Large Selective Kernel Network for Remote Sensing Object Detection

Yuxuan Li†, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang* and Xiang Li*
VCIP, CS, Nankai University
yuxuan.li.17@ucl.ac.uk†, {csjyang, xiang.li.implus}@nankai.edu.cn*

Abstract

Recent research on remote sensing object detection has largely focused on improving the representation of oriented bounding boxes but has overlooked the unique prior knowledge presented in remote sensing scenarios. Such prior knowledge can be useful because tiny remote sensing objects may be mistakenly detected without referencing a sufficiently long-range context, which can vary for different objects. This paper considers these priors and proposes the lightweight Large Selective Kernel Network (LSKNet). LSKNet can dynamically adjust its large spatial receptive field to better model the ranging context of various objects in remote sensing scenarios. To our knowledge, large and selective kernel mechanisms have not been previously explored in remote sensing object detection. Without bells and whistles, our lightweight LSKNet sets new state-of-the-art scores on standard benchmarks, i.e., HRSC2016 (98.46% mAP), DOTA-v1.0 (81.85% mAP), and FAIR1M-v1.0 (47.87% mAP).

1. Introduction

Remote sensing object detection [47, 57, 82] focuses on identifying and locating objects of interest in aerial images, such as vehicles, ships or aircraft. In recent years, one mainstream trend has been to generate bounding boxes that accurately fit the orientation of the objects being detected rather than simply drawing horizontal boxes around them. Consequently, much research has focused on improving the representation of oriented bounding boxes for remote sensing object detection. This has largely been achieved through the development of specialized detection frameworks (*i.e.*, RoI Transformer [12], Oriented R-CNN [69], and R3Det [75]) and oriented box encoding (*i.e.*, gliding vertex [71] and midpoint offset box encoding [69]). Additionally, several loss functions, including GWD [77], KLD [79], and Modulated Loss [54], have been further proposed to enhance the performance of these approaches.

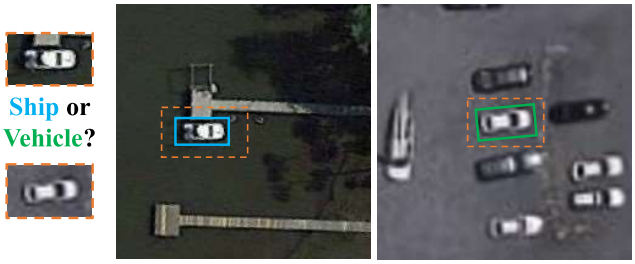


Figure 1. Successfully detecting remote sensing objects requires using a wide range of contextual information. Detectors with a limited receptive field may easily lead to incorrect results.

Despite these advances, relatively few works have considered the strong prior knowledge of remote sensing images. Aerial images are typically captured at high resolutions from a bird's eye view. In particular, most objects in aerial images may be small and difficult to identify based on their appearance alone. Instead, recognizing these objects relies on their context, as the surrounding environment can provide valuable clues about their shape, orientation, and other characteristics. According to an analysis of the remote sensing data, we identify two important priors:

- **Accurate detection often requires a wide range of contextual information.** As illustrated in Fig. 1, the limited context used by object detectors in remote sensing images can often lead to incorrect classifications. Rather than their appearance, the context distinguishes the ship from the vehicle.
- **The contextual information required for different objects is very different.** As shown in Fig. 2, the soccer field requires relatively less contextual information because of the unique distinguishable court borderlines. In contrast, the roundabout may require more context information to distinguish between gardens and ring-like buildings. Intersections, especially those partially covered by trees, require an extremely large receptive field due to the long-range dependencies between the intersecting roads.

To address the challenge of accurately detecting objects

*Corresponding authors.
Team URL: https://github.com/IMPlus-PCALab
Project page: https://github.com/zcablii/LSKNet

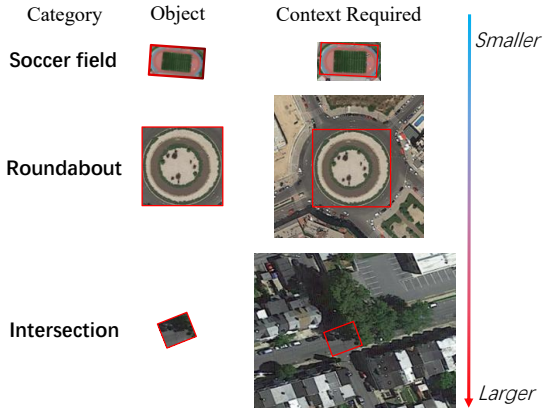


Figure 2. The wide range of contextual information required for different object types is very different by human criteria. The objects with red boxes are the exact ground-truth annotations.

in remote sensing images, which often require a wide and dynamic range of contextual information, we propose a novel lightweight detection backbone called Large Selective Kernel Network (LSKNet). Our approach involves dynamically adjusting the receptive field of the feature extraction backbone to more effectively process the varying wide context of the objects being detected. This is achieved through a spatial selective mechanism, which efficiently weights the features processed by a sequence of large depthwise kernels and then spatially merge them. The weights of these kernels are determined dynamically based on the input, allowing the model to adaptively use different large kernels and adjust the receptive field for each target in space as needed.

To our knowledge, the proposed LSKNet is the first to investigate using large and selective kernels for remote sensing object detection. Despite its simplicity and lightweight nature, our model achieves state-of-the-art performance on three popular datasets: HRSC2016 (98.46% mAP), DOTA-v1.0 (81.85% mAP), and FAIR1M-v1.0 (47.87% mAP), surpassing previously published results. Furthermore, we demonstrate that our model's behaviour aligns with the two priors above, which in turn verifies the effectiveness of the proposed mechanism.

2. Related Work

2.1. Remote Sensing Object Detection Framework

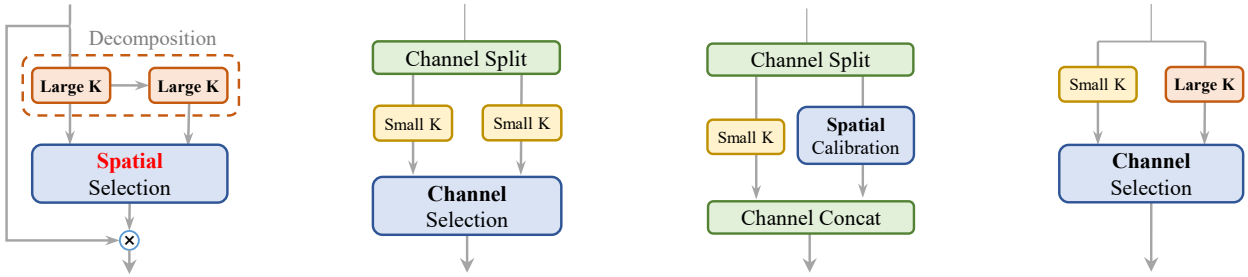
High-performance remote sensing object detectors often rely on the RCNN [56] framework, which consists of a region proposal network and regional CNN detection heads. The RPN proposes high-quality regions of interest (RoIs) from the backbone feature maps, while the regional CNN detection heads are responsible for object classification and bounding box regression. Several variations on the RCNN framework have been proposed in recent years. The two-stage RoI transformer [12] uses fully-connected layers to rotate candidate horizontal anchor boxes in the first stage, and then features within the boxes are extracted for further regression and classification. SCRDet [78] uses an attention mechanism to reduce background noise and improve the modelling of crowded and small objects. Oriented RCNN [69] and Gliding Vertex [71] introduce new box encoding systems to address the instability of training losses caused by rotation angle periodicity. Some approaches [32, 61, 87] treat remote sensing detection as a point detection task [74], providing an alternative way of addressing remote sensing detection problems.

Rather than relying on the proposed anchors, one-stage detection frameworks classify and regress oriented bounding boxes directly from grid densely sampled anchors. The one-stage S^2A network [23] extracts robust object features via oriented feature alignment and orientation-invariant feature extraction. DRN [50], on the other hand, leverages attention mechanisms to dynamically refined the backbone's extracted features for more accurate predictions. In contrast with Oriented RCNN and Gliding Vertex, RSDet [54] addresses the discontinuity of regression loss by introducing a modulated loss. LD [86] enhances the localization quality of oriented bounding boxes by distillation. AOPG [6] and R3Det [75] adopt a progressive regression approach, refining bounding boxes from coarse to fine granularity. In addition to CNN, AO2-DETR [9] introduces a transformer-based detection framework, DETR [4], into remote sensing detection tasks, which brings more research diversity.

While these approaches have achieved promising results in addressing the issue of rotation variance, they do not consider the strong and valuable prior information presented in aerial images. Instead, our approach uses the large kernel and spatial selective mechanism to better model these priors without modifying the current detection framework.

2.2. Large Kernel Networks

Transformer-based [59] models, such as the Vision Transformer (ViT) [1, 11, 14, 53, 60], Swin transformer [25, 39, 51, 64, 70, 83], and pyramid transformer [62, 67], have gained popularity in computer vision. Research [17, 45, 55, 72, 85] has demonstrated that the large receptive field is a key factor in their success. Recent work has shown that well-designed convolutional networks with large receptive fields can also be highly competitive with transformer-based models. For example, ConvNeXt [40] uses 7×7 depth-wise convolutions in its backbone, resulting in significant performance improvements in downstream tasks. In addition, RepLKNet [13] even uses a 31×31 convolutional kernel via re-parameterization, achieving compelling performance. A subsequent work SLaK [38], further expands the kernel size to 51×51 through kernel decomposition and sparse group techniques. RF-Next [18] automatically searches for a fixed



(a) LSK (**ours**) (b) ResNeSt [84] (c) SCNet [37] (d) SKNet [33]

Figure 3. Architectural comparison between our proposed LSK module and other representative selective mechanism modules. K: Kernel.

large kernel for various tasks. VAN [19] introduces an efficient decomposition of large kernels as convolutional attention. Similarly, SegNeXt [20] and Conv2Former [28] demonstrate that large kernel convolution plays an important role in modulating the convolutional features with a richer context.

Although large kernel convolutions have received attention in general object recognition, there has been a lack of research examining their significance in remote sensing detection. As previously noted in Sec. 1, aerial images possess unique characteristics that make large kernels particularly well-suited for remote sensing. As far as we know, our work represents the first attempt to introduce large kernel convolutions for remote sensing and to examine their importance in this field.

2.3. Attention/Selective Mechanism

The attention mechanism [21] is a simple but effective way to enhance neural representations for various tasks. The channel attention SE block [30] uses global average information to reweight feature channels, while spatial attention modules like GENet [29], GCNet [3], and SGE [34] enhance a network's ability to model context information via spatial masks. CBAM [66] and BAM [52] combine both channel and spatial attention.

In addition to channel/spatial attention mechanisms, kernel selection is a self-adaptive and effective technique for dynamic context modelling. CondConv [73] and Dynamic convolution [5] use parallel kernels to adaptively aggregate features from multiple convolution kernels. SKNet [33] introduces multiple branches with different convolutional kernels and selectively combines them along the channel dimension. ResNeSt [84] extends the idea of SKNet by partitioning the input feature map into several groups. Similarly to the SKNet, SCNet [37] uses branch attention to capturing richer information and spatial attention to improve localization ability. Deformable Convnets [8, 89] introduce a flexible kernel shape for convolution units.

Our approach bears the most similarity to SKNet [33]. However, there are **two key distinctions** between the two

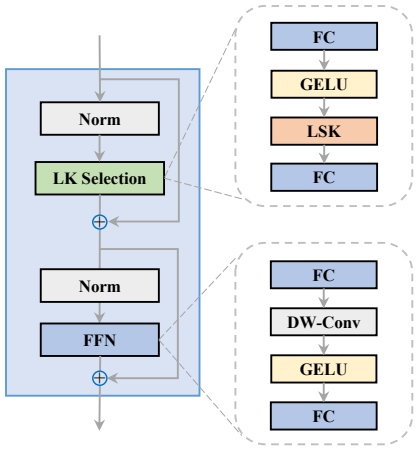


Figure 4. A block of LSKNet.

methods. Firstly, our proposed selective mechanism relies explicitly on a sequence of large kernels via decomposition, a departure from most existing attention-based approaches. Secondly, our method adaptively aggregates information across large kernels in the spatial dimension rather than the channel dimension utilized by SKNet. This design is more intuitive and effective for remote sensing tasks because channel-wise selection fails to model the spatial variance for different targets across the image space. The detailed structural comparisons are listed in Fig. 3.

3. Methods

3.1. LSKNet Architecture

The overall architecture of the LSKNet backbone is simply built upon repeated LSKNet Blocks. Fig 4 illustrates an LSKNet Block, which is inspired by ConvNeXt [41], MetaFormer [81], PVT-v2 [63], Conv2Former [28] and VAN [19]. Each LSKNet block consists of two residual sub-blocks: the Large Kernel Selection (LK Selection) sub-block and the Feed-forward Network (FFN) sub-block.

The LK Selection sub-block dynamically adjusts the network's receptive field as needed. The core LSK module

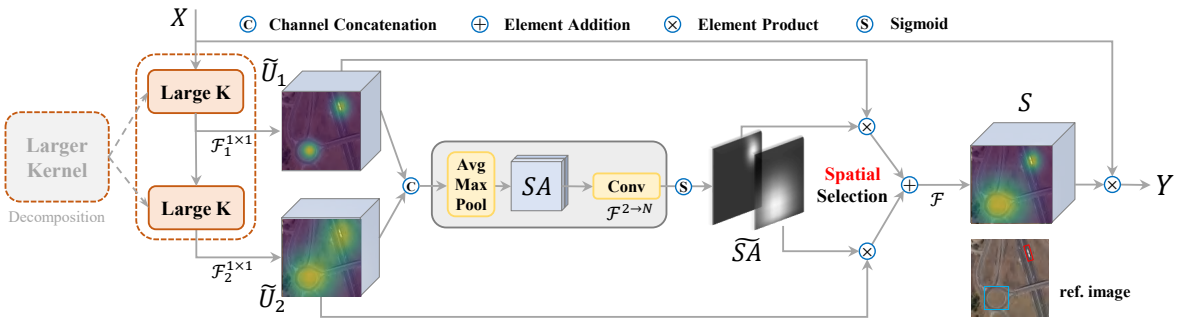


Figure 5. A conceptual illustration of LSK module.

Model	$\{C_1, C_2, C_3, C_4\}$	$\{D_1, D_2, D_3, D_4\}$	#P
⋆ LSKNet-T	{32, 64, 160, 256}	{3, 3, 5, 2}	4.3M
⋆ LSKNet-S	{64, 128, 320, 512}	{2, 2, 4, 2}	14.4M

Table 1. **Variants of LSKNet used in this paper**. C_i: feature channel number; D_i: number of LSKNet blocks of each stage i.

RF	(k, d) sequence	#P	FLOPs
23	(23, 1)	40.4K	42.4G
	(5,1) ⟶ (7, 3)	11.3K	11.9G
29	(29, 1)	60.4K	63.3G
	(3, 1) ⟶ (5, 2) ⟶ (7, 3)	11.3K	13.6G

Table 2. **Theoretical efficiency comparisons of two representative examples** by expanding a single large depth-wise kernel into a sequence, given channels being 64. k: kernel size; d: dilation.

(Fig. 5) is embedded in the LK Selection sub-block. It consists of a sequence of large kernel convolutions and a spatial kernel selection mechanism, which will be elaborated on later. The FFN sub-block is used for channel mixing and feature refinement, which consists of a sequence of a fully connected layer, a depth-wise convolution, a GELU [26] activation, and a second fully connected layer.

The detailed configuration of different variants of LSKNet used in this paper is listed in Tab. 1.

3.2. Large Kernel Convolutions

According to the *prior (2)*, as stated in Sec. 1, it is suggested to model a series of multiple long-range contexts for adaptive selection. Therefore, we propose constructing a larger kernel convolution by *explicitly decomposing* it into a sequence of depth-wise convolutions with a large growing kernel and increasing dilation. Specifically, for the i-th depth-wise convolution, the expansion of the kernel size k, dilation rate d, and the receptive field RF are defined as follows:

$$k_{i-1} \leq k_i;\ d_1 = 1,\ d_{i-1} < d_i \leq RF_{i-1}, \tag{1}$$

$$RF_1 = k_1,\ RF_i = d_i(k_i - 1) + RF_{i-1}. \tag{2}$$

The increasing kernel size and dilation rate ensure that the receptive field expands quickly enough. We set an upper bound on the dilation rate to guarantee that the dilation convolution does not introduce gaps between feature maps. For instance, we can decompose a large kernel into 2 or 3 depth-wise convolutions as in Tab. 2, which have a theoretical receptive field of 23 and 29, respectively.

There are two advantages of the proposed designs. First, it explicitly yields multiple features with various large receptive fields, which makes it easier for the later kernel selection. Second, sequential decomposition is more efficient than simply applying a single larger kernel. As shown in Tab. 2, under the same resulted theoretical receptive field, our decomposition greatly reduces the number of parameters compared to the standard large convolution kernels. To obtain features with rich contextual information from different ranges for input $\mathbf{X}$, a series of decomposed depth-wise convolutions with different receptive fields are applied:

$$\mathbf{U}_0 = \mathbf{X},\quad \mathbf{U}_{i+1} = \mathcal{F}_i^{dw}(\mathbf{U}_i), \tag{3}$$

where $\mathcal{F}_i^{dw}(\cdot)$ are depth-wise convolutions with kernel k_i and dilation d_i. Assuming there are N decomposed kernels, each of which is further processed by a 1×1 convolution layer $\mathcal{F}^{1\times1}(\cdot)$:

$$\widetilde{\mathbf{U}}_i = \mathcal{F}_i^{1\times1}(\mathbf{U}_i),\ \text{for } i \text{ in } [1, N], \tag{4}$$

allowing channel mixing for each spatial feature vector. Then, a selection mechanism is proposed to dynamically select kernels for various objects based on the multi-scale features obtained, which would be introduced next.

3.3. Spatial Kernel Selection

To enhance the network's ability to focus on the most relevant spatial context regions for detecting targets, we use a spatial selection mechanism to spatially select the feature maps from large convolution kernels at different scales. Firstly, we concatenate the features obtained from different kernels with different ranges of receptive field:

$$\widetilde{\mathbf{U}} = [\widetilde{\mathbf{U}}_1; ...; \widetilde{\mathbf{U}}_i], \tag{5}$$

and then efficiently extract the spatial relationship by applying channel-based average and maximum pooling (denoted

as $\mathcal{P}_{avg}(\cdot)$ and $\mathcal{P}_{max}(\cdot)$) to $\widetilde{\mathbf{U}}$:

$$\mathbf{SA}_{avg} = \mathcal{P}_{avg}(\widetilde{\mathbf{U}}),\ \mathbf{SA}_{max} = \mathcal{P}_{max}(\widetilde{\mathbf{U}}), \tag{6}$$

where $\mathbf{SA}_{avg}$ and $\mathbf{SA}_{max}$ are the average and maximum pooled spatial feature descriptors. To allow information interaction among different spatial descriptors, we concatenate the spatially pooled features and use a convolution layer $\mathcal{F}^{2\to N}(\cdot)$ to transform the pooled features (with 2 channels) into N spatial attention maps:

$$\widehat{\mathbf{SA}} = \mathcal{F}^{2\to N}([\mathbf{SA}_{avg}; \mathbf{SA}_{max}]). \tag{7}$$

For each of the spatial attention maps, $\widehat{\mathbf{SA}}_i$, a sigmoid activation function is applied to obtain the individual spatial selection mask for each of the decomposed large kernels:

$$\widetilde{\mathbf{SA}}_i = \sigma(\widehat{\mathbf{SA}}_i), \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The feature maps from the sequence of decomposed large kernels are weighted by their corresponding spatial selection masks and then fused by a convolution layer $\mathcal{F}(\cdot)$ to obtain the attention feature $\mathbf{S}$:

$$\mathbf{S} = \mathcal{F}(\sum_{i=1}^{N} (\widetilde{\mathbf{SA}}_i \cdot \widetilde{\mathbf{U}}_i)). \tag{9}$$

The final output of the LSK module is the element-wise product between the input feature $\mathbf{X}$ and $\mathbf{S}$, similarly in [19, 20, 28]:

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{S}. \tag{10}$$

Fig. 5 shows a detailed conceptual illustration of an LSK module where we intuitively demonstrate how the large selective kernel works by adaptively collecting the corresponding large receptive field for different objects.

4. Experiments

4.1. Datasets

HRSC2016 [42] is a high-resolution remote sensing dataset which is collected for ship detection. It consists of 1,061 images which contain 2,976 instances of ships.

DOTA-v1.0 [68] consists of 2,806 remote sensing images. It contains 188,282 instances of 15 categories: 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).

FAIR1M-v1.0 [58] is a recently published remote sensing dataset that consists of 15,266 high-resolution images and more than 1 million instances. It contains 5 categories and 37 sub-categories objects.

4.2. Implementation Details

In our experiment, we report the results of the oriented object detection models on HRSC2016, DOTA-v1.0 and FAIR1M-v1.0 datasets. To ensure fairness, we follow the same dataset processing approach as other mainstream methods [23, 24, 69]. For DOTA-v1.0 and FAIR1M-v1.0 datasets, we adopt multi-scale training and testing strategy by first rescaling the images into three scales (0.5, 1.0, 1.5), and then cropping each scaled image into 1024×1024 sub-images with a patch overlap of 500 pixels. For the HRSC2016 dataset, we rescale the images by setting the longer side of the image to 800 pixels, without changing their aspect ratios.

During our experiments, the backbones are first pretrained on the ImageNet-1K [10] dataset and then finetuned on the target remote sensing benchmarks. In ablation studies, we adopt the 100-epoch backbone pretraining schedule for experimental efficiency (Tab. 3, 5, 4, 6, 7). We adopt a 300-epoch backbone pretraining strategy to pursue higher accuracy in main results (Tab. 8, 9, 10), similarly to [6, 23, 69, 75]. In main results (Tab. 8, 9), the "Pre." column stands for the dataset on which the networks/backbones are pretrained (IN: Imagenet [10] dataset; CO: Microsoft COCO [36] dataset; MA: Million-AID [43] dataset). Unless otherwise stated, LSKNet is defaulting to be built within the framework of Oriented RCNN [69] due to its compelling performance and efficiency. All the models are trained on the training and validation sets and tested on the testing set. Following [69], we train the models for 36 epochs on the HRSC2016 datasets and 12 epochs on the DOTA-v1.0 and FAIR1M-v1.0 datasets, with the AdamW [44] optimizer. The initial learning rate is set to 0.0004 for HRSC2016, and 0.0002 for the other two datasets, with a weight decay of 0.05. The models are implemented under MMRotate [88] framework. We use 8 RTX3090 GPUs with a batch size of 8 for model training and use a single RTX3090 GPU for testing. All the FLOPs we report in this paper are calculated with a 1024×1024 image input.

4.3. Ablation Study

In this section, we report ablation study results on the DOTA-v1.0 test set to investigate its effectiveness.

Large Kernel Decomposition. Deciding on the number of kernels to decompose is a critical choice for the LSK module. We follow Eq. (1) to configure the decomposed kernels. The results of the ablation study on the number of large kernel decompositions, when the theoretical receptive field is fixed at 29, are shown in Tab. 3. It suggests that decomposing the large kernel into two depth-wise large kernels results in a good trade-off between speed and accuracy, achieving the best performance in terms of both FPS (frames per second) and mAP (mean average precision).

Receptive Field Size and Selection Type. Based on our

(k, d) sequence	RF	Num.	FPS	mAP (%)
(29, 1)	29	1	18.6	80.66
(5, 1) $\longrightarrow$ (7, 4)	29	2	**20.5**	**80.91**
(3, 1) $\longrightarrow$ (5, 2) $\longrightarrow$ (7, 3)	29	3	19.2	80.77

Table 3. **The effects of the number of decomposed large kernels** on the inference FPS and mAP, given the theoretical receptive field being 29. We adopt LSKNet-T backbones pretrained on ImageNet for 100 epochs. Decomposing the large kernel into two depth-wise kernels achieves the best performance of speed and accuracy.

(k_1, d_1)	(k_2, d_2)	CS	SS	RF	FPS	mAP (%)
(3, 1)	(5, 2)	-	-	11	22.1	80.80
(5, 1)	(7, 3)	-	-	23	21.7	80.94
(5, 1)	(7, 4)	-	-	29	20.5	80.91
(7, 1)	(9, 4)	-	-	39	21.3	80.84
(5, 1)	(7, 3)	✓	-	23	19.6	80.57
(5, 1)	(7, 3)	-	✓	23	20.7	**81.31**

Table 4. **The effectiveness of the key design components** of the LSKNet when the large kernel is decomposed into a sequence of two depth-wise kernels. CS: channel selection (likewise in SKNet [33]); SS: spatial selection **(ours)**. We adopt LSKNet-T backbones pretrained on ImageNet for 100 epochs. The LSKNet achieves the best performance when using a reasonably large receptive field with spatial selection.

Pooling		FPS	mAP (%)
Max.	Avg.		
✓		20.7	81.23
	✓	20.7	81.12
✓	✓	20.7	**81.31**

Table 5. Ablation study on the effectiveness of the **maximum and average pooling in spatial selection** of our proposed LSK module. We adopt LSKNet-T backbones pretrained on ImageNet for 100 epochs. The best result is obtained when using both.

evaluations presented in Tab. 3, we find that the optimal solution for our proposed LSKNet is to decompose the large kernel into two depth-wise kernels in series. Furthermore, Tab. 4 shows that excessively small or large receptive fields can hinder the performance of the LSKNet, and a receptive field size of approximately 23 is determined to be the most effective. In addition, our experiments indicate that the proposed spatial selection approach is more effective than channel attention (similarly in SKNet [33]) for remote sensing object detection tasks. It suggests that in detection tasks, spatial information is more critical.

Pooling Layers in Spatial Selection. We conduct experiments to determine the optimal pooling layers for spatial selection, as reported in Tab. 5. The results suggest that using both max and average pooling in the spatial selection component of our LSK module provides the best performance without sacrificing inference speed.

Performance of LSKNet backbone under different detection frameworks. To validate the generality and effectiveness of our proposed LSKNet backbone, we evaluate its performance under various remote sensing detection frameworks, including two-stage frameworks O-RCNN [69] and RoI Transformer [12] as well as one-stage frameworks S^2A-Net [23] and R3Det [75]. The re-

Frameworks	ResNet-18	⋆ LSKNet-T
ORCNN [69]	79.27	81.31 **(+2.04)**
RoI Trans. [12]	78.32	80.89 **(+2.57)**
S^2A-Net [23]	76.82	80.15 **(+3.33)**
R3Det [75]	74.16	78.39 **(+4.23)**
#P (backbone only)	11.2M	4.3M **(-62%)**
FLOPs (backbone only)	38.1G	19.1G **(-50%)**

Table 6. **Comparison of LSKNet-T and ResNet-18** as backbones with different detection frameworks on DOTA-v1.0. The LSKNet-T backbone is pretrained on ImageNet for 100 epochs. The lightweight LSKNet-T achieves significantly higher mAP in various frameworks than ResNet-18.

Group	Model (backbone only)	**#P**	**FLOPs**	**mAP (%)**
Baseline	ResNet-18	11.2M	38.1G	79.27
Large Kernel	VAN-B1 [19]	13.4M	52.7G	81.15
	ConvNeXt V2-N [65]	15.0M	51.2G	80.81
	MSCAN-S [20]	13.1M	45.0G	81.12
Selective Attention	SKNet-26 [33]	14.5M	58.5G	80.67
	ResNeSt-14 [84]	8.6M	57.9G	79.51
	SCNet-18 [37]	14.0M	50.7G	79.69
Ours	⋆ LSKNet-S	14.4M	54.4G	**81.48**
Prev Best	CSPNeXt [46]	26.1M	87.6G	81.33

Table 7. **Comparison on LSKNet-S and other (large kernel/selective attention) backbones** under O-RCNN [69] framework on DOTA-v1.0, except that the Prev Best is under RTMDet [46] framework. All backbones are pretrained on ImageNet for 100 epochs. Our LSKNet achieves the best mAP under similar complexity budgets.

Method	Pre.	**mAP(07)↑**	**mAP(12)↑**	**#P ↓**	**FLOPs ↓**
DRN [50]	IN	-	92.70	-	-
CenterMap [61]	IN	-	92.80	41.1M	198G
Rol Trans. [12]	IN	86.20	-	55.1M	200G
G. V. [71]	IN	88.20	-	41.1M	198G
R3Det [75]	IN	89.26	96.01	41.9M	336G
DAL [48]	IN	89.77	-	36.4M	216G
GWD [77]	IN	89.85	97.37	47.4M	456G
S^2ANet [23]	IN	90.17	95.01	38.6M	198G
AOPG [6]	IN	90.34	96.22	-	-
ReDet [24]	IN	90.46	97.63	31.6M	-
O-RCNN [69]	IN	90.50	97.60	41.1M	199G
RTMDet [46]	CO	90.60	97.10	52.3M	205G
⋆ LSKNet-T	IN	90.54	98.13	**21.0M**	**124G**
⋆ LSKNet-S	IN	**90.65**	**98.46**	31.0M	161G

Table 8. Comparison with state-of-the-art methods on the **HRSC2016** dataset. The LSKNet-S backbone is pretrained on ImageNet for 300 epochs, the same with most compared methods [23, 69, 75]. mAP (07/12): VOC 2007 [15]/2012 [16] metrics.

Method	Pre.	mAP ↑	#P ↓	FLOPs ↓	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
One-stage																			
R3Det [75]	IN	76.47	41.9M	336G	89.80	83.77	48.11	66.77	78.76	83.27	87.84	90.82	85.38	85.51	65.57	62.68	67.53	78.56	72.62
CFA [22]	IN	76.67	-	-	89.08	83.20	54.37	66.87	81.23	80.96	87.17	90.21	84.32	86.09	52.34	69.94	75.52	80.76	67.96
DAFNe [31]	IN	76.95	-	-	89.40	86.27	53.70	60.51	82.04	81.17	88.66	90.37	83.81	87.27	53.93	69.38	75.61	81.26	70.86
SASM [27]	IN	79.17	-	-	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
AO2-DETR [9]	IN	79.22	74.3M	304G	89.95	84.52	56.90	74.83	80.86	83.47	88.47	90.87	86.12	**88.55**	63.21	65.09	79.09	**82.88**	73.46
S^2ANet [23]	IN	79.42	-	-	88.89	83.60	57.74	81.95	79.94	83.19	**89.11**	90.78	84.87	87.81	70.30	68.25	78.30	77.01	69.58
R3Det-GWD [77]	IN	80.23	41.9M	336G	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
RTMDet-R [46]	IN	80.54	52.3M	205G	88.36	84.96	57.33	80.46	80.58	84.88	88.08	**90.90**	86.32	87.57	69.29	70.61	78.63	80.97	**79.24**
R3Det-KLD [79]	IN	80.63	41.9M	336G	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
RTMDet-R [46]	CO	81.33	52.3M	205G	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
Two-stage																			
SCRDet [78]	IN	72.61	-	-	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21
Rol Trans. [12]	IN	74.61	55.1M	200G	88.65	82.60	52.53	70.87	77.93	76.67	86.87	90.71	83.83	82.51	53.95	67.61	74.67	68.75	61.03
G.V. [71]	IN	75.02	41.1M	198G	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32
CenterMap [61]	IN	76.03	41.1M	198G	89.83	84.41	54.60	70.25	77.66	78.32	87.19	90.66	84.89	85.27	56.46	69.23	74.13	71.56	66.06
CSL [76]	IN	76.17	37.4M	236G	**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
ReDet [24]	IN	80.10	-	-	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
DODet [7]	IN	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 [6]	IN	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
O-RCNN [69]	IN	80.87	41.1M	199G	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
KFloU [80]	IN	80.93	58.8M	206G	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
RVSA [60]	MA	81.24	114.4M	414G	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-T (**ours**)	IN	81.37	**21.0M**	**124G**	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
⋆ LSKNet-S (**ours**)	IN	81.64	31.0M	161G	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
⋆ LSKNet-S* (**ours**)	IN	**81.85**	31.0M	161G	89.69	85.70	61.47	83.23	81.37	86.05	88.64	90.88	88.49	87.40	71.67	71.35	79.19	81.77	**80.86**

Table 9. Comparison with state-of-the-art methods on the **DOTA-v1.0** dataset with multi-scale training and testing. The LSKNet backbones are pretrained on ImageNet for 300 epochs, similarly to [23, 69, 75]. *: With EMA finetune similarly to the compared methods [46].

Model	G. V.* [71]	RetinaNet* [35]	C-RCNN* [2]	F-RCNN* [56]	RoI Trans.* [12]	O-RCNN [69]	⋆ LSKNet-T	⋆ LSKNet-S
mAP(%)	29.92	30.67	31.18	32.12	35.29	45.60	46.93	**47.87**

Table 10. Comparison with state-of-the-art methods on the **FAIR1M-v1.0** dataset. The LSKNet backbones are pretrained on ImageNet for 300 epochs, similarly to [23, 69, 75]. *: Results are referenced from FAIR1M paper [58].

sults in Tab. 6 show that our proposed LSKNet-T backbone significantly improves detection performance compared to ResNet-18, while using only 38% of its parameters and with 50% fewer FLOPs.

Comparison with Other Large Kernel/Selective Attention Backbones. We also compare our LSKNet with 6 popular high-performance backbone models with large kernels or selective attention. As shown in Tab. 7, under similar model sizes and complexity budgets, our LSKNet outperforms all other models on the DOTA-v1.0 dataset.

4.4. Main Results

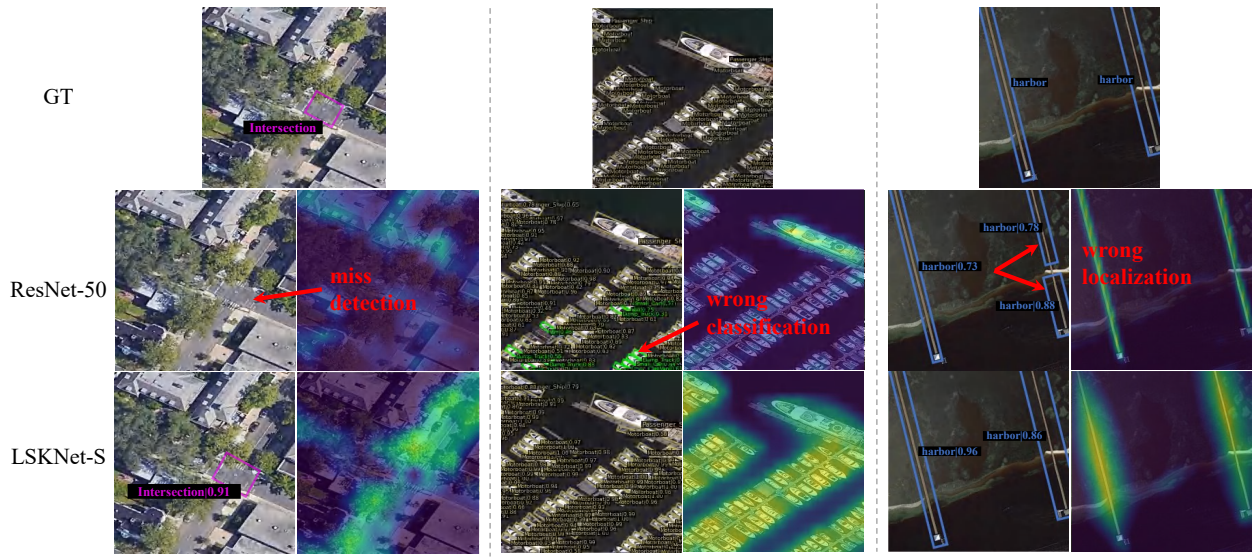
Results on HRSC2016. We evaluated the performance of our LSKNet against 12 state-of-the-art methods on the HRSC2016 dataset. The results presented in Tab. 8 demonstrate that our LSKNet-S outperforms all other methods with an mAP of **90.65%** and **98.46%** under the PASCAL VOC 2007 [15] and VOC 2012 [16] metrics, respectively.

Results on DOTA-v1.0. We compare our LSKNet with 20 state-of-the-art methods on the DOTA-v1.0 dataset, as reported in Tab. 9. Our LSKNet-T, LSKNet-S and LSKNet-S* achieve state-of-the-art with mAP of **81.37%**, **81.64%** and **81.85%** respectively. Notably, our high-performing LSKNet-S reaches an inference speed of **18.1** FPS on 1024x1024 images with a single RTX3090 GPU.

Results on FAIR1M-v1.0. We compare our LSKNet against 6 other models on the FAIR1M-v1.0 dataset, as shown in Tab. 10. The results reveal that our LSKNet-T and LSKNet-S perform exceptionally well, achieving state-of-the-art mAP scores of **46.93%** and **47.87%** respectively, surpassing all other models by a significant margin.

4.5. Analysis

Detection Results Visualization. Visualization examples of detection results and Eigen-CAM [49] are shown in Fig. 6. LSKNet can capture much more context information relevant to the detected targets, leading to better performance in various hard cases, which justifies our *prior (1)*.



(a) Robustness to obstacles (b) Accurate Object Classification (c) Accurate Object Localization

Figure 6. **Eigen-CAM visualization** of Oriented RCNN detection framework with ResNet-50 and LSKNet-S. Our proposed LSKNet can model a much long range of context information, leading to better performance in various hard cases.

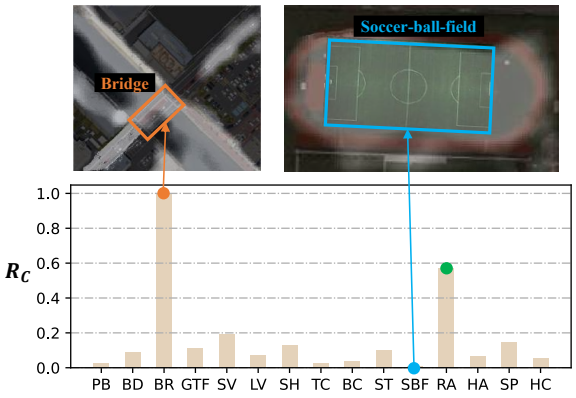


Figure 7. Normalised **Ratio R_c of Expected Selective RF Area and GT Bounding Box Area** for object categories in DOTA-v1.0. The relative range of context required for different object categories varies a lot. Examples of Bridge and Soccer-ball-field are given, where the visualized receptive field is obtained from Eq. (8) (i.e., the spatial activation) of our well-trained LSKNet model.

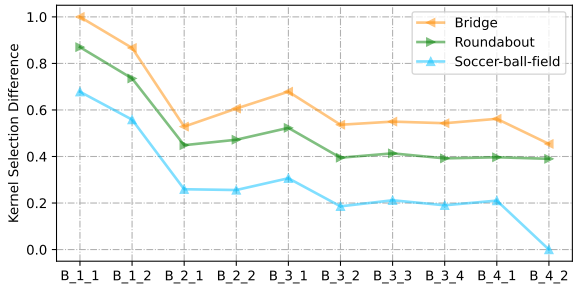


Figure 8. Normalised **Kernel Selection Difference** in the LSKNet-T blocks for Bridge, Roundabout and Soccer-ball-field. B_i_j represents the j-th LSK block in stage i. A greater value is indicative of a dependence on a broader context.

Relative Context Range for Different Objects. To investigate the relative range of receptive field for each object category, we define R_c as the *Ratio of Expected Selective RF Area and GT Bounding Box Area* for category c:

$$R_c = \frac{\sum_{i=1}^{I_c} A_i / B_i}{I_c}, \tag{11}$$

$$A_i = \sum_{d=1}^{D} \sum_{n=1}^{N} |\widetilde{\mathbf{SA}}_n^d \cdot RF_n|, \; B_i = \sum_{j=1}^{J_i} Area(\mathrm{GT}_j), \tag{12}$$

where I_c is the number of images that contain the object category c only. The A_i is the sum of spatial selection activation in all LSK blocks for input image i, where D is the number of blocks in an LSKNet, and N is the number of decomposed large kernels in an LSK module. B_i is the total pixel area of all J_i annotated oriented object bounding boxes (GT). We plot the normalized R_c in Fig. 7 which represents the relative range of context required for different object categories for a better view.

The results suggest that the Bridge category stands out as requiring a greater amount of additional contextual information compared to other categories, primarily due to its similarity in features with roads and the necessity of contextual clues to ascertain whether it is enveloped by water. Conversely, the Court categories, such as Soccer-ball-field, necessitate minimal contextual information owing to their distinctive textural attributes, specifically the court bound-

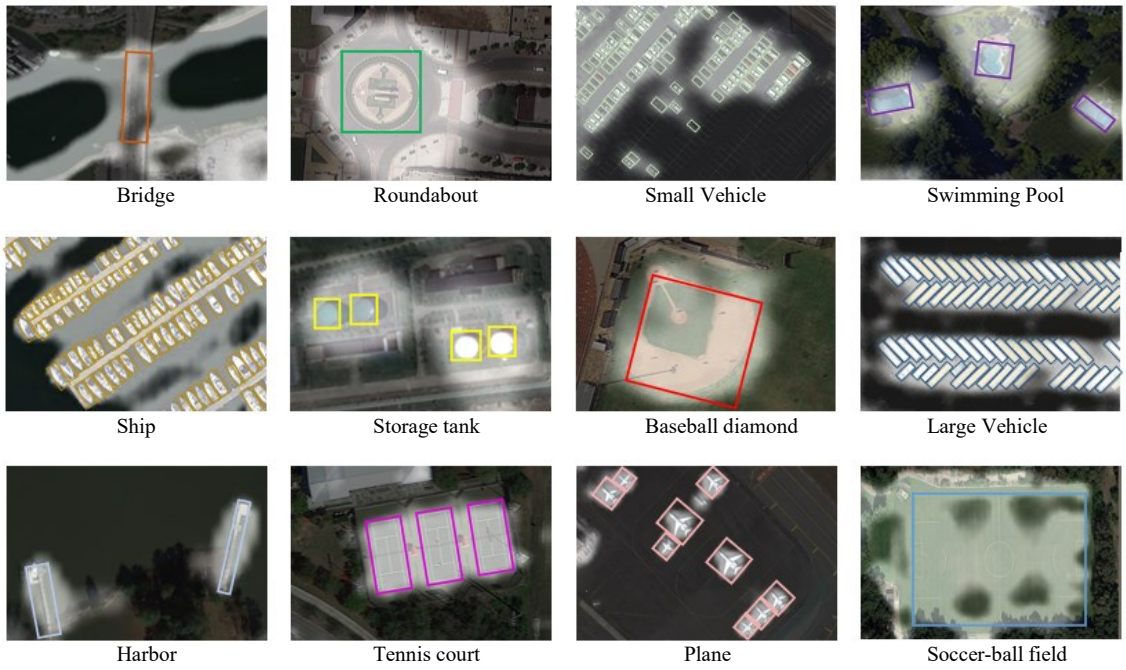


Figure 9. Receptive field activation for more object categories in DOTA-v1.0, where the activation map is obtained from the Eq. (8) (i.e., the spatial activation) of our well-trained LSKNet model.

ary lines. It aligns with our knowledge and further supports *prior (2)* that the relative range of contextual information required for different object categories varies greatly.

Kernel Selection Behaviour. We further investigate the kernel selection behaviour in our LSKNet. For object category c, the *Kernel Selection Difference* ΔA_c (i.e., larger kernel selection - smaller kernel selection) of an LSKNet-T block is defined as:

$$\Delta A_c = |\widetilde{\mathbf{SA}}_{larger} - \widetilde{\mathbf{SA}}_{smaller}| \,. \tag{13}$$

We demonstrate the normalised ΔA_c over all images for three typical categories: Bridge, Roundabout and Soccer-ball-field and for each LSKNet-T block in Fig. 8. As expected, the participation of larger kernels of all blocks for Bridge is higher than that of Roundabout, and Roundabout is higher than Soccer-ball-field. This aligns with the common sense that Soccer-ball-field indeed does not require a large amount of context, since its own texture characteristics are already sufficiently distinct and discriminatory.

We also surprisingly discover another selection pattern of LSKNet across network depth: LSKNet usually utilizes larger kernels in its shallow layers and smaller kernels in higher levels. This indicates that networks tend to quickly focus on capturing information from large receptive fields in low-level layers so that higher-level semantics can contain sufficient receptive fields for better discrimination.

Spatial Activation Visualisations. Spatial activation map examples for more object categories in DOTA-v1.0 are shown in Fig. 9, where the activation map is obtained from Eq. (8) (i.e., the spatial activation) of our well-trained LSKNet model. The object categories are arranged in decreasing order from top left to bottom right based on the *Ratio of Expected Selective RF Area and GT Bounding Box Area* as illustrated in Fig. 7. The spatial activation visualisation results also demonstrate that the model's behaviour aligns with our proposed two priors and the above analysis, which in turn verifies the effectiveness of the proposed mechanism.

5. Conclusion

In this paper, we propose the Large Selective Kernel Network (LSKNet) for remote sensing object detection tasks, which is designed to utilize the inherent characteristics in remote sensing images: the need for a wider and adaptable contextual understanding. By adapting its large spatial receptive field, LSKNet can effectively model the varying contextual nuances of different object types. Extensive experiments demonstrate that our proposed lightweight model achieves state-of-the-art performance on competitive remote sensing benchmarks.

Acknowledgement. This research was supported by the NSFC (NO. 62176130, 62206134, 62225604) and the Fundamental Research Funds for the Central Universities (Nankai University, 070-63233084, 070-63233089). Computation is supported by the Supercomputing Center of Nankai University.

References

[1] Yakoub Bazi, Laila Bashmal, Mohamad M. Al Rahhal, Reham Al Dayil, and Naif Al Ajlan. Vision transformers for remote sensing image classification. *Remote Sensing*, 2021. 2

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018. 7

[3] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In *ICCV Workshops*, 2019. 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[5] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *CVPR*, 2020. 3

[6] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *TGARS*, 2022. 2, 5, 6, 7

[7] Gong Cheng, Yanqing Yao, Shengyang Li, Ke Li, Xingxing Xie, Jiabao Wang, Xiwen Yao, and Junwei Han. Dual-aligned oriented detector. *TGARS*, 2022. 7

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 3

[9] Linhui Dai, Hong Liu, Hao Tang, Zhiwei Wu, and Pinhao Song. AO2-DETR: Arbitrary-oriented object detection transformer. *TCSVT*, 2022. 2, 7

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[11] Peifang Deng, Kejie Xu, and Hong Huang. When cnns meet vision transformer: A joint framework for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 2022. 2

[12] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning RoI transformer for oriented object detection in aerial images. In *CVPR*, 2019. 1, 2, 6, 7

[13] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *CVPR*, 2022. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR 2021*, 2021. 2

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. 6, 7

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. 6, 7

[17] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *Visual Intelligence*, 2023. 2

[18] Shanghua Gao, Zhong-Yu Li, Qi Han, Ming-Ming Cheng, and Liang Wang. Rf-next: Efficient receptive field search for convolutional neural networks. *IEEE TPAMI*, 45(3):2984–3002, 2023. 2

[19] Meng-Hao Guo, Chengrou Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shiyong Hu. Visual attention network. *ArXiv*, 2022. 3, 5, 6

[20] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. In *NeurIPS*, 2022. 3, 5, 6

[21] Meng-Hao Guo, Tianxing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 11 2021. 3

[22] Zonghao Guo, Chang Liu, Xiaosong Zhang, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In *CVPR*, 2021. 7

[23] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *TGARS*, 2020. 2, 5, 6, 7

[24] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. ReDet: A rotation-equivariant detector for aerial object detection. In *CVPR*, 2021. 5, 6, 7

[25] Siyuan Hao, Bin Wu, Kun Zhao, Yuanxin Ye, and Wei Wang. Two-stream swin transformer with differentiable sobel operator for remote sensing image classification. *Remote Sensing*, 2022. 2

[26] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, 2016. 4

[27] Liping Hou, Ke Lu, Jian Xue, and Yuqiu Li. Shape-adaptive selection and measurement for oriented object detection. In *AAAI*, 2022. 7

[28] Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2former: A simple transformer-style ConvNet for visual recognition. *ArXiv*, 2022. 3, 5

[29] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In *NeurPIS*, 2018. 3

[30] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3

[31] Steven Lang, Fabrizio Ventola, and Kristian Kersting. Dafne: A one-stage anchor-free deep model for oriented object detection. *CoRR*, 2021. 7

[32] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 2

[33] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *CVPR*, 2019. 3, 6

[34] Yuxuan Li, Xiang Li, and Jian Yang. Spatial group-wise enhance: Enhancing semantic feature learning in cnn. In *ACCV*, 2022. 3

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 7

[36] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[37] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Changhu Wang, and Jiashi Feng. Improving convolutional networks with self-calibrated convolutions. In *CVPR*, 2020. 3, 6

[38] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Mykola Pechenizkiy, Decebal Mocanu, and Zhangyang Wang. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *ArXiv*, 2022. 2

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[40] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 2

[41] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 3

[42] Zikun Liu, Hongzhen Wang, Lubin Weng, and Yiping Yang. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE Geoscience and Remote Sensing Letters*, 2016. 5

[43] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021. 5

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ArXiv*, 2017. 5

[45] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *NeurIPS*, 2016. 2

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

[47] Jie Mei, Yi-Bo Zheng, and Ming-Ming Cheng. D2anet: Difference-aware attention network for multi-level change detection from satellite imagery. *Computational Visual Media*, 2023. 1

[48] Qi Ming, Zhiqiang Zhou, Lingjuan Miao, Hongwei Zhang, and Linhao Li. Dynamic anchor learning for arbitrary-oriented object detection. *CoRR*, 2020. 6

[49] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. *CoRR*, 2020. 7

[50] Xingjia Pan, Yuqiang Ren, Kekai Sheng, Weiming Dong, Haolei Yuan, Xiaowei Guo, Chongyang Ma, and Changsheng Xu. Dynamic refinement network for oriented and densely packed object detection. In *CVPR*, 2020. 2, 6

[51] Teerapong Panboonyuen, Kulsawasd Jitkajornwanich, Siam Lawawirojwong, Panu Srestasathiern, and Peerapon Vateekul. Transformer-based decoder designs for semantic segmentation on remotely sensed images. *Remote Sensing*, 2021. 2

[52] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In-So Kweon. Bam: Bottleneck attention module. In *British Machine Vision Conference*, 2018. 3

[53] Malsha V. Perera, Wele Gedara Chaminda Bandara, Jeya Maria Jose Valanarasu, and Vishal M. Patel. Transformer-based SAR image despeckling. *CoRR*, 2022. 2

[54] Wen Qian, Xue Yang, Silong Peng, Junchi Yan, and Yue Guo. Learning modulated loss for rotated object detection. In *AAAI*, 2021. 1, 2

[55] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 2

[56] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2, 7

[57] Xian Sun, Yu Tian, Wanxuan Lu, Peijin Wang, Ruigang Niu, Hongfeng Yu, and Kun Fu. From single- to multi-modal remote sensing imagery interpretation: a survey and taxonomy. *Science China Information Sciences*, 2023. 1

[58] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, Martin Weinmann, Stefan Hinz, Cheng Wang, and Kun Fu. FAIR1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022. 5, 7

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2

[60] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer towards remote sensing foundation model. *TGARS*, 2022. 2, 7

[61] Jinwang Wang, Wen Yang, Heng-Chao Li, Haijian Zhang, and Gui-Song Xia. Learning center probability map for detecting objects in aerial images. *TGARS*, 2021. 2, 6, 7

[62] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 2

[63] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *CVM*, 2022. 3

[64] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482, 2023. 2

[65] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In-So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *Arxiv*, 2023. 6

[66] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, 2018. 3

[67] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. P2T: Pyramid pooling transformer for scene understanding. *IEEE TPAMI*, pages 1–12, 2022. 2

[68] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *CVPR*, 2018. 5

[69] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for object detection. In *ICCV*, 2021. 1, 2, 5, 6, 7

[70] Xiangkai Xu, Zhejun Feng, Changqing Cao, Mengyuan Li, Jin Wu, Zengyan Wu, Yajie Shang, and Shubing Ye. An improved swin transformer-based model for remote sensing object detection and instance segmentation. *Remote Sensing*, 2021. 2

[71] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE TPAMI*, 2021. 1, 2, 6, 7

[72] Haotian Yan, Zhe Li, Weijian Li, Changhu Wang, Ming Wu, and Chuang Zhang. Contnet: Why not use convolution and transformer at the same time? *CoRR*, 2021. 2

[73] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *NeurIPS*, 2019. 3

[74] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *CVPR Workshops*, 2017. 2

[75] Xue Yang, Qingqing Liu, Junchi Yan, and Ang Li. R3det: Refined single-stage detector with feature refinement for rotating object. *CoRR*, 2019. 1, 2, 5, 6, 7

[76] Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *ECCV*, 2020. 7

[77] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. In *ICML*, 2021. 1, 6, 7

[78] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. SCRDet: Towards more robust detection for small, cluttered and rotated objects. In *ICCV*, 2019. 2, 7

[79] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. In *NeurIPS*, 2021. 1, 7

[80] Xue Yang, Yue Zhou, Gefan Zhang, Jirui Yang, Wentao Wang, Junchi Yan, Xiaopeng Zhang, and Qi Tian. The KFIoU loss for rotated object detection. In *ICLR*, 2022. 7

[81] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *CVPR*, 2022. 3

[82] Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Asghar, and Brian Lee. A survey of modern deep learning based object detection models. *Digital Signal Processing*, 2022. 1

[83] Cui Zhang, Liejun Wang, Shuli Cheng, and Yongming Li. Swinsunet: Pure transformer network for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2

[84] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. In *CVPR Workshops*, 2022. 3, 6

[85] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 2

[86] Zhaohui Zheng, Rongguang Ye, Qibin Hou, Dongwei Ren, Ping Wang, Wangmeng Zuo, and Ming-Ming Cheng. Localization distillation for object detection. *IEEE TPAMI*, 2023. 2

[87] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *ArXiv*, 2019. 2

[88] Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, et al. Mmrotate: A rotated object detection benchmark using pytorch. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7331–7334, 2022. 5

[89] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019. 3