Spatial Transform Decoupling for Oriented Object Detection

Hongtian Yu*, Yunjie Tian*, Qixiang Ye, Yunfan Liu†

University of Chinese Academy of Sciences
{yuhongtian17, tianyunjie19}@mails.ucas.ac.cn, {qxye, liuyunfan}@ucas.ac.cn

Abstract

Vision Transformers (ViTs) have achieved remarkable success in computer vision tasks. However, their potential in rotation-sensitive scenarios has not been fully explored, and this limitation may be inherently attributed to the lack of spatial invariance in the data-forwarding process. In this study, we present a novel approach, termed Spatial Transform Decoupling (STD), providing a simple-yet-effective solution for oriented object detection with ViTs. Built upon stacked ViT blocks, STD utilizes separate network branches to predict the position, size, and angle of bounding boxes, effectively harnessing the spatial transform potential of ViTs in a divide-and-conquer fashion. Moreover, by aggregating cascaded activation masks (CAMs) computed upon the regressed parameters, STD gradually enhances features within regions of interest (RoIs), which complements the self-attention mechanism. Without bells and whistles, STD achieves state-of-the-art performance on the benchmark datasets including DOTA-v1.0 (82.24% mAP) and HRSC2016 (98.55% mAP), which demonstrates the effectiveness of the proposed method. Source code is available at https://github.com/yuhongtian17/Spatial-Transform-Decoupling.

Introduction

Recent years have witnessed substantial progress and notable breakthroughs in computer vision, which can be primarily attributed to the advent of Vision Transformer (ViT) models. Benefiting from the powerful self-attention mechanism, ViTs consistently achieve new state-of-the-art performance across vision tasks including classification (Dosovitskiy et al. 2020; Liu et al. 2021; Zhang et al. 2022c; Fang et al. 2023), object detection (Li et al. 2022b; Fang et al. 2022; Tian et al. 2023), and semantic segmentation (Xie et al. 2021a; Yuan et al. 2021). Despite the progress made, the capability of ViTs in spatial transform invariance has not been fully explored and understood. In many scenarios, ViTs are treated as a universal approximator, expected to automatically handle various vision data irrespective of their orientations and appearances.

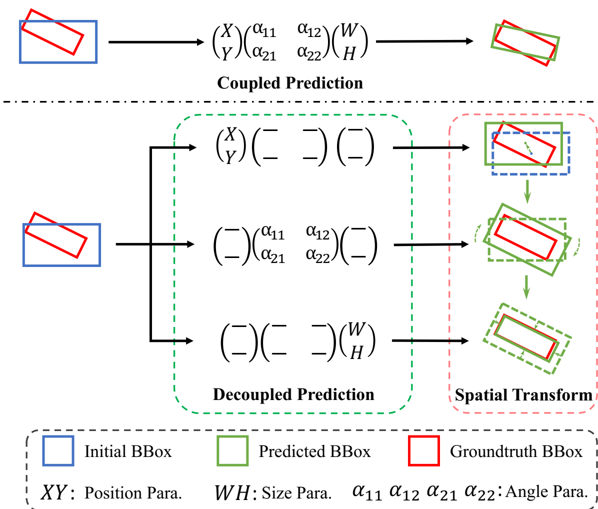


Figure 1: Conventional approaches (upper) estimate the position, size, and angle using a single RoI feature. In contrast, STD (lower) predicts and refines the parameters of bounding boxes in a divide-and-conquer (decoupled) manner.

In this study, we aim to tap into the potential of ViTs in tackling the challenging spatial transform issue of vision tasks, *e.g.*, detecting objects in remote sensing scenarios, where images are captured from a bird's-eye view and target objects may appear in arbitrary orientations. To determine an oriented bounding box, initial research efforts (Ren et al. 2015; Lin et al. 2017b) suggested a direct regression approach for spatial transform parameters, including the spatial coordinates (x and y), object size (w and h), and the angle (α). However, such a straightforward regression strategy often results in discontinuous boundaries due to the inconsistency in angle representation and periodicity, as well as the suboptimal design of loss functions (Yang and Yan 2020; Yang et al. 2021b,c, 2022).

Rather than solely concentrating on developing more sophisticated angle representations or refining training objectives, it is essential to tackle the foundational issue of effectively extracting rotation-related features. In particular, we enhance the conventional structure of the bounding box prediction head by allocating distinct feature maps to predict

*Equal contribution.
†Corresponding author.

parameters associated with diverse semantic interpretations, such as the object's location, shape, and orientation. This approach guides the feature extraction process in a controlled and effective manner. Furthermore, by estimating the parameters associated with a particular spatial transform at each stage, this step-wise strategy facilitates the progressive refinement of estimation results, which in turn can contribute to improving the overall accuracy of the model.

Building upon the insights and discussions presented earlier, we propose Spatial Transform Decoupling (STD), a straightforward yet effective solution for oriented object detection, which decouples the estimation of transformation parameters related to object positions, sizes, and angles, Fig. 1. Concretely, a multi-branch network design is utilized, where each individual branch is designated to predict parameters that correspond to distinct spatial transforms. From another perspective, STD supplements the self-attention mechanism by allocating distinct responsibilities to self-attention modules at different stages of parameter prediction, which effectively utilizes the spatial transform capabilities of ViTs in a divide-and-conquer fashion. Furthermore, STD integrates cascaded activation masks (CAMs) to enhance the features extracted by stacked Transformer blocks, effectively suppressing background information while highlighting foreground objects. By refining features within regions of interest (RoIs) using CAMs, the feature representation for oriented objects is both decoupled and progressively enhanced. As a simple-yet-effective design, STD can be integrated with various ViT-based detectors and achieve significant performance improvements over the state-of-the-art methods. For instance, STD achieves 82.24% mAP on DOTA-v1.0 and 98.55% mAP on HRSC2016, surpassing the accuracy of all existing detectors.

The contributions of this work are summarized as:

- The Spatial Transform Decoupling (STD) approach is introduced to address the challenge of oriented object detection by estimating parameters for spatial transforms through separate network branches. STD demonstrates remarkable generalizability and can seamlessly integrate with a variety of ViT detectors.
- Cascade activation masks (CAMs) are integrated into the self-attention module at each layer of ViT to progressively enhance the features. CAMs offer spatially dense guidance, directing the attention maps to focus more on foreground objects rather than the background.
- Experimental results demonstrate that STD surpasses state-of-the-art methods by a significant margin across a variety of oriented object detection benchmarks.

Related Work

Oriented Object Detection

Existing methods have investigated oriented object detection from the perspectives of feature robustness, region proposal refinement, and target regression enhancement.

Feature Invariance/Equivalence. Invariance or equivalence is an essential problem when designing/learning visual feature representations. During the era of hand-crafted features, SIFT (Lowe 1999) utilizes dominant orientation-based feature alignment to achieve invariance to rotation and robustness to moderate perspective transforms. With the rise of CNNs, STN (Jaderberg et al. 2015) achieves rotation invariance by manipulating the feature maps according to the transformation matrix estimated using a sub-CNN. Group equivariant CNN (Cohen and Welling 2016) proposes a natural generalization of CNNs, enabling them to group objects from the same categories regardless of orientations. ORN (Zhou et al. 2017) introduces Active Rotating Filters (ARFs), which dynamically rotate during the convolution process and thereby produce feature maps with location and orientation explicitly encoded. ReDet (Han et al. 2021) achieves rotation-equivariant convolution (e2cnn (Weiler and Cesa 2019)) by incorporating a rotation-invariant backbone, which normalizes the spatial and orientational information of features.

Region Proposal Refinement. RoI Transformer (Ding et al. 2019) enhances two-stage detectors by iteratively repeating the RPN-RoI head structure (Ren et al. 2015; He et al. 2017). Oriented RCNN (Xie et al. 2021b) streamlines the process of oriented proposal generation and directly predicts oriented proposals based on the features extracted by the backbone and FPN (Feature Pyramid Network) (Lin et al. 2017a) module. Drawing inspiration from a similar concept, R^3Det (Yang et al. 2021a) introduces a feature refinement stage to the orientation regression head.

Target Regression Enhancement. Gliding Vertex (Xu et al. 2020) converts the task of rotated box prediction into regressing the offset for horizontal boxes along the four edges. CSL (Yang and Yan 2020) addresses the potential abrupt change in loss computation by proposing a label-based solution for angle prediction. CFA (Guo et al. 2021) and Oriented RepPoints (Li et al. 2022a) make improvements to the nine-point prediction methods (Yang et al. 2019). GWD (Yang et al. 2021b), KLD (Yang et al. 2021c), and KFIoU (Yang et al. 2022) use two-dimensional Gaussian distributions to solve the angle prediction problem. Despite the progress of various approaches proposed, few of them explore the impact of decoupling spatial transform, $e.g.$, position (x, y), size (w, h), and angle (α), on the hierarchical feature representation.

Vision Transformer

Drawing inspiration from the NLP field (Vaswani et al. 2017; Devlin et al. 2018), ViTs divide the image into multiple patch tokens for feature extraction and processing (Dosovitskiy et al. 2020; Liu et al. 2021; Zhang et al. 2022c; Tian et al. 2023). It has attracted significant attention in recent years owing to its remarkable success in computer vision tasks. DETR (Carion et al. 2020) is a representative work that extends ViTs towards object detection, establishing the fundamental paradigm for applying ViT to this task. MAE (He et al. 2022) proposes a novel pre-training mode that deviates from the classic fully supervised pre-training era of CNNs (He, Girshick, and Dollár 2019). Building upon MAE, ViTDet (Li et al. 2022b) and MIMDet (Fang et al. 2022), *etc*, have made significant advancements in the development of ViT for object detection.

While Vision Transformers have demonstrated promis-

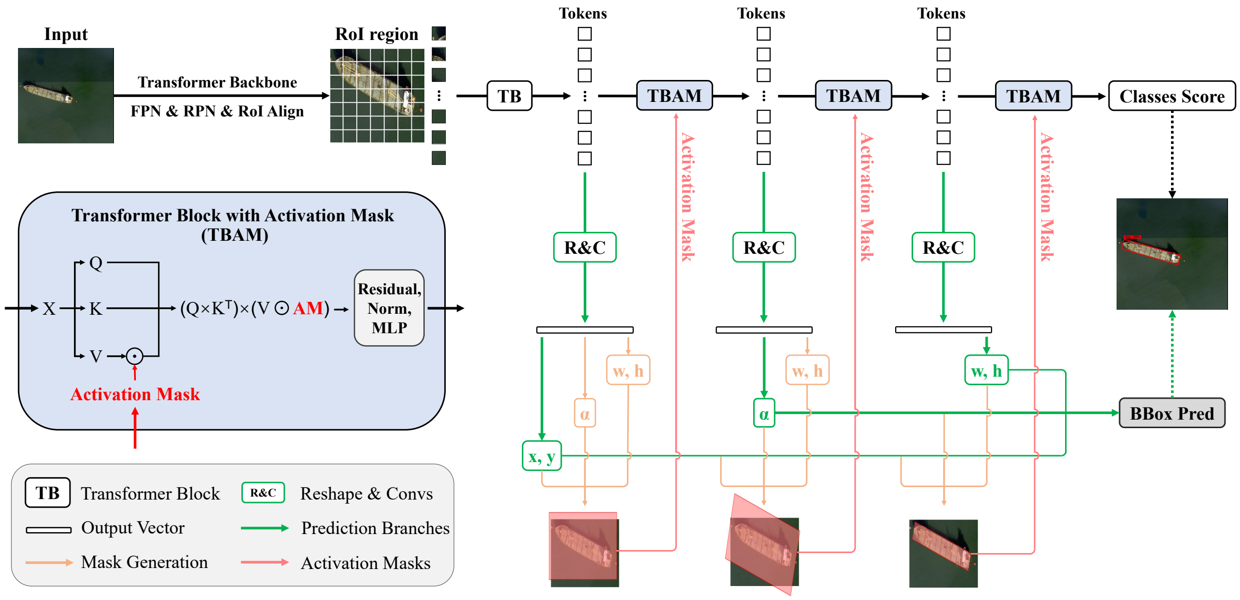


Figure 2: The framework of the proposed Spatial Transform Decoupling (STD) method. The detailed structure of Transformer blocks integrated with activation masks (TBAM) is shown on the left.

ing results in various visual tasks, they still encounter challenges in leveraging their advantages in handling object spatial transform, *e.g.*, oriented object detection. Recently, RVSA (Zhang et al. 2022a; Wang et al. 2022) made an initial attempt to improve the structure of ViT for oriented object detection tasks, which was achieved by updating Window Attention (Liu et al. 2021; Li et al. 2022b; Fang et al. 2022) to Rotated Varied Size Attention. Nevertheless, these methods solely rely on the self-attention mechanism to handle various spatial transformations, without explicitly introducing dense guiding information.

The Proposed Method

This section starts with an elucidation of the motivation behind Spatial Transform Decoupling (STD). Subsequently, a detailed explanation of the overall structure of STD is provided, offering an in-depth understanding of its architectural design and how it functions. Next, we delve into a detailed decoupling structure and introduce the cascaded activation masks (CAMs) for progressive feature refinement. Special emphasis is placed on their significant contribution to the overall performance enhancement of STD.

Overview

The proposed STD can be readily seen as an extension of existing oriented object detectors, and an overview of the architecture is depicted in Figure 2. The primary innovation of STD resides within the detection head module, while for other components, such as the backbone, Region Proposal Network (RPN), and loss functions, we maintain consistency with mainstream detection frameworks (Ren et al. 2015; Xie et al. 2021b). As a result, STD demonstrates significant generalizability, enabling its compatibility with a variety of detectors. Specifically, for the purpose of a clear explanation, we adopt STD within the Faster RCNN framework (Ren et al. 2015) as the default configuration. Throughout the experiments, we will also showcase the performance of STD in combination with other detectors, such as Oriented RCNN (Xie et al. 2021b).

	2FCBBoxHead	MAEBBoxHead (Not Pre-trained)	MAEBBoxHead (Pre-trained)
mAP	69.67	69.16	71.07

Table 1: Performance comparison of Faster RCNN with the same backbone (ViT-small) but different heads. The training is carried out on the DOTA-v1.0 dataset (Xia et al. 2018) for 12 epochs.

ViTs have demonstrated impressive performance across a broad spectrum of visual tasks. However, their utilization in the context of oriented object detection remains relatively unexplored. Nevertheless, existing pre-trained Transformer models are capable of extracting meaningful features, which contributes to establishing a strong foundation for achieving impressive performance in oriented object detection tasks. Therefore, we adopt a design inspired by the imTED (Zhang et al. 2022b) detector and substitute the backbone as well as head modules of the two-stage detector with **Vision Transformer blocks** pre-trained using the MAE method.

Specifically, we employ the ViT-small model as the backbone instead of ResNet-50, and use a 4-layer Transformer block to replace the conventional detection head in Faster

RCNN built with fully connected (FC) layers. Please note that the ViT-small backbone is obtained from the MAE pre-trained encoder, and the 4-layer Transformer block is derived from the pre-trained decoder, which forms the MAEBBoxHead module. Once the regions of interest (RoIs) are obtained, the feature maps are uniformly divided into 7×7 tokens, which are subsequently fed into the parameter regression head, as depicted in Figure 2. Experiments are conducted to validate the effectiveness of this framework in addressing the oriented object detection problem, and the results are presented in Table 1. In subsequent experiments, the pre-trained MAEBBoxHead is used as the baseline method by default.

Afterward, the proposed Spatial Transform Decoupling (STD) module is built upon the aforementioned backbone network. To enhance the performance of decoupling, we employ a hierarchical structure to predict the bounding box parameters in a layer-wise manner, and further enhance it by leveraging the guidance provided by the cascaded activation masks (CAMs). Detailed explanations of these contributions will be provided in the following two subsections.

Decoupled Parameter Prediction

As highlighted in the Introduction Section, different parameters of an oriented bounding box are expected to possess distinct properties (e.g., rotation-variance or rotation-invariance), and therefore, they should be computed based on different feature maps. However, most conventional methods (Ren et al. 2015; Lin et al. 2017b) depend on a single feature map to predict all bounding box parameters, potentially resulting in the issue of coupled features. To solve this problem, we introduce a multi-branch network to achieve hierarchical and disentangled parameter prediction.

As shown in Figure 2, we compute different components of the oriented bounding box based on the feature map at various stages of the Transformer decoder in a cascaded manner. Specifically, $\{x, y\}$, α, $\{w, h\}$ and the class score are obtained based on the feature maps of the $1st$, $2nd$, $3rd$, and $4th$ layer of the Transformer block, respectively (the rationale behind this design will be detailed in ablation study). After obtaining the discrete output from each Transformer block, we first reshape them into 7×7 feature maps and then apply convolutional layers to further enhance the features. Next, after globally averaging the resultant feature maps, FC layers are adopted to make the final predictions, which are then used to produce the bounding box and CAMs (details explained in the next subsection). Please note that the proposed mechanism is highly generalizable, as one can easily adjust the number of estimated parameters by simply adding or removing predicting branches.

Cascaded Activation Masks

To further regulate the decoupling process and improve the accuracy of prediction results, we intend to provide dense guidance for bounding box prediction at each stage. To achieve this goal, cascaded activation masks (CAMs) are introduced to enhance the features generated by the multiple branches.

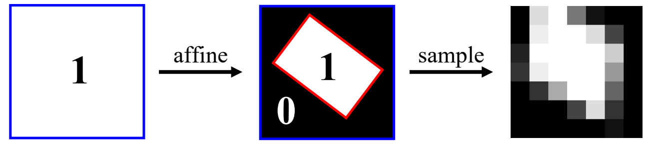


Figure 3: The translation between the predicted bounding box and the activation mask after affine transformation. The blue box represents the proposal region and the red box represents the activation mask.

An ideal activation mask with binary values should have the regions corresponding to foreground objects assigned with a value of 1, and all background locations set to 0. To align the activated regions with the foreground area as much as possible, we propose to generate activation masks by incorporating information from both the proposal and the predicted bounding box.

To be specific, the center point, size, and orientation of the estimated bounding box, $i.e.$, $(x_b, y_b, w_b, h_b, \alpha_b)$, could be expressed as

$$\begin{aligned} x_b &= x_p + w_p \cdot dx, \\ y_b &= y_p + h_p \cdot dy, \\ w_b &= w_p \cdot e^{dw}, \\ h_b &= h_p \cdot e^{dh}, \\ \alpha_b &= d\alpha \end{aligned} \tag{1}$$

where (x_p, y_p) and (w_p, h_p) respectively denote the center coordinates and shape of the proposal, and $(dx, dy, dw, dh, d\alpha)$ are the predicted values related to the oriented bounding box obtained from STD. Then, with the proposal placed in a rectangular coordinate system (x, y) and its four vertices located at $(-1, -1)$, $(1, -1)$, $(1, 1)$, and $(-1, 1)$, the affine transformation against the bounding box $(x^{'}, y^{'})$ could be formulated as:

$$\begin{pmatrix} x^{'} \\ y^{'} \end{pmatrix} = \begin{pmatrix} \cos d\alpha \cdot e^{dw} & -\sin d\alpha \cdot e^{dh} \cdot \frac{h_p}{w_p} \\ \sin d\alpha \cdot e^{dw} \cdot \frac{w_p}{h_p} & \cos d\alpha \cdot e^{dh} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2 \cdot dx \\ 2 \cdot dy \end{pmatrix} \tag{2}$$

As illustrated in Figure 3, the activation mask could be produced by applying the affine transformation in Eq.(2) to a matrix $\boldsymbol{AM}$ with all elements set to 1. After integrating $\boldsymbol{AM}$ into the self-attention module in STD, the mapping function implemented by **Transformer Block with Activation Mask** (TBAM, see Figure 2) could be written as

$$\mathrm{TBAM}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{AM}) = \mathrm{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}})\boldsymbol{V'} \tag{3}$$

where $\boldsymbol{V'}$ is obtained by performing an element-wise multiplication between $\boldsymbol{V}$ and $\boldsymbol{AM}$ ($\boldsymbol{V'} = \boldsymbol{V} \odot \boldsymbol{AM}$). By multiplying with $\boldsymbol{V}$, $\boldsymbol{AM}$ could direct the model's attention by highlighting the foreground while suppressing the background. In the forward propagation process, the utilization of activation maps guides the decoupled predicted values in earlier stages to direct the self-attention mechanism of the subsequent Transformer blocks; while during the backward propagation process, the discrepancies in the decoupled predicted values from later stages are propagated through the activation maps, affecting the feature extraction process of the previously decoupled predicted values. This cascaded architecture enhances the interconnection between decoupled predicted values at various levels.

Experiment

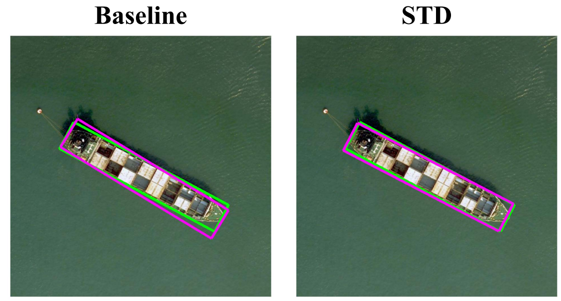
Experimental Setting

Datasets Experiments are conducted on two commonly-used datasets for oriented object detection, namely DOTA-v1.0 (Xia et al. 2018) and HRSC2016 (Liu et al. 2017). **DOTA-v1.0** is a large-scale object detection dataset for optical remote sensing images, which comprises 2,806 images with diverse dimensions, spanning from 800 to 4,000 pixels in width and height. The dataset consists of a total of 188,282 individual instances distributed across 15 different classes, and it is partitioned into training, validation, and test sets containing 1,411, 458, and 937 images, respectively. **HRSC2016** is an optical remote sensing image dataset designed for ship detection. It comprises 1,680 images with diverse widths and heights, ranging from 339 pixels to 1333 pixels. The commonly used training, validation, and test sets consist of 436, 181, and 444 images, respectively.

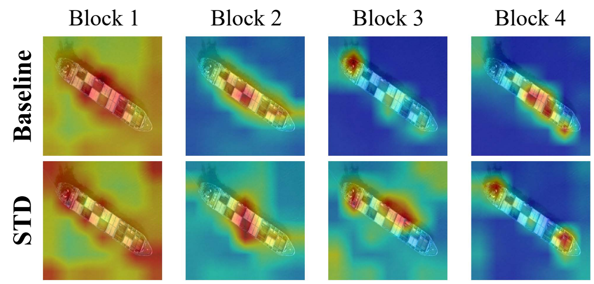
Implementation Details The experimental results are obtained on the MMRotate platform (Zhou et al. 2022). We employ the checkpoints of ViT-small/-base (Dosovitskiy et al. 2020) and HiViT-base (Zhang et al. 2022c), which are all pre-trained using the MAE (He et al. 2022) self-supervised strategy. We pre-train the ViT-small model and directly utilize the open-sourced checkpoints for the other models, wherein all the Transformer blocks of both the encoder and decoder are fully inherited.

For a fair comparison, we adopt a similar experimental configuration as used in the benchmark methods (Wang et al. 2022; Yang et al. 2022; Xie et al. 2021b). The performance evaluation on DOTA-v1.0 follows a *multi-scale* setting, where the model is trained on the trainval-set and tested on the test-set. In contrast, for the ablation study on DOTA-v1.0, a *single-scale* setting is adopted, where the model is trained on the train-set and tested on the val-set. All images would be cropped into patches of size 1024×1024 with an overlap of 500/200 pixels in *multi-scale*/*single-scale* setting. In the *multi-scale* setting, images are resized by $0.5\times$, $1.0\times$, and $1.5\times$ before undergoing the cropping process, and no scale adjustment is adopted in the *single-scale* setting. In the HRSC2016 dataset, images are resized in such a way that the larger dimension of width and height becomes 800, while maintaining their original aspect ratios.

During training, horizontal/vertical flipping and random rotation operations are conducted to increase the scale and diversity of training data. The model is trained for 12 epochs on DOTA-v1.0 and 36 epochs on HRSC2016. We adopt the AdamW optimizer (Kingma and Ba 2014) with an initial learning rate of $1e^{-4}$/$2.5e^{-4}$ for DOTA-v1.0/HRSC2016, a weight decay of 0.05, and a layer decay of 0.75/0.90 for ViT/HiViT. All experiments are conducted on $8\times$A100 GPUs with a batch size of 8.



(a) Prediction Results Comparison



(b) Attention Maps Comparison

Figure 4: Visualization of attention maps. Compare to the baseline Transformer, the attention maps in STD (bk1 to bk4) exhibit a stronger alignment with the semantic interpretation of the parameter estimated at the respective stage.

Ablation Study

Feasibility of Decoupled Parameter Prediction Prior to assessing the feasibility of the decoupling approach, we first investigate the performance of bounding box prediction relying solely on a single feature map in various levels. As shown in Table 2a, a consistent enhancement in performance is evident when employing feature maps from deeper layers for bounding box prediction. This observation suggests that while deep feature maps contribute to improved feature representations for object detection, shallower layers still contain valuable information for bounding box prediction, as their performance is only slightly lower.

We also compare the performance with the decoupled parameter estimation approach. As shown in Fig. 2, we used the feature maps from the first block to predict x and y, the second block for α, the third block for w and h, and the final stage for the class score cls. Without the aid of CAMs, the performance of this decoupled configuration is slightly lower than predicting bounding boxes with the feature maps from the third/fourth block by a margin of 0.09%/0.28% mAP. These results provide evidence that suggests the feasibility and potential of designing a decoupled structure. As previously mentioned, we introduce CAMs to enhance the decoupling process, further reducing the performance gap between decoupled and non-decoupled approaches.

Rationality of Model Design To showcase the rationale behind the detailed architecture of STD, we investigate the impact of both the order of parameter decoupling and the lo-

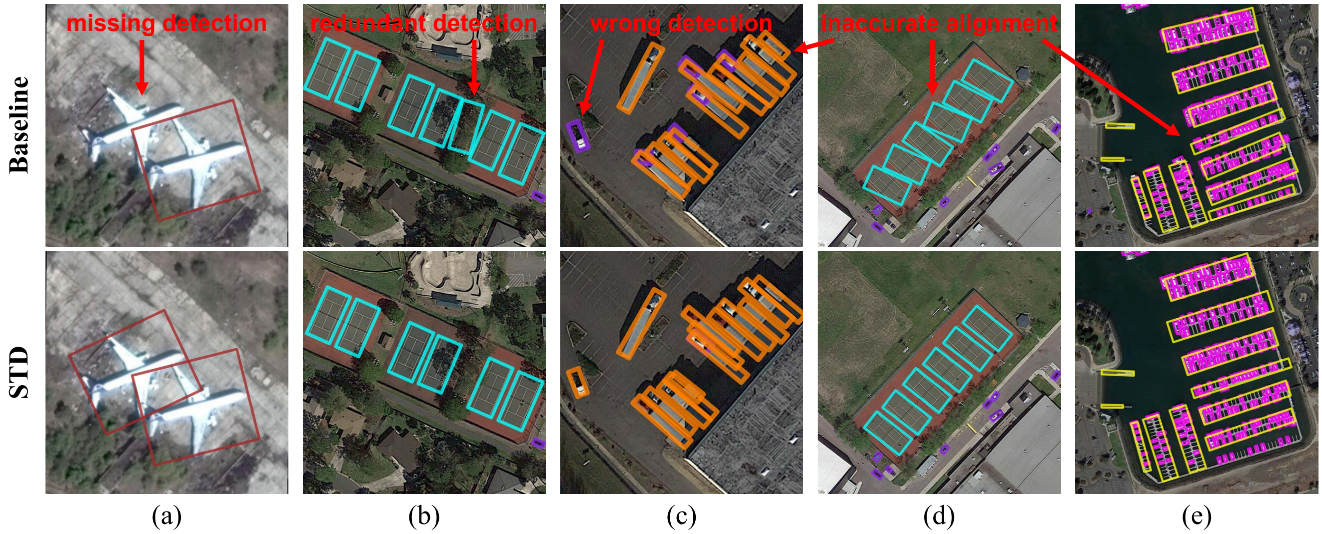


Figure 5: Comparison of detection results. STD demonstrates superior performance in reducing false detections ((a), (b), and (c)), better discerning clustered objects ((c) and (e)), and improving the alignment with oriented objects ((c), (d), and (e)).

Feature Levels				mAP	
bk1	bk2	bk3	bk4	-	+CAMs
□	-	-	♢	69.65	71.85
-	□	-	♢	70.37	72.39
-	-	□	♢	70.88	72.27
-	-	-	□♢	71.07†	72.21
(xy)	α	(wh)	♢	70.79	**72.78**‡

(a) Feature Levels

Decoupling Order			mAP
(xy)	(wh)	α	72.30
(wh)	(xy)	α	72.26
(xy)	α	(wh)	**72.78**
(wh)	α	(xy)	72.13
α	(xy)	(wh)	72.03
α	(wh)	(xy)	71.93

(b) Decoupling Order

Table 2: Results of diagnostic studies. (a) Comparison of detection accuracy achieved using feature maps from various levels of Transformer block (bk1 to bk4). □ denotes coupled bounding box prediction and ♢ refers to class score estimation. † indicates the performance of original MAEBBoxHead while ‡ indicates our STD's. (b) The influence of decoupling order on the overall performance of STD.

Detector	Model	BBoxHead	mAP
Faster RCNN	ViT-S	MAEBBoxHead	71.07
Faster RCNN	ViT-S	STD	72.78
Oriented RCNN	ViT-S	MAEBBoxHead	72.41
Oriented RCNN	ViT-S	STD	73.43

Table 3: Comparison of object detection accuracy achieved by different RoI extraction networks.

cation of activation mask integration. As shown in Table 2b, the order of decoupling has a significant influence on the performance of STD, and the optimal result is achieved under the configuration $\{x, y\} \rightarrow \alpha \rightarrow \{w, h\}$. This phenomenon can be explained by the fact that alternative prediction orders fail to ensure that the RoI could consistently cover the entire foreground object.

Adaptability to Different Backbones As discussed in the preceding section, as long as the RoI fully covers the foreground object, the activation masks can effectively activate the entire foreground region. Hence, our approach is expected to be adaptable to other RoI extraction methodologies. As indicated in Table 3, the decoupling module of STD also demonstrates strong performance when incorporated into the Oriented RCNN object detector (Xie et al. 2021b), which showcases the remarkable generalizability of our method.

Visualization of Attention Maps In Figure 4, we present visualizations of the attention maps from different decoder layers of STD and a baseline Transformer model (Rotated Faster RCNN+ViT-S). In comparison to the baseline Transformer, the attention maps generated by the STD model at each stage exhibit a closer alignment with the semantic meaning of the corresponding predicted parameter. Specifically, when obtaining the positional information x, y, the attention tends to concentrate around the center of the object. Following this, the attention becomes more widespread, targeting one end and one edge of the object to capture information about its orientation α. Finally, the attention predominantly focuses on both ends of the object, aiming to capture details related to its scale. This phenomenon is likely a result of the decoupled bounding box prediction mechanism and the step-wise guidance provided by the activation masks, further confirming the effectiveness of the proposed architectural approach.

Qualitative Comparison We also present a qualitative comparison between the results of STD and the baseline Transformer in Figure 5. STD is capable of mitigating the occurrence of false negatives/positives (as depicted in Figure 5(a), (b)), while also achieving notably improved alignment with oriented foreground objects across different scales (as shown in Figure 5(c), (d), (e)). This observation

Method	Model	Pre.	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
CSL	R152	IN	**90.25**	85.53	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93	76.17
R^3Det	R152	IN	89.80	83.77	48.11	66.77	78.76	83.27	87.84	90.82	85.38	85.51	65.67	62.68	67.53	78.56	72.62	76.47
SASM	RX101	IN	89.54	85.94	57.73	78.41	79.78	84.19	**89.25**	90.87	58.80	87.27	63.82	67.81	78.67	79.35	69.37	79.17
ReDet	ReR50	IN	88.81	82.48	60.83	80.82	78.34	86.06	88.31	90.87	88.77	87.03	68.65	66.90	79.26	79.71	74.67	80.10
GWD	R152	IN	89.66	84.99	59.26	82.19	78.97	84.83	87.70	90.21	86.54	86.85	**73.47**	67.77	76.92	79.22	74.92	80.23
DEA	ReR50	IN	89.92	83.84	59.65	79.88	80.11	**87.96**	88.17	90.31	**88.93**	**88.46**	68.93	65.94	78.04	79.69	75.78	80.37
KLD	R152	IN	89.92	85.13	59.19	81.33	78.82	84.38	87.50	89.80	87.33	87.00	72.57	71.35	77.12	79.34	**78.68**	80.63
O-RCNN	R50	IN	89.84	85.43	61.09	79.82	79.71	85.35	88.82	90.88	86.68	87.73	72.21	70.80	82.42	78.18	74.11	80.87
KFIoU	Swin-T	IN	89.44	84.41	**62.22**	**82.51**	80.10	86.07	88.68	**90.90**	87.32	88.38	72.80	71.95	78.96	74.95	75.27	80.93
RVSA-O	ViTAE-B	$\mathcal{M}^\dagger$	88.97	85.76	61.46	81.27	79.98	85.31	88.30	90.84	85.06	87.50	66.77	**73.11**	**84.75**	**81.88**	77.58	81.24
RTMDet-R	RTM-L	CO	88.01	**86.17**	58.54	82.44	**81.30**	84.82	88.71	90.89	88.77	87.37	71.96	71.18	81.23	81.40	77.13	**81.33**
	Swin-B	IN	89.26	86.48	62.09	82.86	79.97	85.64	88.47	90.70	86.69	87.54	71.84	68.74	79.62	81.11	76.64	81.18
KFIoU	ViT-B	$\mathcal{M}$	89.32	84.52	62.53	81.86	81.55	86.53	89.00	90.76	87.67	88.29	67.31	71.75	79.63	80.25	72.25	80.88
	HiViT-B	$\mathcal{M}$	88.71	86.02	62.45	80.24	80.84	85.19	88.47	90.70	86.64	86.37	67.41	74.63	79.04	80.80	82.28	81.32
RVSA-O	ViT-B	$\mathcal{M}^\dagger$	87.63	85.23	61.73	81.11	80.68	85.37	88.26	90.80	86.38	87.21	67.93	69.81	84.06	81.25	77.76	81.01
STD-O (Ours)	ViT-B	$\mathcal{M}$	88.56	84.53	62.08	81.80	81.06	85.06	88.43	90.59	86.84	86.95	72.13	71.54	84.30	82.05	78.94	81.66
STD-O (Ours)	HiViT-B	$\mathcal{M}$	89.15	85.03	60.79	82.06	80.90	85.76	88.45	90.83	87.71	87.29	**<u>73.99</u>**	71.25	**<u>85.18</u>**	**<u>82.17</u>**	**<u>82.95</u>**	**<u>82.24</u>**

Table 4: Performance comparison on DOTA-v1.0. Classes: PL-plane; BD-baseball diamond; BR-bridge; GTF-ground track field; SV-small vehicle; LV-large vehicle; SH-ship; TC-tennis court; BC-baseball court; ST-storage tank; SBF-soccer ball field; RA-roundabout; HA-harbor; SP-swimming pool; HC-helicopter. Pretraining: IN-supervised pretraining on the ImageNet; CO-supervised pretraining on the MS COCO; $\mathcal{M}$-MAE self-supervised pretraining on the ImageNet; $\mathcal{M}^\dagger$-MAE self-supervised pretraining on the MillionAID (Long et al. 2021), a large remote sensing dataset including about 1 million images.

highlights the capability of STD in effectively modeling the object orientations without compromising the precision in capturing spatial location and shape information.

Performance Comparison

In this section, we present comprehensive experimental results obtained on the DOTA-v1.0 and HRSC2016 datasets.

DOTA-v1.0 Table 4 provides a comprehensive comparison of our method with state-of-the-art approaches on DOTA-v1.0, including CSL (Yang and Yan 2020), R^3Det (Yang et al. 2021a), SASM (Hou et al. 2022), ReDet (Han et al. 2021), GWD (Yang et al. 2021b), DEA (Liang et al. 2022), KLD (Yang et al. 2021c), Oriented RCNN (O-RCNN) (Xie et al. 2021b), KFIoU (Yang et al. 2022), RVSA (Wang et al. 2022), and RTMDet (Lyu et al. 2022). We evaluate STD within Oriented RCNN frameworks (STD-O) and both ViT and HiViT models are used for evaluations. Remarkably, STD achieves new state-of-the-art performance in both frameworks. When coupled with ViT-B and HiViT-B backbones, STD achieves 81.66% and 82.24% mAP, respectively, surpassing the previous best results.

HRSC2016 In Table 5, the experimental results demonstrate that the STD method outperforms all other methods, achieving an impressive mAP of 90.67% and 98.55% under the PASCAL VOC 2007 (Everingham et al. 2010) and VOC 2012 metrics, respectively.

Computational Cost. STD has a small computational cost overhead. While the baseline detector with HiViT-B takes around 288ms on average to process an image using A100 GPUs, the overall STD detector requires an average processing time of 315ms.

Method	Model	Pre.	mAP(07)	mAP(12)
CenterMap (Wang et al. 2021)	R101	IN	-	92.70
RoI Trans. (Ding et al. 2019)	R101	IN	86.20	-
Gilding Vertex (Xu et al. 2020)	R101	IN	88.20	-
R^3Det (Yang et al. 2021a)	R101	IN	89.26	96.01
DAL (Ming et al. 2021)	R101	IN	89.77	-
GWD (Yang et al. 2021b)	R101	IN	89.85	97.37
S^2ANet (Han et al. 2022)	R101	IN	90.17	95.01
AOPG (Cheng et al. 2022)	R101	IN	90.34	96.22
ReDet (Han et al. 2021)	ReR50	IN	90.46	97.63
O-RCNN (Xie et al. 2021b)	R101	IN	90.50	97.60
RTMDet-R (Lyu et al. 2022)	RTM-L	CO	90.60	97.10
STD-O (ours)	ViT-B	$\mathcal{M}$	**<u>90.67</u>**	**<u>98.55</u>**
STD-O (ours)	HiViT-B	$\mathcal{M}$	90.63	98.20

Table 5: Comparison of performance on HRSC2016.

Conclusion

This paper introduces Spatial Transform Decoupling (STD), an oriented object detection method that separates the parameter prediction process into multiple disentangled stages. Such a decoupled process is further enhanced by incorporating cascaded activation masks, which introduce dense guidance into the self-attention mechanism. Extensive experiments have demonstrated the effectiveness of STD on multiple popular benchmarks. To the best of our knowledge, STD is a pioneering method that tackles oriented object detection in remote sensing with a structural perspective. Notably, the Transformer-based nature of STD enables seamless integration with various advanced pre-trained models, providing significant benefits to the research community.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) under Grant 62225208 and 62171431, and by China Postdoctoral Science Foundation under Grant 2023M743442.

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229.

Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; and Han, J. 2022. Anchor-free oriented proposal generator for object detection. *TGRS*, 60: 1–11.

Cohen, T.; and Welling, M. 2016. Group equivariant convolutional networks. In *ICML*, 2990–2999.

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

Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; and Lu, Q. 2019. Learning RoI transformer for oriented object detection in aerial images. In *CVPR*, 2849–2858.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *IJCV*, 88(2): 303–338.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 19358–19369.

Fang, Y.; Yang, S.; Wang, S.; Ge, Y.; Shan, Y.; and Wang, X. 2022. Unleashing vanilla vision transformer with masked image modeling for object detection. *arXiv preprint arXiv:2204.02964*.

Guo, Z.; Liu, C.; Zhang, X.; Jiao, J.; Ji, X.; and Ye, Q. 2021. Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In *CVPR*, 8792–8801.

Han, J.; Ding, J.; Li, J.; and Xia, G.-S. 2022. Align deep features for oriented object detection. *TGRS*, 60: 1–11.

Han, J.; Ding, J.; Xue, N.; and Xia, G.-S. 2021. Redet: A rotation-equivariant detector for aerial object detection. In *CVPR*, 2786–2795.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *CVPR*, 16000–16009.

He, K.; Girshick, R.; and Dollár, P. 2019. Rethinking imagenet pre-training. In *ICCV*, 4918–4927.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*, 2961–2969.

Hou, L.; Lu, K.; Xue, J.; and Li, Y. 2022. Shape-adaptive selection and measurement for oriented object detection. In *AAAI*, volume 36, 923–932.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. *NeurIPS*, 28.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, W.; Chen, Y.; Hu, K.; and Zhu, J. 2022a. Oriented reppoints for aerial object detection. In *CVPR*, 1829–1838.

Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022b. Exploring plain vision transformer backbones for object detection. In *ECCV*, 280–296.

Liang, D.; Geng, Q.; Wei, Z.; Vorontsov, D. A.; Kim, E. L.; Wei, M.; and Zhou, H. 2022. Anchor retouching via model interaction for robust object detection in aerial images. *TGRS*, 60: 1–13.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *CVPR*, 2117–2125.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017b. Focal loss for dense object detection. In *ICCV*, 2980–2988.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 10012–10022.

Liu, Z.; Yuan, L.; Weng, L.; and Yang, Y. 2017. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *ICPRAM*, volume 2, 324–331.

Long, Y.; Xia, G.-S.; Li, S.; Yang, W.; Yang, M. Y.; Zhu, X. X.; Zhang, L.; and Li, D. 2021. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 14: 4205–4230.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *ICCV*, 1150–1157.

Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; and Chen, K. 2022. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*.

Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; and Li, L. 2021. Dynamic anchor learning for arbitrary-oriented object detection. In *AAAI*, volume 35, 2355–2363.

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

Tian, Y.; Xie, L.; Wang, Z.; Wei, L.; Zhang, X.; Jiao, J.; Wang, Y.; Tian, Q.; and Ye, Q. 2023. Integrally Pre-Trained Transformer Pyramid Networks. In *CVPR*, 18610–18620.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; and Zhang, L. 2022. Advancing plain vision transformer toward remote sensing foundation model. *TGRS*, 61: 1–15.

Wang, J.; Yang, W.; Li, H.-C.; Zhang, H.; and Xia, G.-S. 2021. Learning center probability map for detecting objects in aerial images. *TGRS*, 59(5): 4307–4323.

Weiler, M.; and Cesa, G. 2019. General e (2)-equivariant steerable cnns. *NeurIPS*, 32.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *CVPR*, 3974–3983.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021a. SegFormer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34: 12077–12090.

Xie, X.; Cheng, G.; Wang, J.; Yao, X.; and Han, J. 2021b. Oriented R-CNN for object detection. In *ICCV*, 3520–3529.

Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.-S.; and Bai, X. 2020. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *TPAMI*, 43(4): 1452–1459.

Yang, X.; and Yan, J. 2020. Arbitrary-oriented object detection with circular smooth label. In *ECCV*, 677–694.

Yang, X.; Yan, J.; Feng, Z.; and He, T. 2021a. R3det: Refined single-stage detector with feature refinement for rotating object. In *AAAI*, volume 35, 3163–3171.

Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; and Tian, Q. 2021b. Rethinking rotated object detection with gaussian wasserstein distance loss. In *ICML*, 11830–11841.

Yang, X.; Yang, X.; Yang, J.; Ming, Q.; Wang, W.; Tian, Q.; and Yan, J. 2021c. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *NeurIPS*, 34: 18381–18394.

Yang, X.; Zhou, Y.; Zhang, G.; Yang, J.; Wang, W.; Yan, J.; ZHANG, X.; and Tian, Q. 2022. The KFIoU Loss for Rotated Object Detection. In *ICLR*.

Yang, Z.; Liu, S.; Hu, H.; Wang, L.; and Lin, S. 2019. Reppoints: Point set representation for object detection. In *ICCV*, 9657–9666.

Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; and Wang, J. 2021. Hrformer: High-resolution transformer for dense prediction. *arXiv preprint arXiv:2110.09408*.

Zhang, Q.; Xu, Y.; Zhang, J.; and Tao, D. 2022a. Vsa: Learning varied-size window attention in vision transformers. In *ECCV*, 466–483.

Zhang, X.; Liu, F.; Peng, Z.; Guo, Z.; Wan, F.; Ji, X.; and Ye, Q. 2022b. Integral migrating pre-trained transformer encoder-decoders for visual object detection. *arXiv preprint arXiv:2205.09613*.

Zhang, X.; Tian, Y.; Huang, W.; Ye, Q.; Dai, Q.; Xie, L.; and Tian, Q. 2022c. Hivit: Hierarchical vision transformer meets masked image modeling. *arXiv preprint arXiv:2205.14949*.

Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C.; et al. 2022. Mmrotate: A rotated object detection benchmark using pytorch. In *ACM MM*, 7331–7334.

Zhou, Y.; Ye, Q.; Qiu, Q.; and Jiao, J. 2017. Oriented response networks. In *CVPR*, 519–528.