VL-Rotate: Vision-Language Learning for Few-Shot OoD Rotated Object Detection

Anonymous Authors[1]

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-language model leveraging pretrained knowledge from CLIP's text encoder for aligning image and text features in the embedding space. We incorporate two innovative modules: Oriented Object Text Alignment Module (OOTA) and Fine-Grained Region-Text Contrastive Module (FRTC), to yield alignment and similarity scores from image-text pairs, thereby guiding the model's fine-tuning throughout the training phase. Aimed at elevating detection accuracy and bolstering few-shot OoD inference capabilities, we evaluate our method on diverse challenging few-shot OoD datasets (HRSCOoD, DOTAOoD, RSDDOoD, and DIORROoD), constructed from widely utilized public datasets. Compared to prior works, VL-Rotate achieves state-of-the-art results across all datasets, reaching an impressive up to 75.2% mAP on HRSCOoD's out-of-distribution subset, demonstrating the benefits of natural language guidance and text-image alignment. Our results validate the model's effectiveness and potential in advancing ROD.

1. Introduction

Rotated object detection (ROD) has emerged as a prominent research area in computer vision, with recent advancements in novel methods (Lyu et al., 2022; Yu & Da, 2023; Yang et al., 2023a;b) that offer significant progress in this area. This technique is crucial for detecting objects in aerial or remote-sensing imagery and text detection (Han et al., 2021b; Ding et al., 2019; Zonghao Guo & Ye, 2021). where objects often appear dense, elongated, and in various orientations. Oriented bounding boxes (OBBs) have thus become the preferred localization method over traditional horizontal bounding boxes, with many specialized detectors delivering promising results on challenging datasets.

Current research in ROD focuses on refining network architectures, feature extraction methods, and loss functions to improve accuracy. We contend that cross-modal learning, specifically the fusion of vision and language, holds promise for advancing ROD. Although large-scale image-text pairs have been used to pre-train models for robust feature representations, ROD's unique challenges such as complex backgrounds, dense object placements, and rotated frame predictions have limited the effective use of text information for detection. To date, no proposed method has fully harnessed the potential of language to enhance ROD performance. Moreover, rotated object detection, particularly in aerial images, faces distribution shift challenges. For example, in ship detection, captured ships from remote-sensing images are mainly docked in harbors, creating a complex and densely arranged scenario. Training models with ships sailing on the open sea is difficult due to the vastness of the sea, leading to an imbalanced dataset biased towards harbors. Distinct wake trails left by ships traveling on the sea further interfere with accurate predictions. This out-of-distribution (OoD) generalization dilemma significantly degrades the model's detection performance.

To address these limitations, here we propose a pioneering approach, Vision-Language Alignment Learning for Few-Shot OoD Rotated Object Detection (VL-Rotate), leveraging language representations to enhance predictions of rotated objects in OoD scenarios within a few-shot learning framework. Our motivation for considering few-shot learning arises from the limited large-scale datasets in this field, with key datasets like DOTA (Xia et al., 2018) and DIORR (Li et al., 2020a) (1869 and 23463 images) being significantly smaller than widely-used vanilla object detection datasets such as COCO (Lin et al., 2014) (>330K images). Given the computational burden of expanding

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

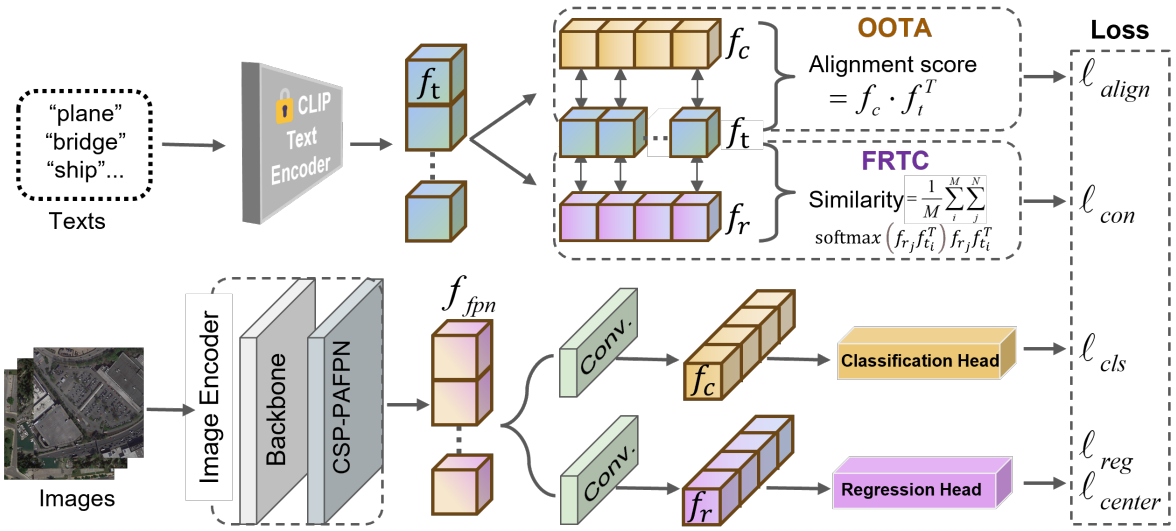

Figure 1. The left lower branch shows the image encoder derived from RTMDet producing f_{fpn}, leading to classification f_{c} and region features f_{r}. The left upper branch illustrates CLIP's text encoder generating text features f_t. OOTA (FRTC) facilitates alignment (contrastive) learning between f_{c} (f_{r}) and f_{t} to generate alignment (similarity) score fine-tuned by ℓ_{align} (ℓ_{con}).

datasets, optimizing performance with minimal training samples is crucial. The main contributions are as follows: (1) we propose novel vision-language methods comprising a OOTA and a FRTC, using alignment and contrastive losses. It enhances feature representations with text during fine-tuning, improving image-text alignment and detection performance. (2) We conduct extensive experiments on diverse OoD datasets using few-shot learning, achieving up to 75.1% higher mAP than prior approaches. (3) VL-Rotate pioneers vision-language integration for ROD, addressing OoD challenges, and achieving state-of-the-art performance.

2. Related Work

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 two main categories: two-stage detector (Yang et al., 2019; Ding et al., 2019; Xu et al., 2021; Han et al., 2021b; Zonghao Guo & Ye, 2021; Xie et al., 2021; Wentong Li, 2022) and one-stage detector (Tian et al., 2019; Yang et al., 2021; Han et al., 2021a; Hou et al., 2022; Yang et al., 2023b;a; Yu & Da, 2023; Lyu et al., 2022). Two-stage detectors, building on the RCNN framework (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015), excel in accuracy by employing deep convolutional networks for feature mapping, a Region Proposal Network (RPN) for proposing regions, and dedicated heads for specific tasks. More recently, Oriented RepPoints (Wentong Li, 2022) introduced an effective adaptive points representation to capture the geometric information of objects and proposed a corresponding quality assessment for adaptive points learning.

On the other hand, one-stage detectors, comprising a backbone, feature pyramid network (FPN), and task-specific head, are recognized for their speed and parameter efficiency, eliminating the need for generating extensive predefined region proposals. Recently, there has been a growing trend of exploring one-stage detectors for ROD. Noteworthy contributions in this area include SASM (Hou et al., 2022), which enhances detection accuracy through shape-adaptive sampling and localization strategies; KFIoU (Yang et al., 2023b), which introduces a novel regression loss addressing boundary issues via Gaussian Wasserstein distance; and RTMDet (Lyu et al., 2022), offering an efficient real-time detection solution with large-kernel depth-wise 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 (Zhang et al., 2022):
(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 (Ye et al., 2021; Gao et al., 2023), transfer learning (Wenzel et al., 2022; Blanchard et al., 2021), and meta-learning (Zhang et al., 2023).
(2) **Invariant representation learning** Exemplified by Invariant Risk Minimization (IRM) (Arjovsky et al., 2019), this approach explores causal relationships in data across different environments based on causal invariant features. Invariant risk minimization games, proposed by (Ahuja et al.,

2020), treat IRM as finding the Nash equilibrium of an ensemble game among multiple environments. Pareto Invariant Risk Minimization (Chen et al., 2022) 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 (Zhou et al., 2022c) 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 (Wang et al., 2022) and Bayesian methods (Kristiadi et al., 2022), etc.

2.3. 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 (Su et al., 2019), LX-MERT (Tan & Bansal, 2019), and UNITER (Chen et al., 2020), have explored efficient approaches to learn representations from image and text pairs. Notably, Contrastive Language-Image Pretraining (CLIP, (Radford et al., 2021)) stands out by effectively learning vision-language representations through maximizing image-text feature similarity, using over 400 million pairs for pre-training. CLIP's framework has further inspired developments in few-shot learning, with models such as CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) innovating in prompt-tuning for better few-shot generalization, and CLIP-Adapter (Gao et al., 2021) introducing modifications for fine-tuning visual backbones in downstream applications.

3. Methodology

In this work, we introduce VL-Rotate, featuring a one-stage anchor-free detector, an Oriented Object Text Alignment Module, and a Fine-grained Region-text Contrastive Module, as depicted in Figure 1. Subsequent sections detail these components.

3.1. Oriented Object Text Alignment Module

Applying Large-scale pre-trained vision-language models to ROD tasks, which require precise region and angle predictions, poses challenges in effectively fusing visual and text information. Here we address this issue by innovatively introducing the Oriented Object Text Alignment Module (OOTA). We employ RTMDet (Lyu et al., 2022) as the backbone for integrating OOTA. RTMDet comprises a backbone, an FPN, and three task-specific heads for classification, region, and angle predictions. To achieve OOTA, we leverage the contrastive language image pre-training framework, which includes an image encoder and a text encoder, enabling the alignment of image features and language features in the embedding space. This approach significantly bolsters the model's capability in out-of-distribution ROD by enhancing robustness against samples diverging from the training set, crucial for accurately detecting test samples with distinct characteristics.

3.1.1. Alignment Architecture

The one-stage detector extracts image features through the backbone and employs the FPN to detect objects of various sizes. The FPN generates output region features that contain image feature representations and rich multi-scale information. Building upon this, we focus on the text feature of each category extracted from the CLIP's text encoder and adjust the dimension of the classification feature from the region feature to match the text feature. This process is termed OOTA. Using the frozen pre-training knowledge of the text encoder, the classification features generated by the model can be aligned with it to obtain more robust OoD inference performance. Alignment scores result from element-wise multiplication of the classification feature and text feature matrices. Given that the text feature from the pre-train model and the classification feature from FPN reside in distinct embedding spaces, we perform fine-tuning for rotated object detection. This process aligns the text feature with the image feature during training. The alignment loss is then utilized to guide the model's fine-tuning through the backward process.

3.1.2. Alignment Loss

To enable the alignment learning guided model to integrate image feature representations with text during fine-tuning and address rotated object detection in OoD scenarios, we introduce the alignment loss, denoted as ℓ_{align}. As shown in Figure 1, given a training sample x_{image} and all categories set $Y_{label} = \{y_{class_1}, y_{class_2}, ..., y_{class_M}\}$, we input x_{image} into the image encoder, i.e., backbone+FPN, and obtain the feature $f_{fpn} \in \mathbf{R}^{N \times D}$, where N is the number of features and D is the feature dimension. The text feature $f_t \in \mathbf{R}^{M \times C}$ is obtained by inputting Y_{label} to the text encoder, where M is the number of categories and C is the dimension of the text feature.

Unlike RTMDet, we adjust the function of the convolution layer in the classification head and modify its output channel dimension to C, allowing the output feature to serve two functions: alignment learning and classification. We apply the convolution layer to f_{fpn} to obtain the classification feature $f_c \in \mathbf{R}^{N \times C}$. For alignment learning, we calculate the inner product between f_c and f_t to obtain the alignment score, which predicts the classification results. The alignment loss ℓ_{align} is then introduced for training using the Generalized Focal Loss proposed by (Li et al., 2020b), an

effective choice for dense object detection, given the prevalence of dense objects in aerial images during rotated object detection.

3.2. Fine-Grained Region-Text Contrastive Module

Remote-sensing images provide rich instance-level annotations that offer valuable text information about the image, aiding the model in learning image-text alignment. To fully exploit the instance-level information, we propose a novel contrastive learning method called Fine-Grained Region-Text Contrastive Module (FRTC). FRTC leverages pairs of images from rotated object detection and associated text, enabling enhanced utilization of instance-level annotations from both modalities. This approach facilitates the learning of fine-grained text-region correspondences, which are mutually reinforced through the alignment learning process. Further details are provided in the subsequent subsections.

3.2.1. Similarity Computation from Text to Region

Firstly, a training image denoted as x_{image} and a category set Y_{label} containing texts of all categories are provided. The combined input of x_{image} and Y_{label} is fed into the model to yield both the region feature f_r and the text feature f_t. The region feature $f_r \in \mathbf{R}^{N\times C}$ is extracted by processing the image feature through the convolution layer of the regression head within the FPN, where the last convolution layer's output size is (B, C, H, W). In this context, $N = H \times W$ represents the number of output regions and C denotes the feature dimension of each region. $f_t \in \mathbf{R}^{M\times C}$ is the text features derived from the label set Y_{label} using the text encoder of CLIP. Here, C signifies the feature dimension of the text feature f_{t_i} corresponding to y_{class_i}, and M pertains to the number of categories within the category set Y_{label}.

For the text feature f_{t_i}, its text-region similarity with all region features f_r is denoted as $\Omega(f_r, f_{t_i})_i$:

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

Subsequently, the total text-region similarity $\Omega(f_r, f_t)$ is calculated by taking the sum of text-region similarity:

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

From Equation (2), the text-region similarity of text feature f_t and region feature f_r is derived, reflecting the similarity between x_{image} and Y_{label}. Note that the total text-region similarity accounts for all region features including some background regions unrelated to text information in x_{image}, particularly in remote-sensing images where objects are typically small. This introduces considerable distracting information to the text-region similarity. 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} corresponding to the i-th category. This leads to the optimal-matching text-region similarity $\bar{\Omega}(f_r, f_t)$:

$$\bar{\Omega}(f_r, f_t) = \frac{1}{M}\sum_{i=1}^{M} \hat{f}_{r_i} f_{t_i}^T \tag{3}$$

Clearly, when considering all text features f_t solely for the most compatible region feature f_r, the total text-region similarity $\Omega(f_r, f_t)$ achieves its maximum: $\Omega(f_r, f_t) \leq \bar{\Omega}(f_r, f_t)$.

Furthermore, it is worth noting that the aforementioned optimal-matching text-region similarity assumes a one-to-one relationship between text features and region features. However, the dense distribution of objects in aerial images often leads to multiple objects of the same category appearing in the image, creating a one-to-many relationship between text features and region features. Consequently, the optimal-matching text-region similarity may not fully meet the requirements for calculating the text-region relationship. In the context of rotated object detection, particularly in scenarios with dense detection, achieving a balance between the desired text-region similarity and the one-to-many relationship becomes even more pivotal. We introduce the softmax-weight-sum method to encode the probability of text features across all region features. Specifically, 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 $f_{r_j} f_{t_i}^T$ can be described as:

$$\text{softmax}(f_{r_j} f_{t_i}^T) = \frac{\exp(f_{r_j} f_{t_i}^T/\beta)}{\sum_r \exp(f_r f_{t_i}^T/\beta)} \tag{4}$$

where β is the hyperparameter that controls the level of softmax probability sharpening.

The softmax probability is then incorporated into Equation (2) to derive the final matching text-region similarity as follows:

$$\bar{\Omega}(f_r, f_t) = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N} \text{softmax}(f_{r_j} f_{t_i}^T) f_{r_j} f_{t_i}^T \tag{5}$$

This refined text-region similarity matching accounts for all region features and appropriately weights each feature, thereby highlighting the significance of region features aligned with the text.

3.2.2. Contrastive Loss

A standard contrastive loss, similar to the one in (Radford et al., 2021), can be formulated based on the refined text-

Figure 2. Visualization of the inference results between proposed VL-Rotate and the baseline model RTMDet on DOTAOoD. The purple, yellow, red, blue, and cyan rectangles respectively represent the class "small vehicle", "harbor", "plane", "helicopter", and "tennis court". More visualization results of other categories can be found in the Appendix.

region similarity matching. It is expressed as:

$$\ell_{con} = -\frac{1}{B}\log\frac{\exp(\bar{\Omega}(f_r, f_{t_i})/\beta)}{\sum_r \exp(\bar{\Omega}(f_r, f_{t_i})/\beta)} \quad (6)$$

Here, the term B represents the batch size in a single iteration. Given a training sample x_{image} and the corresponding category set Y_{label}, the region feature f_r is derived from x_{image} using the image encoder (backbone+FPN), while the text feature f_t is generated from Y_{label} through the text encoder. $\bar{\Omega}(f_r, f_t)$ represents the text-region similarity as depicted in Equation (5), and β serves as a hyperparameter that tempers the sharpness of the contrastive loss. This loss aids in training the model to learn a more refined text-region similarity, thereby enhancing its robustness to distribution shifts arising from background interference in rotated object detection, particularly in out-of-distribution scenarios.

3.3. Overall Training Loss

The comprehensive training loss incorporates the losses ℓ_{cls}, ℓ_{reg}, and ℓ_{center}, which are utilized in RTMDet for predicting classification and regression, respectively. The expression for the overall training loss is given by:

$$\ell = \omega_1\ell_{cls} + \omega_2\ell_{reg} + \omega_3\ell_{center} + \omega_4\ell_{align} + \omega_5\ell_{con} \quad (7)$$

As mentioned, here, ℓ_{align} represents the alignment loss, which follows the Generalized Focal Loss introduced by (Li et al., 2020b), and ℓ_{con} denotes the contrastive loss. The terms ℓ_{cls}, ℓ_{reg}, and ℓ_{center} correspond respectively to the classification loss, region prior loss, and regression cost, derived from RTMDet. The coefficients ω_1, ω_2, ω_3, ω_4 and ω_5 are the weights assigned to these five losses. For ω_1, ω_2, and ω_3, we adopt default values of $\omega_1 = 1$, $\omega_2 = 3$, and $\omega_3 = 1$ in RTMDet. Only ω_4 and ω_5 are treated as hyperparameters during training.

4. Experiments

For fair comparisons, we implement all experiments based on the MMRotate(Zhou et al., 2022d) with Pytorch 1.10 and run on a NVIDIA GeForce RTX 3080 GPU. More details can be found in the Appendix. We will introduce the datasets used in experiments and how OoD distribution shifts are introduced for evaluation. Interestingly, we find under the challenging OoD setting, baseline methods' performances do not improve or get worse with more training data provided. We use the same few-shot setting for all methods for fair comparisons.

4.1. Datasets

4.1.1. HRSCOoD

HRSCOoD, derived from the widely used HRSC2016 dataset (Liu et al., 2017), serves as a dataset that captures intricate real-world nuances of ship image distribution and wake presence. HRSC2016 offers 1061 annotated images with OBB for ship detection, including 2976 objects and is split into training (436), validation (181), and testing (444) subsets. We introduce distribution shifts by segregating images based on whether they are taken in inshore as they demonstrate distinct texture and style information, posing challenges for OoD generalization. Images featuring harbors scenes from the original training set compose the training subset, while offshore images form the testing sub-

Table 1. Comparison of our approach with rotated object detection methods on DOTAOoD. Categories including: plane (PL), baseball-diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC).

METHOD	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	MAP(%)
ROTATED REPPOINTS (YANG ET AL., 2019)	68.9	57.1	5.9	13.7	32.6	11.4	25.7	60.9	22.1	51.8	28.4	29.8	16.2	51.4	6.1	32.1
ROTATED FCOS (TIAN ET AL., 2019)	69.7	35.0	14.5	26.8	28.6	21.8	34.6	79.9	27.6	46.1	22.3	24.1	16.7	49.4	7.6	33.6
ROI TRANSFORMER (DING ET AL., 2019)	71.4	72.7	25.6	49.7	54.8	44.0	58.4	81.5	51.8	59.7	39.0	29.1	41.7	56.1	9.1	49.7
R^3DET (YANG ET AL., 2021)	68.3	57.6	15.7	28.9	40.3	32.1	41.6	80.6	25.9	48.4	22.6	23.0	13.8	43.2	1.0	36.2
REDET (HAN ET AL., 2021B)	79.2	74.5	26.8	45.0	56.6	45.6	67.5	**85.6**	53.3	60.9	40.6	28.2	44.2	44.3	9.1	50.7
CFA (ZONGHAO GUO & YE, 2021)	66.3	56.4	22.1	39.1	52.4	38.0	53.7	80.3	32.3	50.8	24.3	26.1	28.0	56.8	7.7	42.3
ORIENTED R-CNN (XIE ET AL., 2021)	70.5	71.3	23.5	43.6	53.4	46.5	59.9	81.3	42.2	59.4	32.6	28.0	24.5	50.9	9.1	46.5
SASM (HOU ET AL., 2022)	62.0	27.8	11.1	12.3	25.9	25.4	12.2	72.2	10.3	38.4	17.2	11.9	14.5	42.8	0.1	25.6
ORIENTED REPPOINTS (WENTONG LI, 2022)	64.0	60.1	18.4	44.6	48.1	36.8	55.7	80.5	35.2	52.3	27.6	27.5	22.8	56.7	9.1	42.6
KFIOU (YANG ET AL., 2023B)	67.9	52.0	0.1	10.7	11.4	2.4	13.8	75.3	6.0	37.5	20.6	6.4	10.3	20.4	0.1	22.3
H2RBOX (YANG ET AL., 2023A)	72.6	56.4	3.7	28.4	19.0	6.1	29.3	79.9	35.4	53.0	17.7	20.3	20.1	56.3	1.8	35.0
PSC (YU & DA, 2023)	62.6	0.1	0.2	2.8	13.3	0.5	3.8	2.0	0.5	38.2	6.6	0.1	0.1	1.0	7.1	15.8
RTMDET (LYU ET AL., 2022)	77.2	74.6	21.3	43.4	**69.8**	71.9	77.7	84.2	51.8	64.8	39.5	27.4	52.4	61.1	9.1	55.1
RTMDET+VL-ROTATE(OURS)	**79.7**	**75.8**	**30.9**	**53.1**	68.9	**73.1**	**78.8**	84.7	**56.9**	**67.7**	**40.9**	**29.6**	**61.8**	**63.0**	**9.2**	**58.3**

Table 2. Comparison of our approach with baseline rotated object detection methods on HRSCOoD. Reported are mAP values for both the in-domain and out-of-distribution test sets.

METHOD	MAP(%, ID)	MAP(%, OOD)
ROTATED REPPOINTS (YANG ET AL., 2019)	35.73	45.90
ROTATED FCOS (TIAN ET AL., 2019)	60.27	62.57
ROI TRANSFORMER (DING ET AL., 2019)	68.91	67.49
R^3DET (YANG ET AL., 2021)	38.30	35.57
REDET (HAN ET AL., 2021B)	67.65	70.31
CFA (ZONGHAO GUO & YE, 2021)	33.31	38.96
ORIENTED R-CNN (XIE ET AL., 2021)	53.76	50.41
SASM (HOU ET AL., 2022)	65.53	59.30
ORIENTED REPPOINTS (WENTONG LI, 2022)	38.07	27.64
KFIOU (YANG ET AL., 2023B)	14.67	31.70
H2RBOX (YANG ET AL., 2023A)	24.53	47.96
PSC (YU & DA, 2023)	75.29	67.57
RTMDET (LYU ET AL., 2022)	64.33	51.83
RTMDET+VL-ROTATE(OURS)	**76.70**	**75.20**

Table 3. Comparison of our approach with baseline rotated object detection methods on RSDDOoD. Reported are mAP values for both the in-domain and out-of-distribution test sets.

METHOD	MAP(%, ID)	MAP(%, OOD)
ROTATED REPPOINTS (YANG ET AL., 2019)	28.70	50.37
ROTATED FCOS (TIAN ET AL., 2019)	22.93	40.00
ROI TRANSFORMER (DING ET AL., 2019)	38.63	57.30
R^3DET (YANG ET AL., 2021)	36.67	65.37
REDET (HAN ET AL., 2021B)	**54.40**	74.30
CFA (ZONGHAO GUO & YE, 2021)	39.24	66.51
ORIENTED R-CNN (XIE ET AL., 2021)	49.70	73.00
SASM (HOU ET AL., 2022)	42.67	71.93
ORIENTED REPPOINTS (WENTONG LI, 2022)	39.33	63.20
KFIOU (YANG ET AL., 2023B)	34.53	58.23
H2RBOX (YANG ET AL., 2023A)	19.31	38.55
PSC (YU & DA, 2023)	12.40	24.05
RTMDET (LYU ET AL., 2022)	52.27	57.20
RTMDET+VL-ROTATE(OURS)	52.13	**74.97**

set. We randomly selected 64 images from the training set used for few-shot learning.

4.1.2. DOTAOoD

DOTAOoD is derived from DOTA v1.0 (Xia et al., 2018), which includes 2,800 aerial images with 188,000 objects across 15 categories, used for rotated object detection. To emulate real-world test scenarios, we build DOTAOoD by inducing distribution shifts in the DOTA dataset using the k-means clustering algorithm. All aerial images from both training and test sets are flattened by pixel and clustered into two classes based on the pixel characteristics: One for training and the other for testing. This approach equips the model to adapt to diverse data distributions. A few-shot training set is formed by randomly selecting 16 images. All images in both training and test sets are cropped to 1024 × 1024 pixels, with a 256-pixel overlap in training samples.

4.1.3. RSDDOoD

RSDDOoD is a dataset derived from RSDD-SAR (Congan et al., 2022) for OoD generalization. Similar to HRSC2016, RSDD-SAR is a remote-sensing ship detection dataset. It includes 7,000 image slices with various modes, polarizations and resolutions, with over 10,000 objects. We introduce domain shifts in this new dataset named RSDDOoD based on inshore and offshore contexts. We also randomly select 64 samples within RSDDOoD's training set for few-shot learning.

4.1.4. DIORROoD

DIORROoD is a dataset with distribution shifts constructed from DIORR (Li et al., 2020a). DIORR offers 23k images with 190k instances across 20 classes, containing diverse object sizes and distinct acquisition conditions. Similar to DOTAOoD, we employ K-means clustering to induce shifts, partitioning DIORR into training and test sets to create DIORROoD. For few-shot training, 64 images are randomly drawn from the training set. More details for the construction of all datasets are in the Appendix.

4.2. Results

The experiment results on HRSCOoD validate the effectiveness of our method, as depicted in Table 2. Surpassing all counterparts, VL-Rotate achieves an impressive 75.20% mAP in the out-of-distribution scenario, outperforming the leading comparison method by 4.89%, and the baseline (RTMDet) by 23.37%. Additionally, our method outperforms other comparison methods on the in-domain test set,

Table 4. Performance comparison of mAP values using 64-shot learning between our proposed method and previously reported methods for rotated object detection on DIORROoD. Classes including: airplane (C1), airport (C2), baseball field (C3), basketball court (C4), bridge (C5), chimney (C6), expressway service area (C7), expressway toll station (C8), dam (C9), golf field (C10), ground track field (C11), harbor (C12), overpass (C13), ship (C14), stadium (C15), storage tank (C16), tennis court (C17), train station (C18), vehicle (C19), windmill (C20).

METHOD	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	MAP(%)
ROTATED REPPOINTS (YANG ET AL., 2019)	67.5	9.7	77.2	67.9	10.4	76.8	25.7	26.9	11.3	**59.8**	61.5	7.8	12.6	49.7	77.3	53.0	62.6	8.0	19.6	39.9	41.2
ROTATED FCOS (TIAN ET AL., 2019)	70.7	15.4	79.2	70.4	16.2	79.0	41.5	39.2	10.4	52.3	76.2	9.4	28.0	67.2	70.8	58.5	80.8	20.0	24.0	41.8	47.5
ROI TRANSFORMER (DING ET AL., 2019)	74.4	1.2	80.1	77.6	14.9	81.2	30.2	41.4	7.6	8.5	81.0	4.2	26.1	70.3	82.2	60.0	81.3	21.5	29.0	38.7	45.5
R^3DET (YANG ET AL., 2021)	65.6	10.2	76.9	76.1	11.0	80.4	35.9	34.8	5.1	37.1	75.1	8.0	18.6	56.9	80.4	57.4	80.5	15.9	20.2	39.4	44.3
REDET (HAN ET AL., 2021B)	74.8	19.2	80.8	**80.2**	21.7	**81.6**	53.4	55.7	15.0	18.7	**83.8**	11.3	37.6	79.7	**84.3**	60.5	**81.5**	30.3	31.5	**43.1**	52.2
CFA (ZONGHAO GUO & YE, 2021)	14.9	0.2	47.8	29.8	1.1	14.6	0.3	2.9	0.1	1.9	12.5	1.2	1.4	34.2	16.6	33.8	67.9	0.7	6.3	22.2	15.5
ORIENTED R-CNN (XIE ET AL., 2021)	70.6	5.1	80.0	75.5	12.3	80.8	24.9	33.0	8.5	18.2	79.8	2.1	21.2	68.9	80.8	59.0	81.0	20.9	26.5	34.2	44.2
SASM (HOU ET AL., 2022)	1.3	0.5	25.5	10.6	9.2	2.5	0.2	20.0	1.2	2.7	26.3	0.1	0.8	15.1	28.9	12.4	51.5	0.2	3.3	5.8	10.0
ORIENTED REPPOINTS (WENTONG LI, 2022)	29.9	0.6	53.5	44.6	11.0	7.5	1.5	17.1	2.4	1.4	26.8	0.4	6.7	31.4	28.6	29.3	67.9	1.4	13.3	27.5	20.1
KFIOU (YANG ET AL., 2023B)	70.5	11.1	73.3	75.5	13.9	81.1	33.6	30.3	7.1	45.8	75.8	13.3	28.5	65.0	77.0	58.9	80.5	20.6	19.3	37.5	45.9
H2RBOX (YANG ET AL., 2023A)	70.5	7.7	77.5	67.7	6.2	79.4	25.5	32.7	4.2	47.1	72.7	3.9	9.7	53.7	71.7	55.8	80.3	16.6	20.8	36.8	42.0
PSC (YU & DA, 2023)	66.3	8.1	74.7	76.0	13.2	**81.6**	31.9	33.0	10.3	54.1	77.9	14.9	20.5	58.4	79.8	57.3	81.0	22.9	20.4	38.7	42.4
RTMDET (LYU ET AL., 2022)	74.7	16.6	80.7	76.6	18.1	81.0	50.1	51.0	11.2	50.7	80.9	13.8	34.6	78.8	81.7	65.7	**81.5**	27.3	29.8	46.7	52.5
RTMDET+VL-ROTATE(OURS)	**79.0**	**21.5**	**80.9**	78.7	**21.9**	80.7	**58.4**	**56.2**	**17.7**	55.2	83.6	**16.4**	**38.4**	**79.8**	84.0	**66.0**	**81.5**	**30.8**	30.8	**48.1**	**55.5**

achieving the highest mAP of 76.70%. We also conducted experiments on RSDDOoD, a dataset constructed similarly to HRSCOoD, and the performance results are showcased in Table 3. Our VL-Rotate attains a mAP of 74.97% among state-of-the-art methods in out-of-distribution scenarios. Given the inherent challenges of the task, some methods struggle to accurately detect objects. Also, our method exhibits performance on par with the best comparison method (ReDet) on the in-domain test set.

Another experiment was conducted on DOTAOoD, with performance metrics reported in Table 1. Our method achieves an impressive average mAP of 58.3%, outperforming other comparison methods and exhibiting better performances across most categories. As visualized in Figure 2, compared to the baseline, our method exhibits significantly lower occurrences of false positives and missed detection across different classes. We further conducted experiments on DIORROoD with rich distribution shifts. While certain comparison methods fail to detect, our proposed approach maintains consistent and competitive performance across most classes, achieving the highest average mAP of 55.5% among all reference methods. Notably, our method particularly excels in specific categories like harbors, benefiting from its text and image embedding alignment. This alignment helps overcome interference in scenarios where ships impede harbor detection, enhancing accurate detection for both categories in dense scenes. These findings show the superior performance and robustness of our method for detecting rotated objects in few-shot OoD generalization scenarios and highlight its potential for many practical applications when data are expensive to obtain.

4.3. Ablation Study

In this section, we conduct a series of ablation experiments to demonstrate the superiority of our method and exclude potential confounding factors. Note that all other experimental configurations align with the aforementioned implementation details.

Table 5. The comparison of OoD mAP values using different one-stage detectors with VL-Rotate.

METHOD	HRSCOOD	DOTAOOD	RSDDOOD	DIORROOD
R^3DET (YANG ET AL., 2021)	35.6	36.2	65.4	44.3
KFIOU (YANG ET AL., 2023B)	31.7	22.3	58.2	45.9
RTMDET (LYU ET AL., 2022)	51.8	49.9	57.2	52.5
R^3DET+VL-ROTATE	43.9	37.5	69.6	46.1
KFIOU+VL-ROTATE	42.0	27.2	69.1	48.9
RTMDET+VL-ROTATE	**75.2**	**52.2**	**75.0**	**54.0**

Table 6. The comparison of OoD mAP values using different text encoders with VL-Rotate. Note that W2V denotes word2vector.

METHOD	LANGUAGE MODEL	HRSCOOD	RSDDOOD
RTMDET (LYU ET AL., 2022)	-	51.8	57.2
RTMDET+VL-ROTATE	W2V	63.8	72.7
RTMDET+VL-ROTATE	BERT	63.2	74.3
RTMDET+VL-ROTATE	CLIP(VIT/B-32)	**75.2**	**75.0**

4.3.1. VARIOUS ONE-STAGE DETECTOR WITH VL-ROTATE

Integrating VL-Rotate with RTMDet significantly improves its performance on OoD datasets. Moreover, VL-Rotate is also capable of being applied to other one-stage detector architectures such as R3Det (Yang et al., 2021) and KFIoU (Yang et al., 2023b), yielding significant mAP improvements on OoD datasets as well. Specifically, compared to their original backbones, the performance was improved by 8.3% and 10.3% on HRSCOoD, 1.3% and 4.9% on DOTAOoD, 4.2% and 10.9% on RSDDOoD, and 1.8% and 3.0% on DIORROoD, respectively. Detailed performance metrics are available in Table 5.

4.3.2. VARIOUS LANGUAGE MODELS WITH VL-ROTATE

To assess the impact of the CLIP text encoder on VL-Rotate, we substituted it with alternative language models such as W2V (Mikolov et al., 2013) and BERT (Devlin et al., 2018), keeping all other components constant. Results on HRSCOoD and RSDDOoD, detailed in Table 6, show that VL-Rotate with BERT and W2V achieved mAPs of 63.2% and 63.8% on HRSCOoD, and 74.3% and 72.7% on RSDDOoD, respectively, surpassing the baseline RTMDet by

Table 7. Performance comparison of using different backbones in RTMDet and our proposed method.

METHOD	BACKBONE	HRSCOOD (MAP, %)	DOTAOOD (MAP, %)	RSDDOOD (MAP, %)	DIORROOD (MAP, %)
RTMDET (LYU ET AL., 2022)	RTMDET-TINY	39.1	44.0	71.0	31.4
	RTMDET-S	43.0	46.4	75.5	36.4
	RTMDET-M	65.7	50.5	79.6	47.4
	RTMDET-L	51.8	49.9	57.2	52.5
RTMDET+ VL-ROTATE(OURS)	RTMDET-TINY	39.5	45.2	71.6	39.7
	RTMDET-S	43.9	48.3	75.9	44.4
	RTMDET-M	69.2	51.1	**82.3**	50.4
	RTMDET-L	**75.2**	**52.2**	75.0	**54.0**

11.4% to 17.1%. This improvement leverages the rich pre-training of these text encoders, enhanced by our OOTA and FRTC integration in VL-Rotate, especially notable with CLIP's text encoder, which yields the best performance.

4.3.3. VARIOUS BACKBONES

We also performed ablation studies using different RTMDet backbones. Results comparing RTMDet and our approach across various backbones (RTMDet-tiny, RTMDet-s, and RTMDet-m) are presented in Table 7. Irrespective of the backbone employed, our method consistently outperforms or achieves comparable performance to the baseline across all datasets. It underscores that the choice of backbone does not diminish the substantial performance gains our method offers. Also, compared to other smaller backbones, RTMDet-l achieves the highest mAP in most datasets for our approach. This highlights the effectiveness of larger parameterized backbones in enhancing feature representation learning for both vision and language in OoD generalization.

4.3.4. CHANNEL SIZE OF CONVOLUTION IN HEAD

Since we increased the convolutional head's channel number to align with language features in our method, we then evaluated whether the improved few-shot performance is due to channel count. We, therefore, match the RTMDet's channel size with ours, with detailed results shown in Table 10 in the Appendix. Notably, our method still showcases a substantial improvement of 22.9%, 0.5%, 9.4%, and 1.4% mAP over the baseline across all four datasets upon increasing the channel size from default 256 to 512. These findings suggest that while the baseline's OoD generalization performance benefits from a larger channel size, it remains significantly lower than our proposed method's performance.

4.3.5. MODULES ABLATION ANALYSIS

We also evaluated the independent contributions of OOTA and FRTC on HRSCOoD, with results shown in Table 9. Our approach outperforms the baseline (RTMDet, where neither module is used) by 11.8% and 7.6% when OOTA and FRTC are used individually, respectively. Significantly, their combined use exceeds these individual gains, outperforming the baseline by 23.4%. These findings highlight the positive impact of OOTA and FRTC and their synergistic effect in enhancing the model's language utilization for improved few-shot OoD generalization in rotated object detection.

Table 8. Performance comparison of using different learning shots on HRSCOoD.

METHOD	16-SHOT	32-SHOT	64-SHOT	FULL DATA
REDET (HAN ET AL., 2021B)	14.3	30.0	70.3	72.1
RTMDET (LYU ET AL., 2022)	6.3	20.3	51.8	69.7
RTMDET+VL-ROTATE(OURS)	**14.9**	**51.8**	**75.2**	**80.5**

Table 9. Performance of using individual modules on HRSCOoD.

METHOD	OOTA	FRTC	MAP(%)
RTMDET (LYU ET AL., 2022)	-	-	51.8
RTMDET+VL-ROTATE(OURS)	✓		63.6
		✓	59.4
	✓	✓	**75.2**

4.3.6. NUMBER OF SHOTS

We evaluated the performance of our method and the baseline using a different number of shots on HRSCOoD (Table 8). Decreasing shots from 64 to 32 and 16 results in performance decline for both approaches. However, our method consistently outperforms the baseline, highlighting the significant potential of vision-language learning. This guided our decision to use 64 shots instead of a smaller number, striking a balance between cost and performance to ensure adequate real-world performance. This principle guides shot selection for other datasets as well. We also compared our few-shot method to the second-best approach (ReDet) trained with full data on HRSCOoD. Notably, even with a 64-shot training, our method continues to outperform other methods trained with a full dataset, including ReDet, further demonstrating the proposed methods' robustness and stability. More results can be found in Table 11 in the Appendix.

5. Conclusion

In addressing the intricate challenge of few-shot out-of-distribution (OoD) generalized rotated object detection, we introduce VL-Rotate, a vision-language learning framework. Comprising two modules—Oriented Object Text Alignment Module (OOTA) and Fine-Grained Region-Text Contrastive Module (FRTC)—VL-Rotate demonstrates a powerful approach. OOTA aligns region features with text features in a high-dimensional space, leveraging language-rich representations to guide improved regression, enhancing generalization under distribution shifts. FRTC enhances instance-level annotation utilization, fostering fine-grained word-region correspondence learning. Through extensive experiments on publicly available datasets, VL-Rotate showcases state-of-the-art few-shot OoD generalized performance. This work advances the landscape of rotated object detection by successfully addressing its more challenging variant.

Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

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

Blanchard, G., Deshmukh, A. A., Dogan, Ü., Lee, G., and Scott, C. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22 (1):46–100, 2021.

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

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

Congan, X., Hang, S., Jianwei, L., Yu, L., Libo, Y., Long, G., Wenjun, Y., and Taoyang, W. Rsdd-sar: Rotated ship detection dataset in sar images. *Journal of Radars*, 11(4): 581–599, 2022.

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

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

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

Gao, I., Han, R., and Yue, D. Exploring adversarial training for out-of-distribution detection. 2023.

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

Girshick, R. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

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

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

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

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

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

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

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., and Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33: 21002–21012, 2020b.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, Z., Yuan, L., Weng, L., and Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods*, volume 2, pp. 324–331. SciTePress, 2017.

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

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

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

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

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

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

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

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

Wentong Li, Yijie Chen, K. H. J. Z. Oriented reppoints for aerial object detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

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

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

Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.-S., and Bai, X. 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. ISSN 0162-8828. doi: 10.1109/tpami.2020.2974745.

Yang, X., Yan, J., Feng, Z., and He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3163–3171, 2021.

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

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

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

Ye, N., Tang, J., Deng, H., Zhou, X.-Y., Li, Q., Li, Z., Yang, G.-Z., and Zhu, Z. Adversarial invariant learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12441–12449. IEEE, 2021.

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

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

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

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

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

Zhou, X., Lin, Y., Zhang, W., and Zhang, T. Sparse invariant risk minimization. In *International Conference on Machine Learning*, pp. 27222–27244. PMLR, 2022c.

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

Zonghao Guo, Chang Liu, X. Z. J. J. X. J. and Ye, Q. 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.

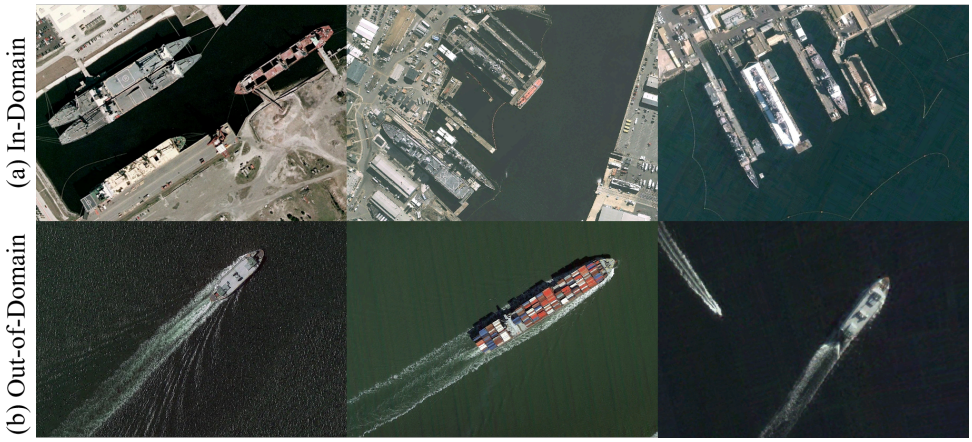

Figure 3. Visualization of training set and test set on HRSCOoD.

A. Implementation Details

Our proposed approach is implemented based on RTMDet-R (Lyu et al., 2022). The baseline employs RTMDet-l with ImageNet pre-trained checkpoint. Note that our method is agnostic to backbone selection. This is further validated with other RTMDet backbones—RTMDet-tiny, RTMDet-s, and RTMDet-m—in ablation study. Experiments are performed with single-scale training across all datasets, without data augmentation except in DOTAOoD where random rotation and flip are applied. The evaluation metric is mAP calculated with rotated IoU, following PASCAL VOC2007 (Everingham et al., 2010) conventions. The training adopts the AdamW optimizer with initial learning rates of 0.002 and 0.05, and weight decays of 0.00025 and 0.05 for HRSCOoD and DOTAOoD, respectively. Maximum epochs are set at 108 for HRSCOoD and 36 for DOTAOoD. RSDDOoD and DIORROoD are similarly constructed and their experiment settings respectively match those of HRSCOoD and DOTAOoD. As for ω_4 and ω_5, we equally select 1e-2, 1e-4, 1e-2, and 1e-3 for HRSCOoD, DOTAOoD, RSDDOoD, and DIORROoD respectively. Batch sizes per GPU are 8 for HRSCOoD, 4 for DOTAOoD, 2 for RSDDOoD, and 2 for DIORROoD. All experiments of our approach utilize mmrotate-1.x (Zhou et al., 2022d), PyTorch 1.10, and run on 2 NVIDIA GeForce RTX 3080Ti GPUs with 12GB memory each, on the Ubuntu 20.04 operating system. For other compared rotated object detection methods, we adhere to the configurations provided by their respective experiments on HRSC, DOTA, RSDD, and DIORR. Some reference methods may not fit well with the mmrotate-1.x branch, prompting their experimentation on the main branch. In cases where these methods lack dataset-specific configurations, default experimental settings from either mmrotate-1.x or mmrotate-main are adopted. The default runtime is used unless otherwise specified in the original configurations. In the context of ship detection datasets like HRSCOoD and RSDDOoD, a 3x schedule is employed. Methods that require more extensive training and a lower learning rate for optimal convergence are adjusted to a 6x schedule. For DOTAOoD and DIORROOD, these methods adhere to the default 1x schedule. To ensure the credibility of our results, each experiment is repeated over 3 times, and the average of stable outcomes is calculated. This rigorous approach guarantees reliable and robust comparisons between our proposed method and other methods.

B. Datasets

HRSCOoD: The original HRSC2016 (Liu et al., 2017) dataset comprises 436 training images, 181 validation images, and 444 test images. We introduce distribution shifts by segregating images based on ship presence in harbors. Specifically, the original training, validation, and test sets contain 410, 116, and 413 images with ships in harbors, while the harbor-absent images in these sets are 26, 15, and 31, respectively. We aggregate images with ships in harbors from the training, validation, and test sets to form the training subset in HRSCOoD. Conversely, images without ships in harbors from the test set of HRSC2016 constitute the test subset in HRSCOoD. As a result, the training subset comprises 436 images with ships in harbors, while the test subset consists of 72 harbor-absent images. Illustrative samples from training and test subsets in HRSCOoD are shown in Figure 3.

RSDDOoD: The initial training dataset comprises 295 inshore images and 4705 offshore images, while the testing dataset

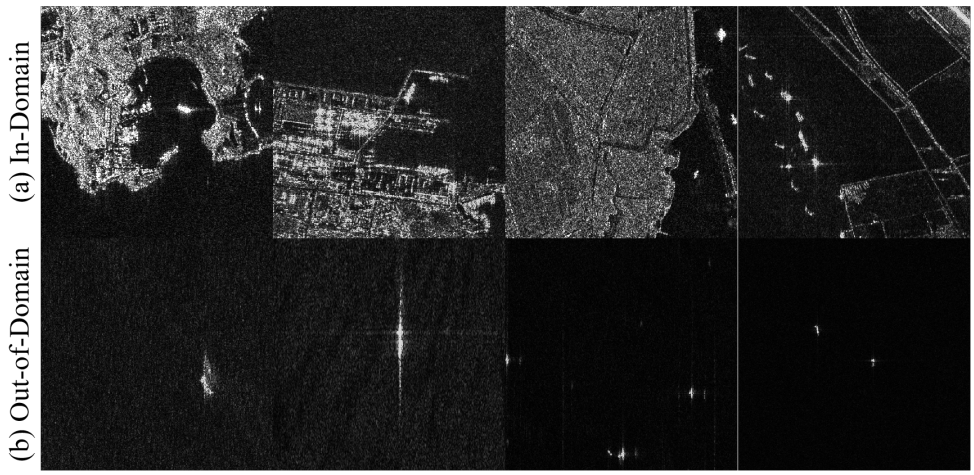

Figure 4. Visualization of training set and test set on RSDDOoD.

includes 159 inshore images and 1841 offshore images. To introduce domain shifts, we merge the original datasets and partition them according to inshore and offshore contexts. Specifically, the inshore images from the original training set constitute the training set (295 images) for the newly created RSDDOoD. Similarly, the inshore images from the original testing set form the in-domain test set (159 images) for RSDDOoD. Additionally, all offshore images from both the original training and testing sets are consolidated to form the new out-of-distribution test set (6546 images) for RSDDOoD. Illustrative images in training subset and test subset in RSDDOoD are shown in Figure 4.

Both HRSCOoD and RSDDOoD involve tasks related to rotated ship detection. However, their acquisition methods differ, as evident in Figure 3 and Figure 4. HRSCOoD employs high-definition aerial images, facilitating relatively clear ship identification. In contrast, RSDDOoD depicts ships as bright spots, posing an identification challenge. Consequently, certain methods like CFA and PSC excel in HRSCOoD but struggle with RSDDOoD. Conversely, methods excelling in RSDDOoD often identify a broad range of bright pixels with high luminance due to ships appearing as bright spots. However, these methods tend to perform poorer on HRSCOoD, which contains diverse vessel types and richer object details, including Rotated RepPoints, Rotated FCOS, R^3det, SASM, Oriented Reppoints, and KFIoU. In contrast, our method consistently demonstrates stable and robust performance across both datasets.

DOTAOoD: In DOTAOoD, we employ a pixel-level K-means clustering algorithm to divide images into two categories, introducing distribution shifts. One category forms the training subset, consisting of 1070 images, while the other forms the test subset, totaling 799 images. Both training and test set images are cropped to 1024×1024 pixels, with a 256-pixel overlap in training samples. This results in 12694 images in the training subset and 9915 images in the test subset. Exemplary images from these subsets are presented in Figure 5. The training set in DOTAOoD is more likely to include images with harbors and black-and-white photos compared to the test set. On the other hand, the test set is expected to feature a higher proportion of images depicting river banks, roundabouts, and airports.

DIORROoD: The construction for DIORROoD is akin to that of DOTAOoD. By utilizing K-means clustering, we partition the images into two categories. The training set comprises 10,279 images, while the test set consists of 13,184 images. Illustrated examples from both the training and test subsets in DIORROoD can be found in Figure 6. In DIORROoD, the training set contains more images featuring storage tanks, complex urban areas and airports, whereas the test set is characterized by a higher occurrence of baseball fields, harbors, and ground track fields.

C. Additional Results

Visualization Results Figure 7 demonstrates additional visualization samples from DOTAOoD. The result distinctly showcases that our method outperforms the baseline in rotated object detection, excelling with a higher detection rate and a reduced occurrence of false positives. Furthermore, our method can adeptly identify complete objects within dense scenes and remain robust against environmental perturbations.

Figure 5. Visualization of training set and test set on DOTAOoD.

Ablation Study: Number of Shots Table 11 presents supplementary results for 16-shot, 32-shot, and full data scenarios for all the methods on HRSCOoD. Performance decline is observed across methods with decreasing shot counts. Remarkably, our method consistently outperforms other comparative approaches across shot variations, highlighting the robust few-shot learning capabilities of VL-Rotate. Notably, even compared to other methods trained with full data, our few-shot learning approach (64 shot) attains the highest mAP, further validating the effectiveness and robustness of our method.

Ablation Study: Channel Size of Convolution in Head We match the RTMDet's channel size with ours, with results shown in Table 10. Notably, our method still showcases a substantial improvement over the baseline across all four datasets upon increasing the channel size from default 256 to 512.

Figure 6. Visualization of training set and test set on DIORROoD.

Table 10. Comparision of the mAP values resulting from the adjustment of channel size in RTMDet-l. The notation "CS" represents the channel size.

METHOD	CS	HRSCOOD (MAP, %)	DOTAOOD (MAP, %)
RTMDET (LYU ET AL., 2022)	256	51.8	49.9
RTMDET (LYU ET AL., 2022)	512	52.3	51.7
RTMDET+VL-ROTATE(OURS)	512	**75.2**	**52.2**
METHOD	**CS**	**RSDDOOD (MAP, %)**	**DIORROOD (MAP, %)**
RTMDET (LYU ET AL., 2022)	256	57.2	52.5
RTMDET (LYU ET AL., 2022)	512	65.6	52.6
RTMDET+VL-ROTATE(OURS)	512	**75.0**	**54.0**

Table 11. Performance comparison of mAP values using different learning shot on HRSCOoD out-of-distribution test set.

METHOD	16-SHOT	32-SHOT	64-SHOT	FULL DATA
ROTATED REPPOINTS (YANG ET AL., 2019)	5.3	23.4	45.9	48.7
ROTATED FCOS (TIAN ET AL., 2019)	7.7	31.2	62.6	65.7
ROI TRANSFORMER (DING ET AL., 2019)	8.0	36.8	67.5	68.4
R^3DET (YANG ET AL., 2021)	4.5	14.4	35.6	56.4
REDET (HAN ET AL., 2021B)	14.3	30.0	70.3	72.1
CFA (ZONGHAO GUO & YE, 2021)	2.0	1.0	39.0	64.3
ORIENTED R-CNN (XIE ET AL., 2021)	9.1	50.4	50.4	52.5
SASM (HOU ET AL., 2022)	9.9	26.5	59.3	62.6
ORIENTED REPPOINTS (WENTONG LI, 2022)	0.7	2.2	27.6	63.4
KFIOU (YANG ET AL., 2023B)	2.6	11.2	31.7	55.0
H2RBOX (YANG ET AL., 2023A)	3.4	19.7	48.0	66.4
PSC (YU & DA, 2023)	13.4	7.1	67.6	67.1
RTMDET (LYU ET AL., 2022)	6.3	20.3	51.8	69.7
VL-ROTATE(OURS)	**14.9**	**51.8**	**75.2**	**80.5**

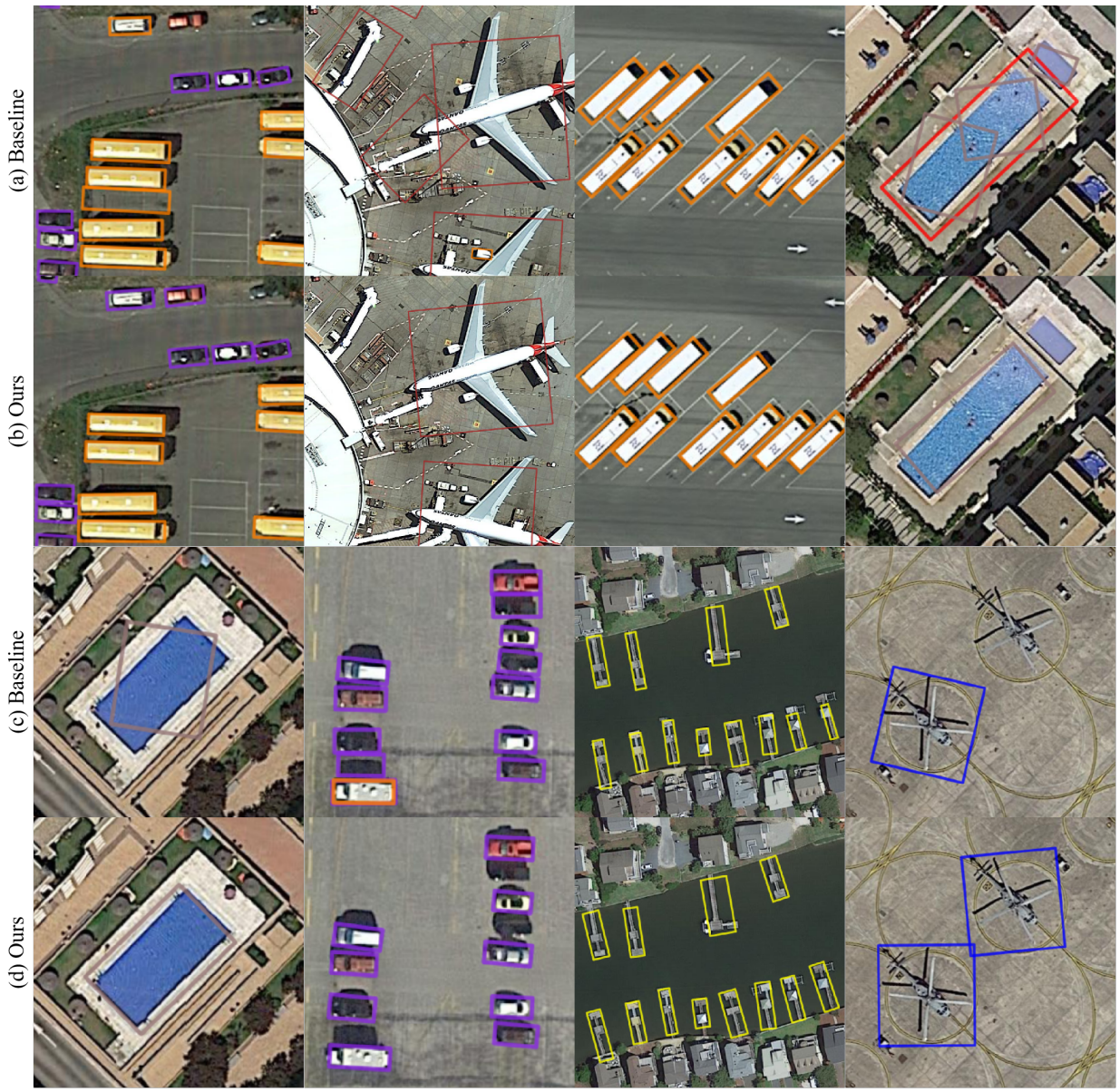

Figure 7. Visualization of training set and test set on DOTAOoD. The purple, orange, red, brown, yellow, and blue rectangles respectively represent the class "small vehicle", "large vehicle", "plane", "swimming pool", "harbor", and "helicopter".