LSKNet: 针对遥感图像分析的轻量级基础骨干网络

李宇轩[1]，李翔[1,4†]，戴一冕[3]，侯淇彬[1,4]，刘丽[2]，刘永祥[2]，程明明[1,4†]，杨健[1†]

[1] 视觉计算与图像处理重点实验室, 南开大学，天津, 中国.
[2] 国防科技大学, 长沙, 中国.
[3] 模式计算与应用实验室, 南京理工大学, 南京, 中国.
[4] 南开国际先进研究院（深圳福田），深圳, 中国.

贡献作者： yuxuan.li.17@ucl.ac.uk; xiang.li.implus@nankai.edu.cn; yimian.dai@gmail.com; andrewhoux@gmail.com; lilyliu_nudt@163.com; lyx_bible@sina.com; cmm@nankai.edu.cn; csjyang@nankai.edu.cn;
† 通讯作者

摘要

遥感图像由于其固有的复杂性，为下游任务带来了独特的挑战。尽管已有大量研究致力于遥感分类、目标检测、语义分割和变化检测，但大多数研究忽视了遥感场景中蕴含的宝贵先验知识。这些先验知识的重要性在于，如果不参考足够长程的上下文，遥感目标可能会被错误识别，而不同目标所需的上下文范围也各不相同。本文考虑了这些先验知识，提出了一种轻量级的自适应大核卷积骨干网络（Large Selective Kernel Network，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不同范围上下文。据作者所知，自适应选择机制和大核卷积结构在遥感图像领域尚未被探索。在不引入额外复杂结构的情况下，本文提出的轻量级 LSKNet 骨干网络在标准遥感分类、目标检测、语义分割和变化检测基准测试中创造了新的最优成绩。本文的全面分析进一步验证了所提出先验知识的重要性以及 LSKNet 的有效性。相关代码可在https://github.com/zcablii/LSKNet获取。

关键词: 遥感图像, 卷积神经骨干网络, 大核卷积, 注意力机制, 目标检测, 语义分割。

1 引言

遥感图像由于其复杂的特性，包括高分辨率、随机方向、类内变化大、多尺度场景和密集小目标等，为下游任务带来了独特的挑战。为应对这些挑战，研究人员进行了广泛的探索，重点关注了各种方法，如用于分类的特征集成技术 [1–4] 和大规模预训练 [5–7]。此外，还提出了处理旋转方差 [8–10] 或采用新的定向框编码 [11, 12] 的方法用于目标检测任务。同时，多尺度特征融合 [13–19] 技术的整合也被用于提高检测和分割任务的性能。随着 SAM [20] 和 LLaVA [21] 等大型模型的快速发展，许多工作利用这些模型强大的通用知识进行下游任务的微调 [22, 23]，取得了显著的性能提升。

尽管取得了这些进展，但考虑到遥感图像的强先验知识来构建高效基础模型的工作相对较少。航空图像通常以鸟瞰视角捕获高分辨率图像。特别是航空图像中的大多数目标可能很小，仅凭外观难以识别。相反，识别这些目标需要依赖其上下文，因为周围环境可以提供有关其形状、方向和其他特征的有价值线索。根据对遥感数据的分析，本文挖掘出两个重要的先验：

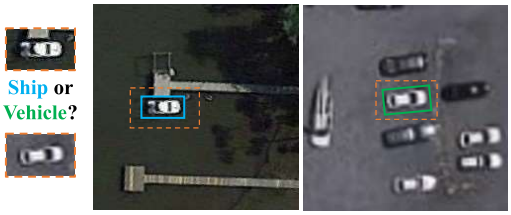


图 1: 成功检测遥感目标需要利用广泛的上下文信息，而感受野有限的检测器容易误判。

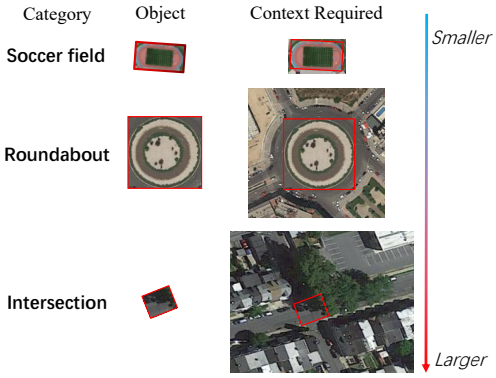


图 2: 根据人类标准，不同类型目标所需的上下文信息范围差异很大。红色框内的目标为精确的真实标注。

1. **准确识别通常需要广泛的上下文信息**。如图 1所示，遥感图像中目标检测器使用的有限上下文常常导致错误分类。区分船舶和车辆的不是它们的外观，而是上下文。
2. **不同目标所需的上下文信息差异很大**。如图 2所示，由于独特可辨识的场地边界线，足球场需要相对较少的上下文信息。相比之下，环岛可能需要更多的上下文信息来区分花园和环形建筑。交叉路口，尤其是被树木部分遮挡的部分，由于交叉道路之间的长程依赖关系，需要极大的感受野。

为解决遥感图像中准确识别目标的挑战，特别是那些通常需要广泛且动态的上下文信息的目标，本文提出了一种新颖的轻量级骨干网络，称为自适应大核卷积网络 (Large Selective Kernel Network, LSKNet)。本文的方法在特征提取骨干网络中动态调制感受野，这使得更高效地适应和处理所需的多样化、广泛的上下文成为可能。具体而言，本文通过空间选择性机制实现这一目标，该机制有效地对一系列大型深度可分离卷积核处理的特征进行加权，然后在空间上合并它们。这些核的权重是基于输入动态确定的，使模型能够自适应地使用不同的大型核，并根据需要调整每个目标在空间中的感受野。

本文是对先前工作 **LSKNet** [24] 的扩展版本。具体而言，本文进行了进一步的实验，以评估所提出的 LSKNet 骨干网络在广泛的遥感应用中的泛化能力，包括在 UCM [25]、AID [26] 和 NWPU [27] 数据集上的遥感场景分类，在合成孔径雷达模态数据集 SAR-Aircraft [28] 上的目标检测，在 Potsdam [29]、Vaihingen [30]、LoveDA [31]、UAVid [32] 和 GID [33] 数据集上的语义分割任务，以及在 LEVIR-CD [34] 和 S2Looking [35] 数据集上的变化检测任务。此外，本文还对 LSKNet 和 SKNet 进行了全面深入的比较，以突出 LSKNet 的差异和优势。综上所述，本文的贡献可归纳为以下**四个**主要方面：

- 挖掘出了遥感数据中存在的两个重要先验。
- 据本文所知，所提出的 LSKNet 骨干网络是首次如何通过大型选择性卷积核来精确利用上述先验完成遥感下游任务的模型。
- 尽管结构简单且轻量化，LSKNet 在 14 个广泛使用的公共数据集上的三个重要遥感任务中达到了最先进的性能，包括遥感场景分类 (UCM [25], AID [26], NWPU [27])、目标检测 (DOTA [36], HRSC2016 [37], FAIR1M [38], SAR-Aircraft [28])、语义分割 (Potsdam [29], Vaihingen [30], LoveDA [31], UAVid [32], GID [33]) 和变化检测 (LEVIR-CD [34], S2Looking [35])。
- 本文对所提方法进行了全面分析，进一步验证了所挖掘先验的重要性以及 LSKNet 模型在解决遥感图像分析挑战方面的有效性。

2 相关工作

2.1 遥感图像分析

遥感场景分类。遥感场景分类 [2, 4–6, 39, 40] 由于复杂背景和显著的类内变化而成为一项极具挑战性的任务。为应对这一挑战，研究者提出了多个模型，如 MGML [2]、ESD [3] 和 KFBNet [4] 等。这些模型旨在利用集成技术，整合多层次特征以提高分类性能。随着视觉 Transformer（ViT）[41] 的出现，基于 ViT 的大型模型 [42, 43] 逐渐兴起。此外，近期高性能的基于 ViT 的模型，如 RSP-ViTAE [5, 44] 和 RVSA [6]，在大规模遥感数据集 millionAID [45] 上进行了预训练，进一步推动了该领域的发展。

然而，特征集成通常会在骨干网络中引入多个分支，这增加了复杂性并降低了计算效率。同样，使用基于 ViT 的骨干网络可能导致模型过于庞大，不适合某些实际应用场景。

遥感目标检测。遥感目标检测 [46–51] 专注于在航空图像中识别和定位受关注的目标。近期的一个主要趋势是生成能准确匹配被检测目标方向的边界框。因此，大量研究致力于改进遥感目标检测中的定向边界框表示。为缓解 CNN 网络固有的旋转方差问题，研究者提出了几个著名的检测框架，包括 RoI Transformer [52]、Oriented RCNN [11]、S^2A Network [53]、DRN [54] 和 R3Det [9]。Oriented RCNN [11] 和 Gliding Vertex [12] 通过引入新的边界框编码系统，为解决旋转角周期性导致的训练损失不稳定问题做出了重要贡献。此外，GWD [10]、KLD [55] 和 LD [56] 等技术被开发用于解决回归损失的不连续性或提高边界框的定位质量。

尽管这些方法在解决旋转方差问题上取得了可喜的成果，但它们并未考虑航空图像中存在的富有价值的先验信息。相比之下，本文的方法利用大核和空间选择机制更好地建模这些先验，而无需修改现有的检测框架。

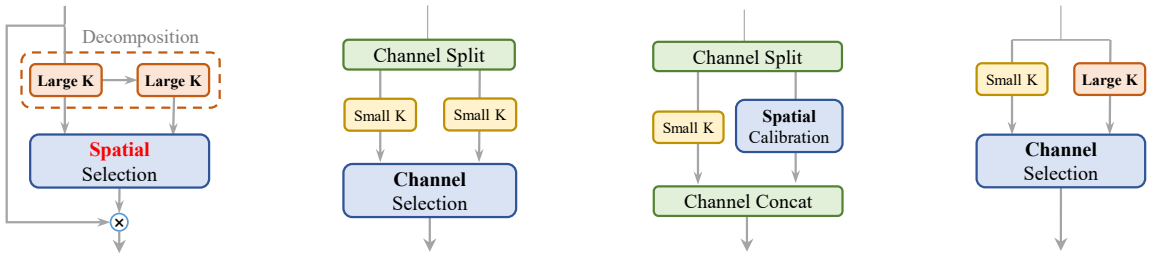
遥感语义分割。近期遥感语义分割模型的主要进展集中在应用注意力机制和多尺度特征融合技术上 [13–17, 60–62]。这些方法有效地聚合了细粒度细节和粗粒度语义，显著提升了分割性能。由此可见，整合大感受野语义进行多尺度特征融合对分割任务起着至关重要的作用。尽管现有方法取得了长足的进步，但它们往往忽视了前文提到的有价值的先验 *2)*。相比之下，本文提出的骨干模型考虑了遥感图像中的宝贵先验，提供了更灵活的多范围感受野特征，以解决这一局限性。

遥感变化检测。遥感变化检测旨在从不同时间获取的同一位置的一对图像中分割出具有语义变化的受关注区域。主流方法将此任务视为一种特殊形式的双输入图像分割。这些方法涉及在模型特征流中融合 [63–67] 或交互 [68–71] 双时相图像的特征，然后使用分割头生成最终的变化图。近期众多变化检测框架 [68, 72] 表明，更强大的骨干网络能显著提升性能，这说明特征提取的有效性和高效率仍是提升变化检测模型的关键因素。

2.2 大核网络

基于 Transformer 的模型 [73]，如视觉 Transformer（ViT） [6, 41]、Swin transformer [74–77] 和金字塔 transformer [78, 79]，在计算机视觉领域日益流行。研究 [80–84] 表明，大感受野是它们成功的关键因素之一。更有近期研究显示，设计良好的具有大感受野的卷积网络也能与基于 transformer 的模型相媲美。例如，ConvNeXt [85] 在其骨干网络中使用 7×7 深度可分离卷积，显著提升了下游任务的性能。此外，RepLKNet [86] 通过重参数化甚至使用了 31×31 的卷积核，取得了令人信服的性能。随后的 SLaK [87] 工作通过核分解和稀疏分组技术将核大小进一步扩展到 51×51。RF-Next [88] 为各种任务自动搜索固定的大核。VAN [89] 引入了一种高效的大核分解作为卷积注意力。同样，SegNeXt [90] 和 Conv2Former [91] 证明了大核卷积在调制具有丰富上下文的卷积特征方面发挥着重要作用。

尽管大核卷积在一般目标识别中受到关注，但在遥感检测中对其重要性的研究仍然不足。如前文 1所述，航空图像具有独特的特征，使得大核



(a) LSK (**ours**) (b) ResNeSt [57] (c) SCNet [58] (d) SKNet [59]

图 3: 本文提出的 LSK 模块与其他选择性机制模块的架构对比。K：核。

特别适合遥感应用。据本文所知，这项工作是首次将大核卷积引入遥感图像领域并研究其重要性。

2.3 注意力/选择机制

注意力机制 [92] 是一种简单而有效的方法，可以增强各种任务的神经表示。通道注意力 SE 块 [93] 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 [94]、GCNet [95]、CTNet [96] 和 SGE [97] 通过空间掩码增强网络建模上下文信息的能力。CBAM [98] 和 BAM [99] 结合了通道和空间注意力。自注意力机制最初在自然语言处理领域流行 [73]，近年来在计算机视觉领域也得到了广泛应用。视觉 Transformer（ViT）[41] 利用自注意力捕捉图像中的全局依赖关系和上下文信息。近年来，使用自注意力机制的模型在自然图像分类 [100]、检测 [101] 和分割 [20] 任务中取得了极具竞争力的性能。然而，在许多遥感图像任务中，如目标检测和分割，全局上下文信息并非总是必要的。例如，在检测汽车时，数百米外的河流信息并无用处。因此，最近的研究致力于将局部先验信息引入 Transformer 模型，如 Swin [102]、PVT [78, 103]、HiViT [104] 和 ViTAE [105]。这些模型在遥感场景中相比于原始 ViT 在计算效率和优化方面具有优势 [6, 106]。

除注意力机制外，核选择是一种自适应且有效的动态上下文建模技术。CondConv [107] 和动态卷积 [108] 并行核自适应地聚合多个卷积核的特征。SKNet [59] 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 [57] 通过将输入特征图划分为多个组，扩展了 SKNet 的理念。类似地，SCNet [58] 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 [109, 110] 为卷积单元引入了灵活的核形状。本文的方法与 SKNet [59] 最为相似。然而，两种方法之间存在**两个关键区别**。首先，本文提出的选择机制明确依赖于分解技术得到的一系列大核，这与大多数现有的基于注意力的方法不同。其次，本方法在空间维度上自适应地聚合大核信息，而不是像 SKNet 那样在通道维度上进行。这种设计对于遥感任务来说更加直观和有效，因为通道维度的选择无法对图像空间中不同目标的空间变化进行建模。详细的结构比较列于图 3中。

3 方法

3.1 LSKNet 网络架构

LSKNet 骨干网络的整体架构主要由重复的 LSK 模块构建而成（详细信息请参阅补充材料）。LSK 模块的设计灵感来源于 ConvNeXt [111]、MetaFormer [112]、PVT-v2 [103]、Conv2Former [91] 和 VAN [89]。每个 LSK 模块由两个残差子模块组成：大核选择（LK Selection）子模块和前馈网络（FFN）子模块。

LK Selection 子模块能够根据需求动态调整网络的感受野。核心的 LSK 模块（如图 4所示）

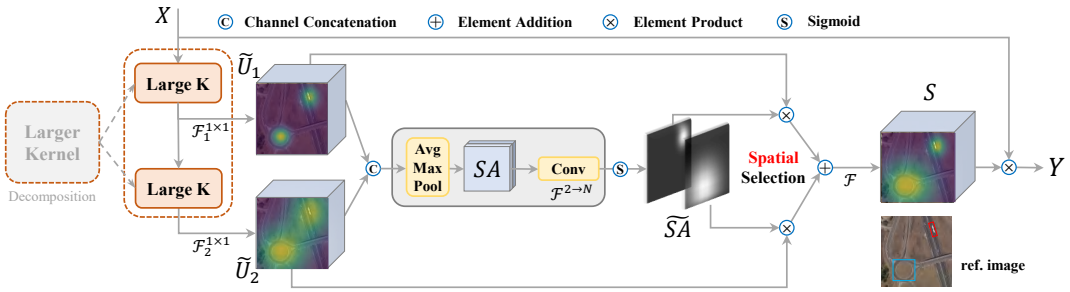


图 4: LSK 模块的概念性示意图。

表 1: **本文中使用的 LSKNet 变体**。C_i：特征通道数；D_i：每个阶段 i 中 LSK 块的数量。

模型	$\{C_1, C_2, C_3, C_4\}$	$\{D_1, D_2, D_3, D_4\}$	#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

表 2: 符号、维度及含义诠释。

符号	维度	含义
X	$C \times H \times W$	输入特征
N	1	选择核数量
i	1	分解核索引
$\widetilde{\mathbf{U}}_i$	$C \times H \times W$	富含上下文的特征
$\mathbf{SA}_{max}$	$1 \times H \times W$	通过最大池化得到的空间注意力
$\mathbf{SA}_{avg}$	$1 \times H \times W$	通过平均池化得到的空间注意力
$\widetilde{\mathbf{SA}}_i$	$N \times H \times W$	空间选择注意力
S	$C \times H \times W$	融合后的注意力特征
Y	$C \times H \times W$	输出特征

嵌入在 LK Selection 子模块中。该模块由一系列大核卷积和空间尺度的核选择机制组成，具体细节将在后文详细阐述。FFN 子模块用于通道混合和特征细化，由全连接层、深度可分离卷积、GELU [113] 激活函数和第二个全连接层依次组成。表 1列出了本文所使用的 LSKNet 不同变体的详细配置。此外，表 2提供了重要符号的完整列表，包括它们对应的维度和含义。这些符号在图 4和后续章节的方程中被广泛引用。

表 3: **两个代表性示例的理论效率比较**，本文将单个大型深度可分离卷积核展开为核序列，假设通道数为 64。k：核大小；d：膨胀率。

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

3.2 大核卷积

第 1节中的先验 *2)* 建议对一系列多尺度长程上下文进行建模以实现自适应选择模型感受野。因此，本文提出通过显式分解的方式，将大核卷积构建为一系列具有逐渐增大核尺寸和扩张率的深度可分离卷积。具体而言，对于第 i 个深度可分离卷积，核大小 k、扩张率 d 和感受野 RF 的扩展定义如下：

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

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

逐渐增大的核大小和扩张率确保了感受野能够快速扩展。本文对扩张率设置了上限，以保证扩张卷积不会在特征图之间引入间隙。例如，如表 3所示，可以将大核分解为 2 个或 3 个深度可分离卷积，理论感受野分别为 23 和 29。这种设计具有两个优势：首先，它显式地产生了具有不同大感受野

的多个特征，便于后续的核选择；其次，顺序分解比直接应用单个更大的核更加高效。如表 3所示，在相同的理论感受野下，本文的分解方法与标准大卷积核相比大大减少了参数数量。

为了从输入 $\mathbf{X}$ 中获取具有不同范围丰富上下文信息的特征，LSKNet 应用了一系列具有不同感受野的被分解的深度可分离卷积：

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

其中 $\mathcal{F}_i^{dw}(\cdot)$ 是具有核 k_i 和扩张率 d_i 的深度可分离卷积。假设有 N 个分解核，每个核都通过 1×1 卷积层 $\mathcal{F}^{1\times1}(\cdot)$ 进行进一步处理：

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

这允许对每个空间特征向量进行通道混合。接下来，本文提出了一种选择机制，基于获得的多尺度特征动态选择适用于不同目标的核，这将在下一节中详细介绍。

3.3 空间尺度的核选择机制

为了增强网络聚焦于检测目标最相关空间上下文区域的能力，本文采用了空间选择机制，对不同尺度的大卷积核所得到的特征图进行空间选择。首先，本文将通过不同感受野范围的卷积核所获得的特征进行拼接：

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

然后，通过对 $\widetilde{\mathbf{U}}$ 应用基于通道的平均池化和最大池化（分别表示为 $\mathcal{P}_{avg}(\cdot)$ 和 $\mathcal{P}_{max}(\cdot)$）来高效提取空间关系：

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

其中，$\mathbf{SA}_{avg}$ 和 $\mathbf{SA}_{max}$ 分别为平均池化和最大池化后的空间特征描述符。为了实现不同空间描述符之间的信息交互，本文将空间池化特征进行拼接，并使用卷积层 $\mathcal{F}^{2\to N}(\cdot)$ 将池化特征（具有 2 个通道）转换为 N 个空间注意力图：

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

对于每个空间注意力图 $\widehat{\mathbf{SA}}_i$，本文应用 sigmoid 激活函数以获得每个被分解大卷积核的单独空间选择掩码：

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

其中 $\sigma(\cdot)$ 表示 sigmoid 函数。分解大卷积核序列的特征图通过其对应的空间选择掩码进行加权，然后通过卷积层 $\mathcal{F}(\cdot)$ 融合，得到注意力特征 $\mathbf{S}$：

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

LSK 模块的最终输出是输入特征 $\mathbf{X}$ 与 $\mathbf{S}$ 的逐元素乘积，类似于 [89–91] 中的方法：

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

图 4展示了 LSK 模块的详细概念图，直观地演示了大选择性卷积核如何自适应地选择不同目标对应的大感受野来工作。

4 实验论证

在主要结果中，本文采用了在 Imagenet-1K [125] 上进行 300 轮的主干网络预训练策略以追求更高的性能，这与 [9, 11, 53] 的做法类似。然而，对于场景分类任务，本文遵循 [5] 中概述的预训练设置，在 millionAID 数据集 [45] 上进行 300 轮预训练。本文直接使用官方或者默认的训练、验证和测试集划分，并遵循每个基准测试的主流设置以确保公平性。在消融研究中，为了实验效率，本文采用了在 Imagenet-1K 上进行 100 轮的主干网络预训练策略。表格中，最佳得分用**粗体**表示，次佳得分用下划线标注。本节中的"FLOPs" 是通过将 1024×1024 像素的图像输入网络计算得出的。有关实验实施的更多细节（如训练计划和数据预处理）以及结果可视化，可参见补充材料。

表 4: 不同模型在场景分类上的性能表现。

模型	#P ↓	FLOPs ↓	UCM-82	AID-28	AID-55	NWPU-19	NWPU-28
MSANet [114]	>42.3M	>164.3	98.96	93.53	96.01	90.38	93.52
ViT-B [41]	86.0M	118.9G	99.28	93.81	96.08	90.96	93.96
SCCov [115]	13.0M	-	99.05	93.12	96.10	89.30	92.10
MA-FE [116]	>25.6M	>86.3G	99.66	-	95.98	-	93.21
MG-CAP [117]	>42.3M	>164.3G	99.00	93.34	96.12	90.83	92.95
LSENet [118]	25.9M	>86.3G	99.78	94.41	96.36	92.23	93.34
IDCCP [119]	25.6M	86.3G	99.05	94.80	96.95	91.55	93.76
F^2BRBM [120]	25.6M	86.3G	99.58	96.05	96.97	92.74	94.87
EAM [121]	>42.3M	>164.3	98.98	94.26	97.06	91.91	94.29
MBLANet [1]	-	-	99.64	95.60	97.14	92.32	94.66
GRMANet [122]	54.1M	171.4G	99.19	95.43	97.39	93.19	94.72
KFBNet [4]	-	-	99.88	95.50	97.40	93.08	95.11
CTNet [42]	-	-	-	96.25	97.70	93.90	95.40
RSP-R50 [5]	25.6M	86.3G	99.48	96.81	97.89	93.93	95.02
RSP-Swin [5]	27.5M	37.7G	99.52	96.83	98.30	94.02	94.51
RSP-ViTAE [5]	19.3M	119.1G	99.90	96.91	98.22	**94.41**	95.60
RVSA [6]	114.4M	301.3G	-	97.01	98.50	93.92	95.66
ConvNext [85]	28.0M	93.7G	99.81	95.43	97.40	94.07	94.76
FSCNet [123]	28.8M	166.1G	**100**	95.56	97.51	93.03	94.76
UPetu [124]	87.7M	>322.2G	99.05	96.29	97.06	92.13	93.79
MBENet [3]	23.9M	108.5G	99.81	96.00	98.54	92.50	95.58
FENet [2]	23.9M	92.0G	99.86	96.45	**98.60**	92.91	95.39
⋆ LSKNet-T	**4.3M**	**19.2G**	99.81	96.80	98.14	94.07	95.75
⋆ LSKNet-S	14.4M	54.4G	99.81	**97.05**	98.22	94.27	**95.83**

4.1 场景分类

4.1.1 分类数据集

遥感图像分类的主流方法 [1, 5, 120, 122] 通常在三个标准场景识别数据集上进行实验，包括 UC Merced 土地利用（UCM） [25] 数据集、航空图像数据集（AID） [26] 和西北工业大学收集的图像场景分类（NWPU） [27] 数据集。

UCM 是一个相对较小的数据集，仅包含 2,100 张图像和 21 个类别，每个类别有 100 张图像。所有图像的尺寸为 256 × 256。

AID 包含 10,000 张图像，分为 30 个类别，所有图像的尺寸为 600 × 600。

NWPU 是一个相对较大的数据集，包含 31,500 张图像和 45 个类别，每个类别有 700 张图像。所有图像的尺寸为 256 × 256。

遵循遥感分类工作的主流方法 [1, 5, 120, 122]，本文在五个标准基准上进行实验，即 UCM-82、AID-28、AID-55、NWPU-19 和 NWPU-28。

4.1.2 分类结果

表 4展示了各种对比方法的分类结果。本文将所提出的 LSKNets 与其他 22 种最先进的遥感场景分类方法进行了比较。值得注意的是，在不使用任何技巧（如 MBENet [3] 和 FENet [2] 中的特征集成）的情况下，本文提出的轻量级模型 LSKNet-T 和 LSKNet-S 在多个数据集上都展现出了具有竞争力的性能。这些结果表明，LSKNet 在各种场景下进行准确场景分类方面具有良好的效果，同时也展示了其作为骨干网络进行特征提取的潜力。

表 5: 在 **DOTA-v1.0** 数据集上与最先进模型的比较，采用多尺度训练和测试。*：与比较方法类似，使用 EMA 微调 [126]。

模型	Pre.	mAP↑	#P↓	FLOPs↓	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
单阶段																			
R3Det [9]	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 [127]	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 [128]	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 [129]	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 [130]	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 [53]	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 [10]	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 [126]	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 [55]	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 [126]	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
两阶段																			
SCRDet [131]	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
ViTDet [132]	IN	74.41	103.2M	502G	88.38	75.86	52.24	74.42	78.52	83.22	88.47	90.86	77.18	86.98	48.95	62.77	76.66	72.97	57.48
RoI Trans. [52]	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. [12]	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 [133]	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 [134]	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 [8]	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 [135]	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 [136]	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 [11]	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
KFIoU [137]	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 [6]	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	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	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*	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**

表 6: 在 **FAIR1M-v1.0** 数据集上与最先进模型的比较。*：结果引用自 FAIR1M 论文 [38]。

模型	G. V.* [12]	RetinaNet* [138]	C-RCNN* [139]	F-RCNN* [140]	RoI Trans.* [52]	O-RCNN [11]	LSKNet-T	LSKNet-S
mAP(%)	29.92	30.67	31.18	32.12	35.29	45.60	46.93	**47.87**

4.2 定向目标和合成孔径雷达目标检测

4.2.1 目标检测数据集

为评估所提出模型在遥感检测任务中的适用性，本文在 4 个具有挑战性的数据集上进行了实验。这些数据集包括 3 个广泛使用的定向目标检测数据集：HRSC2016 [37]、DOTA-v1.0 [36] 和 FAIR1M-v1.0 [38]，以及一个复杂且具有挑战性的合成孔径雷达（SAR）数据集 SAR-Aircraft [28]。

DOTA-v1.0 [36] 由 2,806 张遥感图像组成，包含 188,282 个实例，涵盖 15 个类别：飞机（PL）、棒球场（BD）、桥梁（BR）、田径场（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）、环岛（RA）、港口（HA）、游泳池（SP）和直升机（HC）。

HRSC2016 [37] 是一个专门用于船舶检测的高分辨率遥感数据集，由 1,061 张图像组成，包含 2,976 个船舶实例。

FAIR1M-v1.0 [38] 是一个近期发布的遥感数据集，包含 15,266 张高分辨率图像和超过 100 万个实例。该数据集涵盖 5 个主类别和 37 个子类别的目标。

SAR-Aircraft 数据集 [28] 是一个专为 SAR 模态目标检测收集的最新遥感数据集。与前述 3 个 RGB 模态数据集不同，SAR 数据集中的图像

表 7: 在 **HRSC2016** 数据集上与最先进模型的比较。mAP (07/12): VOC 2007 [142]/2012 [143] 评价指标。

模型	Pre.	mAP(07)	mAP(12)	#P	FLOPs
DRN [54]	IN	-	92.70	-	-
CenterMap [133]	IN	-	92.80	41.1M	198G
Rol Trans. [52]	IN	86.20	-	55.1M	200G
G. V. [12]	IN	88.20	-	41.1M	198G
R3Det [9]	IN	89.26	96.01	41.9M	336G
DAL [141]	IN	89.77	-	36.4M	216G
GWD [10]	IN	89.85	97.37	47.4M	456G
S^2ANet [53]	IN	90.17	95.01	38.6M	198G
AOPG [136]	IN	90.34	96.22	-	-
ReDet [8]	IN	90.46	97.63	31.6M	-
O-RCNN [11]	IN	90.50	97.60	41.1M	199G
RTMDet [126]	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

都为灰度图像。该数据集包含 7 个不同类别，分别是 A220、A320/321、A330、ARJ21、Boeing737、Boeing787 和其他。数据集由 3,489 张训练图像和 879 张测试图像组成，总计 16,463 个飞机实例。

4.2.2 检测结果

在定向目标检测实验中，考虑到其出色的性能和效率，本文默认将 LSKNets 构建在 Oriented RCNN [11] 框架内。

DOTA-v1.0 数据集结果。本文将 LSKNet 与 20 种最先进的方法在 DOTA-v1.0 数据集上进行了比较，结果如表 5所示。本文提出的 LSKNet-T、LSKNet-S 和 LSKNet-S* 分别达到了 **81.37%**、**81.64%** 和 **81.85%** 的最优 mAP。值得注意的是，性能优异的 LSKNet-S 在单个 RTX3090 GPU 上处理 1024x1024 图像时，推理速度可达 **18.1** FPS。

HRSC2016 数据集结果。本文在 HRSC2016 数据集上评估了 LSKNet 与 12 种最先进方法的性能。表 7中的结果表明，本文提出的 LSKNet-S 在 PASCAL VOC 2007 [142] 和 VOC 2012 [143] 指标下分别达到了 **90.65%** 和 **98.46%** 的 mAP，优于所有其他方法。

表 8: **SAR-Aircraft** 测试集上的 mAP 结果。

RetinaNet [138] 2x	#P	mAP_{50}	mAP_{75}
ResNet-50 [144]	25.6M	0.469	0.324
PVT-Tiny [78]	13.2M	0.498	0.335
Res2Net-50 [145]	25.7M	0.528	0.339
Swin-T [102]	28.3M	0.586	0.346
ConvNeXt V2-N [146]	15.0M	0.589	0.350
VAN-B1 [89]	13.4M	0.603	0.375
⋆ LSKNet-T	**4.3M**	0.582	0.354
⋆ LSKNet-S	14.4M	**0.624**	**0.387**

Cascade Mask RCNN [147] 2x	#P	mAP_{50}	mAP_{75}
ResNet-50 [144]	25.6M	0.483	0.339
PVT-Tiny [78]	13.2M	0.502	0.344
Res2Net-50 [145]	25.7M	0.544	0.372
ConvNeXt V2-N [146]	15.0M	0.581	0.428
Swin-T [102]	28.3M	0.596	0.416
VAN-B1 [89]	13.4M	0.604	0.457
⋆ LSKNet-T	**4.3M**	0.586	0.435
⋆ LSKNet-S	14.4M	**0.614**	**0.458**

FAIR1M-v1.0 数据集结果。本文将 LSKNet 与其他 6 种模型在 FAIR1M-v1.0 数据集上进行了比较，结果如表 6所示。结果表明，本文提出的 LSKNet-T 和 LSKNet-S 表现出色，分别达到了 **46.93%** 和 **47.87%** 的最优 mAP 得分，显著超越了其他所有模型。细粒度类别结果可参见补充材料。

SAR-Aircraft 数据集结果。本文评估了所提出的 LSKNets 与 5 种最先进的骨干网络在 Cascade Mask RCNN [147] 和 RetinaNet [138] 检测框架下的性能。结果如表 8所示，清楚地表明本文提出的 LSKNets 在 SAR 目标检测任务中提供了显著且实质性的性能改进。

定量分析。 在比较的模型中，使用原始 ViT 骨干网络的 ViTDet 具有最大的计算复杂度（相比 LSKNet-T 高 4.0 倍的 FLOPs）和第二大的模型规模（相比 LSKNet-T 多 4.9 倍的参数），但在 DOTA-v1.0 数据集的目标检测任务上表现不佳。另一种基于 ViT 的模型变体 RVSA，以 ViTAE 为基础，融合了多尺度和二维局部性归纳偏置，在

建模图像特征方面比原始 ViT 骨干网络更为有效。尽管 RVSA 效果显著，但仍存在模型规模庞大（相比 LSKNet-T 多 5.4 倍的参数）和计算复杂度高（相比 LSKNet-T 高 3.3 倍的 FLOPs）的问题。这两种基于 ViT 的模型均无法超越轻量级的 LSKNet-T。

LSKNet 的优势还体现在 DOTA-v1.0 数据集中容易混淆的类别上，如小型车辆（+2.49%）和船舶（+3.59%）（表 5），以及 FAIR1M 数据集中需要大量上下文信息的类别上，如交叉路口（+2.08%）、环岛（+6.53%）和桥梁（+6.11%）（补充材料中的表 S4）。这些结果进一步验证了本文提出的*先验 1* 和*先验 2* 的有效性，并证实了所提出的基础骨干模型的有效性。

4.3 语义分割

4.3.1 分割数据集

遵循主流分割研究的做法 [13, 60]，本文通过在五个标准数据集上进行评估来验证所提出模型在遥感分割任务中的有效性：Potsdam [29]、Vaihingen [30]、LoveDA [31]、UAVid [32] 和 GID [33] 数据集。

Potsdam [29] 是一个高分辨率语义分割数据集，包含 38 张高分辨率图像。它由 6 个语义类别组成：不透水表面、建筑物、低矮植被、树木、汽车和一个背景类别（杂波）。

Vaihingen [30] 同样是一个高分辨的语义分割数据集，由 33 张高分辨率图像组成。其语义类别与 Potsdam 相同。

LoveDA [31] 是一个多尺度且复杂的遥感语义分割数据集，包含 5,987 张 1024×1024 像素的图像。其中，2522 张用于训练，1,669 张用于验证，1,796 张用于在线测试。该数据集包含 7 个语义类别：建筑物、道路、水体、裸地、森林、农田和背景。

UAVid [32] 是一个高分辨率且复杂的无人机（UAV）语义分割数据集。它包含 200 张训练图像、70 张验证图像和 150 张在线测试图像。该数据集由 8 个不同类别组成：建筑物、道路、树木、植被、移动车辆、静止车辆、人类和其他。

表 9: Potsdam 测试集上的定量比较结果。OA：总体精度

模型	mF1 ↑	OA ↑	mIOU ↑
ERFNet [149]	85.8	84.5	76.2
DABNet [150]	88.3	86.7	79.6
BiSeNet [151]	89.8	88.2	81.7
EaNet [15]	90.6	88.7	83.4
MARESU-Net [61]	90.5	89.0	83.9
DANet [14]	88.9	89.1	80.3
SwiftNet [152]	91.0	89.3	83.8
FANet [16]	91.3	89.8	84.2
ShelfNet [153]	91.3	89.9	84.4
ABCNet [17]	92.7	91.3	86.5
Segmenter [154]	89.2	88.7	80.7
BANet [60]	92.5	91.0	86.3
SwinUperNet [102]	92.2	90.9	85.8
UNetFormer [13]	92.8	91.3	86.8
⋆ LSKNet-T	92.9	91.7	86.7
⋆ LSKNet-S	**93.1**	**92.0**	**87.2**

GID [33] 数据集是一个中等分辨率的土地覆盖分割数据集，地面采样距离（GSD）为 4m，包含 150 张 7,200×6,800 像素的图像。按照 [148] 的方法，本研究从原始 GID 数据集中选择了 15 张预定义图像，并将所有图像裁剪为 256×256 像素，最终得到 7,830 张训练图像和 3,915 张测试图像。该数据集包含六个语义类别：建成区、农田、森林、草地、水体和其他。

4.3.2 分割结果

本文在上述 5 个数据集上对所提出的 LSKNet-T 和 LSKNet-S 模型与多个近期提出的高水平模型进行了全面比较。对于 Potsdam、Vaihingen、LoveDA 和 UAVid 数据集，由于 UNetFormer [13] 框架具有令人信服的性能且开源可用，LSKNet 被集成到该框架中。对于 GID 数据集，本文使用 SegFormer 框架比较了各种骨干网络模型。具体而言，本研究在 Potsdam 数据集上与 14 个模型进行了比较（表 9），在 Vaihingen 数据集上与 16 个模型进行了比较（表 10），在 LoveDA

表 10: Vaihingen 测试集上的定量比较结果。

模型	mF1 ↑	OA ↑	mIOU ↑
PSPNet [155]	79.0	87.7	68.6
ERFNet [149]	78.9	85.8	69.1
DANet [14]	79.6	88.2	69.4
DABNet [150]	79.2	84.3	70.2
Segmenter [154]	84.1	88.1	73.6
BOTNet [156]	84.8	88.0	74.3
FANet [16]	85.4	88.9	75.6
BiSeNet [151]	84.3	87.1	75.8
DeepLabV3+ [157]	87.4	89.0	-
ShelfNet [153]	87.5	89.8	78.3
MARESU-Net [61]	87.7	90.1	78.6
EaNet [15]	87.7	89.7	78.7
SwiftNet [152]	88.3	90.2	79.6
ABCNet [17]	89.5	90.7	81.3
BANet [60]	89.6	90.5	81.4
UNetFormer [13]	90.4	91.0	82.7
⋆ LSKNet-T	91.7	**93.6**	84.9
⋆ LSKNet-S	**91.8**	**93.6**	**85.1**

数据集上与 13 个模型进行了比较（表 11），在 UAVid 数据集上与 16 个模型进行了比较（表 12），在 GID 数据集上与 6 个骨干网络模型进行了比较（表 13）。值得注意的是，本文提出的 LSKNet-T 和 LSKNet-S 模型表现出色，在所有数据集的大多数主要指标上均超越了其他最先进的方法。

4.4 变化检测

4.4.1 变化检测数据集

遵循主流变化检测研究的做法 [68, 71, 176]，本文在以下两个标准数据集上进行评估来验证所提出模型在遥感变化检测任务中的有效性：LEVIR-CD [34] 和 S2Looking [35]。

LEVIR-CD [34] 包含 637 对来自 Google Earth 的双时相图像，每张图像的尺寸为 1024×1024 像素，地面采样距离（GSD）为 0.5 米。该数据集标注了 31,333 个二元变化实例。

S2Looking [35] 由全球光学卫星拍摄的 5,000 对双时相图像组成。每张图像的尺寸为 1024×1024 像素，GSD 范围在 0.5 至 0.8 米之间。该数据集标注超过 65,920 个二元变化实例。

4.4.2 变化检测结果

在变化检测实验中，由于 Changer [68] 框架具有令人信服的性能且开源可用，LSKNet 默认构建于该框架之上。本文在 LEVIR-CD 和 S2Looking 数据集上对所提出的 LSKNet-T 和 LSKNet-S 模型与 17 个近期高性能模型进行了全面比较。表 14中的结果证实，所提出的 LSKNet-T 和 LSKNet-S 模型表现出色，在所有数据集的主要指标（F1 和 IoU）上均超越了其他最先进的方法。

4.5 消融分析

本节将报告在 DOTA-v1.0 测试集上进行的消融实验结果。选择 DOTA-v1.0 数据集进行消融研究主要基于两个因素：首先，目标检测是一项实用且具有挑战性的任务，而 DOTA-v1.0 数据集提供了多样化且复杂的目标和场景用于评估；其次，众多可用模型的存在使得全面比较成为可能，从而能够对本文提出方法的有效性进行深入评估。在消融研究中，为了提高实验效率，本文采用了 100 轮的骨干网络预训练计划（表 15、 16、 17、 19、 18）。

大核分解。确定分解的核数量是 LSK 模块的一个关键选择。本文遵循公式(1)来配置分解后的核。表 15展示了在理论感受野固定为 29 的情况下，对大核分解数量进行消融研究的结果。结果表明，将大核分解为两个深度可分离大核可以在速度和精度之间取得良好的平衡，在 FPS（每秒帧数）和 mAP（平均精度均值）方面都达到了最佳性能。

核感受野大小。基于表 15中的评估结果，本文发现将大核分解为两个**串联**的深度可分离卷积核是最优的策略。此外，表 16显示，过小或过大的感受野都会影响 LSKNet 的性能，而约为 23 的感受野大小被确定为最有效的选择。

SKNet 和不同注意力选择类型的比较。LSKNet 与 SKNet 有两个关键区别。首先，本文提出的选择机制依赖于通过核分解实现的一**序列**

表 11: LoveDA 测试集上的定量比较结果。

模型	**mIoU** ↑	背景	建筑物	道路	水体	裸地	森林	农田
Segmenter [154]	47.1	38.0	50.7	48.7	77.4	13.3	43.5	58.2
SegFormer [158]	47.4	43.1	52.3	55.0	70.7	10.7	43.2	56.8
DeepLabV3+ [157]	47.6	43.0	50.9	52.0	74.4	10.4	44.2	58.5
UNet [159]	47.6	43.1	52.7	52.8	73.0	10.3	43.1	59.9
UNet++ [160]	48.2	42.9	52.6	52.8	74.5	11.4	44.4	58.8
SemanticFPN [161]	48.2	42.9	51.5	53.4	74.7	11.2	44.6	58.7
FarSeg [162]	48.2	43.1	51.5	53.9	76.6	9.8	43.3	58.9
PSPNet [155]	48.3	44.4	52.1	53.5	76.5	9.7	44.1	57.9
FactSeg [163]	48.9	42.6	53.6	52.8	76.9	16.2	42.9	57.5
TransUNet [164]	48.9	43.0	56.1	53.7	78.0	9.3	44.9	56.9
BANet [60]	49.6	43.7	51.5	51.1	76.9	16.6	44.9	62.5
HRNet [165]	49.8	44.6	55.3	57.4	78.0	11.0	45.3	60.9
SwinUperNet [102]	50.0	43.3	54.3	54.3	78.7	14.9	45.3	59.6
DC-Swin [166]	50.6	41.3	54.5	56.2	78.1	14.5	**47.2**	62.4
UNetFormer [13]	52.4	44.7	58.8	54.9	79.6	20.1	46.0	62.5
Hi-ResNet [167]	52.5	**46.7**	58.3	55.9	80.1	17.0	46.7	**62.7**
⋆ LSKNet-T	53.2	46.4	59.5	57.1	79.9	21.8	46.6	61.4
⋆ LSKNet-S	**54.0**	**46.7**	**59.9**	**58.3**	**80.2**	**24.6**	46.4	61.8

表 12: UAVid 测试集上的定量比较结果。

模型	**mIoU** ↑	其他	建筑物	道路	树木	植被	移动车辆	静止车辆	人类
MSD [32]	57.0	57.0	79.8	74.0	74.5	55.9	62.9	32.1	19.7
CANet [168]	63.5	66.0	86.6	62.1	79.3	**78.1**	47.8	**68.3**	19.9
DANet [14]	60.6	64.9	85.9	77.9	78.3	61.5	59.6	47.4	9.1
SwiftNet [152]	61.1	64.1	85.3	61.5	78.3	76.4	51.1	62.1	15.7
BiSeNet [151]	61.5	64.7	85.7	61.1	78.3	77.3	48.6	63.4	17.5
MANet [61]	62.6	64.5	85.4	77.8	77.0	60.3	67.2	53.6	14.9
ABCNet [17]	63.8	67.4	86.4	81.2	79.9	63.1	69.8	48.4	13.9
Segmenter [154]	58.7	64.2	84.4	79.8	76.1	57.6	59.2	34.5	14.2
SegFormer [158]	66.0	66.6	86.3	80.1	79.6	62.3	72.5	52.5	28.5
BANet [60]	64.6	66.7	85.4	80.7	78.9	62.1	69.3	52.8	21.0
BOTNet [156]	63.2	64.5	84.9	78.6	77.4	60.5	65.8	51.9	22.4
CoaT [169]	65.8	69.0	**88.5**	80.0	79.3	62.0	70.0	59.1	18.9
UNetFormer [13]	67.8	68.4	87.4	81.5	80.2	63.5	73.6	56.4	31.0
⋆ LSKNet-T	69.3	**69.6**	87.9	82.8	80.6	64.8	**77.3**	60.2	31.3
⋆ LSKNet-S	**70.0**	**69.6**	84.8	**82.9**	**80.9**	65.5	76.8	64.9	**31.8**

大核的显式特征流动，这与大多数现有基于注意力方法的做法不同。相比之下，SKNet 采用了**并行**分解技术。其次，LSKNet 在空间维度上自适应地聚合大核信息，而非 SKNet 或 LSKNet-CS 所使用的通道维度。这种设计对遥感任务而言更为直观和有效，因为通道选择无法捕捉图像空间中不同目标的空间尺度变化。此外，本文还评估了一种同时利用空间和通道选择的 LSKNet 变体。表 16中的实验结果表明，在检测任务中，空间信息起着更为关键的作用。然而，同时包含空间和通

表 13: GID 测试集上的定量比较结果。

骨干网络	mF1 ↑	OA ↑	mIoU ↑
ConvNext-v2-N [146]	75.1	78.9	62.5
ResNet-50 [144]	75.3	80.0	64.1
Swin-T [102]	77.8	80.8	65.6
ResNest-50 [57]	79.7	80.3	67.2
VAN-S [89]	80.2	82.1	68.2
MSCAN-S [90]	80.4	81.4	68.4
⋆ LSKNet-T	79.4	81.5	67.2
⋆ LSKNet-S	**83.2**	**82.3**	**69.6**

道选择可能会增加模型优化的难度，导致性能略有下降。补充材料中详细比较了 SKNet、LSKNet、LSKNet-CS（通道选择版本）和LSKNet-SCS（空间和通道选择版本）的模块架构概念。

空间选择中的池化层。本文进行了实验以确定空间选择的最佳池化层选择，结果如表 17所示。实验表明，在 LSK 模块的空间选择模块中同时使用最大池化和平均池化可以在不牺牲推理速度的情况下获得最佳性能。

LSKNet 骨干网络在不同检测框架下的性能。为验证所提出的 LSKNet 骨干网络的通用性和有效性，本文在多种遥感检测框架下评估了其性能，包括两阶段框架 O-RCNN [11] 和 RoI Transformer [52]，以及单阶段框架 S^2A-Net [53] 和 R3Det [9]。表 19中的结果显示，与 ResNet-18 相比，本文提出的 LSKNet-T 骨干网络显著提高了检测性能，同时仅使用了 38% 的参数量和 50% 的 FLOPs。这些发现凸显了所提出的 LSKNet 骨干网络轻量化且功能强大的通用性特征。

与其他大核/选择性注意力骨干网络的比较。本文还将 LSKNet 与 9 种流行的大核或选择性注意力骨干网络进行了比较。如表 18所示，使用原始 ViT [41] 骨干网络的 ViTDet [132] 在所比较的模型中具有最大的模型规模和计算复杂度，但在所有任务中表现均不佳。表 5中的观察结果显示，它在具有明显细粒度特征的目标（如球场和直升机）上表现尤其差。这表明全局上下文信息建模对遥感场景并不高效。在相似或更少的模型规模和复杂度预算下，本文提出的 LSKNet 在遥感目标检测（DOTA-v1.0）、分割（Vaihingen）和变化检测（LEVIR-CD）任务上均优于其他所有模型，突显了其在捕获和处理遥感图像语义特征方面的有效性。

5 分析

本节针对目标检测任务进行分析，因为实例级信息对理解模型的整体行为具有重要意义。

检测结果可视化。图 5展示了检测结果和 Eigen-CAM [178] 的可视化示例。LSKNet 能够捕获与检测目标相关的合理范围的上下文信息，从而在各种困难情况下表现更好，这验证了本文的先验假设 *1)*。相比之下，ResNet 通常只能捕获有限范围的上下文信息，而 ViTDet 虽然能捕获大范围但粗糙的空间信息，在目标小而密集时难以建模细粒度细节。这两种模型在具有挑战性的场景中均表现有限。

不同目标的相对上下文范围。为研究每个目标类别的感受野相对范围，本文定义了 R_c 作为类别 c 的预期选择性感受野面积与真实边界框面积之比：

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

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

其中，I_c 是仅包含目标类别 c 的图像数量。A_i 是输入图像 i 在所有 LSK 块中空间选择激活的总和，D 是 LSKNet 中的块数，N 是 LSK 模块中分解大核的数量。B_i 是所有 J_i 个标注的定向目标边界框（真实值）的总像素面积。图 6中归一化的 R_c 直观地展示了不同目标类别所需的相对上下文范围。结果表明，桥梁类别相比其他类别需要更多的额外上下文信息，这主要是由于其特征与道路相似，且需要上下文线索来确定其是否被水包围。同样，环岛类别也有相对较高的 R_c 值，为 0.57。相反，球场类别的 R_c 值相对较低，均低于 0.1。由于其独特的纹理属性，特别是场地边界

表 14: LEVIR-CD 和 S2Looking 数据集上变化检测的定量比较结果。

模型	LEVIR-CD [34]				S2Looking [35]			
	Precision ↑	**Recall ↑**	**F1 ↑**	**IoU ↑**	**Precision ↑**	**Recall ↑**	**F1 ↑**	**IoU ↑**
FC-EF [63]	86.91	80.17	83.40	71.53	81.36	8.95	7.65	8.77
FC-Siam-Conc [63]	91.99	76.77	83.69	71.96	**83.29**	15.76	13.19	15.28
FC-Siam-Di [63]	89.53	83.31	86.31	75.92	68.27	18.52	13.54	17.05
STANet [34]	83.81	91.00	87.26	77.40	38.75	56.49	45.97	29.84
DTCDSCN [170]	88.53	86.83	87.67	78.05	68.58	49.16	57.27	40.12
HANet [171]	91.21	89.36	90.28	82.27	61.38	55.94	58.54	41.38
CDNet [172]	91.60	86.50	89.00	80.14	67.48	54.93	60.56	43.43
CDMC [173]	93.09	88.07	90.51	82.67	64.88	58.15	61.34	44.23
IFNet [67]	91.17	90.51	90.83	83.22	66.46	61.95	64.13	47.19
SNUNet [64]	92.45	90.17	91.30	83.99	71.94	56.34	63.19	46.19
BiT [70]	91.97	88.62	90.26	82.26	74.80	55.56	63.76	46.80
HCGMNet [174]	92.96	90.61	91.77	84.79	72.51	57.06	63.87	46.91
ChangeFormer [65]	92.59	89.68	91.11	83.67	72.82	56.13	63.39	46.41
C2FNet [175]	93.69	89.47	91.83	84.89	74.84	54.14	62.83	45.80
CGNet [176]	93.15	90.90	92.01	85.21	70.18	59.38	64.33	47.41
DiFormer [71]	**93.75**	90.59	92.15	85.44	72.39	61.19	66.31	49.60
Changer-MiT_b0 [68]	93.61	90.56	92.06	85.29	73.01	62.04	67.08	50.47
⋆ LSKNet-T	92.56	**91.83**	92.19	85.51	70.44	**64.46**	67.32	50.74
⋆ LSKNet-S	93.34	91.23	**92.27**	**85.65**	71.90	63.64	**67.52**	**50.96**

表 15: **分解大型卷积核数量对推理 FPS 和 mAP 的影响**，理论感受野为 29。将大型卷积核分解为两个深度可分离卷积核可以在速度和精度方面达到最佳性能。

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

线，它们只需要最少的上下文信息。 这与认知相符，进一步支持了本文的先验假设 *2)*，即不同目标类别所需的上下文信息相对范围差异很大。

核选择机制。本文进一步研究了 LSKNet 中的核选择机制。对于目标类别 c，LSKNet-T 结构块的核选择差异 ΔA_c（即大核选择 − 小核选择）定义如下：

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

表 16: **LSKNet 关键设计组件的有效性**，大型卷积核被分解为两个深度可分离卷积核序列。CS：通道选择；SS：空间选择 **(本文方法)**。当使用具有空间选择的合理大感受野时，LSKNet 达到最佳性能。

(k_1, d_1)	(k_2, d_2)	Flow	CS	SS	RF	FPS	mAP	
(3, 1)	(5, 2)	Series	-	-	11	22.1	80.80	
(5, 1)	(7, 3)	Series	-	-	23	21.7	80.94	
(5, 1)	(7, 4)	Series	-	-	29	20.5	80.91	
(7, 1)	(9, 4)	Series	-	-	39	21.3	80.84	
(3, 1)	(5, 1)	Parallel	✓	-	5	23.3	80.19	(SKNet [59])
(5, 1)	(7, 3)	Series	✓	-	23	19.6	80.57	(LSKNet-CS)
(5, 1)	(7, 3)	Series	✓	✓	23	18.6	80.82	(LSKNet-SCS)
(5, 1)	(7, 3)	Series	-	✓	23	20.7	**81.31**	(LSKNet)

图 7展示了三个典型类别（桥梁、环岛和足球场）在所有图像上的归一化 ΔA_c，以及每个 LSKNet-T 块的 ΔA_c。如预期所示，桥梁类别在所有块中的 ΔA_c 平均比环岛高约 30%，而环岛又比足球场高约 70%。这与常识相符，即足球场确实不需要

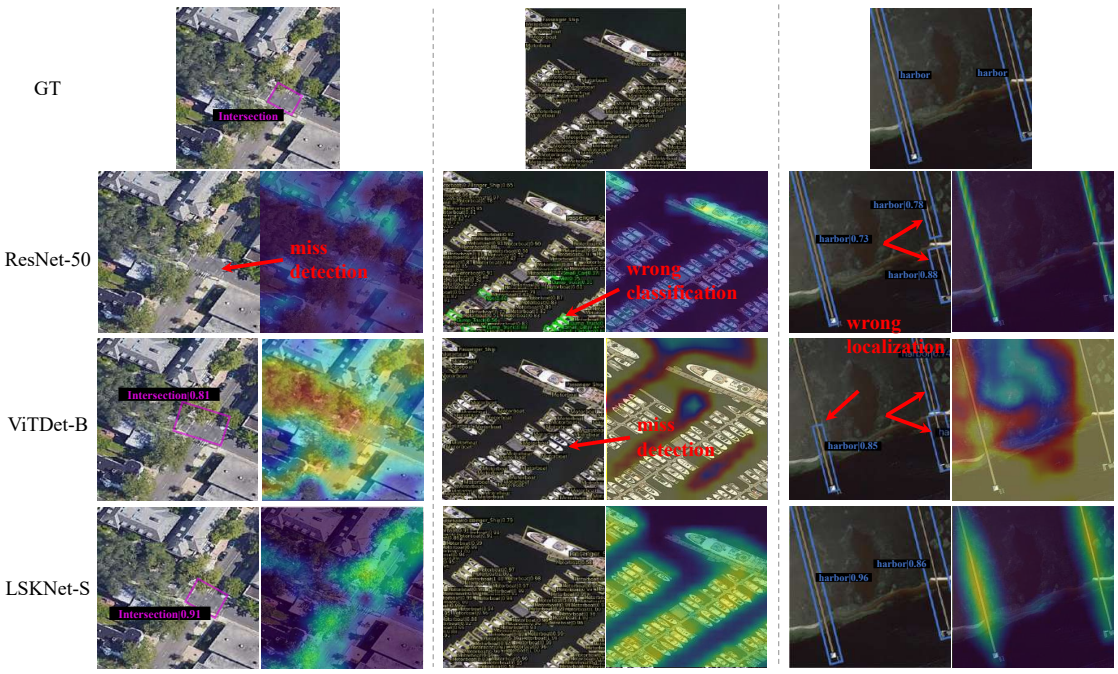


图 5: **Eigen-CAM 可视化**：基于 ResNet-50、ViTDet 和 LSKNet-S 的 Oriented RCNN 检测框架。本文提出的 LSKNet 能够建模合理长程的上下文信息，在各种困难情况下表现更佳。

表 17: 关于本文提出的 LSK 模块中**最大和平均池化对空间选择有效性**的消融研究。结果表明，同时使用两种池化方法可获得最佳效果。

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

大量上下文信息，因为其自身的纹理特征已经足够独特和具有辨识度。

本文还意外发现了 LSKNet 在网络深度上的另一种选择模式：LSKNet 通常在浅层使用较大的核，而在高层使用较小的核。第一层块的平均 ΔA_c 为 0.78，第二和第三层块为 0.40，最后一层块仅为 0.33。这表明网络倾向于在低层快速聚焦于捕获大感受野的信息，以便高层语义能包含足够的感受野，从而实现更好的区分。

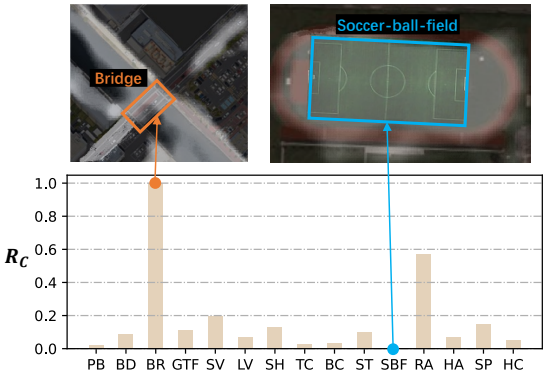


图 6: DOTA-v1.0 数据集中各目标类别的**预期选择性感受野面积与真实边界框面积的归一化比率** R_c。不同目标类别所需的相对上下文范围差异显著。本文使用公式 (8)（即空间激活）可视化训练后的 LSKNet 模型的感受野。

空间激活图可视化。图 8展示了 DOTA-v1.0 数据集中更多目标类别的空间激活图示例，其中激活图是利用训练后的 LSKNet 模型通过公式(8)

表 18: LSKNet-S 与其他（大型卷积核或动态/选择性注意力）骨干网络在遥感目标检测（DOTA-v1.0）、分割（Vaihingen）和变化检测（LEVIR-CD）任务上的比较。在相似或更低的复杂度预算下，本文提出的 LSKNet 达到了最佳的 mAP。

类别	模型 骨干网络	#P	Flops	DOTA-v1.0			Vaihingen			LEVIR-CD			
				mAP	@50	@75	F1	OA	mIoU	P.	R.	F1	IoU
基准模型	ResNet-18	11.2M	38.1G	50.54	79.27	55.33	90.15	92.62	82.47	92.97	90.61	91.77	84.80
大核	ViTDet [132]	86.6M	394.9G	45.60	74.41	49.39	81.01	83.74	54.91	80.72	90.59	85.37	74.48
	ConvNeXt v2-N [146]	15.0M	51.2G	52.91	80.81	58.58	89.13	92.15	81.17	93.12	89.73	91.39	84.15
	Swin-T [102]	28.3M	91.1G	51.54	80.81	56.71	90.74	93.01	83.40	93.04	90.25	91.63	84.55
	MSCAN-S [90]	13.1M	45.0G	52.52	81.12	57.92	91.16	93.04	84.10	93.39	91.14	92.25	85.62
	VAN-B1 [89]	13.4M	52.7G	52.69	81.15	58.11	91.30	93.12	84.41	93.31	91.20	92.24	85.60
动态/选择性注意力	ResNeSt-14 [57]	8.6M	57.9G	49.79	79.51	53.41	90.31	92.84	82.72	92.47	90.38	91.41	84.18
	SCNet-18 [58]	14.0M	50.7G	49.91	79.69	53.55	90.50	92.97	83.04	92.03	91.27	91.65	84.58
	DCN-Res50 [177]	26.2M	121.2G	49.26	79.74	52.97	90.93	93.07	83.72	92.84	90.67	91.74	84.74
	SKNet-26 [59]	14.5M	58.5G	51.53	80.67	56.51	90.83	93.01	83.56	93.09	91.09	92.08	85.32
本文	⋆ LSKNet-S	14.4M	54.4G	**53.32**	**81.48**	**58.83**	**91.81**	**93.61**	**85.12**	**93.44**	**91.13**	**92.27**	**85.65**

表 19: LSKNet-T 与 ResNet-18 作为骨干网络在 DOTA-v1.0 数据集上不同检测框架中的比较。LSKNet-T 在各种框架中均显著优于 ResNet-18，实现了更高的 mAP。

框架	ResNet-18	⋆ LSKNet-T
ORCNN [11]	79.27	81.31 **(+2.04)**
RoI Trans. [52]	78.32	80.89 **(+2.57)**
S^2A-Net [53]	76.82	80.15 **(+3.33)**
R3Det [9]	74.16	78.39 **(+4.23)**
#P (backbone only)	11.2M	4.3M **(-62%)**
FLOPs (backbone only)	38.1G	19.1G **(-50%)**

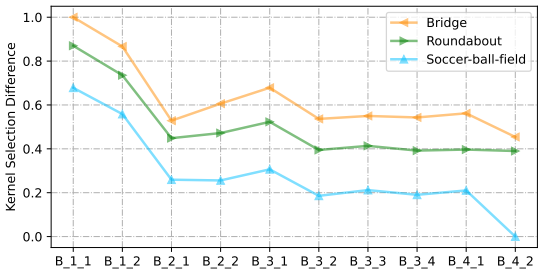


图 7: LSKNet-T 块中桥梁、环岛和足球场的归一化**核选择差异**。B_i_j 表示第 i 阶段的第 j 个 LSK 块。较大的值表示对更广泛上下文的依赖。

（即空间激活）计算得到。目标类别按照图 6所示的预期选择性感受野面积与真实边界框面积之比从左上到右下依次递减排列。空间激活图可视化结果进一步证实了模型的行为与本文挖掘出的两个先验假设和上述分析相一致，从而验证了所提出机制的有效性。

6 结论

本文提出了轻量级自适应大核网络（LSKNet），作为遥感图像分析下游任务（如场景分类、目标检测和语义分割）的新骨干网络。LSKNet 专门设计用于建模遥感图像的固有特征：更广泛和可适应的上下文理解。通过采用大空间感受野，LSKNet 能够有效捕获和建模遥感图像中不同目标类型所呈现的多样化上下文细节。大量实验表明，本文提出的轻量级模型在竞争激烈的遥感基准测试中达到了最先进的性能。本文进行的大量综合分析验证了所提出轻量级模型的有效性和重要性。

致谢

本研究得到了国家自然科学基金青年科学基金（批准号：62206134、62361166670、62276145、62176130、62225604、62301261）、中央高校基本科

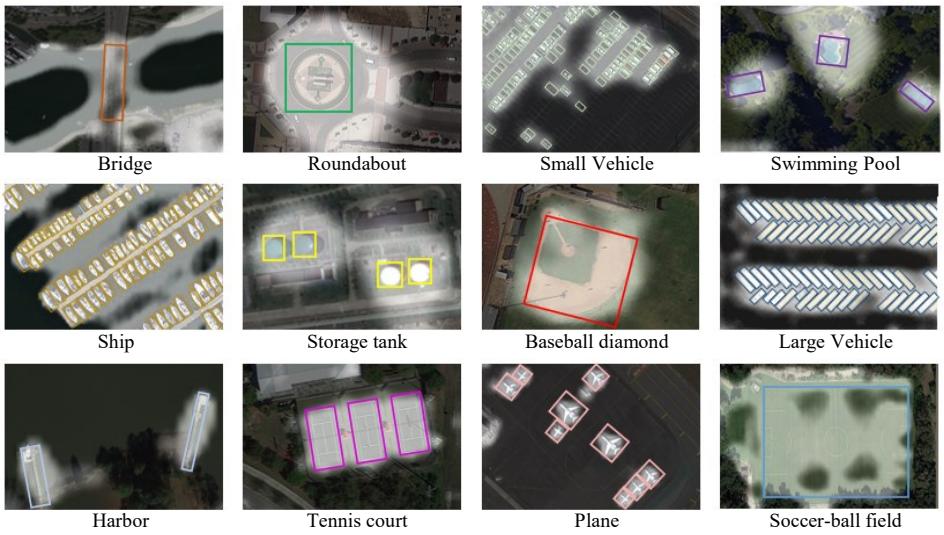


图 8: DOTA-v1.0 数据集中更多目标类别的感受野激活图，其中本文使用公式 (8)（即空间激活）可视化训练后的 LSKNet 模型的激活图。

研业务费(南开大学,070-63233084、070-63233089)以及天津市视觉计算与图像处理重点实验室的支持。计算资源由南开大学超级计算中心提供支持，同时得到了中国博士后科学基金（批准号：2021M701727）的资助。

数据可用性声明

公开存储库中的可用数据：

Imagenet 数据集可在以下网址获取：https://www.image-net.org/

UCM 数据集可在以下网址获取：http://weegee.vision.ucmerced.edu/datasets/landuse.html

AID 数据集可在以下网址获取：https://captain-whu.github.io/AID/

NWPU 数据集可在以下网址获取：https://www.tensorflow.org/datasets/catalog/resisc45

MillionAID 数据集可在以下网址获取：https://captain-whu.github.io/DiRS/

DOTA 数据集可在以下网址获取：https://captain-whu.github.io/DOTA/dataset.html

FAIR1M-v1.0 数据集可在以下网址获取：https://www.gaofen-challenge.com/benchmark

SAR-Aircraft 数据集可在以下网址获取：https://radars.ac.cn/web/data/getData?dataType=SARDataset_en

Potsdam 和 Vaihingen 数据集可在以下网址获取：https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx

LoveDA 数据集可在以下网址获取：https://codalab.lisn.upsaclay.fr/competitions/421

UAVid 数据集可在以下网址获取：https://uavid.nl/

GID 数据集可在以下网址获取：https://x-ytong.github.io/project/GID.html

LEVIR-CD 数据集可在以下网址获取：https://justchenhao.github.io/LEVIR/

S2Looking 数据集可在以下网址获取：https://github.com/S2Looking/Dataset

参考文献

[1] Chen, S.-B., Wei, Q.-S., Wang, W.-Z., Tang, J., Luo, B., Wang, Z.-Y.: Remote sensing

scene classification via multi-branch local attention network. TIP (2022)

[2] Zhao, Q., Lyu, S., Li, Y., Ma, Y., Chen, L.: Mgml: Multigranularity multilevel feature ensemble network for remote sensing scene classification. IEEE Transactions on Neural Networks and Learning Systems (2022)

[3] Zhao, Q., Ma, Y., Lyu, S., Chen, L.: Embedded self-distillation in compact multibranch ensemble network for remote sensing scene classification. TGRS (2022)

[4] Li, F., Feng, R., Han, W., Wang, L.: High-resolution remote sensing image scene classification via key filter bank based on convolutional neural network. TGRS (2020)

[5] Wang, D., Zhang, J., Du, B., Xia, G.-S., Tao, D.: An empirical study of remote sensing pretraining. TGRS (2022)

[6] Wang, D., Zhang, Q., Xu, Y., Zhang, J., Du, B., Tao, D., Zhang, L.: Advancing plain vision transformer towards remote sensing foundation model. TGRS (2022)

[7] Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., Li, J., Rong, X., Yang, Z., Chang, H., He, Q., Yang, G., Wang, R., Lu, J., Fu, K.: Ringmo: A remote sensing foundation model with masked image modeling. TGRS (2023)

[8] Han, J., Ding, J., Xue, N., Xia, G.-S.: ReDet: A rotation-equivariant detector for aerial object detection. In: CVPR (2021)

[9] Yang, X., Liu, Q., Yan, J., Li, A.: R3Det: Refined single-stage detector with feature refinement for rotating object. CoRR (2019)

[10] Yang, X., Yan, J., Ming, Q., Wang, W., Zhang, X., Tian, Q.: Rethinking rotated object detection with Gaussian Wasserstein distance loss. In: ICML (2021)

[11] Xie, X., Cheng, G., Wang, J., Yao, X., Han, J.: Oriented R-CNN for object detection. In: ICCV (2021)

[12] Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.-S., Bai, X.: Gliding vertex on the horizontal bounding box for multi-oriented object detection. TPAMI (2021)

[13] Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M.: UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing (2022)

[14] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR (2019)

[15] Zheng, X., Huan, L., Xia, G.-S., Gong, J.: Parsing very high resolution urban scene images by learning deep convnets with edge-aware loss. ISPRS Journal of Photogrammetry and Remote Sensing (2020)

[16] Hu, P., Perazzi, F., Heilbron, F.C., Wang, O., Lin, Z., Saenko, K., Sclaroff, S.: Real-time semantic segmentation with fast attention. IEEE Robotics and Automation Letters (2020)

[17] Li, R., Zheng, S., Zhang, C., Duan, C.,

Wang, L., Atkinson, P.M.: ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. ISPRS Journal of Photogrammetry and Remote Sensing (2021)

[18] Chen, Y., Yuan, X., Wu, R., Wang, J., Hou, Q., Cheng, M.-M.: YOLO-MS: Rethinking multi-scale representation learning for real-time object detection. arXiv (2023)

[19] Zhang, W., Jiao, L., Li, Y., Huang, Z., Wang, H.: Laplacian feature pyramid network for object detection in vhr optical remote sensing images. TGRS (2022)

[20] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., *et al.*: Segment anything. In: ICCV (2023)

[21] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS (2024)

[22] Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z.: Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. TGRS (2024)

[23] Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat: Grounded large vision-language model for remote sensing. arXiv (2023)

[24] Li, Y., Hou, Q., Zheng, Z., Cheng, M.-M., Yang, J., Li, X.: Large selective kernel network for remote sensing object detection. In: ICCV (2023)

[25] Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the International Conference on Advances in Geographic Information Systems (2010)

[26] Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X.: AID: A benchmark data set for performance evaluation of aerial scene classification. TGRS (2017)

[27] Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE (2017)

[28] Zhirui, W., Sun, X.: SAR-AIRcraft-1.0: High-resolution SAR Aircraft Detection and Recognition Dataset. https://radars.ac.cn/web/data/getData?dataType=SARDataset_en (2023)

[29] Photogrammetry, T.I.S., (ISPRS), R.S.: 2D Semantic Labeling Contest - Potsdam. https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx (2022)

[30] ISPRS: 2D Semantic Labeling - Vaihingen. https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx (2022)

[31] Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y.: LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. arXiv (2021)

[32] Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A., Yang, M.Y.: UAVid: A semantic segmentation dataset for uav imagery. ISPRS Journal of Photogrammetry and Remote

Sensing (2020)

[33] Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L.: Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sensing of Environment (2020)

[34] Chen, H., Shi, Z.: A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sensing (2020)

[35] Shen, L., Lu, Y., Chen, H., Wei, H., Xie, D., Yue, J., Chen, R., Lv, S., Jiang, B.: S2looking: A satellite side-looking dataset for building change detection. Remote Sensing (2021)

[36] Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: DOTA: A large-scale dataset for object detection in aerial images. In: CVPR (2018)

[37] Liu, Z., Wang, H., Weng, L., Yang, Y.: Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. TGRS Letters (2016)

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

[39] Su, Z., Zhang, J., Wang, L., Zhang, H., Liu, Z., Pietikäinen, M., Liu, L.: Lightweight pixel difference networks for efficient visual representation learning. TPAMI (2023)

[40] Sun, S., Zhi, S., Liao, Q., Heikkilä, J., Liu, L.: Unbiased scene graph generation via two-stage causal modeling. TPAMI (2023)

[41] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)

[42] Deng, P., Xu, K., Huang, H.: When CNNs meet vision transformer: A joint framework for remote sensing scene classification. TGRS Letters (2022)

[43] Bazi, Y., Bashmal, L., Rahhal, M.M.A., Dayil, R.A., Ajlan, N.A.: Vision transformers for remote sensing image classification. Remote Sensing (2021)

[44] Zhang, Q., Xu, Y., Zhang, J., Tao, D.: Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. IJCV (2023)

[45] Long, Y., Xia, G.-S., Li, S., Yang, W., Yang, M.Y., Zhu, X.X., Zhang, L., Li, D.: 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)

[46] Zaidi, S.S.A., Ansari, M.S., Aslam, A.,

Kanwal, N., Asghar, M., Lee, B.: A survey of modern deep learning based object detection models. Digital Signal Processing (2022)

[47] Mei, J., Zheng, Y.-B., Cheng, M.-M.: D2ANet: Difference-aware attention network for multi-level change detection from satellite imagery. Computational Visual Media (2023)

[48] Sun, X., Tian, Y., Lu, W., Wang, P., Niu, R., Yu, H., Fu, K.: From single- to multi-modal remote sensing imagery interpretation: a survey and taxonomy. Science China Information Sciences (2023)

[49] Zhang, W., Deng, W., Cui, Z., Liu, J., Jiao, L.: Object knowledge distillation for joint detection and tracking in satellite videos. TGRS (2024)

[50] Zhang, W., Jiao, L., Liu, F., Yang, S., Liu, J.: Dfat: Dynamic feature-adaptive tracking. IEEE Transactions on Circuits and Systems for Video Technology (2023)

[51] Li, Y., Li, X., Li, W., Hou, Q., Liu, L., Cheng, M.-M., Yang, J.: Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. arXiv (2024)

[52] Ding, J., Xue, N., Long, Y., Xia, G.-S., Lu, Q.: Learning RoI transformer for oriented object detection in aerial images. In: CVPR (2019)

[53] Han, J., Ding, J., Li, J., Xia, G.-S.: Align deep features for oriented object detection. TGRS (2020)

[54] Pan, X., Ren, Y., Sheng, K., Dong, W., Yuan, H., Guo, X., Ma, C., Xu, C.: Dynamic refinement network for oriented and densely packed object detection. In: CVPR (2020)

[55] Yang, X., Yang, X., Yang, J., Ming, Q., Wang, W., Tian, Q., Yan, J.: Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence. In: NeurIPS (2021)

[56] Zheng, Z., Ye, R., Hou, Q., Ren, D., Wang, P., Zuo, W., Cheng, M.-M.: Localization distillation for object detection. TPAMI (2023)

[57] Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., Smola, A.: ResNeSt: Split-attention networks. In: CVPRW (2022)

[58] Liu, J.-J., Hou, Q., Cheng, M.-M., Wang, C., Feng, J.: Improving convolutional networks with self-calibrated convolutions. In: CVPR (2020)

[59] Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: CVPR (2019)

[60] Wang, L., Li, R., Wang, D., Duan, C., Wang, T., Meng, X.: Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. Remote Sensing (2021)

[61] Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., Atkinson, P.M.: Multiattention network for semantic segmentation of fine-resolution remote sensing images.

TGRS (2021)

[62] Zhang, D., Zhang, H., Tang, J., Hua, X.-S., Sun, Q.: Causal intervention for weakly-supervised semantic segmentation. NeurIPS (2020)

[63] Daudt, R.C., Le Saux, B., Boulch, A.: Fully convolutional siamese networks for change detection. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 4063–4067 (2018). IEEE

[64] Fang, S., Li, K., Shao, J., Li, Z.: Snunet-cd: A densely connected siamese network for change detection of vhr images. IEEE Geoscience and Remote Sensing Letters (2021)

[65] Bandara, W.G.C., Patel, V.M.: A transformer-based siamese network for change detection. In: IEEE International Geoscience and Remote Sensing Symposium (2022)

[66] Codegoni, A., Lombardi, G., Ferrari, A.: Tinycd: A (not so) deep learning model for change detection. Neural Computing and Applications (2023)

[67] Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., Liu, G.: A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing (2020)

[68] Fang, S., Li, K., Li, Z.: Changer: Feature interaction is what you need for change detection. TGRS (2023)

[69] Zhao, S., Zhang, X., Xiao, P., He, G.: Exchanging dual-encoder–decoder: A new strategy for change detection with semantic guidance and spatial localization. TGRS (2023)

[70] Chen, H., Qi, Z., Shi, Z.: Remote sensing image change detection with transformers. TGRS (2021)

[71] Lin, H., Hang, R., Wang, S., Liu, Q.: Diformer: A difference transformer network for remote sensing change detection. IEEE Geoscience and Remote Sensing Letters (2024)

[72] Wang, D., Zhang, J., Xu, M., Liu, L., Wang, D., Gao, E., Han, C., Guo, H., Du, B., Tao, D., et al.: Mtp: Advancing remote sensing foundation model via multi-task pretraining. arXiv (2024)

[73] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS (2017)

[74] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)

[75] Zhang, C., Wang, L., Cheng, S., Li, Y.: SwinSUNet: Pure transformer network for remote sensing image change detection. TGRS (2022)

[76] Panboonyuen, T., Jitkajornwanich, K., Lawawirojwong, S., Srestasathiern, P., Vateekul, P.: Transformer-based decoder designs for semantic segmentation on remotely sensed images. Remote Sensing

(2021)

[77] Wang, X., Chen, G., Qian, G., Gao, P., Wei, X.-Y., Wang, Y., Tian, Y., Gao, W.: Large-scale multi-modal pre-trained models: A comprehensive survey. Machine Intelligence Research (2023)

[78] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: ICCV (2021)

[79] Wu, Y.-H., Liu, Y., Zhan, X., Cheng, M.-M.: P2T: Pyramid pooling transformer for scene understanding. TPAMI (2022)

[80] Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)

[81] Yan, H., Li, Z., Li, W., Wang, C., Wu, M., Zhang, C.: ConTNet: Why not use convolution and transformer at the same time? CoRR (2021)

[82] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR (2021)

[83] Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: NeurIPS (2016)

[84] Fan, D.-P., Ji, G.-P., Xu, P., Cheng, M.-M., Sakaridis, C., Gool, L.V.: Advances in deep concealed scene understanding. Visual Intelligence (2023)

[85] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR (2022)

[86] Ding, X., Zhang, X., Han, J., Ding, G.: Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs. In: CVPR (2022)

[87] Liu, S., Chen, T., Chen, X., Chen, X., Xiao, Q., Wu, B., Pechenizkiy, M., Mocanu, D., Wang, Z.: More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. ArXiv (2022)

[88] Gao, S., Li, Z.-Y., Han, Q., Cheng, M.-M., Wang, L.: RF-Next: Efficient receptive field search for convolutional neural networks. TPAMI (2023)

[89] Guo, M.-H., Lu, C., Liu, Z.-N., Cheng, M.-M., Hu, S.: Visual attention network. Computational Visual Media (2022)

[90] Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z.-N., Cheng, M.-M., Hu, S.-M.: SegNeXt: Rethinking convolutional attention design for semantic segmentation. In: NeurIPS (2022)

[91] Hou, Q., Lu, C.-Z., Cheng, M.-M., Feng, J.: Conv2Former: A simple transformer-style ConvNet for visual recognition. ArXiv (2022)

[92] Guo, M.-H., Xu, T., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R., Cheng, M.-M., Hu, S.-M.: Attention mechanisms in computer vision: A survey. Computational Visual Media (2021)

[93] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)

[94] Hu, J., Shen, L., Albanie, S., Sun, G., Vedaldi, A.: Gather-Excite: Exploiting feature context in convolutional neural networks. In: NeurPIS (2018)

[95] Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: GCNet: Non-local networks meet squeeze-excitation networks and beyond. In: ICCVW (2019)

[96] Li, Z., Sun, Y., Zhang, L., Tang, J.: Ctnet: Context-based tandem network for semantic segmentation. TPAMI (2022)

[97] Li, Y., Li, X., Yang, J.: Spatial group-wise enhance: Enhancing semantic feature learning in cnn. In: ACCV (2022)

[98] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: ECCV (2018)

[99] Park, J., Woo, S., Lee, J.-Y., Kweon, I.-S.: BAM: Bottleneck attention module. In: British Machine Vision Conference (2018)

[100] Srivastava, S., Sharma, G.: Omnivec: Learning robust representations with cross modal sharing. In: Winter Conference on Applications of Computer Vision (2024)

[101] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020)

[102] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: CVPR (2021)

[103] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: PVT v2: Improved baselines with pyramid vision transformer. Computational Visual Media (2022)

[104] Zhang, X., Tian, Y., Xie, L., Huang, W., Dai, Q., Ye, Q., Tian, Q.: Hivit: A simpler and more efficient design of hierarchical vision transformer. In: ICLR (2022)

[105] Xu, Y., Zhang, Q., Zhang, J., Tao, D.: Vitae: Vision transformer advanced by exploring intrinsic inductive bias. NeurIPS (2021)

[106] Yu, H., Tian, Y., Ye, Q., Liu, Y.: Spatial transform decoupling for oriented object detection. In: AAAI (2024)

[107] Yang, B., Bender, G., Le, Q.V., Ngiam, J.: CondConv: Conditionally parameterized convolutions for efficient inference. NeurIPS (2019)

[108] Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: Attention over convolution kernels. In: CVPR (2020)

[109] Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: CVPR (2019)

[110] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: ICCV (2017)

[111] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR (2022)

[112] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y.,

Wang, X., Feng, J., Yan, S.: MetaFormer is actually what you need for vision. In: CVPR (2022)

[113] Hendrycks, D., Gimpel, K.: Bridging nonlinearities and stochastic regularizers with gaussian error linear units. CoRR (2016)

[114] Zhang, G., Xu, W., Zhao, W., Huang, C., Yk, E.N., Chen, Y., Su, J.: A multiscale attention network for remote sensing scene images classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2021)

[115] He, N., Fang, L., Li, S., Plaza, J., Plaza, A.: Skip-connected covariance network for remote sensing scene classification. IEEE Transactions on Neural Networks and Learning Systems (2020)

[116] Liu, C., Dai, H., Wang, S., Chen, J.: Remote sensing image scene classification based on multidimensional attention and feature enhancement. IAENG International Journal of Computer Science (2023)

[117] Wang, S., Guan, Y., Shao, L.: Multigranularity canonical appearance pooling for remote sensing scene classification. TIP (2020)

[118] Bi, Q., Qin, K., Zhang, H., Xia, G.-S.: Local semantic enhanced convnet for aerial scene recognition. TIP (2021)

[119] Wang, S., Ren, Y., Parr, G.P., Guan, Y., Shao, L.: Invariant deep compressible covariance pooling for aerial scene categorization. TGRS (2020)

[120] Zhang, X., An, W., Sun, J., Wu, H., Zhang, W., Du, Y.: Best representation branch model for remote sensing image scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2021)

[121] Zhao, Z., Li, J., Luo, Z., Li, J., Chen, C.: Remote sensing image scene classification based on an enhanced attention module. TGRS Letters (2020)

[122] Li, B., Guo, Y., Yang, J., Wang, L., Wang, Y., An, W.: Gated recurrent multiattention network for VHR remote sensing image classification. TGRS (2021)

[123] Wang, W., Sun, Y., Li, J., Wang, X.: Frequency and spatial based multi-layer context network (fscnet) for remote sensing scene classification. International Journal of Applied Earth Observation and Geoinformation (2024)

[124] Dong, Z., Gu, Y., Liu, T.: Upetu: A unified parameter-efficient fine-tuning framework for remote sensing foundation model. TGRS (2024)

[125] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)

[126] Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., Chen, K.: RTMDet: An empirical study of designing real-time object detectors. CoRR (2022)

[127] Guo, Z., Liu, C., Zhang, X., Jiao, J., Ji, X., Ye, Q.: Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection. In: CVPR (2021)

[128] Lang, S., Ventola, F., Kersting, K.: DAFNe: A one-stage anchor-free deep model for oriented object detection. CoRR (2021)

[129] Hou, L., Lu, K., Xue, J., Li, Y.: Shape-adaptive selection and measurement for oriented object detection. In: AAAI (2022)

[130] Dai, L., Liu, H., Tang, H., Wu, Z., Song, P.: AO2-DETR: Arbitrary-oriented object detection transformer. IEEE Transactions on Circuits and Systems for Video Technology (2022)

[131] Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., Fu, K.: SCRDet: Towards more robust detection for small, cluttered and rotated objects. In: ICCV (2019)

[132] Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: ECCV (2022)

[133] Wang, J., Yang, W., Li, H.-C., Zhang, H., Xia, G.-S.: Learning center probability map for detecting objects in aerial images. TGRS (2021)

[134] Yang, X., Yan, J.: Arbitrary-oriented object detection with circular smooth label. In: ECCV (2020)

[135] Cheng, G., Yao, Y., Li, S., Li, K., Xie, X., Wang, J., Yao, X., Han, J.: Dual-aligned oriented detector. TGRS (2022)

[136] Cheng, G., Wang, J., Li, K., Xie, X., Lang, C., Yao, Y., Han, J.: Anchor-free oriented proposal generator for object detection. TGRS (2022)

[137] Yang, X., Zhou, Y., Zhang, G., Yang, J., Wang, W., Yan, J., Zhang, X., Tian, Q.: The KFIoU loss for rotated object detection. In: ICLR (2022)

[138] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)

[139] Cai, Z., Vasconcelos, N.: Cascade R-CNN: Delving into high quality object detection. In: CVPR (2018)

[140] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)

[141] Ming, Q., Zhou, Z., Miao, L., Zhang, H., Li, L.: Dynamic anchor learning for arbitrary-oriented object detection. CoRR (2020)

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

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

[144] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

[145] Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., Torr, P.: Res2Net: A new multi-scale backbone architecture. TPAMI (2021)

[146] Woo, S., Debnath, S., Hu, R., Chen, X., Liu,

Z., Kweon, I.-S., Xie, S.: ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. Arxiv (2023)

[147] Cai, Z., Vasconcelos, N.: Cascade R-CNN: High quality object detection and instance segmentation. TPAMI (2019)

[148] Li, R., Duan, C., Zheng, S., Zhang, C., Atkinson, P.M.: Macu-net for semantic segmentation of fine-resolution remotely sensed images. IEEE Geoscience and Remote Sensing Letters **19** (2022)

[149] Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R.: ERFNet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems (2017)

[150] Li, G., Yun, I., Kim, J., Kim, J.: DABNet: Depth-wise Asymmetric Bottleneck for Real-time Semantic Segmentation (2019)

[151] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In: ECCV (2018)

[152] Oršić, M., Šegvić, S.: Efficient semantic segmentation with pyramidal fusion. Pattern Recognition (2021)

[153] Zhuang, J., Yang, J., Gu, L., Dvornek, N.: ShelfNet for fast semantic segmentation. In: ICCVW (2019)

[154] Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: ICCV (2021)

[155] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)

[156] Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. In: CVPR (2021)

[157] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)

[158] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS (2021)

[159] Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV (2018)

[160] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (2018)

[161] Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: CVPR (2019)

[162] Zheng, Z., Zhong, Y., Wang, J., Ma, A.: Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In: CVPR (2020)

[163] Ma, A., Wang, J., Zhong, Y., Zheng, Z.: FactSeg: Foreground activation-driven small

object semantic segmentation in large-scale remote sensing imagery. TGRS (2021)

[164] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv (2021)

[165] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. TPAMI (2019)

[166] Wang, L.-L., Lui, S.S., Chan, R.C.: The past and future of mapping the biomarkers of psychosis. Current Opinion in Behavioral Sciences (2022)

[167] Sun, L., Zou, H., Wei, J., Cao, X., He, S., Li, M., Liu, S.: Semantic segmentation of high-resolution remote sensing images based on sparse self-attention and feature alignment. Remote Sensing (2023)

[168] Yang, M.Y., Kumaar, S., Lyu, Y., Nex, F.: Real-time semantic segmentation with context aggregation network. ISPRS Journal of Photogrammetry and Remote Sensing (2021)

[169] Xu, W., Xu, Y., Chang, T., Tu, Z.: Co-scale conv-attentional image transformers. In: ICCV (2021)

[170] Liu, Y., Pang, C., Zhan, Z., Zhang, X., Yang, X.: Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. IEEE Geoscience and Remote Sensing Letters (2020)

[171] Han, C., Wu, C., Guo, H., Hu, M., Chen, H.: Hanet: A hierarchical attention network for change detection with bi-temporal very-high-resolution remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2023)

[172] Chen, H., Li, W., Shi, Z.: Adversarial instance augmentation for building change detection in remote sensing images. TGRS (2021)

[173] Zhang, C.-j., Liu, J.-w.: Change detection with incorporating multi-constraints and loss weights. Engineering Applications of Artificial Intelligence (2024)

[174] Han, C., Wu, C., Du, B.: Hcgmnet: A hierarchical change guiding map network for change detection. In: IEEE International Geoscience and Remote Sensing Symposium (2023)

[175] Han, C., Wu, C., Hu, M., Li, J., Chen, H.: C2f-semicd: A coarse-to-fine semi-supervised change detection method based on consistency regularization in high-resolution remote-sensing images. TGRS (2024)

[176] Han, C., Wu, C., Guo, H., Hu, M., Li, J., Chen, H.: Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2023)

[177] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G.,

Hu, H., Wei, Y.: Deformable convolutional networks. In: ICCV (2017)

[178] Muhammad, M.B., Yeasin, M.: Eigen-CAM: Class activation map using principal components. CoRR (2020)