CVPR #14079

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

Anonymous CVPR submission

Paper ID 14079

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.

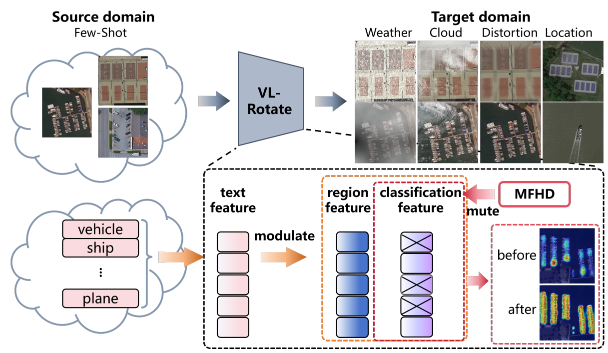

Figure 1. An overview of our work. VL-Rotate aims to learn from a k-shot source domain and generalize to the target domain with unseen data. Our approach integrates text-based prior knowledge to modulate object features and mutes classification features with Masked Feature Heuristics Dropout (MFHD) to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

1. Introduction

Rotated Object Detection (ROD) is a rapidly advancing area in computer vision, with recent innovations [32, 56, 57, 61] 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. 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 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:

- 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.
- 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.
- 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.

2. Related Work

In this section, we will review related works. The complete related work section can be found in the Appendix due to space limitations.

2.1. Rotated Object Detection

Rotated object detection is a challenging task involving dense object prediction and rotated bounding box prediction. Novel methods have been proposed to address this problem, falling into three main categories: two-stage detector [7, 15, 52], refine-stage detector [16, 20, 48, 55, 58, 72] and single-stage detector [32, 56, 61, 65]. In the context of refine-stage detectors, Oriented RepPoints [48] introduced an adaptive points representation to capture the geometric information of objects and proposed a corresponding quality assessment for adaptive points learning. Recently, there has been a growing trend of exploring single-stage detectors. Noteworthy contributions in this area include PSC [61] provides a unified framework to resolve various periodic fuzzy problems and RTMDet [32], offering an efficient real-time detection solution with large-kernel depthwise convolutions.

2.2. 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 [66]:
(1) **Domain generalization-based method** These methods train models on source domains to achieve generalization on unseen target domains. Common approaches include domain adversarial learning [9, 60], transfer learning [3, 49], and meta-learning [64].
(2) **Invariant representation learning** Exemplified by Invariant Risk Minimization (IRM) [2], this approach explores causal relationships in data across different environments based on causal invariant features. Recently, Pareto Invariant Risk Minimization [4] and parse Invariant Risk Minimization [70] have been proposed to further investigate the generalization ability of IRM .
(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 [47] and Bayesian methods [22], etc.

2.3. Vision-Language Pre-trained Models

Recent advancements in large-scale vision-language pretraining have notably enhanced downstream task performance. Contrastive Language-Image Pretraining (CLIP) [35] stands out by effectively learning vision-language representations. CLIP's framework has inspired developments in vison-language learning, with models such as CoOp [69], CoCoOp [68], and CLIP-Adapter [10]. CLIP has also been adapted for various tasks, including DetCLIP [59] for object detection, DenseCLIP [36] for pixel-text matching, and CLIP-ReID [25] for image re-identification, demonstrating its versatility in fine-tuning applications.

3. 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 ori-

CVPR
#14079

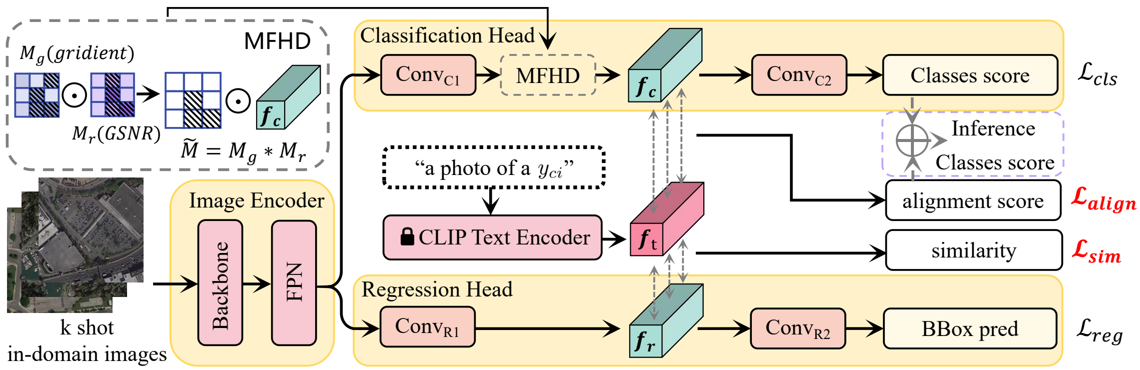

Figure 2. The overall framework of the proposed VL-Rotate. RetinaNet is shown as the baseline, encompassing an Image Encoder and task-specific heads. VL-Rotate includes Masked Feature Heuristics Dropout (MFHD) and CLIP-Guided Fine-Tuning (CFT). During training, MFHD utilizes gradient and GSNR to mute 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.

entations. 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 [28] 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) [27], and task-specific heads for classification and regression.

3.1. 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 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.

3.1.1. 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.

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

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

where $\lambda = 0.5$ balances the two components, merging the model's intrinsic classification ability with text-based prior knowledge for more stable predictions.

3.1.2. 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:

$$\Omega(f_r, f_{t_i})_i = \frac{1}{N} \sum_{j=1}^{N} f_{r_j} f_{t_i}^T \tag{2}$$

The total text-region similarity $\Omega(f_r, f_t)$ is calculated by summing these individual similarities in Eq. (2):

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

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 $\bar{\Omega}(f_r, f_t)$:

$$\bar{\Omega}(f_r, f_t) = \frac{1}{M} \sum_{i=1}^{M} \hat{f}_{r_j} f_{t_i}^T \tag{4}$$

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:

$$\text{softmax}(f_{r_j}, f_{t_i}^T) = \frac{\exp(f_{r_j} f_{t_i}^T / \gamma)}{\sum_r \exp(f_r f_{t_i}^T / \gamma)} \tag{5}$$

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

The softmax probability is then incorporated into Eq. (3) to derive the final matching text-region similarity:

$$\bar{\Omega}(f_r, f_t) = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N} \text{softmax}(f_{r_j}, f_{t_i}^T) f_{r_j} f_{t_i}^T \tag{6}$$

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 $\mathcal{L}_{sim}$:

$$\mathcal{L}_{sim} = -\frac{1}{B} \log \frac{\exp(\bar{\Omega}(f_r, f_{t_i}) / \gamma)}{\sum_r \exp(\bar{\Omega}(f_r, f_{t_i}) / \gamma)} \tag{7}$$

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.

3.1.3. Overall Training Loss

Following RetinaNet, the total loss is calculated as:

$$\mathcal{L} = \omega_1 \mathcal{L}_{cls} + \omega_2 \mathcal{L}_{reg} + \omega_3 \mathcal{L}_{align} + \omega_4 \mathcal{L}_{sim} \tag{8}$$

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 [28] for $\mathcal{L}_{cls}$, $\mathcal{L}_{align}$, and $\mathcal{L}_{sim}$, and GIoU loss [38] for $\mathcal{L}_{reg}$. The weights ω_1, ω_2, ω_3 and ω_4 are empirically set to 1:1:2:2.

3.2. 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 [40], 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 [21, 33], 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) and high Gradient Signal-to-Noise Ratio (GSNR) [29] 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.

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:

$$\tilde{M} = M_g \odot M_r \tag{9}$$

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

$$r_c = \frac{\mathrm{E}^2_{(x,y_c)\sim\mathbb{D}}(g_c)}{\mathrm{Var}_{(x,y_c)\sim\mathbb{D}}(g_c)} \tag{10}$$

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.

4. Experiment

4.1. Experiment Settings

Adhering to the few-shot settings of CoOp [67] 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.

4.1.1. Domain Adaptation

We focus on evaluating performance under domain shifts. While datasets like DOTA-C [17] and DOTA-Cloudy [17] 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 [31] and DIOR-Cloudy [31]. DIOR-C includes 19 different types of corruptions from ImageNet-C [19] 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 [24] 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.

4.1.2. Domain Generalization

We use the original DOTA [51] 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 [26, 45, 50], 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.

4.1.3. Competitors

To conduct comprehensive experiments and provide valuable insights, we explored various methods in few-shot

	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 [8]	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 [53]	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 [46]	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 [30]	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 [23]	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 [11]	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 [52]	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 [7]	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 [15]	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 [28]	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 [56]	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 [32]	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 [61]	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 [44]	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 [65]	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 [16]	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 [55]	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 [58]	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 [20]	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 [72]	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 [48]	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

Table 1. 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 (**Ga**ussian, **Sh**ot, **Im**pulse, **Sp**eckle), Blur (**De**focus, **Gl**ass, **Mo**tion, **Zo**om, **Ga**ussian), Weather (**Sn**ow, **Fr**ost, **Fo**g, **Br**ightness, **Sp**atter), and Digital (**Co**ntrast, **El**astic transform, **Pi**xelate, **JP**EG compression, **Sa**turate). 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.

OoD RoD scenarios.

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 [53] and CD-ViTO [8], 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 [30], IRG-SFDA [46], and OA-DG [23] were utilized to evaluate their few-shot performance under ROD.

4.1.4. Experiment Details

Our experiments were conducted on MMRotate [71]. For fair evaluation, all methods in ROD use ResNet-50 [18] pre-trained on ImageNet as the backbone and follow the default setup on MMRotate. CD-FSOD and DA & DG methods are followed their default settings. VL-Rotate is trained with 3x schedule, 0.005 learning rate, 0.9 momentum, and 0.0001 weight decay. Random flipping is employed to avoid over-fitting without any additional tricks. **Further details are provided in Appendix.**

4.2. Main Results

4.2.1. Domain Adaptation

We deploy VL-Rotate in some of the ROD methods and compare with all competitors on DA tasks, as shown in Tab. 1. 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 Rep-Points, 1.79% with SASM and 1.99% with Oriented Rep-Points. 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.

4.2.2. Domain Generalization

Tab. 2 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

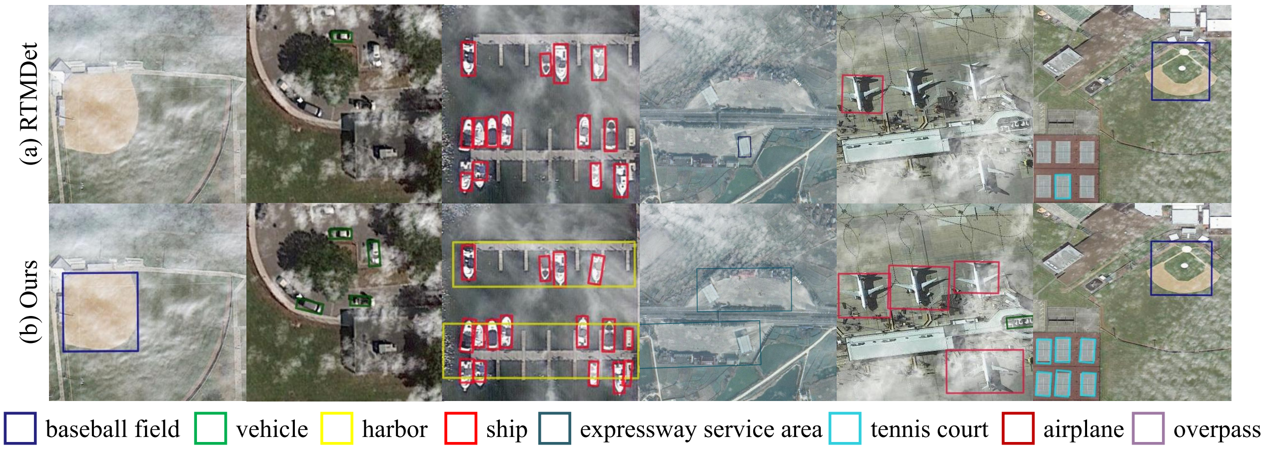

Figure 3. Qualitative comparisons of the inference results between proposed VL-Rotate and the baseline model RTMDet on DIOR-Cloudy.

	Method	ID mAP	OoD mAP
CF	CD-ViTO[8]	13.64	13.00
	Distill-FSOD[53]	31.96	28.29
DADG	SFOD[30]	-	-
	IRG-SFDA[46]	23.20	19.43
	OA-DG[23]	46.40	32.31
Typical ROD	*two-stage*:		
	Faster RCNN OBB [11]	48.21	44.64
	Oriented RCNN [52]	51.69	45.36
	RoI Transformer [7]	54.08	47.83
	ReDet [15]	54.23	48.39
	refine-stage:		
	Reppoints OBB [58]	50.08	46.05
	R^3Det [55]	50.80	45.19
	S^2A-Net [16]	51.07	47.12
	CFA [72]	51.82	47.01
	SASM [20]	53.89	50.02
	Oriented Reppoints [48]	54.96	49.00
	single-stage:		
	H2RBox [56]	36.99	37.34
	RetinaNet OBB [28]	44.10	42.19
	FCOS OBB-PSC [61]	50.11	44.40
	FCOS OBB [44]	50.84	47.33
	Rotated ATSS [65]	51.70	46.39
	RTMDet-l [32]	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

Table 2. 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.

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.

4.3. Ablation Study

We conduct a series of ablation experiments to evaluate the effectiveness of VL-Rotate and exclude potential confounding factors. Unless otherwise specified, the experimental settings align with those described in the experiment details.

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 3. 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.

Method	Language	ID mAP	Impv	OoD mAP	Impv
baseline	-	23.22		14.26	
VL-Rotate	W2V[34]	12.37	-10.85	8.19	-6.07
VL-Rotate	BERT[6]	30.20	+6.98	18.01	+3.75
VL-Rotate	CLIP-Text[35]	**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[42]	28.93	+5.71	17.95	+3.69
VL-Rotate	Long-CLIP[63]	33.61	+10.39	19.49	+5.23
VL-Rotate	SigLIP[62]	34.14	+10.92	20.55	+6.29
VL-Rotate	CLIP[35]	**35.34**	+12.12	**20.69**	+6.43

Table 4. Top: Ablation study results for VL-Rotate using different language models. Bottom: Ablation study results for VL-Rotate using variant CLIP text encoder.

Method	Params	GFLOPs	FPS	OoD mAP
RetinaNet OBB[28]	36.52 M	133.35	699.2	14.26
w/ VL-Rotate	41.62 M	201.36	681.6	20.69
RTMDet-l[32]	52.27 M	124.66	692.8	22.79
w/ VL-Rotate	55.88 M	171.36	676.8	26.12

Table 5. Ablation study results for VL-Rotate inference information on DA task.

Similarly, unless specified, VL-Rotate was implemented in RetinaNet with RetinaNet serving as the baseline.

CVPR #14079

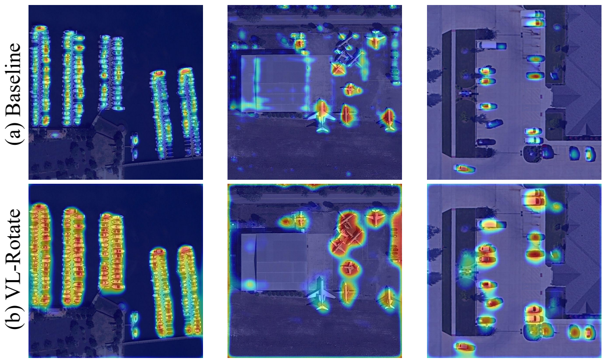

Figure 4. Visualization of VL-Rotate and the baseline.

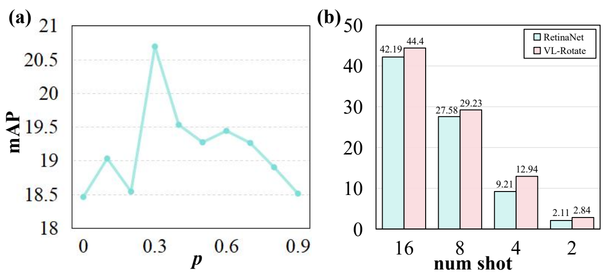

Figure 5. Ablation study results of target domain for VL-Rotate about (a) different p on DA task; (b) different shot on DG task.

4.3.1. 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. When combined, these components work synergistically within VL-Rotate, leading to a collective improvement of 12.12/6.43% ID/OoD mAP. Additionally, the MFHD module is evaluated separately using high-gradient masked and high-GSNR masked conditions. The results demonstrate that the best performance is achieved by combining both gradient and GSNR masks.

4.3.2. Various Language Models

Tab. 4 shows the impact of different language models on VL-Rotate. Using W2V [34] leads to a 6.07% mAP drop while BERT [6] 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.

4.3.3. Varient CLIP Text Encoder

Tab. 4 reports the performance of using different CLIP variants as text encoders. Compared to EVA02-CLIP [42], which explores CLIP through feature distillation, Long-CLIP [63], which enhances short text capabilities and supports long text input, and SigLIP [62], 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.

4.3.4. 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.

4.3.5. 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.

4.3.6. Feature Space Visualization

Fig. 4 shows the visualization results of VL-Rotate and baseline using GradCam [39]. Compared to the baseline, VL-Rotate focuses more object regions.

4.3.7. Inference Efficiency

Tab. 5 presents the inference performance and efficiency of our method on the DA task. Compared to the baseline, our method improves mAP by 45.09% and 14.61%, with only a slight reduction in FPS by 2.52% and 2.31%, respectively.

5. 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] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020. 1

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2, 1

[3] Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1):46–100, 2021. 2, 1

[4] Yongqiang Chen, Kaiwen Zhou, Yatao Bian, Binghui Xie, Kaili Ma, Yonggang Zhang, Han Yang, Bo Han, and James Cheng. Pareto invariant risk minimization. *arXiv preprint arXiv:2206.07766*, 2022. 2, 1

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 1

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 7, 8

[7] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2849–2858, 2019. 2, 6, 7, 1, 3

[8] Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qiu, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. In *Computer Vision – ECCV 2024*, pages 247–264, Cham, 2025. Springer Nature Switzerland. 6, 7

[9] Irena Gao, Ryan Han, and David Yue. Exploring adversarial training for out-of-distribution detection. 2023. 2, 1

[10] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 2

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 6, 7, 1, 2, 3

[12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1

[14] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 1

[15] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2786–2795, 2021. 2, 6, 7, 1, 3

[16] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 2, 6, 7, 3

[17] Haodong He, Jian Ding, and Gui-Song Xia. On the robustness of object detection models in aerial images, 2023. 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6, 2

[19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. 5

[20] Liping Hou, Ke Lu, Jian Xue, and Yuqiu Li. Shape-adaptive selection and measurement for oriented object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 2, 6, 7, 1, 3

[21] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer Vision – ECCV 2020*, pages 124–140, Cham, 2020. Springer International Publishing. 5

[22] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being a bit frequentist improves bayesian neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 529–545. PMLR, 2022. 2, 1

[23] Wooju Lee, Dasol Hong, Hyungtae Lim, and Hyun Myung. Object-aware domain generalization for object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):2947–2955, 2024. 6, 7

[24] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 5, 2

[25] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1405–1413, 2023. 2

[26] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8771–8780, 2021. 5

[27] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 1

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 5, 6, 7, 2

[29] Jinlong Liu, Guoqing Jiang, Yunzhi Bai, Ting Chen, and Huayan Wang. Understanding why neural networks generalize well through gsnr of parameters, 2020. 5

[30] Nanqing Liu, Xun Xu, Yongyi Su, Chengxin Liu, Peiliang Gong, and Heng-Chao Li. Clip-guided source-free object detection in aerial images, 2024. 6, 7

[31] Nanqing Liu, Xun Xu, Yongyi Su, Chengxin Liu, Peiliang Gong, and Heng-Chao Li. Clip-guided source-free object detection in aerial images, 2024. 5, 2

[32] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors, 2022. 1, 2, 6, 7, 3

[33] Mateusz Michalkiewicz, Masoud Faraki, Xiang Yu, Manmohan Chandraker, and Mahsa Baktashmotlagh. Domain generalization guided by gradient signal to noise ratio of parameters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6177–6188, 2023. 5

[34] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 7, 8

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 7, 1

[36] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18082–18091, 2022. 2

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[38] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[39] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 8, 3

[40] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. 5

[41] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1

[42] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. 7, 8

[43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 1

[44] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *arXiv preprint arXiv:1904.01355*, 2019. 6, 7, 1, 2, 3

[45] Vidit Vidit, Martin Engilberge, and Mathieu 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)*, pages 3219–3229, 2023. 5

[46] Vibashan VS, Poojan Oza, and Vishal 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)*, pages 3520–3530, 2023. 6, 7

[47] Chenyang Wang, Junjun Jiang, Xiong Zhou, and Xianming Liu. Resmooth: Detecting and utilizing ood samples when training with data augmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2, 1

[48] Kaixuan Hu Jianke Zhu Wentong Li, Yijie Chen. Oriented reppoints for aerial object detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6, 7, 1, 3

[49] Florian Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. *Advances in Neural Information Processing Systems*, 35:7181–7198, 2022. 2, 1

[50] Aming Wu and Cheng 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)*, pages 847–856, 2022. 5

[51] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5, 2

[52] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3520–3529, 2021. 2, 6, 7, 1, 3

[53] Wuti 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)*, pages 1–5, 2023. 6, 7

[54] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1452–1459, 2021. 1

[55] Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3163–3171, 2021. 2, 6, 7, 1, 3

[56] Xue Yang, Gefan Zhang, Wentong Li, Xuehui Wang, Yue Zhou, and Junchi Yan. H2rbox: Horizontal box annotation is all you need for oriented object detection. 2023. 1, 2, 6, 7, 3

[57] Xue Yang, Yue Zhou, Gefan Zhang, Jirui Yang, Wentao Wang, Junchi Yan, Xiaopeng Zhang, and Qi Tian. The kfiou loss for rotated object detection, 2023. 1

[58] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 6, 7, 1, 3

[59] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing XU, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *Advances in Neural Information Processing Systems*, pages 9125–9138. Curran Associates, Inc., 2022. 2

[60] Nanyang Ye, Jingxuan Tang, Huayu Deng, Xiao-Yun Zhou, Qianxiao Li, Zhenguo Li, Guang-Zhong Yang, and Zhanxing Zhu. Adversarial invariant learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12441–12449. IEEE, 2021. 2, 1

[61] Yi Yu and Feipeng 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. 1, 2, 6, 7, 3

[62] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023. 7, 8

[63] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *Computer Vision – ECCV 2024*, pages 310–325, Cham, 2025. Springer Nature Switzerland. 7, 8

[64] Min Zhang, Zifeng Zhuang, Zhitao Wang, Donglin Wang, and Wenbin Li. Rotogbml: Towards out-of-distribution generalization for gradient-based meta-learning. *arXiv preprint arXiv:2303.06679*, 2023. 2, 1

[65] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan 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)*, 2020. 2, 6, 7, 3

[66] Xingxuan Zhang, Zekai Xu, Renzhe Xu, Jiashuo Liu, Peng Cui, Weitao Wan, Chong Sun, and Chen Li. Towards domain generalization in object detection. *arXiv preprint arXiv:2203.14387*, 2022. 2, 1

[67] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 5

[68] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2

[69] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[70] Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. Sparse invariant risk minimization. In *International Conference on Machine Learning*, pages 27222–27244. PMLR, 2022. 2, 1

[71] Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, Wenwei Zhang, and Kai Chen. Mmrotate: A rotated object detection benchmark using pytorch. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 7331–7334, 2022. 6, 2

[72] Xiaosong Zhang Jianbin Jiao Xiangyang Ji Zonghao Guo, Chang Liu and Qixiang 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)*, 2021. 2, 6, 7, 1, 3