Strip R-CNN: Large Strip Convolution for Remote Sensing Object Detection

Xinbin Yuan[1], **Zhaohui Zheng**[1], **Yuxuan Li**[1], **Xialei Liu**[1], **Li Liu**[2], **Xiang Li**[1,3]
Qibin Hou[1,3*], **Ming-Ming Cheng**[1,3*]
[1]VCIP, School of Computer Science, NKU
[2]Academy of Advanced Technology Research of Hunan, Changsha, China
[3]NKIARI, Futian, Shenzhen, China
yxb@mail.nankai.edu.cn, {houqb, cmm}@nankai.edu.cn

Abstract

In this paper, we show that current approaches using large square kernels or transformer-based global modeling aggregate contextual information uniformly across spatial dimensions, leading to feature dilution and localization errors for elongated targets. To mitigate this issue, we propose Strip R-CNN, the first work to systematically explore large strip convolutions for remote sensing object detection. Our key insight is that strip convolutions enable directional feature aggregation along the dominant spatial dimension of slender objects, reducing background interference while preserving essential geometric information. We design two core components: (i) StripNet, a backbone network employing sequential orthogonal large strip convolutions to capture anisotropic spatial patterns, and (ii) Strip Head, which enhances localization precision by incorporating strip convolutions into the detection head. Unlike previous large-kernel approaches that suffer from computational redundancy and isotropic limitations, our method achieves superior performance with remarkable efficiency. Extensive experiments on multiple benchmarks (DOTA, FAIR1M, HRSC2016, and DIOR) demonstrate significant improvements, with our 30M parameter model achieving 82.75% mAP on DOTA-v1.0, establishing a new state-of-the-art record while providing new insights into anisotropic feature learning for remote sensing applications.

Code — https://github.com/HVision-NKU/Strip-R-CNN

Introduction

Remote sensing object detection has gained significant attention in recent years due to its application in aerial images captured by drones and satellites (Li et al. 2024c,b,e,f,a). A popular pipeline is built on the basis of rotated boxes to cover objects of interest. Due to boundary discontinuity and square-like problems (Yang et al. 2021b,c; Yang and Yan 2020) and the urgent need to capture long-range information (Li et al. 2023; Cai et al. 2024), many research breakthroughs have been made to develop stronger rotated object detectors, including object representations (Xie et al. 2021; Xu et al. 2021; Li et al. 2022; Xiao et al. 2024), IoU-simulated loss functions (Yang et al. 2021b,c; Yang and Yan 2020; Yang et al. 2022), and foundation models (Li et al. 2023; Pu et al. 2023; Cai et al. 2024; Li et al. 2025), etc.

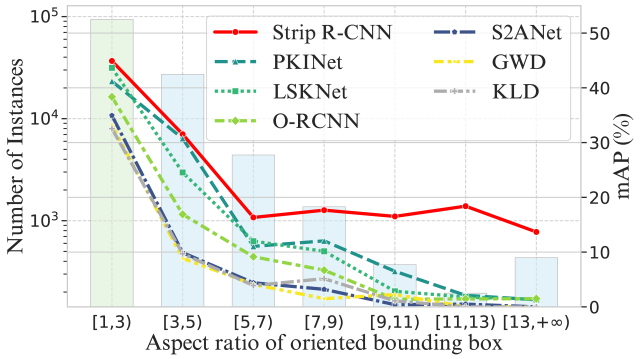


Figure 1: Statistics of the DOTA dataset (Xia et al. 2018) and the detection performance of several recent state-of-the-art detectors. We can see that slender objects (aspect ratio $>$ 3) occupy a non-negligible proportion and detection performance of previous state-of-the-art models declines as aspect ratio increases.

Despite the great progress made by previous work, successfully detecting high aspect ratio objects, which are prevalent in remote sensing object detection, is still a challenging problem. As illustrated in Fig. 1, the statistics of the widely used DOTA dataset (Xia et al. 2018) show that slender objects are quite common in remote sensing scenarios and usually occupy a large proportion of the data. However, existing object detection methods (Xie et al. 2021; Li et al. 2023; Cai et al. 2024; Yang et al. 2021b,c; Han et al. 2020) often struggle with slender objects, and the detection performance decreases as the aspect ratio of objects increases.

We argue that the difficulties in detecting these slender objects arise from two primary challenges. First, *high aspect ratio objects contain rich feature information along one spatial dimension, while exhibiting relatively sparse feature in the other.* Traditional detectors based on convolutional neural networks mostly extract input feature maps within square windows. This design greatly restricts their ability to effectively capture the anisotropic context, which can be commonly found in remote sensing images, leading to the excessive extraction of irrelevant information from surrounding areas. Second, *high aspect ratio objects pose consider-*

*Corresponding authors.

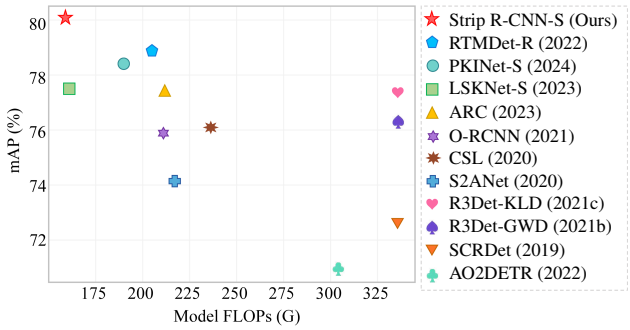


Figure 2: A comprehensive comparison of detection performance on the DOTA dataset of various remote sensing object detectors.

able challenges in regression tasks due to their unique geometric properties. In remote sensing object detection, unlike general object detection, an additional angle regression is required. For high aspect ratio objects, even a small error in angle estimation can lead to significant deviations from the ground truth.

To date, there are few works considering how to deal with challenging high aspect ratio objects. A generic approach widely used in previous methods is to enlarge the receptive field of models with large-kernel convolutions (Ding et al. 2022, 2024). A typical example should be LSKNet (Li et al. 2023), which introduces large-kernel convolutions with a spatial selection mechanism to capture long-range contextual information. PKINet (Cai et al. 2024) further extends LSKNet with a parallel large square conv structure to enhance the performance for large variation of object scales. However, the parallel paradigm of using multiple large-kernel convolutions exacerbates the computational burden, and the complicated block design further restricts model efficiency. For the high variation of the object aspect ratio, how to efficiently use large-kernel convolutions is still an open question.

In this paper, we propose Strip R-CNN to handle the challenging slender objects, which can efficiently combine the advantage of the square convolution and the strip convolution with less feature redundancy. Our design principles are two-fold. First, the new network architecture should be simple and efficient. Second, it should be good at handling objects of different aspect ratios even when they are high. Given the characteristics of the objects in remote sensing images, we propose using orthogonal large strip convolutions as the main spatial filters, which comprise the core component of our **StripNet** backbone, called strip module. Furthermore, to conquer the second challenge mentioned above, we strengthen the localization branch with our strip module in the newly proposed **Strip Head**, thereby achieving precise object localization. Combining the two new components, our method is quite simple but can generalize well to objects of especially high aspect ratios as shown in Fig. 1.

To our knowledge, Strip R-CNN is the first work to explore how to take advantage of large strip convolutions for remote sensing object detection. Despite simplicity and the lightweight nature, our Strip R-CNN achieves state-of-the-art performance on the standard DOTA benchmark without bells and whistles, as shown in Fig. 2. Notably, our model Strip R-CNN-S with only 30M learnable parameters achieves *the best results on DOTA leaderboard with 82.75% mAP*. We also conduct extensive experiments on several other remote sensing datasets, including FAIR1M, HRSC2016, and DIOR, and show the superiority of our Strip R-CNN over other methods. We hope that our design principles could provide new research insights for the remote sensing imagery community.

Related Work

Remote Sensing Object Detection. Generic object detection typically relies on horizontal bounding boxes to detect objects (Chen et al. 2025; Zheng et al. 2022b; Ren et al. 2015; Zheng et al. 2020, 2022a). However, in remote sensing scenarios, where objects are arbitrarily oriented, horizontal boxes often fail to precisely localize objects and tend to include background information or other objects (Xia et al. 2018). Therefore, rotated bounding boxes are generally adopted for object representation. Early representations of rotated bounding boxes use five parameters (x, y, w, h, θ) (Ding et al. 2019; Yang et al. 2019; Yu and Da 2023). However, due to the periodicity of the angle, training models based on this representation often face boundary discontinuity problems in regression (Yang and Yan 2020; Yang et al. 2022, 2021b). To address this issue, several approaches propose improved representations (Xie et al. 2021; Xiao et al. 2024; Li et al. 2022; Xu et al. 2021; Wang et al. 2019; Yi et al. 2021; Fu et al. 2020; Yang et al. 2021a) for rotated bounding boxes. For instance, Oriented R-CNN (Xie et al. 2021) replaces the angle with midpoint offsets, leading to a six-parameter representation $(x, y, w, h, \triangle\alpha, \triangle\beta)$, which significantly enhances detector performance. COBB (Xiao et al. 2024) introduces a continuous representation of rotated boxes with nine parameters based on the aspect ratio of the minimum enclosing rectangle to the rotated bounding box areas. There are also approaches focusing on mitigating the discontinuity problem in boundary regression through loss functions (Yang et al. 2021b,c; Hou et al. 2023; Cheng et al. 2022b; Qian et al. 2021). For example, GWD (Yang et al. 2021b) and KLD (Yang et al. 2021c) convert rotated bounding boxes into 2D Gaussian distributions and use Gaussian Wasserstein Distance and Kullback-Leibler Divergence as the loss functions. KFIoU (Yang et al. 2022) uses Gaussian modeling and Kalman filtering and propose the SKewIoU Loss.

Despite the great progress made by previous work, successfully detecting high aspect ratio objects, which are prevalent in remote sensing object detection, is still a challenging problem. Our method carefully considers the difficulties posed by these objects and take advantage of large strip convolutions to make our network generalize well to the challenging slender objects, which to our knowledge has not been explored before in this research field.

Large Kernel Networks for Remote Sensing Object Detection. Convolution with large kernel has been an emerging

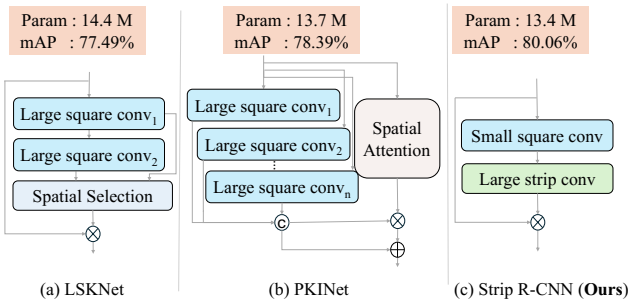


Figure 3: Structural concept comparison between our proposed StripNet and other representative methods using large-kernel convolutions, including LSKNet (Li et al. 2023) and PKINet (Cai et al. 2024).

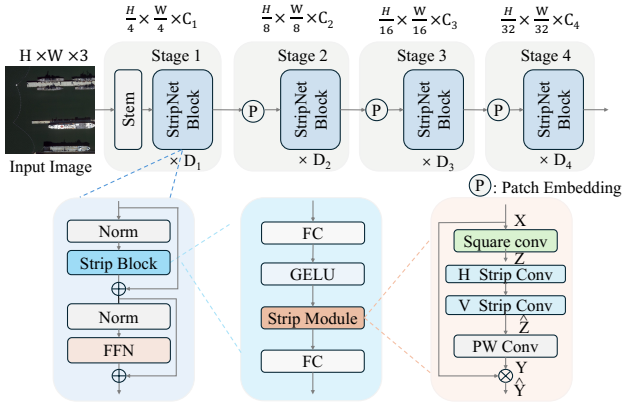


Figure 4: Structure of our basic block and StripNet backbone.

Model	$\{C_1, \ldots, C_4\}$	$\{D_1, \ldots, D_4\}$	#Params	FLOPs
⋆ StripNet-T	{32, 64, 160, 256}	{3, 3, 5, 2}	3.8M	18.2G
⋆ StripNet-S	{64, 128, 320, 512}	{2, 2, 4, 2}	13.3M	52.3G

Table 1: Variants of StripNet backbone. C_i: feature channel number; D_i: number of strip blocks in each stage i.

and promising solution to remote sensing object detection, which has been validated to have highly competitive performance against the Transformer-based methods in image classification and segmentation (Guo et al. 2022a,b; Ding et al. 2022). In remote sensing object detection, some approaches put efforts on employing large-kernel convolutions to get long-range contextual information (Li et al. 2024d; Cai et al. 2024). For example, LSKNet (Li et al. 2024d) utilizes large kernel convolutions and a selection mechanism to model the contextual information needed for different object categories. PKINet (Cai et al. 2024) arranges multiple large-kernel convolutions in parallel to extract dense texture features across diverse receptive fields, and introduces a context anchor attention mechanism to capture relationships between distant pixels. However, the parallel paradigm of leveraging multiple large-kernel convolutions exacerbates the computational burden, and the complicated block design makes the model not efficient. Regarding the high variation of the object aspect ratio, how to efficiently make use of large-kernel convolutions is still an open question. To our knowledge, Strip R-CNN is the first work to explore how to take advantage of large strip convolutions for remote sensing object detection.

Strip R-CNN

In this section, we describe the architecture of the proposed Strip R-CNN. Our goal is to advance rotated object detectors with large strip convolutions so that the resulting model can perform well on objects of different aspect ratios. This is different from previous work that emphasizes the importance of large square kernel convolutions as shown in Fig. 3.

Overall Architecture

Based on the O-RCNN framework (Xie et al. 2021), our Strip R-CNN replace the backbone and detection head with our **StripNet backbone** and **Strip Head**, respectively. Specifically, the backbone is mainly composed of basic blocks proposed as illustrated in Fig. 4, which consists of two residual sub-blocks: the strip sub-block and the feed-forward network sub-block. The strip sub-block is built upon a small-kernel standard convolution and two convolutions with large strip-shaped kernels to capture robust features for objects of different aspect ratios. For the feed-forward network sub-block, we simply follow LSKNet (Li et al. 2023), which is used to facilitate channel mixing and feature refinement. For the stem layers, we keep them the same to LSKNet (Li et al. 2023). For the detection head, we strengthen the localization branch with our proposed strip module, resulting in our Strip Head.

StripNet

Large square kernel convolutions provide essential long-range contextual information for remote sensing applications. PKINet (Cai et al. 2024) introduces parallel large square kernel convolutions and spatial attention mechanism to extract spatial information. However, the parallel paradigm of leveraging multiple large-kernel convolutions exacerbates the computational burden, and the complicated block design makes the model not efficient. Our objective is to efficiently extract essential features for objects of varying aspect ratios. The outcome is a sequential paradigm that efficiently combines the advantages of both standard and strip convolutions without requiring additional information fusion module. In what follows, we provide a detailed description of our basic block: **strip module**.

Given an input tensor $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ with C channels, a depthwise convolution with square kernels $\mathbf{K} \in \mathbb{R}^{C\times k_H\times k_W}$ is first applied to extract local contextual features, yielding $\mathbf{Z}$, where $H \times W$ and $k_H \times k_W$ are the feature size and kernel size, respectively. In practical use, we set $k_H \times k_W$ to 5×5. After the initial depthwise convolution, we use two sequential depthwise convolution of large strip-shaped kernels to better capture objects of high aspect ratios. The output is denoted as $\hat{\mathbf{Z}}$. Unlike standard convolutions that extract features from a square region for each time,

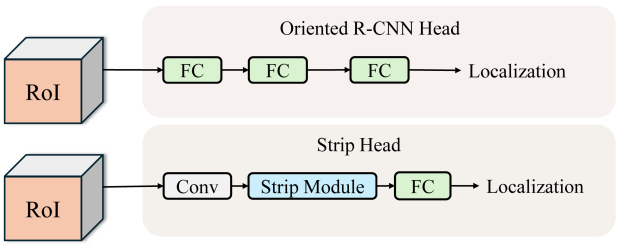


Figure 5: Comparison of Oriented R-CNN head and Strip Head. Our Strip Head incorporates the strip module into the localization head.

large strip convolutions allow the network to focus more on features along either the horizontal or the vertical axis. The combined use of horizontal and vertical large strip convolutions enables the network to collect directional features across both spatial axes, enhancing the representations of elongated or narrow structures in spatial dimension.

To further enhance the interaction of the features across the channel dimension, a simple point-wise convolution is applied to transform $\hat{\mathbf{Z}}$ to $\mathbf{Y}$. In this way, each position of the resulting feature map $\mathbf{Y}$ encodes both horizontal and vertical features across a wide spatial area. Finally, following (Guo et al. 2022b; Hou et al. 2024), we regard the feature map $\mathbf{Y}$ as attention weights to reweigh the input X, which can be formulated as $\hat{\mathbf{Y}} = \mathbf{X} \cdot \mathbf{Y}$, where '$\cdot$' denotes the element-wise multiplication operation. Fig. 4 provides an illustration of our strip module. In our experiments, we will discuss how to choose the kernel size of the strip convolutions.

Our backbone network StripNet consists of four stages, each comprising repeated StripNet blocks. The detailed configurations of different variants of our StripNet backbone are shown in Tab. 1. It is important to emphasize that our StripNet is much simpler than previous remote sensing object detectors using large-kernel convolutions as shown in Fig. 3. We do not utilize any spatial or channel attention mechanisms in our basic block design nor compound fusion operations with different types of large-kernel convolutions. This makes our StripNet quite simple but has great performance on different remote sensing detection benchmarks. Moreover, we found that there is no significant difference in applying horizontal strip convolutions before vertical ones or vice versa. Both approaches are effective.

Strip Head

In localization tasks, models should be sensitive to transformations, as the accuracy of localization depends on the positions of input objects. Previous strong rotated object detectors (Cai et al. 2024; Li et al. 2023) adopt the Oriented R-CNN framework (Xie et al. 2021), whose detection head shares the same linear layers for classification and localization. However, linear layers have limited spatial correlation as demonstrated in (Wu et al. 2020), making them transformation-insensitive and unsuitable for precise localization.

A better solution might be to decouple the classification and localization tasks and using small kernel convolutions

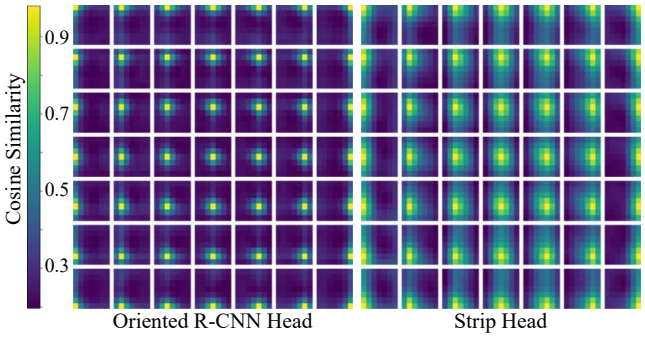


Figure 6: Spatial correlation map comparison of the Oriented R-CNN head and our Strip Head. Strip Head has more spatial correlations in the output feature maps.

in the localization branch as suggested in (Wu et al. 2020). However, our analysis of spatial correlation maps shows that small kernel convolutions capture only short-range spatial correlations, as illustrated on the left of Fig. 6. These short-range correlations are inadequate for accurately localizing slender objects, which require long-range dependencies. To effectively localize objects of varying aspect ratios, we argue that the localization head should be able to capture long-range dependencies, similar to those handled by the backbone network. Large strip convolutions capture both horizontal and vertical features across a broad spatial area, which provide the extended spatial correlations necessary for better localization. Therefore, we propose Strip Head, which leverages large strip convolutions to enhance the detector's localization capabilities. As illustrated at the bottom of Fig. 5, the localization branch begins with a standard 3×3 convolution to extract local features. Then we add a strip module, followed by a fully connected layer to collect long-range spatial dependencies.

Benefiting from the proposed strip module, our Strip Head produces more spatial correlations in the output feature maps, as shown in Fig. 6. In the experiments section, we will show that this design can notably improves the detection capability compared to the Oriented R-CNN head.

Experiments

Experiment Setup

Datasets. We conduct extensive experiments on five popular remote sensing object detection datasets. For detailed usage of the datasets, kindly refer to Supplementary Materials.

- **DOTA-v1.0** (Xia et al. 2018) is a large-scale dataset for remote sensing detection which contains 2,806 images, 188,282 instances, and 15 categories.
- **DOTA-v1.5** (Xia et al. 2018) is a more challenging dataset which contains 403,318 instances and 16 categories.
- **FAIR1M-v1.0** (Sun et al. 2022) is a remote sensing dataset consisting of 15,266 images, 1 million instances, and 37 categories.
- **HRSC2016** (Liu et al. 2016) is a remote sensing dataset that contains 1061 aerial images and 2,976 instances.

Method	mAP	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
Single-Scale																
EMO2-DETR (2023)	70.91	87.99	79.46	45.74	66.64	78.90	73.90	73.30	90.40	80.55	85.89	55.19	63.62	51.83	70.15	60.04
AO2-DETR (2022)	72.15	86.01	75.92	46.02	66.65	79.70	79.93	89.17	90.44	81.19	76.00	56.91	62.45	64.22	65.80	58.96
SASM (2022)	74.92	86.42	78.97	52.47	69.84	77.30	75.99	86.72	90.89	82.63	85.66	60.13	68.25	73.98	72.22	62.37
O-RCNN (2021)	75.87	89.46	82.12	54.78	70.86	78.93	83.00	88.20	90.90	87.50	84.68	63.97	67.69	74.94	68.84	52.28
ReDet (2021)	76.25	88.79	82.64	53.97	74.00	78.13	84.06	88.04	90.89	87.78	85.75	61.76	60.39	75.96	68.07	63.59
R3Det-GWD (2021b)	76.34	88.82	82.94	55.63	72.75	78.52	83.10	87.46	90.21	86.36	85.44	64.70	61.41	73.46	76.94	57.38
R3Det-KLD (2021c)	77.36	88.90	84.17	55.80	69.35	78.72	84.08	87.00	89.75	84.32	85.73	64.74	61.80	76.62	78.49	70.89
ARC (2023)	77.35	89.40	82.48	55.33	73.88	79.37	84.05	88.06	90.90	86.44	84.83	63.63	70.32	74.29	71.91	65.43
LSKNet-S (2023)	77.49	89.66	85.52	57.72	75.70	74.95	78.69	88.24	90.88	86.79	86.38	66.92	63.77	77.77	74.47	64.82
PKINet-S (2024)	78.39	89.72	84.20	55.81	77.63	80.25	84.45	88.12	90.88	87.57	86.07	66.86	70.23	77.47	73.62	62.94
RTMDet-R (2022)	78.85	89.43	84.21	55.20	75.06	80.81	84.53	88.97	90.90	87.38	87.25	63.09	67.87	78.09	80.78	69.13
⋆ Strip R-CNN-S	**80.06**	88.91	86.38	57.44	76.37	79.73	84.38	88.25	90.86	86.71	87.45	69.89	66.82	79.25	82.91	75.58
Multi-Scale																
R3Det-GWD (2021b)	80.23	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
DODet (2022c)	80.62	89.96	85.52	58.01	81.22	78.71	85.46	88.59	90.89	87.12	87.80	70.50	71.54	82.06	77.43	74.47
AOPG (2022a)	80.66	89.88	85.57	60.90	81.51	78.70	85.29	88.85	90.89	87.60	87.65	71.66	68.69	82.31	77.32	73.10
R3Det-KLD (2021c)	80.63	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
KFloU (2022)	80.93	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
PKINet-S (2024)	81.06	89.02	86.73	58.95	81.20	80.41	84.94	88.10	90.88	86.60	87.28	67.10	74.81	78.18	81.91	70.62
RVSA (2022)	81.24	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
LSKNet-S (2023)	81.64	89.57	86.34	63.13	83.67	82.20	86.10	88.66	90.89	88.41	87.42	71.72	69.58	78.88	81.77	76.52
⋆ Strip R-CNN-T	81.40	89.14	84.90	61.78	83.50	81.54	85.87	88.64	90.89	88.02	87.31	71.55	70.74	78.66	79.81	78.16
⋆ Strip R-CNN-S	**82.28**	89.17	85.57	62.40	83.71	81.93	86.58	88.84	90.86	87.97	87.91	72.07	71.88	79.25	82.45	82.82
⋆ Strip R-CNN-S†	**82.75**	88.99	86.56	61.35	83.94	81.70	85.16	88.57	90.88	88.62	87.36	75.13	74.34	84.58	81.49	82.56

Table 2: Comparisons with SOTA methods on the DOTA-v1.0 dataset with single-scale and multi-scale training and testing. The StripNet-S backbone is pretrained on ImageNet for 300 epochs. †: Model ensemble as in MoCAE (Oksuz et al. 2023).

- **DIOR-R** (Cheng et al. 2022a) contains 23,463 images and 192,518 instances.

Implementation details. Our training process is divided into two stages: pretraining on ImageNet (Deng et al. 2009) and fine-tuning on downstream remote sensing datasets. For ablation experiments, we train all the models for 100 epochs. To achieve higher performance, we train the models for 300 epochs for the final results. The number of training epochs for DOTA, DOTAv1.5, HRSC2016, FAIR1M-v1.0, and DIOR-R are set to 12, 12, 36, 12, and 12, respectively, following previous methods. Learning rates are set to 0.0001, 0.0001, 0.0004, 0.0001, and 0.0001, respectively. The input sizes for HRSC2016 and DIOR-R are 800×800, while for the DOTA-v1.0, DOTA-v1.5 and FAIR1M-v1.0 datasets, the input sizes are 1024×1024. During training, we employ the AdamW (Loshchilov and Hutter 2017) optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, and a weight decay of 0.05. All the models are trained on 8 NVIDIA 3090 GPUs with a batch size of 8, and test is conducted on a single NVIDIA 3090 GPU.

Main Results

We first compare our Strip R-CNN with recent state-of-the-art methods with strong backbones implemented within the Oriented R-CNN (Xie et al. 2021) framework on the DOTA v1.0 dataset. As shown in Tab. 3, Strip R-CNN-S achieves an improvement of 1.67% while using 0.4% fewer parameters and only 74.3% of the computations required by PKINet-S (Cai et al. 2024). Additionally, Strip R-CNN-S shows a 2.57% enhancement over LSKNet-S (Li et al. 2023) utilizing 1.1% fewer parameters and 2.2% less computations.

Model (Backbone)	#P↓	FLOPs↓	FPS	mAP (%)
ResNet-50 (He et al. 2016)	23.3M	86.1G	21.8	75.87
LSKNet-S (Li et al. 2023)	14.4M	54.4G	20.7	77.49
PKINet-S (Cai et al. 2024)	13.7M	70.2G	12.0	78.39
⋆StripNet-S	13.3M	52.3G	17.7	80.06

Table 3: Comparisons with different backbone models on DOTA-v1.0. Params and FLOPs are computed for backbone only. All backbones are pretrained on ImageNet for 300 epochs. Our StripNet-S achieves higher mAP than previous popular backbones.

Results on DOTA-v1.0 (Xia et al. 2018). We conduct a comparative analysis of different models and present detailed results for mean Average Precision (mAP) and Average Precision (AP) across categories on the DOTA dataset(refer to more model results in Supplementary Materials). As shown in Tab. 2, our single-scale evaluation demonstrates a 1.67% improvement over PKINet-S. Furthermore, with multi-scale training and testing, we achieve

Method	FR-O (2015)	Mask RCNN (2017)	HTC (2019)	ReDet (2021)	DCFL (2023)	LSKNet-S (2023)	PKINet-S (2024)	**Strip R-CNN-S**
mAP (%)	62.00	62.67	63.40	66.86	67.37	70.26	71.47	**72.27**

Table 4: Comparisons with SOTA methods on the DOTA-v1.5 dataset with single-scale training and testing. The StripNet-S backbone is pretrained on ImageNet for 300 epochs.

Method	RetinaNet* (2017)	C-RCNN* (2018)	F-RCNN* (2015)	RoI Trans.* (2019)	O-RCNN (2021)	LSKNet-S (2023)	**Strip R-CNN-S**
mAP (%)	30.67	31.18	32.12	35.29	45.60	47.87	**48.26**

Table 5: Comparisons with SOTA methods on the FAIR1M-v1.0 dataset. The StripNet-S backbone is pretrained on ImageNet for 300 epochs. *: Results are referenced from the FAIR1M paper (Sun et al. 2022).

Method	**Pre.**	**mAP(07)**	**mAP(12)**	**#P**	**FLOPs**
DAL (2020)	IN	89.77	-	36.4M	216G
GWD (2021b)	IN	89.85	97.37	47.4M	456G
AOPG (2022a)	IN	90.34	96.22	-	-
RTMDet (2022)	CO	90.60	97.10	52.3M	205G
LSKNet-S (2023)	IN	90.65	98.46	31.0M	161G
PKINet-S (2024)	IN	90.65	98.54	30.8M	190G
⋆ Strip R-CNN-S	IN	90.60	**98.70**	30.5M	159G

Table 6: Performance comparisons with SOTA methods on HRSC2016. mAP (07/12): VOC 2007 (Everingham et al. 2007)/2012 (Everingham et al. 2012) metrics. The StripNet-S backbone is pretrained on ImageNet for 300 epochs.

82.28% mAP for a single model. By ensembling the results of RTMDet and Strip R-CNN, following the model ensemble strategy in MoCAE (Oksuz et al. 2023), we achieve 82.75% mAP, setting a new state-of-the-art record.

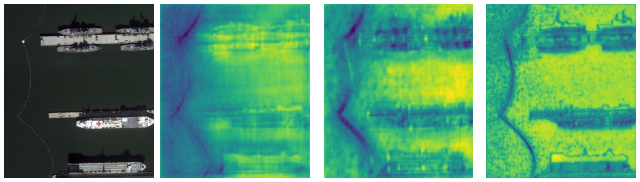
Method	**Pre.**	**#P ↓**	**FLOPs ↓**	**mAP (%)**
Faster RCNN-O (2015)	IN	41.1M	198G	59.54
TIOE-Det (2023)	IN	41.1M	198G	61.98
ARS-DETR (2024)	IN	41.1M	198G	66.12
O-RepPoints (2022)	IN	36.6M	-	66.71
DCFL (2023)	IN	-	-	66.80
LSKNet-S (2023)	IN	31M	161G	65.90
PKINet-S (2024)	IN	30.8M	190G	67.03
⋆ Strip R-CNN-S	IN	30.5M	159G	**68.70**

Table 7: Performance comparisons with SOTA methods on DIOR-R. The StripNet-S backbone is pretrained on ImageNet for 300 epochs.

Results on DOTA-v1.5 (Xia et al. 2018). In this dataset with minuscule instances, as shown in Tab. 4, our approach achieves outstanding performance, demonstrating its efficacy and generalization ability to small objects. Our Strip R-CNN outperforms the former state-of-the-art method, achieving an improvement of 0.8%.

Results on FAIR1M-v1.0 (Sun et al. 2022). The results in Tab. 5 reveal that our Strip R-CNN reaches a highly competitive mAP score of 48.26%. Our method could improve 0.39% mAP for LSKNet (Li et al. 2023) and surpassing all other methods by a significant margin.

Results on DIOR-R (Cheng et al. 2022a). As shown in Tab. 7, we observe 2.80% improvement over LSKNet (Li et al. 2023) and 1.67% improvement over PKINet (Cai et al. 2024).



Input Image Kernel size: 11 Kernel size: 15 Kernel size: 19

Figure 7: Feature map visulization of different kernel sizes. Kernel size 19 leads to more accurate features of high aspect ratio objects.

Results on HRSC2016 (Liu et al. 2016). We achieve 98.70% mAP under the VOC2012 (Everingham et al. 2012) metric, which is the state-of-the-art performance with 0.16% mAP improvement over PKINet (Cai et al. 2024) and 0.24% mAP over LSKNet (Li et al. 2023). The results are shown in Tab. 6.

Across multiple datasets, our method consistently surpasses previous state-of-the-art approaches, demonstrating its generalizability and effectiveness.

Ablation Studies

Kernel size of strip convolutions. The kernel size in strip convolutions is critical for our proposed strip module. We experiment with large kernel sizes, starting from 11 and increasing in increments of 4, as shown in Table. 8. It is observed that a large kernel with size of 19 yields satisfactory results among all the options. In Fig. 7, we further present a visual comparison of the learned feature representations when different kernel sizes are adopted. One can see that strip convs with a larger kernel size can learn the features of slender objects with more precise localization information and well sharpen the object boundaries. By default, we adopt 19 as the kernel size in all stages of the StripNet backbone, which is fixed in all other experiments.

Ablation on strip module design. We perform ablation experiments to analyze the design choices in the strip module as shown in Tab. 9. First, we assess the role of depthwise square convolution. Removing this component leads to a large performance drop, emphasizing the importance of square convolution for capturing features of square-shaped objects. Next, we examine the integration of horizontal and

Kernel Size	#P↓	FLOPs↓	mAP (%)
(19,19,19,19)	13.30M	52.34G	**81.75**
(15,15,15,15)	13.28M	52.19G	81.64
(11,11,11,11)	13.26M	52.03G	81.22
(15,17,19,21)	13.31M	52.26G	81.37
(21,19,17,15)	13.29M	52.34G	81.72

Table 8: Ablation study on the kernel size of our proposed strip module at four stages of the StripNet backbone network. We adopt the StripNet-S backbone pretrained on ImageNet for 100 epochs. The best result is obtained when using kernel size 19 at all stages.

5×5 Square Conv	Large Strip Conv		#P ↓	FLOPs ↓	mAP (%)
	Sequential	Parallel			
✗	✓	✗	13.23M	51.84G	81.38
✓	✓	✗	13.30M	52.34G	**81.75**
✓	✗	✓	13.33M	52.52G	81.54
✓	19×19 Square Conv		14.17M	58.41G	81.44
✓	7×7 Square Conv d=3		13.30M	52.34G	81.55

Table 9: Ablation study on the design of our proposed strip module. We adopt the StripNet-S backbone pretrained on ImageNet for 100 epochs. d: dilation rate.

vertical large strip convolutions, comparing parallel and sequential arrangements. The sequential configuration outperforms the parallel one, as the latter may lacks effective two-dimensional modeling, merely combining the two strip convolutions without capturing the overall object structure. Furthermore, substituting the sequential large strip convolutions with either a 19×19 large kernel convolution or a 7×7 dilated convolution with a dilation rate of 3 results in noticeable performance loss, further validating the effectiveness of the large strip convolutions.

Effectiveness of the Strip Head. In Tab. 10, we compare the performance of two different detectors before and after incorporating the Strip Head. The results show that our Strip Head consistently improves the performance of the two object detectors. We achieve a remarkable AP increase of 4.75% on ROI Transformer (Ding et al. 2019) and a noticable AP increase of 1.71% on Faster RCNN-O (Ren et al. 2015). This demonstrates the effectiveness and generalizability of our Strip Head.

Visual analysis. We present detection results and Eigen-CAM (Muhammad and Yeasin 2020) visualizations in Fig. 8. One can see that for high aspect ratio objects, previous methods such as LSKNet and PKINet have issues like missed detections and significant localization errors. In contrast, our method can successfully detect the high aspect ratio objects. The Eigen-CAM visualizations also show strong activations of our method for these objects, validating the effectiveness of our approach.

Strip R-CNN for improving slender objects. Lastly, to further substantiate the superiority of our method in detecting high aspect ratio objects, we conduct additional experiments on the DOTA dataset. Inspired by the zone evaluation method (Zheng et al. 2024), we evaluate the detection performance for objects within a certain aspect ratio range alone. As shown in Fig. 1, the results show that our method is better than all other object detectors for every aspect ratio range. For the simple case of aspect ratios in the range of $[1, 5]$, our method achieves slightly better results than the state-of-the-art method PKINet. However, with the increase of aspect ratio range, our method outperforms the competitors by a large margin. Intriguingly, for the extreme case of aspect ratios in the range of $[13, +\infty]$, our Strip R-CNN still produces > 10 mAP, while the other detectors **approach 0**! That is to say our method can generally perform better than the other detectors even without carefully considering the objects are slender or not.

Method	Head	mAP (%)
Faster RCNN-O (2015)	Original Head	76.17
	Strip Head	77.88
RoI Trans. (2019)	Original Head	74.61
	Strip Head	79.37

Table 10: Effectiveness of Strip Head on other rotated detectors.

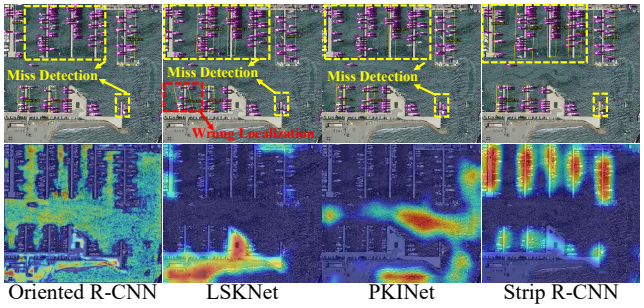


Figure 8: **Top:** Detection results. Our method can successfully capture the high aspect ratio objects. **Bottom:** Eigen-CAM visualizations. Our method shows strong activations for high aspect ratio objects.

Conclusions

In this paper, we alleviate the challenge of detecting slender objects in remote sensing scenarios by leveraging large strip convolutions to better extract features and improve localization of such objects. Based on large strip convolutions, we propose the simple yet highly effective Strip R-CNN. Extensive experiments demonstrate that our method exhibits strong generalization capability and achieves state-of-the-art performance on several remote sensing benchmarks. We hope this research could facilitate the development of object detection in the remote sensing field.

Acknowledgments

This work was funded by National Key Research and Development Project of China (No. 2024YFE0100700), NSFC (No. 62495061, 62276145).

References

Cai, X.; Lai, Q.; Wang, Y.; Wang, W.; Sun, Z.; and Yao, Y. 2024. Poly kernel inception network for remote sensing detection. In *CVPR*, 27706–27716.

Cai, Z.; and Vasconcelos, N. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. In *CVPR*.

Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; Loy, C. C.; and Lin, D. 2019. Hybrid Task Cascade for Instance Segmentation. In *CVPR*.

Chen, Y.; Yuan, X.; Wang, J.; Wu, R.; Li, X.; Hou, Q.; and Cheng, M.-M. 2025. YOLO-MS: rethinking multi-scale representation learning for real-time object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Cheng, G.; Yao, Y.; Li, S.; Li, K.; Xie, X.; Wang, J.; Yao, X.; and Han, J. 2022b. Dual-Aligned Oriented Detector. *IEEE TGRS*.

Cheng, G.; Yao, Y.; Li, S.; Li, K.; Xie, X.; Wang, J.; Yao, X.; and Han, J. 2022c. Dual-Aligned Oriented Detector. *IEEE TGRS*.

Dai, L.; Liu, H.; Tang, H.; Wu, Z.; and Song, P. 2022. AO2-DETR: Arbitrary-Oriented Object Detection Transformer. *IEEE TCSVT*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.

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*.

Ding, X.; Zhang, X.; Han, J.; and Ding, G. 2022. Scaling Up Your Kernels to 31×31: Revisiting Large Kernel Design in CNNs. In *CVPR*.

Ding, X.; Zhang, Y.; Ge, Y.; Zhao, S.; Song, L.; Yue, X.; and Shan, Y. 2024. UniRepLKNet: A Universal Perception Large-Kernel ConvNet for Audio Video Point Cloud Time-Series and Image Recognition. In *CVPR*, 5513–5524.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn; and Zisserman, A. 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

Fu, K.; Chang, Z.; Zhang, Y.; and Sun, X. 2020. Point-based estimator for arbitrary-oriented object detection in aerial images. *IEEE TGRS*, 59(5): 4370–4387.

Guo, M.-H.; Lu, C.; Liu, Z.-N.; Cheng, M.-M.; and Hu, S. 2022a. Visual Attention Network. *Computational Visual Media*.

Guo, M.-H.; Lu, C.-Z.; Hou, Q.; Liu, Z.-N.; Cheng, M.-M.; and Hu, S.-M. 2022b. SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation. In *NeurIPS*.

Han, J.; Ding, J.; Li, J.; and Xia, G.-S. 2020. Align Deep Features for Oriented Object Detection. *IEEE TGRS*.

Han, J.; Ding, J.; Xue, N.; and Xia, G.-S. 2021. ReDet: A Rotation-equivariant Detector for Aerial Object Detection. In *CVPR*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hou, L.; Lu, K.; Xue, J.; and Li, Y. 2022. Shape-Adaptive Selection and Measurement for Oriented Object Detection. In *AAAI*.

Hou, L.; Lu, K.; Yang, X.; Li, Y.; and Xue, J. 2023. G-rep: Gaussian representation for arbitrary-oriented object detection. *Remote Sensing*, 15(3): 757.

Hou, Q.; Lu, C.-Z.; Cheng, M.-M.; and Feng, J. 2024. Conv2former: A simple transformer-style convnet for visual recognition. *IEEE TPAMI*.

Hu, Z.; Gao, K.; Zhang, X.; Wang, J.; Wang, H.; Yang, Z.; Li, C.; and Li, W. 2023. EMO2-DETR: Efficient-matching oriented object detection with transformers. *IEEE TGRS*.

Li, K.; Wang, D.; Hu, Z.; Zhu, W.; Li, S.; and Wang, Q. 2024a. Unleashing channel potential: Space-frequency selection convolution for SAR object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17323–17332.

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

Li, W.; Yang, W.; Hou, Y.; Liu, L.; Liu, Y.; and Li, X. 2024b. SARATR-X: Towards Building A Foundation Model for SAR Target Recognition. *arXiv preprint*.

Li, W.; Yang, W.; Liu, T.; Hou, Y.; Li, Y.; Liu, Z.; Liu, Y.; and Liu, L. 2024c. Predicting gradient is better: Exploring self-supervised learning for SAR ATR with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218: 326–338.

Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.-M.; Yang, J.; and Li, X. 2023. Large Selective Kernel Network for Remote Sensing Object Detection. In *ICCV*.

Li, Y.; Li, X.; Dai, Y.; Hou, Q.; Liu, L.; Liu, Y.; Cheng, M.-M.; and Yang, J. 2024d. LSKNet: A Foundation Lightweight Backbone for Remote Sensing. *IJCV*.

Li, Y.; Li, X.; Li, W.; Hou, Q.; Liu, L.; Cheng, M.-M.; and Yang, J. 2024e. Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. *NeurIPS*.

Li, Y.; Li, X.; Li, Y.; Zhang, Y.; Dai, Y.; Hou, Q.; Cheng, M.-M.; and Yang, J. 2024f. SM3Det: A Unified Model for Multi-Modal Remote Sensing Object Detection. *arXiv preprint arXiv:2412.20665*.

Li, Y.; Zhang, Y.; Tang, W.; Dai, Y.; Cheng, M.-M.; Li, X.; and Yang, J. 2025. Visual Instruction Pretraining for Domain-Specific Foundation Models. *arXiv*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *ICCV*.

Liu, Z.; Wang, H.; Weng, L.; and Yang, Y. 2016. Ship Rotated Bounding Box Space for Ship Extraction From High-Resolution Optical Satellite Images With Complex Backgrounds. *TGRS Letters*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *ArXiv*.

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. *CoRR*.

Ming, Q.; Miao, L.; Zhou, Z.; Song, J.; Dong, Y.; and Yang, X. 2023. Task interleaving and orientation estimation for high-precision oriented object detection in aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196: 241–255.

Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; and Li, L. 2020. Dynamic Anchor Learning for Arbitrary-Oriented Object Detection. *CoRR*.

Muhammad, M. B.; and Yeasin, M. 2020. Eigen-CAM: Class Activation Map using Principal Components. *CoRR*.

Oksuz, K.; Kuzucu, S.; Joy, T.; and Dokania, P. K. 2023. MoCaE: Mixture of Calibrated Experts Significantly Improves Object Detection. *ArXiv*.

Pu, Y.; Wang, Y.; Xia, Z.; Han, Y.; Wang, Y.; Gan, W.; Wang, Z.; Song, S.; and Huang, G. 2023. Adaptive rotated convolution for rotated object detection. In *ICCV*, 6589–6600.

Qian, W.; Yang, X.; Peng, S.; Yan, J.; and Guo, Y. 2021. Learning Modulated Loss for Rotated Object Detection. In *AAAI*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*.

Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; Weinmann, M.; Hinz, S.; Wang, C.; and Fu, K. 2022. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; and Zhang, L. 2022. Advancing Plain Vision Transformer Towards Remote Sensing Foundation Model. *IEEE TGRS*.

Wang, J.; Ding, J.; Guo, H.; Cheng, W.; Pan, T.; and Yang, W. 2019. Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sensing*, 11(24): 2930.

Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; and Fu, Y. 2020. Rethinking classification and localization for object detection. In *CVPR*, 10186–10195.

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*.

Xiao, Z.; Yang, G.; Yang, X.; Mu, T.; Yan, J.; and Hu, S. 2024. Theoretically Achieving Continuous Representation of Oriented Bounding Boxes. In *CVPR*, 16912–16922.

Xie, X.; Cheng, G.; Wang, J.; Yao, X.; and Han, J. 2021. Oriented R-CNN for Object Detection. In *ICCV*.

Xu, C.; Ding, J.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; and Xia, G.-S. 2023. Dynamic coarse-to-fine learning for oriented tiny object detection. In *CVPR*, 7318–7328.

Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.-S.; and Bai, X. 2021. Gliding Vertex on the Horizontal Bounding Box for Multi-Oriented Object Detection. *IEEE TPAMI*.

Yang, X.; Hou, L.; Zhou, Y.; Wang, W.; and Yan, J. 2021a. Dense label encoding for boundary discontinuity free rotation detection. In *CVPR*, 15819–15829.

Yang, X.; and Yan, J. 2020. Arbitrary-Oriented Object Detection with Circular Smooth Label. In *ECCV*.

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*.

Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; and Fu, K. 2019. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In *ICCV*.

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. In *NeurIPS*.

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*.

Yi, J.; Wu, P.; Liu, B.; Huang, Q.; Qu, H.; and Metaxas, D. 2021. Oriented object detection in aerial images with box boundary-aware vectors. In *WACV*, 2150–2159.

Yu, Y.; and Da, F. 2023. Phase-shifting coder: Predicting accurate orientation in oriented object detection. In *CVPR*, 13354–13363.

Zeng, Y.; Chen, Y.; Yang, X.; Li, Q.; and Yan, J. 2024. ARS-DETR: Aspect Ratio-Sensitive Detection Transformer for Aerial Oriented Object Detection. *IEEE TGRS*, 62: 1–15.

Zheng, Z.; Chen, Y.; Hou, Q.; Li, X.; Wang, P.; and Cheng, M.-M. 2024. Zone Evaluation: Revealing Spatial Bias in Object Detection. *IEEE TPAMI*.

Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In *AAAI*, 12993–13000.

Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; and Zuo, W. 2022a. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *IEEE Transactions on Cybernetics*, 52(8): 8574–8586.

Zheng, Z.; Ye, R.; Wang, P.; Ren, D.; Zuo, W.; Hou, Q.; and Cheng, M.-M. 2022b. Localization distillation for dense object detection. In *CVPR*, 9407–9416.