# Large Selective Kernel Network for Remote Sensing Object Detection

# 大选择性核网络用于遥感目标检测

Yuxuan Li, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang and Xiang Li\* IMPlus@PCALab & TMCC, CS, Nankai University

李宇轩，侯启彬，郑兆辉，程明明，杨健，李祥\* IMPlus@PCALab & TMCC，计算机科学与技术，南开大学

yuxuan.li.17@ucl.ac.uk, andrewhoux@gmail.com,

yuxuan.li.17@ucl.ac.uk, andrewhoux@gmail.com,

{zh\_zheng, cmm, csjyang, xiang.li.implus }@nankai.edu.cn

{zh\_zheng, cmm, csjyang, xiang.li.implus}@nankai.edu.cn

# Abstract

# 摘要

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, and the long-range context required by different types of objects can vary. In this paper, we take these priors into account and propose the 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 the best of our knowledge, this is the first time that large and selective kernel mechanisms have been explored in the field of remote sensing object detection. Without bells and whistles, LSKNet sets new state-of-the-art scores on standard benchmarks, i.e., HRSC2016 (98.46% mAP), DOTA-v1.0 (81.85% mAP) and FAIRIM-v1.0 (47.87% mAP). Based on a similar technique, we rank 2nd place in 2022 the Greater Bay Area International Algorithm Competition. Code is available at https://github.com/zcablii/Large-Selective-Kernel-Network.

最近关于遥感目标检测的研究主要集中在提高定向边界框的表示，但忽略了遥感场景中呈现的独特先验知识。这种先验知识可能很有用，因为如果没有参考足够长距离的上下文，微小的遥感目标可能会被错误地检测到，而且不同类型的目标所需的长距离上下文可能会有所不同。在本文中，我们考虑这些先验知识，并提出了大型选择性核网络（LSKNet）。LSKNet能够动态调整其大空间感受野，以更好地模拟遥感场景中各种目标的范围上下文。据我们所知，这是第一次在遥感目标检测领域探索大型和选择性核机制。LSKNet没有使用任何花哨的技术，就在标准基准测试中设置了新的最先进分数，即HRSC2016（98.46%mAP），DOTA-v1.0（81.85%mAP）和FAIRIM-v1.0（47.87%mAP）。基于类似技术，我们在2022年粤港澳大湾区国际算法竞赛中排名第2。代码可在 https://github.com/zcablii/Large-Selective-Kernel-Network 找到。

# 1. Introduction

# 1. 引言

Remote sensing object detection [75] is a field of computer vision that focuses on identifying and locating objects of interest in aerial images, such as vehicles or aircraft. In recent years, one mainstream trend is to generate bounding boxes that accurately fit the orientation of the objects being detected, rather than simply drawing horizontal boxes around them. Consequently, a significant amount of 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, such as RoI Transformer [12], Oriented R-CNN [62] and R3Det [68], as well as techniques for oriented box encoding, such as gliding vertex [64] and midpoint offset box encoding [62]. Additionally, a number of loss functions, including GWD [70], KLD [72] and Modulated Loss [50], have been proposed to further enhance the performance of these approaches.

遥感目标检测 [75] 是计算机视觉领域的一个分支，专注于在航空图像中识别和定位感兴趣的目标，如车辆或飞机。近年来，一种主流趋势是生成与被检测物体方向准确匹配的边界框，而不仅仅是围绕它们绘制水平框。因此，大量研究集中在改善遥感目标检测中定向边界框的表现。这一目标在很大程度上通过开发专门的检测框架实现，如RoI Transformer [12]、Oriented R-CNN [62] 和 R3Det [68]，以及用于定向框编码的技术，如滑动顶点 [64] 和中点偏移框编码 [62]。此外，还提出了多种损失函数，包括GWD [70]、KLD [72] 和调制损失 [50]，以进一步提高这些方法的性能。

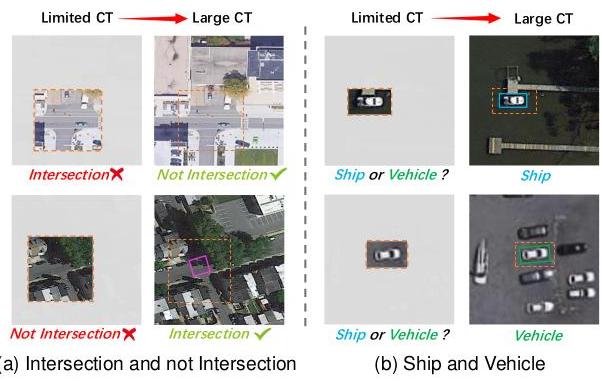


Figure 1: Successfully detecting remote sensing objects requires the use of a wide range of contextual information. Detectors with a limited receptive field may easily lead to incorrect detection results. "CT" stands for Context.

图1：成功检测遥感物体需要使用广泛的环境信息。具有有限感受野的检测器可能会导致错误的检测结果。"CT" 代表上下文。

However, despite these advances, relatively few works have taken into account the strong prior knowledge that exists in remote sensing images. Aerial images are typically captured from a bird’s eye view at high resolutions. In particular, most objects in aerial images may be small in size and difficult to identify based on their appearance alone. Instead, the successful recognition of these objects often relies on their context, as the surrounding environment can provide valuable clues about their shape, orientation, and other characteristics. According to an analysis of mainstream remote sensing datasets, we identify two important priors:

然而，尽管取得了这些进展，相对较少的工作考虑到了遥感图像中存在的强烈先验知识。航空图像通常以鸟瞰视角高分辨率捕获。特别是，航空图像中的大多数物体可能尺寸较小，仅凭外观难以识别。相反，成功识别这些物体通常依赖于它们的上下文，因为周围环境可以提供关于它们的形状、方向和其他特征的宝贵线索。根据对主流遥感数据集的分析，我们确定了两个重要的先验知识：

(1) Accurate detection of objects in remote sensing images often requires a wide range of contextual information. As illustrated in Fig. 1(a), the limited context used by object detectors in remote sensing images can often lead to incorrect classifications. In the upper image, for example, the detector may classify the junction as an intersection due to its typical characteristics, but in reality, it is not an intersection. Similarly, in the lower image, the detector may classify the junction as not being an intersection due to the presence of large trees, but again, this is incorrect. These errors can occur because the detector is only considering a limited amount of contextual information in the immediate vicinity of the objects. A similar scenario can be also observed in the example of ships and vehicles in Fig. 1(b).

(1) 在遥感图像中准确检测目标通常需要大量的上下文信息。如图1(a)所示，遥感图像中目标检测器使用的有限上下文信息常常会导致错误的分类。例如，在上方图像中，检测器可能因为典型的特征将交叉口分类为交叉路口，但实际上它并不是交叉路口。同样，在下方图像中，由于有大树的存在，检测器可能将交叉口分类为非交叉路口，但这也是错误的。这些错误可能是因为检测器仅考虑了目标附近有限的上下文信息。类似的情况也可以在图1(b)中的船舶和车辆的例子中观察到。

\*Corresponding author. Team site: https://github.com/IMPlus-PCALab

\*通讯作者。团队网站：https://github.com/IMPlus-PCALab

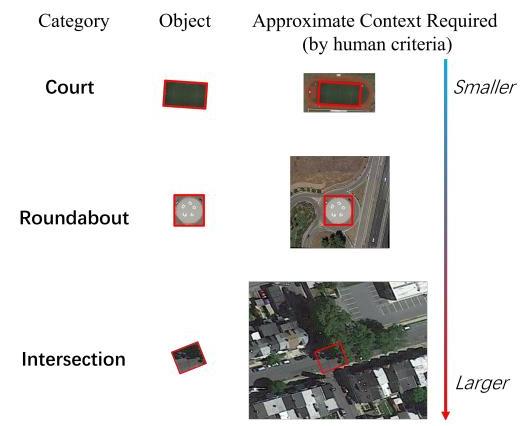


Figure 2: The wide range of contextual information required for different object types is very different by human criteria. The objects with red boxes are the exact ground-truth annotations.

图2：根据人类标准，不同类型的目标所需的上下文信息范围差异很大。带有红色框的物体是确切的地面真实标注。

(2) The wide range of contextual information required for different object types is very different. As shown in Fig. 2, the amount of contextual information required for accurate object detection in remote sensing images can vary significantly depending on the type of object being detected. For example, Soccer-ball-field may require relatively less extra contextual information because of the unique distinguishable court borderlines. In contrast, Roundabouts may require a larger range of context information in order to distinguish between gardens and ring-like buildings. Intersections, especially those that are partially covered by trees, often require an extremely large receptive field due to the long-range dependencies between the intersecting roads. This is because the presence of trees and other obstructions can make it difficult to identify the roads and the intersection itself based on appearance alone. Other object categories, such as bridges, vehicles, and ships, may also require different scales of the receptive field in order to be accurately detected and classified.

(2) 不同类型的对象所需的大量上下文信息差异很大。如图2所示，在遥感图像中进行准确的目标检测所需的上下文信息量可能因检测的对象类型而显著变化。例如，足球场可能需要相对较少的额外上下文信息，因为其独特的场地边界线易于辨识。相比之下，环形交通岛可能需要更广泛的上下文信息，以区分花园和环状建筑物。交叉口，尤其是部分被树木覆盖的交叉口，通常需要一个极大的感受野，因为相交道路之间存在长距离依赖性。这是因为树木和其他障碍物的存在可能使得仅凭外观难以识别道路和交叉口本身。其他对象类别，如桥梁、车辆和船舶，也可能需要不同尺度的感受野才能被准确检测和分类。

To address the challenge of accurately detecting objects in remote sensing images, which often require a wide and dynamic range of contextual information, we propose a novel approach called Large Selective Kernel Network (LSKNet). Our approach involves dynamically adjusting the receptive field of the feature extraction backbone in order to more effectively process the varying wide context of the objects being detected. This is achieved through a spatial selective mechanism, which weights the features processed by a sequence of large depth-wise kernels efficiently and then spatially merges 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.

为了解决在遥感图像中准确检测目标的问题，这些问题通常需要广泛且动态的上下文信息，我们提出了一种名为大型选择性核网络（LSKNet）的新方法。我们的方法涉及动态调整特征提取基础网络的特征感受野，以更有效地处理检测对象变化的广泛上下文。这是通过空间选择性机制实现的，该机制有效地对一系列大深度核处理后的特征进行加权，并在空间上进行合并。这些核的权重是基于输入动态确定的，允许模型自适应地使用不同的大型核，并根据需要为每个目标在空间中调整感受野。

To the best of our knowledge, our proposed LSKNet is the first to investigate and discuss the use of large and selective kernels for remote sensing object detection. Despite its simplicity, our model achieves state-of-the-art performance on three popular datasets: HRSC2016 (98.46% mAP), DOTA-v1.0 (81.64% mAP), and FAIR1M-v1.0 (47.87% mAP), surpassing previously published results. Furthermore, we demonstrate that our model’s behaviour exactly aligns with the aforementioned two priors, which in turn verifies the effectiveness of the proposed mechanism.

就我们的知识所及，我们提出的LSKNet是首个研究并讨论在遥感目标检测中使用大尺寸和选择性核的。尽管模型简单，我们的模型在三个流行的数据集上实现了最先进的性能：HRSC2016（98.46%mAP）、DOTA-v1.0（81.64%mAP）和FAIR1M-v1.0（47.87%mAP），超过了之前发表的结果。此外，我们证明我们的模型的行为与上述两个先验完全一致，这反过来验证了所提出机制的有效性。

# 2. Related Work

# 2. 相关工作

# 2.1. Remote Sensing Object Detection Framework

# 2.1. 遥感目标检测框架

High-performance remote sensing object detectors often rely on the RCNN [52] framework, which consists of a region proposal network and regional CNN detection heads. 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 [71] uses an attention mechanism to reduce background noise and improve the modelling of crowded and small objects. Oriented RCNN [62] and Gliding Vertex [64] introduce new box encoding systems to address the instability of training losses caused by rotation angle periodicity. Some approaches treat remote sensing detection as a point detection task [67], providing an alternative way of addressing remote sensing detection problems.

高性能的遥感目标检测器通常依赖于RCNN[52]框架，该框架包括一个区域提议网络和区域CNN检测头。近年来，已经提出了RCNN框架的几种变体。两阶段的RoI变换器[12]在第一阶段使用全连接层来旋转候选的水平锚框，然后提取框内的特征以进行进一步的回归和分类。SCRDet[71]使用注意力机制来减少背景噪声并改善对拥挤和小物体的建模。Oriented RCNN[62]和Gliding Vertex[64]引入了新的框编码系统来解决由于旋转角度周期性引起的训练损失不稳定性问题。一些方法 将遥感检测视为点检测任务[67]，为解决遥感检测问题提供了另一种方法。

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 network [20] extracts robust object features via oriented feature alignment and orientation-invariant feature extraction. DRN [46], 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 [50] addresses the discontinuity of regression loss by introducing a modulated loss. AOPG [6] and R3Det [68] adopt a progressive regression approach, refining bounding boxes from coarse to fine granularity. In addition to CNN-based frameworks, AO2-DETR [9] introduces a transformer-based detection framework, DETR [4], into remote sensing detection tasks, which brings more research diversity.

与依赖于提议的锚点的方法不同，单阶段检测框架直接从密集采样的网格锚点对定向边界框进行分类和回归。单阶段 网络 [20] 通过定向特征对齐和方向不变特征提取来提取鲁棒的对象特征。另一方面，DRN [46] 利用注意力机制动态细化主干网络提取的特征，以进行更准确的预测。与 Oriented RCNN 和 Gliding Vertex 相比，RSDet [50] 通过引入调制损失来解决回归损失的间断性。AOPG [6] 和 R3Det [68] 采用逐步回归方法，从粗糙到精细的粒度细化边界框。除了基于 CNN 的框架外，AO2-DETR [9] 将基于变换器的检测框架 DETR [4] 引入到遥感检测任务中，这带来了更多研究多样性。

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

虽然这些方法在解决旋转方差问题方面取得了有希望的结果，但它们并没有考虑到航空图像中呈现的强烈且有价值的前验信息。相反，我们的方法专注于利用大核和空间选择机制来更好地建模这些先验信息，而不修改当前的检测框架。

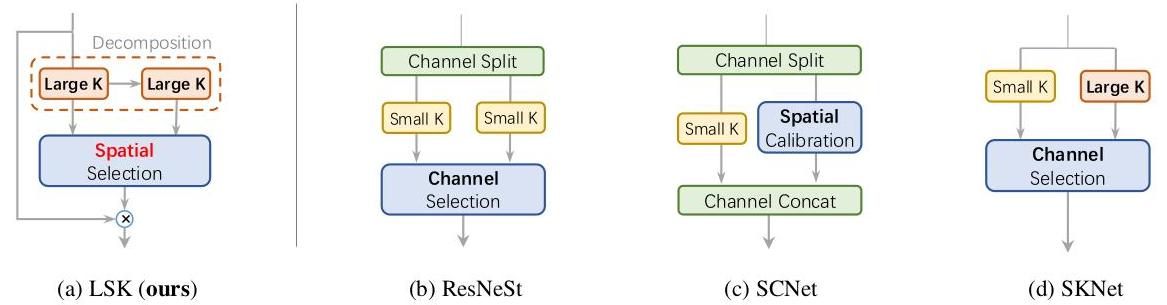


Figure 3: Architectural comparison between our proposed LSK module and other representative selective mechanism modules. K: Kernel.

图 3：我们提出的 LSK 模块与其他代表性选择机制模块的架构比较。K：核。

# 2.2. Large Kernel Networks

# 2.2. 大核网络

Transformer-based [54] models, such as the Vision Transformer (ViT) [14, 49, 55, 11, 1], Swin transformer [36, 22, 63, 76, 47] and PVT [57] have gained popularity in computer vision due to their effectiveness in image recognition tasks. Research has demonstrated that the large receptive field is a key factor in their success. In light of this, 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 [37] uses depth-wise convolutions in their backbone, resulting in significant performance improvements in downstream tasks. In addition, Re-pLKNet [13] even uses a convolutional kernel via re-parameterization, achieving compelling performance. A subsequent work SLaK [35], further expands the kernel size to through kernel decomposition and sparse group techniques. VAN [17] introduces an efficient decomposition of large kernels as convolutional attention. Similarly, SegNeXt [18] and Conv2Former [25] demonstrate that large kernel convolution plays an important role in modulating the convolutional features with a richer context.

基于Transformer的 [54] 模型，如视觉Transformer（ViT）[14, 49, 55, 11, 1]、Swin transformer [36, 22, 63, 76, 47] 和 PVT [57]，由于在图像识别任务中的有效性而广泛应用于计算机视觉领域。研究 表明，大感受野是它们成功的关键因素。基于这一点，近期的工作表明，设计良好的具有大感受野的卷积网络也可以与基于Transformer的模型高度竞争。例如，ConvNeXt [37] 在其主干网络中使用 深度可分卷积，导致在下游任务中性能显著提升。此外，Re-pLKNet [13] 甚至通过参数重置使用 卷积核，实现了令人信服的性能。后续工作 SLaK [35] 进一步通过核分解和稀疏组技术将核大小扩展到 。VAN [17] 引入了一种高效的大核分解作为卷积注意力。同样，SegNeXt [18] 和 Conv2Former [25] 证明了大型卷积核在调制具有更丰富上下文的卷积特征中发挥着重要作用。

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

尽管大核卷积在通用对象识别领域受到了关注，但在遥感检测的特定领域中，对它们重要性的研究仍然不足。正如引言中提到的，航空图像具有独特的特征，使得大核特别适合遥感任务。据我们所知，我们的工作是首次尝试引入大核卷积用于遥感，并探讨它们在这一领域的重要性。

# 2.3. Attention/Selective Mechanism

# 2.3. 注意力/选择机制

The attention mechanism is a simple and effective way to enhance neural representations for various tasks. The channel attention SE block [27] uses global average information to reweight feature channels, while spatial attention modules like GENet [26], GCNet [3], and SGE [31] enhance a network’s ability to model context information via spatial masks. CBAM [60] and BAM [48] combine both channel and spatial attention to make use of the advantages of both.

注意力机制是一种简单而有效的方法，用于增强各种任务的神经表示。通道注意力 SE 块 [27] 使用全局平均信息重新加权特征通道，而空间注意力模块如 GENet [26]、GCNet [3] 和 SGE [31] 通过空间掩码增强网络建模上下文信息的能力。CBAM [60] 和 BAM [48] 结合了通道和空间注意力，以利用两者的优势。

In addition to channel/spatial attention mechanisms, kernel selections are also a self-adaptive and effective technique for dynamic context modelling. CondConv [66] and Dynamic convolution [5] use parallel kernels to adaptively aggregate features from multiple convolution kernels. SKNet [30] introduces multiple branches with different convolutional kernels and selectively combines them along the channel dimension. ResNeSt [77] extends the idea of SKNet by partitioning the input feature map into several groups. Similarly to the SKNet, SCNet [34] uses branch attention to capture richer information and spatial attention to improve localization ability. Deformable Convnets [80, 8] introduce a flexible kernel shape for convolution units.

除了通道/空间注意力机制，核选择也是一种自适应且有效的技术，用于动态上下文建模。CondConv [66] 和 Dynamic convolution [5] 使用并行核自适应地从多个卷积核中聚合特征。SKNet [30] 引入了具有不同卷积核的多个分支，并在通道维度上选择性地组合它们。ResNeSt [77] 通过将输入特征图划分为几组来扩展 SKNet 的想法。与 SKNet 类似，SCNet [34] 使用分支注意力捕获更丰富的信息，并使用空间注意力提高定位能力。可变形卷积网络 [80, 8] 为卷积单元引入了灵活的核形状。

Our approach bears the most similarity to SKNet [30], however, there are two key distinctions between the two 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 as 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.

我们的方法与 SKNet [30] 最为相似，然而，两种方法之间有两个关键区别。首先，我们提出的选择机制明确依赖于通过分解的一系列大核，这与大多数现有的基于注意力的方法不同。其次，我们的方法在空间维度上而不是 SKNet 所使用的通道维度上自适应地聚合大核中的信息。这种设计对于遥感任务来说更为直观和有效，因为通道选择无法对图像空间中不同目标之间的空间变化进行建模。详细的结构比较列在图 3 中。

# 3. Methods

# 3. 方法

# 3.1. LSKNet Architecture

# 3.1. LSKNet 架构

The overall architecture is built upon the recent popular structures (refer to the details in Supplementary Materials (SM)) with a repeated building block.

整体架构基于最近流行的结构 （具体细节请参考补充材料 (SM)）并具有重复的构建块。

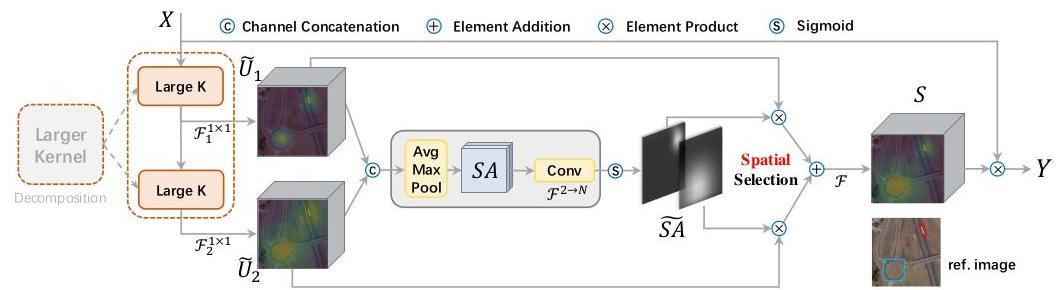


Figure 4: A conceptual illustration of LSK module.

图4：LSK模块的概念性示意图。

| Model |  |  | #P |
| --- | --- | --- | --- |
| \* LSKNet-T |  |  | 4.3M |
| \* LSKNet-S |  |  | 14.4M |

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

表1：本文中使用的LSKNet变体。 ：特征通道数； ：每个阶段的LSKNet块数 。

| RF | sequence | #P | FLOPs |
| --- | --- | --- | --- |
| 23 | (23, 1) |  | 42.4G |
|  | 11.3K | 11.9G |
| 29 | (29, 1) | 60.4K | 63.3G |
|  | 11.3K | 13.6G |

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

表2：将单个大深度卷积核展开为序列的两个代表性示例的理论效率比较，给定通道数为64。 ：核大小； ：扩张率。

The detailed configuration of different variants of LSKNet used in this paper is listed in Tab. 1. 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 core LSK module (Fig. 4) is embedded in the LK Selection sub-block. It consists of a sequence of large kernel convolutions and a spatial kernel selection mechanism, which would be elaborated on later.

本文使用的LSKNet不同变体的详细配置列在表1中。每个LSKNet块由两个残差子块组成：大核选择（LK Selection）子块和前馈网络（FFN）子块。核心LSK模块（图4）嵌入在LK Selection子块中。它由一系列大核卷积和一个空间核选择机制组成，这将稍后详细阐述。

# 3.2. Large Kernel Convolutions

# 3.2 大核卷积

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

根据引言中提到的先前研究(2)，建议为自适应选择建模一系列多个长距离上下文。因此，我们提出通过显式地将大增长核的深度卷积分解为一系列卷积，并增加扩张率来构建更大的核卷积。具体来说，序列中第 个深度卷积的核大小 的扩展、扩张率 和感受野 定义如下：

The increasing of 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.

核心尺寸的增加和扩张率的提高确保了感受野能够快速地扩展。我们为扩张率设定了一个上限，以保证扩张卷积不会在特征图之间引入间隙。例如，我们可以将一个大的核分解为2或3个深度可分卷积，如表2所示，它们分别具有理论上的23和29的感受野。

There are two advantages of the proposed designs. First, it explicitly yields multiple features with various large receptive fields, which makes it easier for 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 , a series of decomposed depth-wise convolutions with different receptive fields are applied:

所提出设计有两个优点。首先，它明确地生成了具有不同大感受野的多个特征，这使得后续的核选择更为容易。其次，顺序分解比简单地应用一个更大的核更为高效。如表2所示，在相同的结果理论感受野下，我们的分解与标准的大卷积核相比大大减少了参数的数量。为了从不同的范围获取输入 的具有丰富上下文信息的特征，应用了一系列分解的深度可分卷积，它们具有不同的感受野：

where are depth-wise convolutions with kernel and dilation . Assuming there are decomposed kernels, each of which is further processed by a convolution layer :

其中 是具有核 和扩张 的深度可分卷积。假设有 个分解的核，每个核进一步通过一个 卷积层 进行处理：

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.

为每个空间特征向量允许通道混合。接着，提出了一个选择机制，根据获得的多尺度特征动态地为不同的目标选择核，接下来将介绍这一机制。

# 3.3. Spatial Kernel Selection

# 3.3. 空间核选择

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:

为了提高网络关注检测目标最相关空间上下文区域的能力，我们使用一个空间选择机制，在不同的尺度上从大卷积核中选择特征图。首先，我们将来自不同核、具有不同感受野范围的特征进行拼接：

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

然后通过应用基于通道的平均和最大池化（分别表示为 和 对 ）有效地提取空间关系：

where and 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 to transform the pooled features (with 2 channels) into spatial attention maps:

其中 和 分别是平均和最大池化的空间特征描述符。为了允许不同空间描述符之间的信息交互，我们将空间池化的特征连接起来，并使用一个卷积层 将池化特征（具有2个通道）转换为 空间注意力图：

For each of the spatial attention maps, , a sigmoid activation function is applied to obtain the individual spatial selection mask for each of the decomposed large kernels:

对于每一个空间注意力图 ，应用sigmoid激活函数以获得每个分解的大核的独立空间选择掩码：

where denotes the sigmoid function. The features from the sequence of decomposed large kernels are then weighted by their corresponding spatial selection masks and fused by a convolution layer to obtain the attention feature :

其中 表示sigmoid函数。来自分解的大核序列的特征随后被它们相应的空间选择掩码加权，并通过一个卷积层 融合以获得注意力特征 ：

The final output of the LSK module is the element-wise product between the input feature and , similarly

LSK模块的最终输出是输入特征 与 的逐元素乘积，类似地

in :

在 中：

Fig. 4 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.

图4展示了LSK模块的详细概念图示，我们直观地演示了如何通过适应性地收集不同物体的相应大感受野来使大选择核工作。

# 4. Experiments

# 4. 实验

# 4.1. Datasets

# 4.1. 数据集

HRSC2016 [39] is a high-resolution remote sensing images which is collected for ship detection. It consists of 1,061 images which contains 2,976 instances of ships.

HRSC2016 [39] 是一个高分辨率遥感图像数据集，用于船舶检测。它包含1,061张图像，其中包含2,976个船舶实例。

DOTA-v1.0 [61] 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).

DOTA-v1.0 [61] 包含2,806张遥感图像。它包含15个类别的188,282个实例：飞机（PL）、棒球钻石（BD）、桥梁（BR）、地面轨道场（GTF）、小型车辆（SV）、大型车辆（LV）、船舶（SH）、网球场（TC）、篮球场（BC）、储罐（ST）、足球场（SBF）、环形交通岛（RA）、港口（HA）、游泳池（SP）和直升机（HC）。

FAIR1M-v1.0 [53] 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.

FAIR1M-v1.0 [53] 是最近发布的一个遥感数据集，包含 15,266 张高分辨率图像和超过一百万个实例。它包含 5 个类别和 37 个子类别的物体。

# 4.2. Implementation Details

# 4.2. 实施细节

In our experiment, we report the results of the detection model on HRSC2016, DOTA-v1.0 and FAIR1M-v1.0 datasets. To ensure fairness, we follow the same dataset processing approach as other mainstream methods . More details can be found in SM. 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

在我们的实验中，我们报告了检测模型在 HRSC2016、DOTA-v1.0 和 FAIR1M-v1.0 数据集上的结果。为确保公平性，我们遵循与其他主流方法相同的数据集处理方式 。更多细节可以在补充材料中找到。在我们的实验过程中，主干网络首先在 ImageNet-1K [10] 数据集上进行预训练，然后针对目标遥感基准进行微调。在消融研究中，我们为了实验效率，采用了 100 个训练周期的主干网络预训练计划。

| sequence | RF | Num. | FPS | mAP (%) |
| --- | --- | --- | --- | --- |
|  | 29 | 1 | 18.6 | 80.66 |
|  | 29 | 2 | 20.5 | 80.91 |
|  | 29 | 3 | 19.2 | 80.77 |

Table 3: The effects of the number of decomposed large kernels on the inference FPS and mAP, given 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.

表 3：给定理论感受野为 29 的情况下，分解大核数量对推理帧率（FPS）和 mAP 的影响。我们采用在 ImageNet 上预训练 100 个周期的 LSKNet-T 主干网络。将大核分解为两个深度可分离核在速度和准确性上取得了最佳性能。

(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 . In main results (Tab. 8, 9), the "Pre." column stands for the dataset on which the networks/backbones are pretrained (IN: Im-agenet [10] dataset; CO: Microsoft COCO [33] dataset; MA: Million-AID [40] dataset). Unless otherwise stated, LSKNet is defaulting to be built within the framework of Oriented RCNN [62] 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 [62], 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 [41] 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 . 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 image input.

(表3、5、4、6、7)。我们采用300个周期的主干网络预训练策略，以追求主要结果（表8、9、10）中的更高准确度，类似于 。在主要结果（表8、9）中，“预训练”列代表网络/主干网络预训练的数据库（IN：ImageNet [10] 数据库；CO：Microsoft COCO [33] 数据库；MA：Million-AID [40] 数据库）。除非另有说明，LSKNet默认在Oriented RCNN [62] 的框架内构建，因为其出色的性能和效率。所有模型都在训练和验证集上进行训练，并在测试集上进行测试。遵循 [62]，我们在HRSC2016数据集上训练模型36个周期，在DOTA-v1.0和FAIR1M-v1.0数据集上训练12个周期，使用AdamW [41] 优化器。HRSC2016的初始学习率设置为0.0004，其他两个数据集的初始学习率设置为0.0002，权重衰减为0.05。我们使用8个RTX3090 GPU和8的批量大小进行模型训练，并使用单个RTX3090 GPU进行测试。本文中我们报告的所有FLOPs都是使用1024 图像输入计算的。

# 4.3. Ablation Study

# 4.3. 剔除研究

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

在本节中，我们报告在DOTA-v1.0测试集上的剔除研究结果，以研究其有效性。

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 the speed and accuracy, achieving the best performance in terms of both FPS (frames per second) and mAP (mean average precision).

大核分解。确定分解的核数是LSK模块的关键选择。我们遵循公式（1）来配置分解后的核。当理论感受野固定为29时，关于大核分解数量的消融研究结果显示在表3中。它表明将大核分解为两个深度可分大核在速度和精度之间取得了良好的平衡，同时在FPS（每秒帧数）和mAP（平均精度均值）方面都取得了最佳性能。

Receptive Field Size and Selection Type. Based on our 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

感受野大小与选择类型。基于我们在表3中呈现的评估结果，我们发现对于我们提出的LSKNet，最优解是将大核分解为两个串联的深度可分核。此外，表4显示，过小或过大的感受野都可能阻碍LSKNet的性能，大约23的感受野大小被确定为最有效的。另外，我们的实验表明，所提出的空间选择方法更为有效。

|  |  | CS | SS | RF | FPS | mAP (%) |
| --- | --- | --- | --- | --- | --- | --- |
|  | (5, 2) | - |  | 11 | 22.1 | 80.80 |
|  |  |  | - | 23 | 21.7 | 80.94 |
| (7, 1) | (9, 4) |  | - | 39 | 21.3 | 80.84 |
| (5, 1) | (7, 3) |  | - | 23 | 19.6 | 80.57 |
|  |  | - |  | 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 [30]); SS: spatial selection (ours). We adopt LSKNet-T backbones pretrained on ImageNet for 100 epochs. The LSKNet achieves best performance when using a reasonably large receptive field with spatial selection.

表4：当大核被分解为两个深度可分核的序列时，LSKNet关键设计组件的有效性。CS：通道选择（与SKNet[30]类似）；SS：空间选择（我们的）。我们采用在ImageNet上预训练100个周期的LSKNet-T主干网络。当使用合理大的感受野和空间选择时，LSKNet达到最佳性能。

| Pooling | |  | mAP (%) |
| --- | --- | --- | --- |
|  | | FPS |  |
| 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. Best result is obtained when using both.

表5：关于我们提出的LSK模块中空间选择的最大池化和平均池化有效性的消融研究。我们采用在ImageNet上预训练100个周期的LSKNet-T主干网络。当同时使用两者时，获得最佳结果。

| Framework mAP (%) | ResNet-18 | \* LSKNet-T |
| --- | --- | --- |
| Oriented RCNN [62] | 79.27 | 81.31 (+2.04) |
| RoI Transformer [12] | 78.32 |  |
| -Net [20] | 76.82 | 80.15 (+3.33) |
| R3Det [68] | 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 significant higher mAP in various frameworks than ResNet-18. than channel attention (similarly in SKNet [30]) for remote sensing object detection tasks.

表6：LSKNet-T和ResNet-18作为不同检测框架的基干网络在DOTA-v1.0上的比较。LSKNet-T基干网络在ImageNet上预训练了100个周期。轻量级的LSKNet-T在各种框架中实现了比ResNet-18显著更高的mAP。比通道注意力（如在SKNet [30]中类似）对于遥感目标检测任务。

Pooling Layers in Spatial Selection. We conduct experiments to determine the optimal pooling layers for spatial selection in remote sensing object detection, 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.

空间选择中的池化层。我们进行了实验以确定遥感目标检测中空间选择的最优池化层，如表5所示。结果显示，在LSK模块的空间选择部分同时使用最大池化和平均池化能够在不牺牲推理速度的情况下提供最佳性能。

Performance of LSKNet backbone under different detection framworks. 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 [62] and RoI Transformer [12] as well as one-stage frameworks - Net [20] and R3Det [68]. The results in Tab. 6 show that our proposed LSKNet backbone significantly improves detection performance compared to ResNet-18, while using only of its parameters and with fewer FLOPs.

不同检测框架下LSKNet基干网络的表现。为了验证我们提出的LSKNet基干网络的通用性和有效性，我们在包括两阶段框架O-RCNN [62] 和 RoI Transformer [12] 以及单阶段框架 - Net [20] 和 R3Det [68] 在内的各种遥感检测框架下评估了其性能。表6的结果显示，我们提出的LSKNet基干网络与ResNet-18相比，在仅使用其 参数和 更少的FLOPs的情况下，显著提高了检测性能。

| Group | Model (backbone only) | #P | FLOPs | mAP (%) |
| --- | --- | --- | --- | --- |
| Baseline | ResNet-18 | 11.2M | 38.1G | 79.27 |
| Large Kernel | VAN-B1 [17] | 13.4M | 52.7G | 81.15 |
| ConvNeXt V2-N [59] | 15.0M | 51.2G | 80.81 |
| MSCAN-S [18] | 13.1M | 45.0G | 81.12 |
| Selective Attention | SKNet-26 [30] | 14.5M | 58.5G | 80.67 |
| ResNeSt-14 [77] | 8.6M | 57.9G | 79.51 |
| SCNet-18 [34] | 14.0M | 50.7G | 79.69 |
| Ours | \* LSKNet-S | 14.4M | 54.4G | 81.48 |
| Prev Best | CSPNeXt [43] | 26.1M | 87.6G | 81.33 |

Table 7: Comparison on LSKNet-S and other (large kernel/selective attention) backbones under O-RCNN [62] framework on DOTA-v1.0, except that the Prev Best is under RT-MDet [43] framework. All backbones are pretrained on ImageNet for 100 epochs. Our LSKNet achieves the best mAP under similar complexity budgets, whilst surpassing the previous best public records [43].

表7：在O-RCNN [62] 框架下，LSKNet-S与其他（大核/选择性注意力）基干网络在DOTA-v1.0上的比较，除了之前的最佳结果是在RT-MDet [43] 框架下。所有基干网络都在ImageNet上预训练了100个周期。我们的LSKNet在相似的复杂度预算下实现了最佳的mAP，同时超过了之前的最佳公开记录 [43]。

| Method | Pre. | mAP (07) | mAP (12) | #P | FLOPs |
| --- | --- | --- | --- | --- | --- |
| DRN [46] | IN | - | 92.70 | - | - |
| CenterMap [56] | IN | - | 92.80 | 41.1M | 198G |
| Rol Trans. [12] | IN | 86.20 | - | 55.1M | 200G |
| G. V. [64] | IN | 88.20 | - | 41.1M | 198G |
| R3Det [68] | IN | 89.26 | 96.01 | 41.9M | 336G |
| DAL [44] | IN | 89.77 | - | 36.4M | 216G |
| GWD [70] | IN | 89.85 | 97.37 | 47.4M | 456G |
| [20] | IN | 90.17 | 95.01 | 38.6M | 198G |
| AOPG [6] | IN | 90.34 | 96.22 | - | - |
| ReDet [21] | IN | 90.46 | 97.63 | 31.6M | - |
| O-RCNN [62] | IN | 90.50 | 97.60 | 41.1M | 199G |
| RTMDet [43] | CO | 90.60 | 97.10 | 52.3M | 205G |
| \* LSKNet-S (ours) | 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 [68, 20, 62]. mAP (07/12): VOC 2007 [15]/2012 [16] metrics.

表8：与最先进方法在HRSC2016数据集上的比较。LSKNet-S的主干网络在ImageNet上预训练了300个周期，与大多数比较的方法 [68, 20, 62] 相同。mAP (07/12)：VOC 2007 [15]/2012 [16] 指标。

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

与其他大核/选择性注意主干网络比较。我们还把我们的LSKNet与6个流行的高性能主干网络模型进行了比较，这些模型具有大核或选择性注意机制。如表7所示，在相似的模型大小和复杂度预算下，我们的LSKNet在DOTA-v1.0数据集上超过了所有其他模型。

# 4.4. Main Results

# 4.4. 主要结果

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 and under the PASCAL VOC 2007 [15] and VOC 2012 [16] metrics, respectively.

HRSC2016上的结果。我们在HRSC2016数据集上评估了我们的LSKNet的性能，与12种最先进的方法进行了比较。表8中呈现的结果表明，我们的LSKNet-S在PASCAL VOC 2007 [15] 和VOC 2012 [16] 指标下，以 和 的mAP超过了所有其他方法。

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 and LSKNet-S achieve state-of-the-art performance with mAP of and respectively. Notably, our high-performing

DOTA-v1.0上的结果。我们在DOTA-v1.0数据集上与20种最先进的方法进行了比较，如表9所示。我们的LSKNet-T和LSKNet-S分别以 和 的mAP达到了最先进的性能。值得注意的是，我们性能高的

| Method | Pre. | mAP | #P | FLOPs | PL | BD | BR | GTF | SV | LV | SH | TC | BC | ST | SBF | RA | HA | SP |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| One-stage | | | | | | | | | | | | | | | | | | | |
| R3Det [68] | 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 [19] | IN | 76.67 | 36.6M | 194G | 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 [28] | 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 [24] | IN | 79.17 |  | 194G | 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 |  | 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 |
| [20] | IN | 79.42 |  | 198G | 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 [70] | 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 [43] | 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 [72] | 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 [43] | 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 [71] | 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. [64] | 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 [56] | 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 [69] | 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 [21] | 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 [62] | 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 [73] | IN | 80.93 |  | 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 [55] | 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 the compared methods [68, 20, 62]. \*: With EMA finetune similarly to the compared methods [43].

表9：与最先进方法在DOTA-v1.0数据集上采用多尺度训练和测试的比较。LSKNet主干网络在ImageNet上预训练了300个周期，与比较的方法 [68, 20, 62] 相同。\*: 与比较方法 [43] 类似地采用EMA微调。

| Model | G. V.\* [64] | RetinaNet\* [32] | C-RCNN\* [2] | F-RCNN\* [52] | RoI Trans.\* [12] | O-RCNN [62] | \* 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 [68, 20, 62]. \*: Results are referenced from FAIR1M paper [53].

表10：与最先进方法在FAIR1M-v1.0数据集上的比较。LSKNet的主干网络在ImageNet上预训练了300个周期，与[68, 20, 62]类似。\*: 结果参考自FAIR1M论文[53]。

| Team Name | Pre-stage | Final-stage |
| --- | --- | --- |
| nust\_milab | 81.16 | 74.16 |
| Secret;Weapon (ours) | 81.11 | 73.94 |
| JiaNeng | 79.07 | 72.90 |
| ema.ai.paas | 78.65 | 72.75 |
| SanRenXing | 78.06 | 71.39 |

Table 11: 2022 the Greater Bay Area International Algorithm Competition results. The dataset is based on FAIR1M-v2.0 [53].

表11：2022年大湾区国际算法竞赛结果。数据集基于FAIR1M-v2.0 [53]。

# LSKNet-S reaches an inference speed of 18.1 FPS on 1024x1024 images with a single RTX3090 GPU.

# LSKNet-S在单个RTX3090 GPU上对1024x1024图像达到18.1 FPS的推理速度。

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 and respectively, surpassing all other models by a significant margin.

FAIR1M-v1.0上的结果。我们在FAIR1M-v1.0数据集上与表10所示的6个其他模型进行了比较。结果显示我们的LSKNet-T和LSKNet-S表现优异，分别达到了最先进的mAP得分 和 ，超过了所有其他模型一大截。

2022 the Greater Bay Area International Algorithm Competition. Our team implemented a model similar to LSKNet for the 2022 the Greater Bay Area International Algorithm Competition and achieved second place, with a minor margin separating us from the first-place winner. The dataset used during the competition is a subset of FAIR1Mv2.0 [53], and the competition results are illustrated in Tab. 11. More details refer to SM.

2022年大湾区国际算法竞赛。我们的团队为2022年大湾区国际算法竞赛实现了一个类似于LSKNet的模型，并获得了第二名，与第一名仅有微弱差距。竞赛中使用的数据集是FAIR1Mv2.0 [53]的一个子集，竞赛结果在表11中有所说明。更多细节请参考补充材料。

# 4.5. Analysis

# 4.5. 分析

Visualization examples of detection results and Eigen-CAM [45] are shown in Fig. 5. It highlights that LSKNet-S can capture much more context information relevant to the detected targets, leading to better performance in various hard cases, which justifies our prior (1).

检测结果和Eigen-CAM [45]的可视化示例在图5中显示。它突出了LSKNet-S能够捕获与检测目标相关的更多上下文信息，这在各种困难情况下导致更好的性能，这证明了我们的先前的假设(1)。

To investigate the range of receptive field for each object category, we define as the Ratio of Expected Selective RF Area and GT Bounding Box Area for category :

为了研究每个物体类别的感受野范围，我们定义 为类别 的期望选择感受野面积与真实边界框面积的比值：

where is the number of images that contain the object category only. The is the sum of spatial selection activation in all LSK blocks for input image , where is the number of blocks in an LSKNet, and is the number of decomposed large kernels in an LSK module. is the total pixel area of all annotated oriented object bounding boxes (GT). We plot the normalized in Fig. 6 which represents the relative range of context required for different object categories for a better view.

其中 是仅包含物体类别 的图像数量。 是输入图像 在所有LSK块中的空间选择激活之和，其中 是LSKNet中的块数， 是LSK模块中分解的大核数。 是所有 标注的定向物体边界框（GT）的总像素面积。我们在图6中绘制了归一化的 ，它表示不同物体类别所需的相对上下文范围，以便更好地观察。

The results suggest that the Bridge category stands out as requiring a greater amount of additional contextual in-

结果表明，桥梁类别突出地需要更多的额外上下文信息。

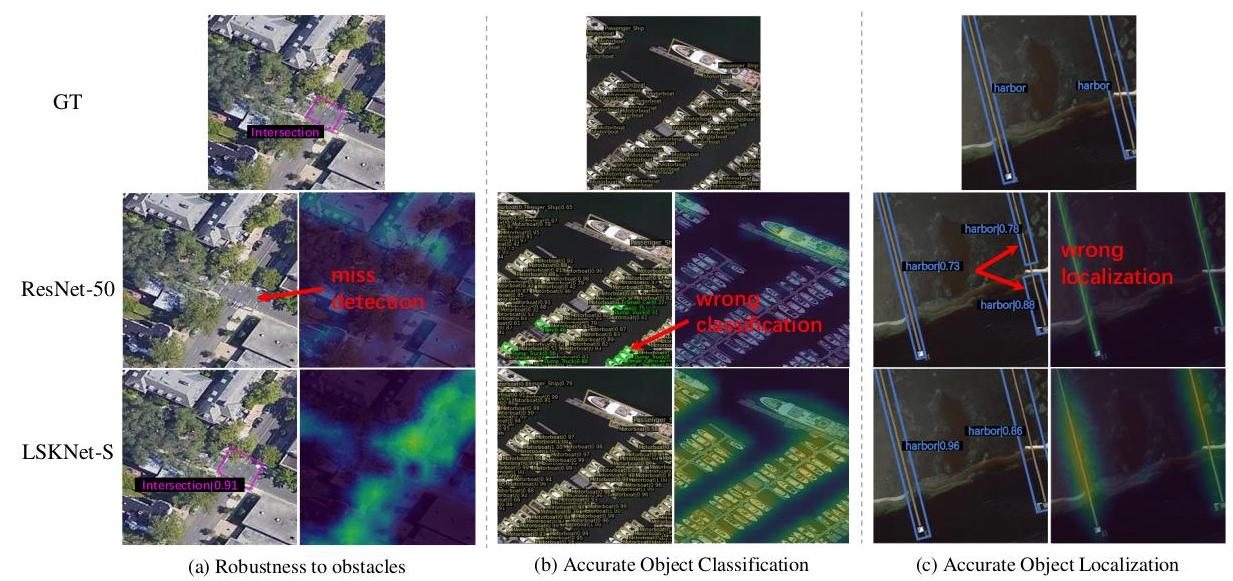


Figure 5: Eigen-CAM visualization of Oriented RCNN detection framework with ResNet-50 and LSKNet-S. Our proposed LSKNet can model a much long range of context information, leading to better performance in various hard cases.

图5：使用ResNet-50和LSKNet-S的定向RCNN检测框架的Eigen-CAM可视化。我们提出的LSKNet能够模拟更长的上下文信息范围，在各种困难情况下表现更好。

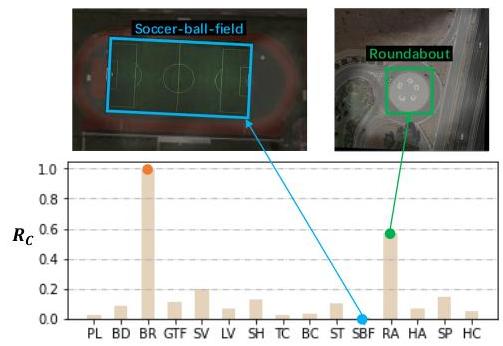


Figure 6: Normalised Ratio of Expected Selective RF Area and GT Bounding Box Area for object categories in DOTA-v1.0. The relative range of context required for different object categories varies a lot. Examples of Bridge and Soccer-ball-field are given, where the visualized receptive field is obtained from Eq. (8) (i.e., the spatial activation) of our well-trained LSKNet model.

图6：DOTA-v1.0中物体类别的归一化期望选择感受野面积与真实边界框面积比值 。不同物体类别所需的相对上下文范围差异很大。给出了桥梁和足球场的例子，其中可视化的感受野是通过我们训练良好的LSKNet模型的公式（8）（即空间激活）获得的。

formation 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 boundary 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.

与其他类别相比，形态学特征的相似性以及需要上下文线索来判断其是否被水包围，导致该类别的形成更为复杂。相反，如足球场这样的法庭类别，由于其独特的纹理特征，特别是法庭边界线，需要的最小上下文信息。这与我们的知识相符，并进一步支持先前研究（2）的结论，即不同对象类别所需相对范围的上下文信息差异很大。

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

我们进一步研究了LSKNet中的核选择行为。对于对象类别 ，LSKNet-T块中的核选择差异 （即较大核选择 - 较小核选择）定义为：

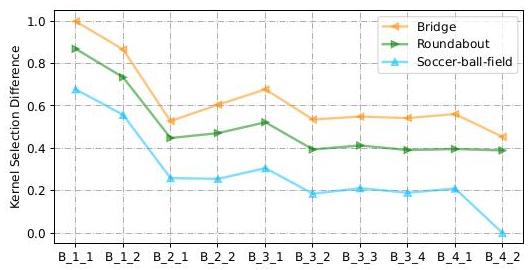


Figure 7: Normalised Kernel Selection Difference in the LSKNet-T blocks for Bridge, Roundabout and Soccer-ball-field. B\_i\_j represents the j-th LSK block in stage i. A greater value is indicative of a dependence on a broader context.

图7：Bridge、Roundabout和Soccer-ball-field在LSKNet-T块中的标准化核选择差异。B\_i\_j代表第i阶段中的第j个LSK块。较大的值表明依赖于更广泛的上下文。

We demonstrate the normalised over all images for three typical categories: Bridge, Roundabout and Soccer-ball-field and for each LSKNet-T block in Fig. 7. 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.

我们展示了图7中三种典型类别：Bridge、Roundabout和Soccer-ball-field以及每个LSKNet-T块在所有图像上的标准化 。如预期，Bridge类别的所有块中较大核的参与度高于Roundabout，而Roundabout又高于Soccer-ball-field。这与常识相符，因为足球场的纹理特征本身就足够独特和具有辨识度，因此不需要大量的上下文。

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.

我们还惊讶地发现了LSKNet在网络深度上的另一种选择模式：LSKNet通常在其浅层使用较大核，在更高层次使用较小核。这表明网络倾向于在低层次快速关注捕获来自大感受野的信息，以便高层次语义包含足够大的感受野以实现更好的辨识。

# 5. Conclusion

# 5. 结论

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 the competitive remote sensing benchmarks.

在本文中，我们提出了用于遥感对象检测任务的大型选择性内核网络（LSKNet），该网络旨在利用遥感图像的固有特性：需要更广泛和适应性的上下文理解。通过调整其大空间感受野，LSKNet能够有效地模拟不同对象类型变化的上下文细微差别。大量实验表明，我们提出的轻量级模型在具有竞争力的遥感基准测试中实现了最先进的性能。

# References

# 参考文献

[1] Yakoub Bazi, Laila Bashmal, Mohamad M. Al Rahhal, Re-ham Al Dayil, and Naif Al Ajlan. Vision transformers for remote sensing image classification. Remote Sensing, 2021. 3

[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, Xingx-ing 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 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. 3

[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 : Revisiting large kernel design in cnns. In CVPR, 2022. 3

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-vain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv, 2020. 3

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 6

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 6

[17] Meng-Hao Guo, Chengrou Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shiyong Hu. Visual attention network. ArXiv, 2022. 3, 5, 6, 12

[18] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. SegNeXt: Rethinking convolutional attention design for semantic segmentation. ArXiv, 2022. 3, 5, 6

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

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

[21] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Re-Det: A rotation-equivariant detector for aerial object detection. In CVPR, 2021. 5, 6, 7, 14

[22] 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. 3

[23] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. CoRR, 2016. 12

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

[25] Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2former: A simple transformer-style ConvNet for visual recognition. ArXiv, 2022. 3, 5, 12

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

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

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

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

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

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

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

[33] 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

[34] 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

[35] 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. 3

[36] 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. 3

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

[38] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feicht-enhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In CVPR, 2022. 12

[39] 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

[40] 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

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

[42] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In NeurIPS, 2016. 3

[43] 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

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

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

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

[47] Teerapong Panboonyuen, Kulsawasd Jitkajornwanich, Siam Lawawirojwong, Panu Srestasathiern, and Peerapon Va-teekul. Transformer-based decoder designs for semantic segmentation on remotely sensed images. Remote Sensing, 2021. 3

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

[49] Malsha V. Perera, Wele Gedara Chaminda Bandara, Jeya Maria Jose Valanarasu, and Vishal M. Patel. Transformer-based SAR image despeckling. CoRR, 2022. 3

[50] Wen Qian, Xue Yang, Silong Peng, Junchi Yan, and Yue Guo. Learning modulated loss for rotated object detection. In

[51] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In ICCV, 2021. 3

[52] 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

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

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. NeurIPS, 2017. 3

[55] 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. 3, 7

[56] 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

[57] 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. 3

[58] 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, 12

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

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

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

[62] 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, 14, 16

[63] 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. 3

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

[65] 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. 3

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

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

[68] 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, 16

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

[70] 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

[71] 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

[72] 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

[73] 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

[74] 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, 12

[75] 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

[76] 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. 3

[77] 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

[78] 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. 3

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

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

# A. Appendix

# 附录A

# A.1. LSKNet Block

# A.1. LSKNet模块

An illustration of an LSKNet Block is shown in Fig 8. The figure illustrates a repeated block in the backbone network, which is inspired by ConvNeXt [38], PVT-v2 [58], VAN [17], Conv2Former [25], and MetaFormer [74]. 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 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 [23] activation, and a second fully connected layer.

图8展示了LSKNet模块的示例。该图说明了主干网络中的重复模块，该模块受到ConvNeXt [38]、PVT-v2 [58]、VAN [17]、Conv2Former [25]和MetaFormer [74]的启发。每个LSKNet模块由两个残差子模块组成：大型内核选择（LK Selection）子模块和前馈网络（FFN）子模块。LK Selection子模块根据需要动态调整网络的感受野。FFN子模块用于通道混合和特征精炼，包括一个全连接层、一个深度卷积、一个GELU [23]激活函数和第二个全连接层的一系列操作。

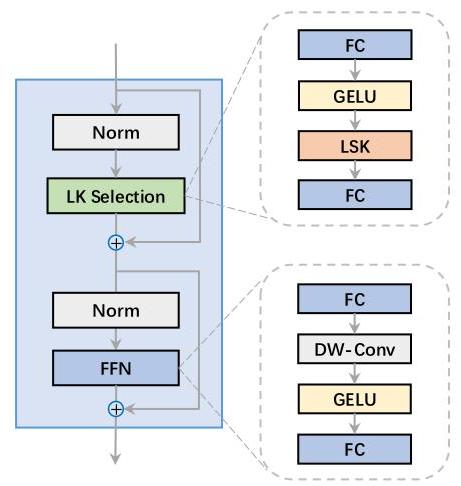


Figure 8: A block of LSKNet.

图8：LSKNet的一个模块。

# A.2. 2022 the Greater Bay Area International Al- gorithm Competition

# A.2. 2022粤港澳大湾区国际算法竞赛

The competition requires participants to train a remote sensing image object detection model using the Jittor framework and produce rotated bounding boxes of objects and their respective types in test images. The dataset used for the competition is a subset of FAIR1M-v2.0 and is provided by the Chinese Academy of Sciences’ Institute of Air and Space Information Innovation. It comprises 5000 training images, 576 preliminary test images, and 577 final test images. Example images of FAIR1M-v2.0 are shown in Fig. 9. The competition evaluates object detection performance based on ten object types: Airplane, Ship, Vehicle, Basketball\_Court, Tennis\_Court, Football\_field, Baseball\_field, Intersection, Roundabout, and Bridge. The mean Average Precision (mAP) evaluation metric is used, calculated based on the Pascal VOC 2012 Challenge. The Pre-stage and Final-stage are using the same finetuned model but with different test sets. The full competition scoreboard can be found at https://www.cvmart.net/race/ 10345/rank.

比赛要求参赛者使用 Jittor 框架训练一个遥感图像目标检测模型，并在测试图像中生成物体的旋转边界框及其相应的类型。比赛使用的数据集是 FAIR1M-v2.0 的子集，由中国科学院航空航天信息创新研究院提供。该数据集包括 5000 张训练图像，576 张初赛测试图像和 577 张决赛测试图像。FAIR1M-v2.0 的示例图像如图 9 所示。比赛根据十种物体类型评估目标检测性能：飞机、船舶、车辆、篮球场、网球场、足球场、棒球场、路口、环形交通岛和桥梁。使用 Pascal VOC 2012 挑战赛的基础上的平均精度均值（mAP）评估指标。预赛阶段和决赛阶段使用的是同一个微调模型，但是测试集不同。完整的比赛排行榜可以在 https://www.cvmart.net/race/ 10345/rank 找到。

| Model (Pre-stage) | mAP (%) |
| --- | --- |
| Single model | 80.29 |
| Single model | 80.42 |
| Output Ensemble | 80.51 |
| Weight Ensemble | 80.81 |
| Multi-level Ensemble (ours) | 81.11 |

Table 12: Multi-level model ensemble strategy results.

表 12：多层次模型集成策略结果。

In this competition, we employ model ensemble strategies to further enhance the performance of our single detection model. Two common methods for model ensemble in object detection are model output ensemble and model weight ensemble. Model output ensemble involves merging the outputs of different detectors using non-maximal suppression (NMS), while model weight ensemble merges the weights of multiple models into a single merged model through weighted averaging. In order to achieve better results, we propose a multi-level ensemble strategy that combines both of these approaches. This strategy consists of two levels of ensembles. In the first level, we merge the weights of the two models with the best performance during training through weight averaging. In the second level, we merge the inference results of the two models using NMS. This forms a multi-layer ensemble mechanism that can produce the final ensembled inference results with high efficiency, using only two models. By employing this multilevel ensemble strategy, we have achieved significant improvements in the performance of our object detection models in this competition as shown in Tab. 12. Some visualisation results of our proposed model on the FAIR1M-v2.0 test set are given in Fig. 9.

在本次竞赛中，我们采用了模型融合策略来进一步提升单个检测模型的性能。目标检测中模型融合的两种常见方法是模型输出融合和模型权重融合。模型输出融合涉及使用非最大值抑制（NMS）合并不同检测器的输出，而模型权重融合则通过加权平均将多个模型的权重合并为一个合并模型。为了获得更好的结果，我们提出了一种多级融合策略，该策略结合了上述两种方法。这一策略包含两个级别的融合。在第一级中，我们通过权重平均合并训练过程中性能最好的两个模型的权重。在第二级中，我们使用NMS合并这两个模型的推理结果。这样形成了一个多层融合机制，可以仅使用两个模型高效地产生最终的融合推理结果。通过采用这种多级融合策略，我们在本次竞赛中显著提高了我们的目标检测模型的性能，如表12所示。我们提出模型在FAIR1M-v2.0测试集上的部分可视化结果如图9所示。

# A.3. SKNet v.s. LSKNet v.s. LSKNet-CS (channel selection version)

# A.3. SKNet 与 LSKNet 及 LSKNet-CS（通道选择版本）的比较

A detailed conceptual comparison of SKNet, LSKNet and LSKNet-CS (channel selection version) module architecture is illustrated in Fig 10. There are two key distinctions between SKNet and LSKNet. 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 as 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.

SKNet、LSKNet和LSKNet-CS（通道选择版本）模块架构的详细概念比较如图10所示。SKNet和LSKNet之间存在两个关键区别。首先，我们提出的选择性机制明确依赖于通过分解的一系列大核，这与大多数现有的基于注意力的方法不同。其次，我们的方法在空间维度上而不是SKNet所使用的通道维度上自适应地聚合大核上的信息。这种设计对于遥感任务更为直观和有效，因为通道选择无法对图像空间中不同目标之间的空间变化进行建模。

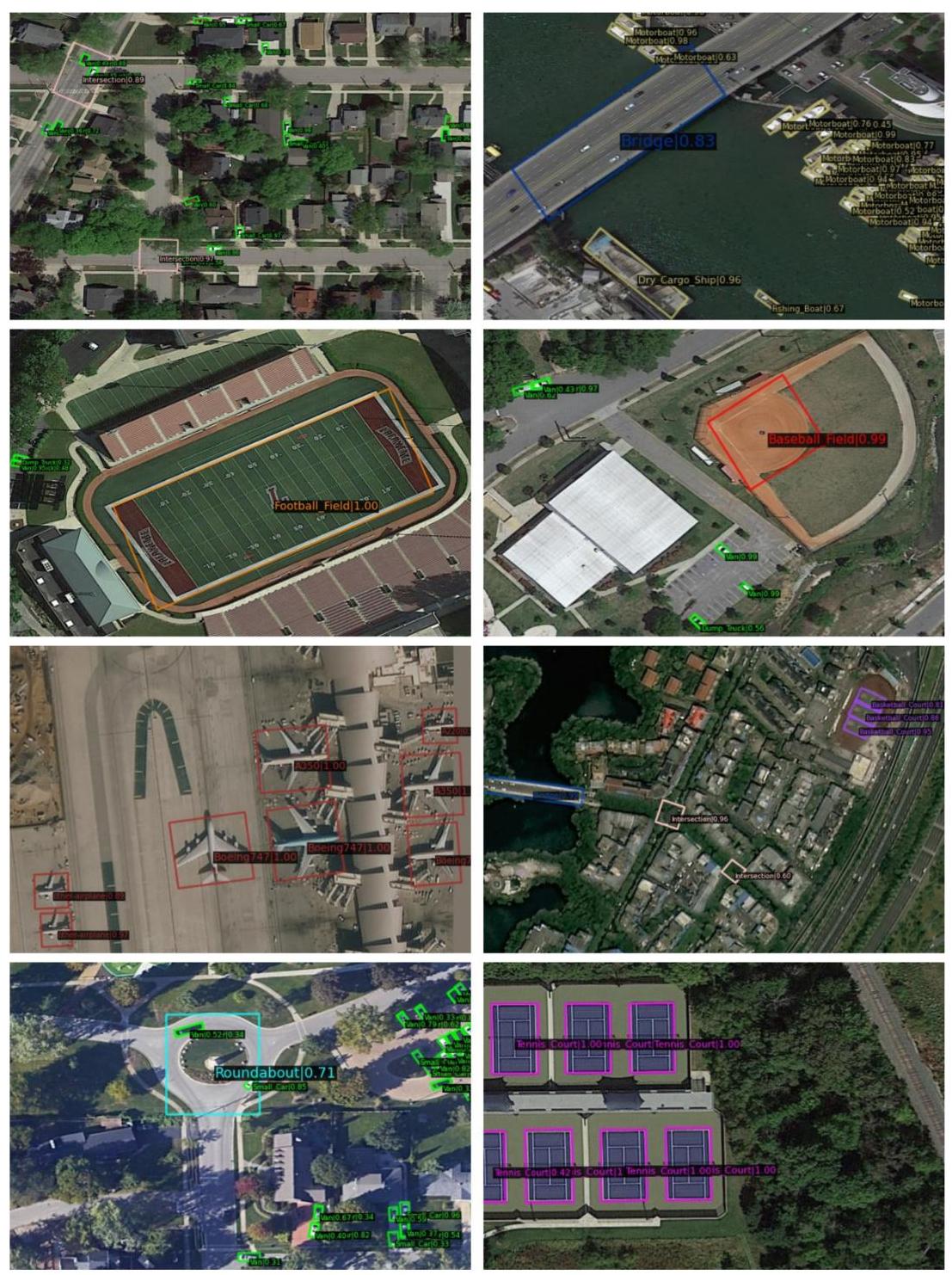
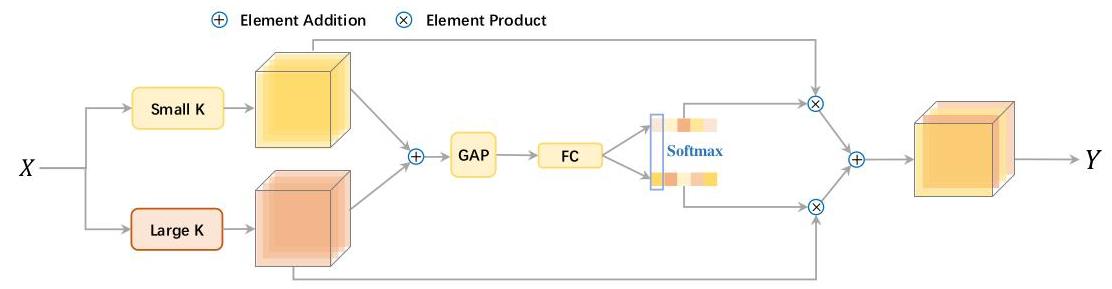


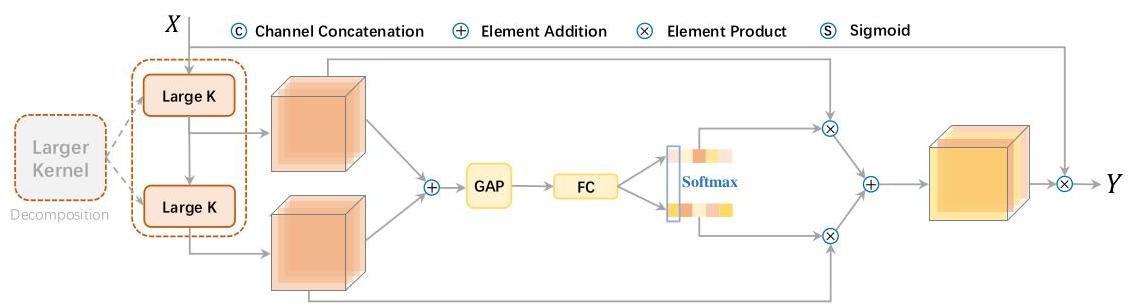
Figure 9: Examples of FAIR1M-v2.0 dataset test results with our LSKNet.

图9：使用我们的LSKNet在FAIR1M-v2.0数据集测试结果示例。



(a) A conceptual illustration of SK module in SKNet.

(a) SKNet中SK模块的概念图示。



(b) A conceptual illustration of LSK module (with Channel Selection) in LSKNet-CS, which is corresponding to "CS" configuration in main paper Tab. 4.

(b) LSKNet-CS中LSK模块（带通道选择）的概念图示，对应主论文表4中的“CS”配置。

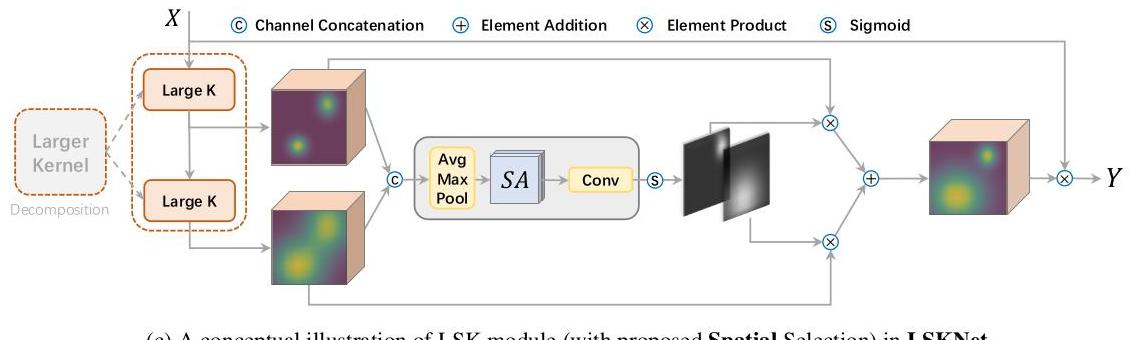


Figure 10: Detailed conceptual comparisons between our proposed SKNet, LSKNet and LSKNet-CS.

图10：我们提出的SKNet、LSKNet和LSKNet-CS之间的详细概念比较。

# A.4. Experiment Implementation Details

# A.4. 实验实施细节

To ensure fairness, we follow the same dataset processing approach as other mainstream methods . 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 , and then cropping each scaled image into 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.

为了确保公平性，我们遵循与其他主流方法相同的数据库处理方法 。对于DOTA-v1.0和FAIR1M-v1.0数据集，我们采用多尺度训练和测试策略，首先将图像缩放到三个尺度 ，然后将每个缩放后的图像裁剪成 子图像，图像块的重叠为500像素。对于HRSC2016数据集，我们将图像缩放，使图像的长边设置为800像素，同时保持其宽高比不变。

# A.5. Spatial Activation Visualisations

# A.5. 空间激活可视化

Receptive field activation examples for more object categories in DOTA-v1.0 are shown in Fig. 11, where the activation map is obtained from Eq. (8) (i.e., the spatial activation) of our well-trained LSKNet model. It demonstrates 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. Similarly, roundabouts also require a larger receptive field in order to distinguish between gardens and ring-like buildings. In order to accurately classify small objects such as ships and vehicles, a large receptive field is necessary to reference the surrounding context (i.e., whether it is in the sea or on land). Conversely, the Plane category and Court categories, such as Soccer-ballfield, necessitate minimal contextual information owing to their distinctive textural attributes, specifically the unique shapes and court boundary lines.

DOTA-v1.0中更多物体类别的感受野激活示例显示在图11中，其中激活图是通过我们训练有素的LSKNet模型的方程式(8)（即空间激活）获得的。它表明，与其它类别相比，Bridge类别需要更多的额外上下文信息，主要是因为其特征与道路相似，并且需要上下文线索来确定它是否被水包围。同样，环岛也需要更大的感受野来区分花园和环状建筑物。为了准确分类如船只和车辆等小物体，需要一个大感受野来参考周围的上下文（即它是在海上还是在陆地上）。相反，Plane类别和Court类别，如足球场，由于它们独特的纹理特征，特别是独特的形状和场地边界线，需要最少的上下文信息。

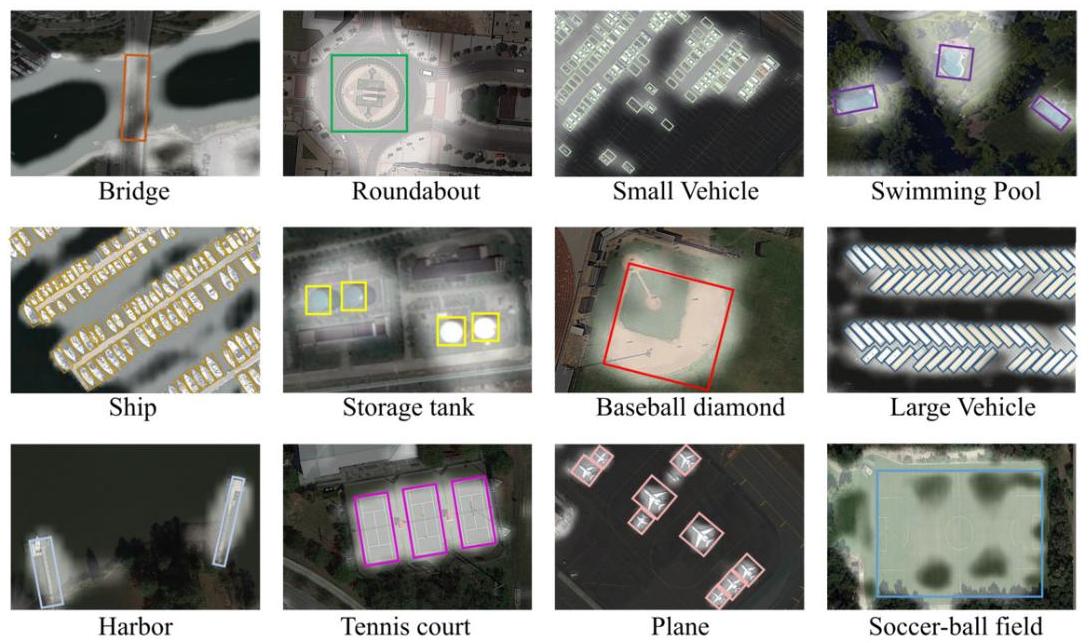


Figure 11: Receptive field activation for more object categories in DOTA-v1.0, where the activation map is obtained from the main paper Eq. (8) (i.e., the spatial activation) of our well-trained LSKNet model. The object categories are arranged in decreasing order from top left to bottom right based on the Ratio of Expected Selective RF Area and GT Bounding Box Area as illustrated in the main paper Fig. 6.

图11：DOTA-v1.0中更多物体类别的感受野激活，其中激活图是通过主论文中的方程式(8)（即空间激活）获得的，使用的是我们训练有素的LSKNet模型。物体类别按照主论文图6中所示的比例（预期选择感受野区域与真实边界框区域的比率）从左上到右下递减排列。

# A.6. FAIR1M benchmark results

# A.6. FAIR1M基准测试结果

Fine-grained category result comparisons with state-of-the-art methods on the FAIR1M-v1.0 dataset are given in Tab. 13.

在FAIR1M-v1.0数据集上，与最先进方法的细粒度类别结果比较如表13所示。

| Coarse Category | Sub Category | Gliding Vertex\* | RetinaNet\* | Cascade RCNN\* | Faster RCNN\* | ROI Trans\* | Oriented RCNN | \* LSKNet-T | \* LSKNet-S |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Airplane | Boeing737 | 35.43 | 38.46 | 40.42 | 36.43 | 39.58 | 42.84 | 45.12 | 39.84 |
| Boeing747 | 47.88 | 55.36 | 52.86 | 50.68 | 73.56 | 87.61 | 84.97 | 86.63 |
| Boeing777 | 15.67 | 24.75 | 29.07 | 22.50 | 18.32 | 18.83 | 20.16 | 24.21 |
| Boeing787 | 48.32 | 51.81 | 52.47 | 51.86 | 56.43 | 62.92 | 56.00 | 56.48 |
| C919 | 0.01 | 0.81 | 0.00 | 0.01 | 0.00 | 22.17 | 25.77 | 24.17 |
| A220 | 40.11 | 40.5 | 44.37 | 47.81 | 47.67 | 47.87 | 50.05 | 52.20 |
| A321 | 39.31 | 41.06 | 38.35 | 43.83 | 49.91 | 70.25 | 71.63 | 73.31 |
| A330 | 16.54 | 18.02 | 26.55 | 17.66 | 27.64 | 73.34 | 67.94 | 72.82 |
| A350 | 16.56 | 19.94 | 17.54 | 19.95 | 31.79 | 77.19 | 74.04 | 75.83 |
| ARJ21 | 0.01 | 1.70 | 0.00 | 0.13 | 0.00 | 32.49 | 40.24 | 46.39 |
| Ship | Passenger Ship | 9.12 | 9.57 | 12.10 | 9.81 | 14.31 | 20.21 | 19.23 | 20.43 |
| Motorboat | 23.34 | 22.55 | 28.84 | 28.78 | 28.07 | 72.13 | 71.08 | 71.38 |
| Fishing Boat | 1.23 | 1.33 | 0.71 | 1.77 | 1.03 | 13.53 | 14.70 | 15.81 |
| Tugboat | 15.67 | 16.37 | 15.35 | 17.65 | 14.32 | 35.50 | 37.09 | 32.84 |
| Engineering Ship | 15.43 | 19.11 | 18.53 | 16.47 | 15.97 | 16.23 | 16.60 | 14.79 |
| Liquid Cargo Ship | 15.32 | 14.26 | 14.63 | 16.19 | 18.04 | 26.49 | 24.74 | 25.37 |
| Dry Cargo Ship | 25.43 | 24.70 | 25.15 | 27.06 | 26.02 | 38.43 | 40.57 | 41.29 |
| Warship | 13.56 | 15.37 | 14.53 | 13.16 | 12.97 | 34.74 | 38.70 | 36.20 |
| Vehicle | Small Car | 66.23 | 65.20 | 68.19 | 68.42 | 68.80 | 74.25 | 75.73 | 76.34 |
| Bus | 23.43 | 22.42 | 28.25 | 28.37 | 37.41 | 47.02 | 46.27 | 55.54 |
| Cargo Truck | 46.78 | 44.17 | 48.62 | 51.24 | 53.96 | 50.22 | 54.06 | 55.84 |
| Dump Truck | 36.56 | 35.37 | 40.40 | 43.60 | 45.68 | 57.56 | 59.52 | 61.57 |
| Van | 53.78 | 52.44 | 58.00 | 57.51 | 58.39 | 75.22 | 75.57 | 76.71 |
| Trailer | 14.32 | 19.17 | 13.66 | 15.03 | 16.22 | 20.91 | 19.30 | 21.46 |
| Tractor | 16.39 | 1.28 | 0.91 | 3.04 | 5.13 | 2.99 | 3.68 | 7.19 |
| Excavator | 16.92 | 17.03 | 16.45 | 17.99 | 22.17 | 19.95 | 28.40 | 25.73 |
| Truck Tractor | 28.91 | 28.98 | 30.27 | 29.36 | 46.71 | 1.77 | 5.66 | 4.74 |
| Court | Basketball Court | 48.41 | 50.58 | 38.81 | 58.26 | 54.84 | 55.35 | 59.74 | 61.78 |
| Tennis Court | 80.31 | 81.09 | 80.29 | 82.67 | 80.35 | 82.96 | 87.07 | 81.06 |
| Football Field | 53.46 | 52.50 | 48.21 | 54.50 | 56.68 | 64.62 | 69.67 | 70.39 |
| Baseball Field | 66.93 | 66.76 | 67.90 | 71.71 | 69.07 | 90.36 | 90.03 | 89.94 |
| Road | Intersection | 59.41 | 60.13 | 55.67 | 59.86 | 58.44 | 60.82 | 60.58 | 62.90 |
| Roundabout | 16.25 | 17.41 | 20.35 | 16.92 | 18.58 | 20.47 | 23.20 | 27.00 |
| Bridge | 10.39 | 12.58 | 12.62 | 11.87 | 31.81 | 33.40 | 38.57 | 39.51 |
| mAP | | 29.92 | 30.67 | 31.18 | 32.12 | 35.29 | 45.60 | 46.93 | 47.87 |

Table 13: Comparisons of fine-grained category results with state-of-the-art methods on the FAIR1M-v1.0 dataset. The LSKNet backbones are pretrained on ImageNet for 300 epochs, similarly to the compared methods R3Det [68], S2ANet [20] and Oriented RCNN [62]. \*: Results are referenced from FAIR1M [53] paper.

表13：在FAIR1M-v1.0数据集上，细粒度类别结果与最新方法的比较。LSKNet主干网络在ImageNet上预训练了300个周期，与比较的方法R3Det [68]、S2ANet [20] 和 Oriented RCNN [62] 类似。\*: 结果来源于FAIR1M [53] 论文。