objects across 15 categories. The categories are plane, baseball diamond, bridge, ground track field, small vehicle, large vehicle, ship, tennis court, basketball court, storage tank, soccer ball field, roundabout, harbor, swimming pool, and helicopter. All images in the training and validation sets were cropped into patches of size 1024×1024, with a padding of 512 pixels.

2) DIOR: DIOR [64] offers 23k images with 190k instances across 20 categories, containing diverse object sizes and distinct acquisition conditions. The categories are airplane, airport, baseball field, basketball court, bridge, chimney, expressway service area, expressway toll station, dam, golf field, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, windmill. In DIOR, DIOR-C [65], and DIOR-Cloudy [65], images were resized to 800x800.

B. Experiment Settings

Adhering to the few-shot settings of CoOp [47] and the ROD settings, we focus on evaluating the fine-tuning performance of the methods in few-shot OoD ROD scenarios. The experiments include two parts: domain adaptation (DA) task and domain generalization (DG) task.

1) Domain Adaptation: We focus on evaluating performance under domain shifts. While datasets like DOTA-C [66] and DOTA-Cloudy [66] contain various domain shifts are available, the high experimental cost of evaluating these datasets—due to the need for individual assessments of different corruption types on servers—remains a significant challenge. To address this, we propose using alternative aerial remote sensing image datasets: DIOR-C [65] and DIOR-Cloudy [65]. DIOR-C includes 19 different types of corruptions from ImageNet-C [67] with a severity level of 3. DIOR-Cloudy is constructed using publicly available cloud images from DOTA-Cloudy through image synthesis. For our experiments, we use the original training set of DIOR [64] with 20 classes as the source data and randomly select 64 images to create a 64-shot training set. The test sets of DIOR-C and DIOR-Cloudy then serve as the unseen target data for evaluation.

2) Domain Generalization: We use the original DOTA [52] training set as source data, randomly selecting 16 images to create a 16-shot training set. The model's performance is evaluated on the DIOR test set to gauge its ability to transfer knowledge between different data distributions. We also use the DOTA validation set as the source test data. Following established protocols in domain-generalized object detection [68]–[70], we focus on the shared object categories between DOTA and DIOR, which include 10 classes: airplane, baseball field, bridge, ground track field, vehicle, ship, tennis court, basketball court, storage tank, and harbor.

3) Competitors: To conduct comprehensive experiments and provide valuable insights, we explored various methods in few-shot OoD RoD scenarios.

TABLE III
THE IMPLEMENTATION DETAILS FOR ALL METHODS. LR DENOTES THE LEARNING RATE, MO DENOTES MOMENTUM, WD DENOTES WEIGHT DECAY, AND LR SCHEDULE DENOTES THE LEARNING RATE SCHEDULE. SPECIFICALLY, MULTISTEPLR REFERS TO A SCHEDULE IN WHICH THE LEARNING RATE IS MULTIPLIED BY 0.1 AT THE 25TH AND 33RD EPOCHS, WHILE COSINEANNEALINGLR IS THE LEARNING RATE SCHEDULE PROPOSED BY RTMDET.

	Method	Schedule	Optimizer	LR	Mo	WD	LR Schedule
CF	CD-ViTO [56]	3x	SGD	0.001	-	-	-
	Distill-FSOD [57]	1x	SGD	0.01	-	-	WarmupMultiStepLR
DADG	IRG-SFDA [58]	1x	SGD	0.001	0.9	0.0001	-
	SFOD [53]	1x	SGD	0.0025	0.9	0.0001	WarmupStepLR
	OA-DG [59]	1x	SGD	0.01	0.9	0.0001	WarmupStepLR
Typical ROD	*two-stage*:						
	Faster RCNN OBB [20]	3x	SGD	0.005	0.9	0.0001	MultiStepLR
	Oriented RCNN [7]	3x	SGD	0.005	0.9	0.0001	MultiStepLR
	RoI Transformer [5]	3x	SGD	0.005	0.9	0.0001	MultiStepLR
	Redet [6]	3x	SGD	0.005	0.9	0.0001	MultiStepLR
	FRCNN OBB+VL-Rotate	3x	SGD	0.005	0.9	0.0001	MultiStepLR
	ReDet+VL-Rotate	3x	SGD	0.005	0.9	0.0001	MultiStepLR
	refine-stage:						
	S^2A-Net [60]	3x	SGD	0.0025	0.9	0.0001	MultiStepLR
	R^3Det [12]	3x	SGD	0.005	0.9	0.0001	MultiStepLR
	Reppoints OBB [13]	3x	SGD	0.008	0.9	0.0001	MultiStepLR
	SASM [10]	3x	SGD	0.008	0.9	0.0001	MultiStepLR
	CFA [14]	3x	SGD	0.008	0.9	0.0001	MultiStepLR
	Oriented Reppoints [11]	3x	SGD	0.008	0.9	0.0001	MultiStepLR
	RepPoints OBB+VL-Rotate	3x	SGD	0.008	0.9	0.0001	MultiStepLR
	SASM+VL-Rotate	3x	SGD	0.008	0.9	0.0001	MultiStepLR
	ORP+VL-Rotate	3x	SGD	0.008	0.9	0.0001	MultiStepLR
	single-stage:						
	RetinaNet OBB [54]	3x	SGD	0.0025	0.9	0.0001	MultiStepLR
	H2RBox [3]	3x	SGD	0.0025	0.9	0.0001	MultiStepLR
	RTMDet-l [1]	3x	AdamW	0.00025	-	0.05	CosineAnnealingLR
	FCOS OBB-PSC [2]	3x	SGD	0.0025	0.9	0.0001	MultiStepLR
	FCOS OBB [15]	3x	SGD	0.0025	0.9	0.0001	MultiStepLR
	Rotated ATSS [16]	3x	SGD	0.0025	0.9	0.0001	MultiStepLR
	RetinaNet OBB+VL-Rotate	3x	SGD	0.005	0.9	0.0001	MultiStepLR
	RTMDet-l+VL-Rotate	3x	AdamW	0.001	-	0.05	CosineAnnealingLR

Typical ROD Methods: We categorized ROD methods into single-stage detectors, refine-stage detectors, and two-stage detectors, examining their performance in tackling the significant challenges posed by few-shot OoD scenarios.

CD-FSOD Methods: Distill-FSOD [57] and CD-ViTO [56], two state-of-the-art Cross-Domain Few-Shot Object Detection (CD-FSOD) approaches, are introduced to explore whether they can address DG and DA tasks under ROD.

DA&DG Object Detection Methods: SFOD [53], IRG-SFDA [58], and OA-DG [59] were utilized to evaluate their few-shot performance under ROD.

4) Experiment Details: We address a more practical and challenging problem and draw attention to the real-world challenges of detecting rotated objects under few-shot and OoD conditions. To ensure a fair comparison, all baselines will be fine-tuned under both DA and DG tasks. Table III shows the experiment details on these two tasks. Our experiments were implemented using PyTorch 1.10 with MMRotate 1.x [71] on two NVIDIA GeForce RTX 4090. For fair evaluation, all methods use ResNet-50 [72] pre-trained on ImageNet as the backbone and follow the default experimental setup on MMRotate with 3x schedule, 2 GPUs $*$ 8 batch sizes per GPU, and SGD optimizer (except for RTMDet [1] with AdamW optimizer). VL-Rotate is trained with a learning rate of 0.005, momentum of 0.9, and weight decay of 0.0001. We set $\omega_1 : \omega_2 : \omega_3 : \omega_4$ to $1 : 1 : 2 : 2$, p to 0.3 (30%) and mask dropout schedule to post-1/2 of the maximum epochs. To avoid overfitting, Random flipping is employed without any additional tricks.

Notably, we also use ResNet-50 pre-trained on ImageNet for VL-Rotate, instead of the pre-trained backbone from CLIP, to eliminate the influence of CLIP's prior vision knowledge. This choice will be further discussed in ablation studies.

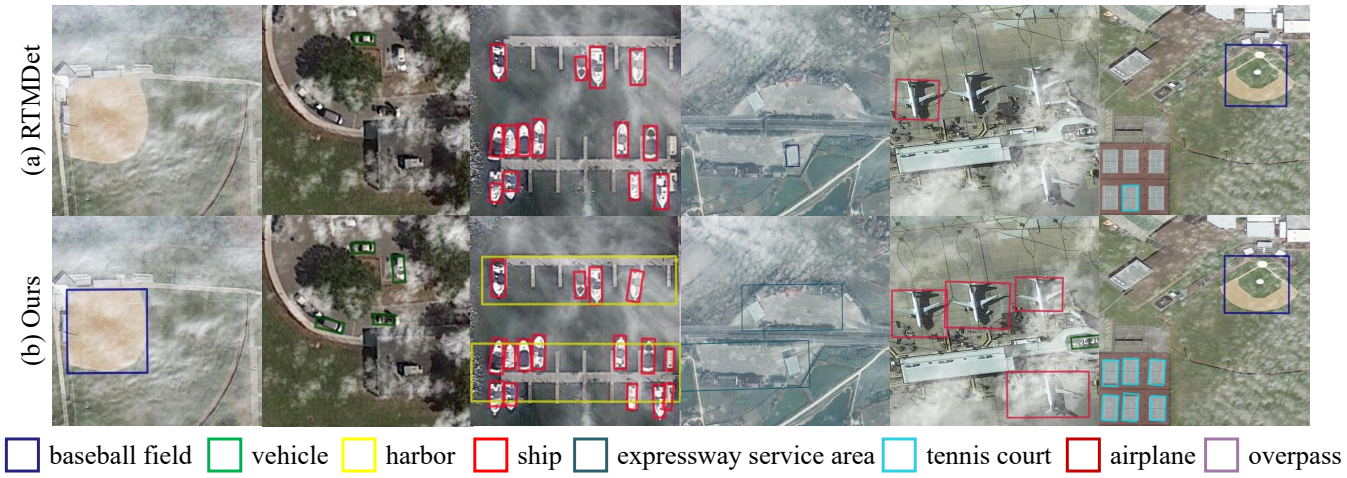

Fig. 3. Qualitative comparisons of the inference results between proposed VL-Rotate and the baseline model RTMDet on DIOR-Cloudy.

TABLE IV
ABLATION STUDY RESULTS OF EACH COMPONENT BASED ON RETINANET ON DOMAIN ADAPTATION TASK. "SCOREM" DENOTES THE SCORE MERGE IN CFT DURING INFERENCE. "IMPV" DENOTES THE OVERALL IMPROVEMENT COMPARED TO RETINANET.

Components					ID mAP	Impv	OoD mAP	Impv
CFT			MFHD					
TCHA	ScoreM	TRFS	Grad	GSNR				
					23.22		14.26	
✓					30.64	+7.42	17.44	+3.18
✓	✓				32.56	+9.34	18.92	+4.66
✓	✓	✓			32.74	+9.52	19.49	+5.23
✓	✓	✓	✓		15.32	-7.9	10.11	-4.15
✓	✓	✓		✓	33.84	+10.62	19.99	+5.73
✓	✓	✓	✓	✓	**35.34**	+12.12	**20.69**	+6.43

TABLE V
TOP: ABLATION STUDY RESULTS FOR VL-ROTATE USING DIFFERENT LANGUAGE MODELS. BOTTOM: ABLATION STUDY RESULTS FOR VL-ROTATE USING VARIANT CLIP TEXT ENCODER.

Method	Language	ID mAP	Impv	OoD mAP	Impv
baseline	-	23.22		14.26	
VL-Rotate	W2V [73]	12.37	-10.85	8.19	-6.07
VL-Rotate	BERT [50]	30.20	+6.98	18.01	+3.75
VL-Rotate	CLIP-Text [40]	**35.34**	+12.12	**20.69**	+6.43
Method	**CLIP Enc. Type**	**ID mAP**	**Impv**	**OoD mAP**	**Impv**
baseline	-	23.22		14.26	
VL-Rotate	EVA02-CLIP [74]	28.93	+5.71	17.95	+3.69
VL-Rotate	Long-CLIP [75]	33.61	+10.39	19.49	+5.23
VL-Rotate	SigLIP [76]	34.14	+10.92	20.55	+6.29
VL-Rotate	CLIP [40]	**35.34**	+12.12	**20.69**	+6.43

TABLE VI
ABLATION STUDY RESULTS FOR VL-ROTATE USING IMAGENET AND CLIP PRE-TRAINED RESNET-50 BACKBONE BASED ON RETINANET ON DOMAIN ADAPTATION TASK.

Method	Pre-Trained	ID mAP	Impv	OoD mAP	Impv
baseline	ImageNet	23.22		14.26	
VL-Rotate	CLIP	22.13	-1.09	12.23	-2.03
VL-Rotate	ImageNet	**35.34**	+12.12	**20.69**	+6.43

TABLE VII
ABLATION STUDY RESULTS FOR VL-ROTATE WITH MASKED FEATURE HEURISTICS DROPOUT (MFHD) ACTIVATED AT DIFFERENT SCHEDULES ON DOMAIN ADAPTATION TASK. THE SCHEDULES FOR ACTIVATING MFHD ARE SELECTED AS FULL TIME (FT), PRE-1/2 AND POST-1/2 OF THE MAXIMUM EPOCHS (PRE-1/2, POST-1/2), PRE-1/3, MID-1/3, POST-1/3 OF THE MAXIMUM EPOCHS (PRE-1/3, MID-1/3, POST-1/3), AND RANDOM ACTIVATION WITH 50% PROBABILITY (RANDOM).

Schedule	ID mAP	OoD mAP
FT	32.71	19.95
pre-1/2	32.59	19.58
post-1/2	**35.34**	**20.69**
pre-1/3	32.04	19.19
mid-1/3	32.94	19.54
post-1/3	32.72	18.91
random	31.18	18.88

C. Main Results

1) Domain Adaptation: We deploy VL-Rotate in some of the ROD methods and compare with all competitors on DA tasks, as shown in Table I. Our method consistently outperforms others, with the most notable improvement observed with RetinaNet. Specifically, VL-Rotate achieves average OoD mAP gains of 1.61% with Faster RCNN, 1.17% with ReDet, 6.43% with RetinaNet, 3.33% with RTMDet, 0.69% with RepPoints, 1.79% with SASM and 1.99% with Oriented RepPoints. Among all the tested methods—whether CD-FSOD, DG & DA, or typical ROD method—the ReDet-based method of VL-Rotate achieves the highest performance, with 31.25% mAP on the target domain, setting a new state-of-the-art. A qualitative comparison of VL-Rotate and the baseline RTMDet is shown in Fig. 3.

2) Domain Generalization: Table II presents the DG results, where our method consistently improves the selected baselines. Notably, it increases mAP by 2.21% for RetinaNet and 1.02% for RTMDet-l on the DIOR test set. RTMDet achieves the best OoD mAP among all baselines, and with VL-Rotate, it further improves, reaching a new SOTA mAP of 56.24% on the source domain and 51.89% on the target domain. Note that for SFOD, a method used for DA tasks, training requires test sets with corruption as the unseen target domain, which leads to the lack of a reference. The source domain results are derived from the DOTA validation set for reference only.

D. Ablation Study

We conduct a series of ablation experiments to evaluate the effectiveness of VL-Rotate and exclude potential confounding

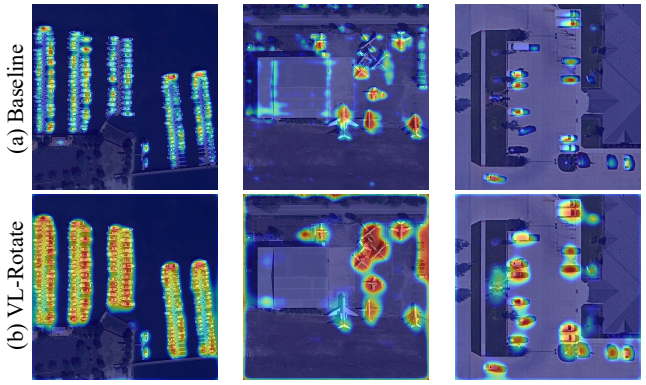

Fig. 4. Visualization of VL-Rotate and the baseline.

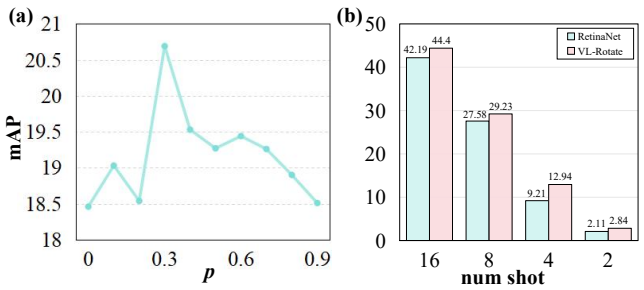

Fig. 5. Ablation study results of target domain for VL-Rotate about (a) different p on DA task; (b) different shot on DG task.

factors. Unless otherwise specified, the experimental settings align with those described in the experiment details. Similarly, unless specified, VL-Rotate was implemented in RetinaNet with RetinaNet serving as the baseline.

1) Performance Analysis of Components: To evaluate the impact of VL-Rotate, we conduct a series of controlled experiments on DA task. We divide CFT into three parts: TCHA, score merge, and TRFS. The results show that each of the components achieved different degrees of improvement in detection accuracy. By introducing TCHA, which guides to align classification features with invariant text prior features by CLIP, the model's accuracy is improved by 7.42/3.18% mAP. Score merge fusing alignment scores and classification scores during inference, leads to an improvement of 1.92/1.48% mAP. MFHD computes refined text-region similarity, guiding regression features to distinguish foreground from background, resulting in a 0.18/0.57% mAP improvement. Additionally, the MFHD module is evaluated separately using high-gradient masked and high-GSNR masked conditions. MFHD selectively forgets the most predictive classification features, enhancing the model's accuracy by 0.72/0.48% mAP. Gradients denote the contribution of feature elements, whereas GSNR measures the reliability of these contributions in OoD setting. Noise impacts features, making low-contribution elements more likely to cause error predictions with high reliability. Gradients alone fail to eliminate these high-confidence, low-contribution elements. The results demonstrate that the best performance is achieved by combining both gradient and GSNR masks. When combined, these components work synergistically within VL-Rotate, leading to a collective improvement of 12.12/6.43% ID/OoD mAP.

2) Various Language Models: Table V shows the impact of different language models on VL-Rotate. Using W2V [73] leads to a 6.07% mAP drop while BERT [50] causes 3.75% mAP gains in unseen data. In contrast, VL-Rotate using CLIP's text encoder can more effectively leverage the rich prior knowledge, outperforming W2V and BERT by 12.5% and 2.68% OoD mAP.

3) Varient CLIP Text Encoder: Table V reports the performance of using different CLIP variants as text encoders. Compared to EVA02-CLIP [74], which explores CLIP through feature distillation, LongCLIP [75], which enhances short text capabilities and supports long text input, and SigLIP [76], which reduces the number of tokens and uses Sigmoid loss for training, the original CLIP achieves the best performance on VL-Rotate. For fair comparison, all models were experimented with the same setting and the base scale weights. The result indicates that the performance of the text encoder does not improve with iterative CLIP architecture updates but is highly dependent on task-specific characteristics. The optimization direction of Variant CLIP may compromise the text encoding capability required for OoD ROD, whereas the original CLIP, with its straightforward contrastive learning objective and broader data coverage, is more suitable for this task.

4) Pre-Trained Backbones: Table VI reports the backbone performance of ResNet-50 pre-trained on ImageNet versus CLIP. The ImageNet pretrained backbone outperforms the CLIP-pretrained backbone by 13.21/8.46% ID/OoD mAP. This outcome is expected since, despite CLIP's ability to learn from large-scale image-text pairs, the subsequent FPN and task-specific heads in the detector disrupt these learned relationships. Also, the backbone decouples classification and regression features, requiring realignment with CLIP's text features during fine-tuning. Thus, our method effectively reduces the impact of CLIP's prior visual knowledge.

5) Mask Dropout Schedule: We present different mask dropout schedules in Table VII. The schedules for activating MFHD are selected as full-time epochs, pre-1/2 and post-1/2 of the maximum epochs, pre-1/3, mid-1/3, post-1/3 of the maximum epochs, and random activation with 50% probability. The experimental results support our hypothesis, showing that activating MFHD later in training is better than doing so earlier, as early activation can hinder the learning of object representations. In the post-training stage, MFHD promotes detectors focusing on learning generalization capabilities. Additionally, full-time activation shows similar results to pre-1/2 activation, suggesting that it also disrupts learning. The poorest performance occurs with a 50% random activation, which introduces instability and disrupts the model's ability to learn consistent feature representations during critical learning phases.

6) Mask Dropout Elements: Fig. 5(a) shows the performance on the target data when muting the top-p largest elements of the classification features. The results indicate that the selection of p should not be too large or too small. A suitable p enables the model to generalize better on unseen target domains.

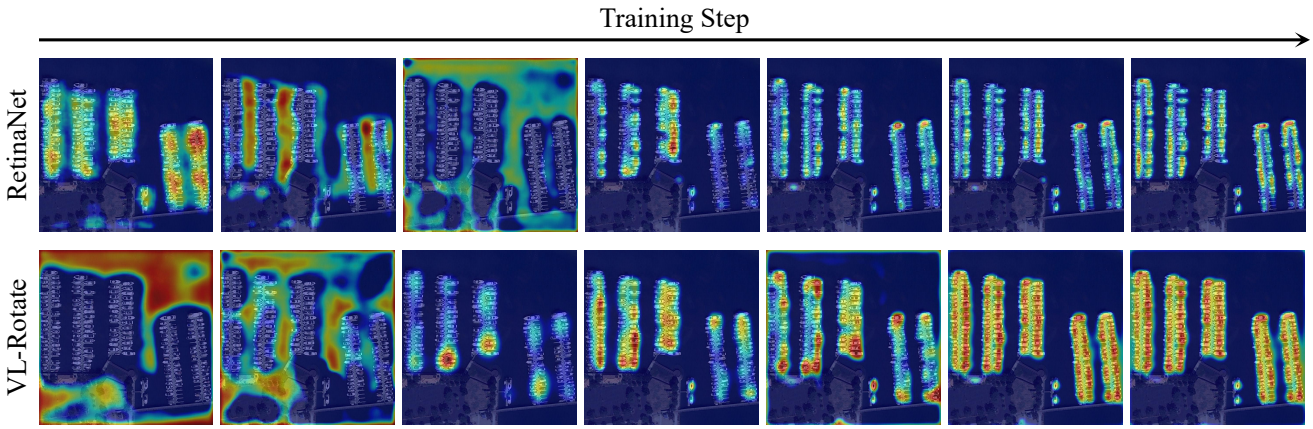

Fig. 6. Visualization of VL-Rotate and RetinaNet during fine-tuning.

7) Number of Shot: Fig. 5(b) shows the performance of using different shot numbers for training in VL-Rotate and RetinaNet on DG task. Our method consistently outperforms the baseline, demonstrating VL-Rotate's robustness and stability.

8) Feature Space Visualization: Fig. 4 shows the visualization results of VL-Rotate and baseline using GradCam [77]. Compared with the baseline, VL-Rotate demonstrates more focused attention on the key object regions (ships, planes, and cars) and enhanced robustness against background interference. The result validates that the joint optimization of CFT and MFHD enables the model to more accurately capture small and densely distributed objects in remote sensing images, highlighting the robustness of VL-Rotate.

9) Visualization of Training Process: We generate a set of visualizations of classification feature gradients during training using GradCam [77], as illustrated in Fig. 6, which demonstrates how a language model modulates the detector. As shown in the figure, RetinaNet struggles to focus on all objects, particularly in scenes with densely packed objects. The results in sparse and uneven attention patterns indicate that RetinaNet is unable to effectively learn the key representations of the objects. In contrast, VL-Rotate, by leveraging text prior knowledge of CLIP extracted from "a photo of a ship/harbor" to modulate object representations and forcing object features to align with text features in the high-dimension embedding space, guides the model to predict, demonstrating the robust learning capability of our approach. Additionally, it is noteworthy that in the early stages of training, VL-Rotate focuses primarily on salient local regions (as seen in the third image from left), which aligns with our description that leaders in the features have limitations in focusing on objects in MFHD. As training progresses, MFHD is activated to encourage all features to participate in predictions, significantly enhancing the performance of the detector.

10) Visualization of Features: Fig. 7 illustrates the feature correlation between classification features and region features in the baseline and VL-Rotate. Given that remote sensing images often contain dense small objects, we show the first stage of multi-stage features with a smaller receptive field

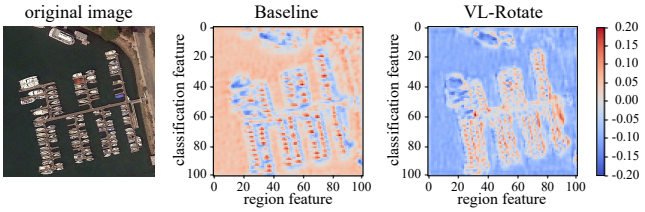

Fig. 7. Visualization of the feature correlations between classification features and region features in baseline and VL-Rotate. Compared with the baseline, our method focuses solely on the categories, establishing correlations between classification and regression on foreground objects.

and compute their correlation using the Pearson correlation coefficient. As shown in the figure, in the baseline, the classification features and region features exhibit high correlation even in the background areas, while showing low correlation for the roads at the harbor that need to be detected. In contrast, text features modulate the classification and region features in VL-Rotate, compelling these features to establish correlations on the categories of interest and ignore background features, thereby demonstrating the effectiveness of our model in remote sensing object detection.

V. Conclusion and Future Work

In this study, we tackled the complex challenge of few-shot out-of-distribution (OoD) generalized rotated object detection by introducing VL-Rotate, a versatile vision-language framework. VL-Rotate comprises two key modules: CLIP-guided Fine-Tuning (CFT) and Masked Feature Heuristics Dropout (MFHD), each contributing to robust performance under domain shifts. CFT enhances generalization by integrating text features into high-dimensional object representations, thereby improving the model's ability to adapt to distribution shifts and making better use of instance-level annotations for fine-grained learning. MFHD selectively deactivates classification features based on feature gradients and GSNR, promoting more stable predictions on unseen data. Extensive experiments on domain adaptation and generalization tasks confirm VL-Rotate's state-of-the-art performance in few-shot OoD scenarios, advancing

the field of rotated object detection by addressing its most challenging variants. We currently focus on the few-shot setting following CoOp and Out-of-Distribution setting. In the future, we will investigate VL-Rotate's performance in open-vocabulary rotated object detection, further exploring novel classes and zero-shot learning.

REFERENCES

[1] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, "Rtmdet: An empirical study of designing real-time object detectors," 2022.

[2] Y. Yu and F. Da, "Phase-shifting coder: Predicting accurate orientation in oriented object detection," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [Online]. Available: https://arxiv.org/abs/2211.06368

[3] X. Yang, G. Zhang, W. Li, X. Wang, Y. Zhou, and J. Yan, "H2rbox: Horizontal box annotation is all you need for oriented object detection," 2023.

[4] X. Yang, Y. Zhou, G. Zhang, J. Yang, W. Wang, J. Yan, X. Zhang, and Q. Tian, "The kfiou loss for rotated object detection," 2023.

[5] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2849–2858.

[6] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: A rotation-equivariant detector for aerial object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2786–2795.

[7] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 3520–3529.

[8] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, and X. Bai, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1452–1459, 2021.

[9] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[10] L. Hou, K. Lu, J. Xue, and Y. Li, "Shape-adaptive selection and measurement for oriented object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[11] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[12] X. Yang, J. Yan, Z. Feng, and T. He, "R3det: Refined single-stage detector with feature refinement for rotating object," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3163–3171.

[13] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[14] Z. Guo, C. Liu, X. Zhang, J. Jiao, X. Ji, and Q. Ye, "Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[15] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," *arXiv preprint arXiv:1904.01355*, 2019.

[16] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[18] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[20] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[21] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[22] X. Zhang, Z. Xu, R. Xu, J. Liu, P. Cui, W. Wan, C. Sun, and C. Li, "Towards domain generalization in object detection," *arXiv preprint arXiv:2203.14387*, 2022.

[23] A. Dayal, V. KB, L. R. Cenkeramaddi, C. Mohan, A. Kumar, and V. N Balasubramanian, "Madg: margin-based adversarial learning for domain generalization," *Advances in Neural Information Processing Systems*, vol. 36, pp. 58 938–58 952, 2023.

[24] N. Ye, J. Tang, H. Deng, X.-Y. Zhou, Q. Li, Z. Li, G.-Z. Yang, and Z. Zhu, "Adversarial invariant learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 12 441–12 449.

[25] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 124–140.

[26] H. Bai, R. Sun, L. Hong, F. Zhou, N. Ye, H.-J. Ye, S.-H. G. Chan, and Z. Li, "Decaug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6705–6713.

[27] M. Zhang, Z. Zhuang, Z. Wang, D. Wang, and W. Li, "Rotogbml: Towards out-of-distribution generalization for gradient-based meta-learning," *arXiv preprint arXiv:2303.06679*, 2023.

[28] S. Liu, T. Chen, Z. Atashgahi, X. Chen, G. Sokar, E. Mocanu, M. Pechenizkiy, Z. Wang, and D. C. Mocanu, "Deep ensembling with no overhead for either training or testing: The all-round blessings of dynamic sparsity," *arXiv preprint arXiv:2106.14568*, 2021.

[29] G. Blanchard, A. A. Deshmukh, Ü. Dogan, G. Lee, and C. Scott, "Domain generalization by marginal transfer learning," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 46–100, 2021.

[30] F. Wenzel, A. Dittadi, P. Gehler, C.-J. Simon-Gabriel, M. Horn, D. Zietlow, D. Kernert, C. Russell, T. Brox, B. Schiele *et al.*, "Assaying out-of-distribution generalization in transfer learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7181–7198, 2022.

[31] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[32] K. Ahuja, K. Shanmugam, K. Varshney, and A. Dhurandhar, "Invariant risk minimization games," in *International Conference on Machine Learning*. PMLR, 2020, pp. 145–155.

[33] Y. Chen, K. Zhou, Y. Bian, B. Xie, K. Ma, Y. Zhang, H. Yang, B. Han, and J. Cheng, "Pareto invariant risk minimization," *arXiv preprint arXiv:2206.07766*, 2022.

[34] X. Zhou, Y. Lin, W. Zhang, and T. Zhang, "Sparse invariant risk minimization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 27 222–27 244.

[35] C. Wang, J. Jiang, X. Zhou, and X. Liu, "Resmooth: Detecting and utilizing ood samples when training with data augmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[36] A. Kristiadi, M. Hein, and P. Hennig, "Being a bit frequentist improves bayesian neural networks," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 529–545.

[37] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[38] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[39] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[41] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[42] ——, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 816–16 825.

[43] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *arXiv preprint arXiv:2110.04544*, 2021.

[44] L. Yao, J. Han, Y. Wen, X. Liang, D. Xu, W. Zhang, Z. Li, C. XU, and H. Xu, "Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 9125–9138. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/3ba960559212691be13fa81d9e5e0047-Paper-Conference.pdf

[45] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 18 082–18 091.

[46] S. Li, L. Sun, and Q. Li, "Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1405–1413.

[47] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[48] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li *et al.*, "Regionclip: Region-based language-image pretraining," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 793–16 803.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[51] Q. Bi, B. Zhou, J. Yi, W. Ji, H. Zhan, and G.-S. Xia, "Good: Towards domain generalized oriented object detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 223, pp. 207–220, 2025.

[52] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[53] N. Liu, X. Xu, Y. Su, C. Liu, P. Gong, and H.-C. Li, "Clip-guided source-free object detection in aerial images," 2024. [Online]. Available: https://arxiv.org/abs/2401.05168

[54] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[55] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[56] Y. Fu, Y. Wang, Y. Pan, L. Huai, X. Qiu, Z. Shangguan, T. Liu, Y. Fu, L. Van Gool, and X. Jiang, "Cross-domain few-shot object detection via enhanced open-set object detector," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 247–264.

[57] W. Xiong, "Cd-fsod: A benchmark for cross-domain few-shot object detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[58] V. VS, P. Oza, and V. M. Patel, "Instance relation graph guided source-free domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 3520–3530.

[59] W. Lee, D. Hong, H. Lim, and H. Myung, "Object-aware domain generalization for object detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, pp. 2947–2955, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28076

[60] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[61] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[62] M. Michalkiewicz, M. Faraki, X. Yu, M. Chandraker, and M. Baktashmotlagh, "Domain generalization guided by gradient signal to noise ratio of parameters," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 6177–6188.

[63] J. Liu, G. Jiang, Y. Bai, T. Chen, and H. Wang, "Understanding why neural networks generalize well through gsnr of parameters," 2020. [Online]. Available: https://arxiv.org/abs/2001.07384

[64] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS journal of photogrammetry and remote sensing*, vol. 159, pp. 296–307, 2020.

[65] N. Liu, X. Xu, Y. Su, C. Liu, P. Gong, and H.-C. Li, "Clip-guided source-free object detection in aerial images," 2024. [Online]. Available: https://arxiv.org/abs/2401.05168

[66] H. He, J. Ding, and G.-S. Xia, "On the robustness of object detection models in aerial images," 2023.

[67] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019. [Online]. Available: https://arxiv.org/abs/1903.12261

[68] C. Lin, Z. Yuan, S. Zhao, P. Sun, C. Wang, and J. Cai, "Domain-invariant disentangled network for generalizable object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 8771–8780.

[69] V. Vidit, M. Engilberge, and M. Salzmann, "Clip the gap: A single domain generalization approach for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 3219–3229.

[70] A. Wu and C. Deng, "Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 847–856.

[71] Y. Zhou, X. Yang, G. Zhang, J. Wang, Y. Liu, L. Hou, X. Jiang, X. Liu, J. Yan, C. Lyu, W. Zhang, and K. Chen, "Mmrotate: A rotated object detection benchmark using pytorch," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, p. 7331–7334.

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[73] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[74] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," 2023. [Online]. Available: https://arxiv.org/abs/2303.15389

[75] B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang, "Long-clip: Unlocking the long-text capability of clip," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 310–325.

[76] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 11 975–11 986.

[77] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Weihan Yin received the MS degree in electronic information from Shanghai Jiao Tong University, Shanghai and the BS degree in information engineering from South China University of Technology, Guangzhou. His research interests include out-of-distribution generalization, few-shot learning and rotated object detection.

Qinying Gu received the master's and PhD degrees from the Cavendish Laboratory, University of Cambridge, in 2022. Currently, she is a young research scientist with the Shanghai Artificial Intelligence Laboratory. Her research is primarily focused on braininspired computing, brain-inspired optoelectronics, bayesian optimization, and its applications across interdisciplinary fields. She has published papers in prestigious journals, including nature materials, nature communications, and communications engineering.

Yaoyun Zhang received the BS degree in Computer Science from Cheng Du University of Technology. He is currently working toward the MS degree with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. His research interests include multimodal learning and artificial intelligence generative content.

Lin Zhu received the BS degree in statistics from Central South University, China, in 2019 and the MS degree in statistics from Shandong University, China, in 2022. She is currently working toward the PhD degree in computer science and rechnology with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China. Her research interests include out-of-distribution generalization, few-shot learning and invariant representation learning.

Qiao Sun received his BS degree in Civil Engineering from Tianjin University, China. He is currently pursuing an MS degree at the Academy for Engineering and Technology, Fudan University (FAET), China. His research interests include natural language processing, robotic learning, and spatial intelligence.

Liujia Yang received his BS degree in Artificial Intelligence from Shanghai Jiao Tong University. He is currently pursuing his MS degree in the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. His research interests include machine learning, neuromorphic computing, and embodied intelligence.

Xinbing Wang (Senior Member, IEEE) received the BS degree (with Hons.) from the Department of Automation, Shanghai Jiaotong University, Shanghai, China, in 1998, the MS degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2001, and the PhD degree, major from the Department of Electrical and Computer Engineering, minor in the Department of Mathematics, North Carolina State University, Raleigh, in 2006. Currently, he is a professor with the Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai, China. He has been an associate editor for IEEE/ACM Transactions on Networking and IEEE Transactions on Mobile Computing, and a member of the Technical Program Committees of several conferences including ACM MobiCom 2012, 2018-2019, ACM MobiHoc 2012-2019, and IEEE INFOCOM 2009-2020.

Nanyang Ye received the PhD degree from the University of Cambridge. He is now an associate professor with Shanghai Jiao Tong University. His current research interests include but are not limited to OoD generalization, Bayesian deep learning, causal inference, etc. He serves as a programme committee member and reviewer for several key machine-learning journals and conferences. He has published several papers on top machine learning and artificial intelligence journals and conferences, such as T-PAMI, IJCV, NeurIPS, CVPR, ICCV, AAAI, IJCAI, etc.