SegNeXt: 重新思考用于语义分割的卷积注意力设计 *

国孟昊 [1], 陆承泽 [2], 侯淇彬 [2], 刘政宁 [3], 程明明 [2], 胡事民 [1]
[1] 清华大学 [2] 南开大学 [3] 北京非十科技有限公司

我们提出了 SegNeXt，一个用于语义分割的简单卷积网络架构。由于自注意力在编码空间信息方面的效率，最近基于 transformer 的模型在语义分割领域占主导地位。在本文中，我们表明卷积注意力是一种比 transformer 中的自注意力机制更有效率的编码上下文信息的方式。通过重新审视成功的分割模型所拥有的特征，我们发现了导致分割模型性能提高的几个关键部分。这促使我们设计了一个新的卷积注意力网络，它使用廉价的卷积操作。在没有任何附加条件的情况下，我们的 SegNeXt 显著提高了以前最先进的方法在流行基准上的性能，包括 ADE20K、Cityscapes、COCO-Stuff、Pascal VOC、Pascal Context 和 iSAID. 值得注意的是，SegNeXt 超过了 EfficientNet-L2 w/ NAS-FPN，在 Pascal VOC 2012 的测试排行榜上只用了它的1/10 的参数就达到了 90.6 % 的 mIoU。与最先进的方法相比，SegNeXt 在 ADE20K 数据集上以相同或更少的计算量平均实现了约 2.0% 的 mIoU 提升。代码公开在https://github.com/uyzhang/JSeg（Jittor）和 https://github.com/Visual-Attention-Network/SegNeXt(Pytorch)。

1. 引言

作为计算机视觉中最基本的研究课题之一，语义分割旨在为每个像素分配一个语义类别，在过去十年中引起了极大的关注。从早期基于 CNN 的模型，如 FCN [54] 和 DeepLab 系列 [5, 7, 9]，到最近基于 transformer 的方法，如 SETR[96] 和 SegFormer [81]，语义分割模型在网络架构方面经历了重大变革。

通过重新审视以前成功的语义分割工作，我们总结了不同模型所拥有的几个关键属性，如表 1所示。基于上述观察，我们认为一个成功的语义分割模型应该具有以下特征：（i）强大的骨干网络作为编码器。与以前基于 CNN 的模型相比，基于 transformer 的模型的性能提高主要来自于更强的骨干网络。(ii) 多尺度信息

*本文是 NeurIPS 2022 论文 [24] 的中译版。

表 1. 我们从成功的语义分割方法中观察到有利于提高模型性能的特性。这里，n 指的是 pixel 或 token 的数量。Strong encoder 表示强骨干网络, 例如 ViT [17] and VAN [25].

Properties	DeepLabV3+	HRNet	SETR	SegFormer	SegNeXt
Strong encoder	✗	✗	✓	✓	✓
Multi-scale interaction	✓	✓	✗	✗	✓
Spatial attention	✗	✗	✓	✓	✓
Computational complexity	$\mathcal{O}(n)$	$\mathcal{O}(n)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n)$

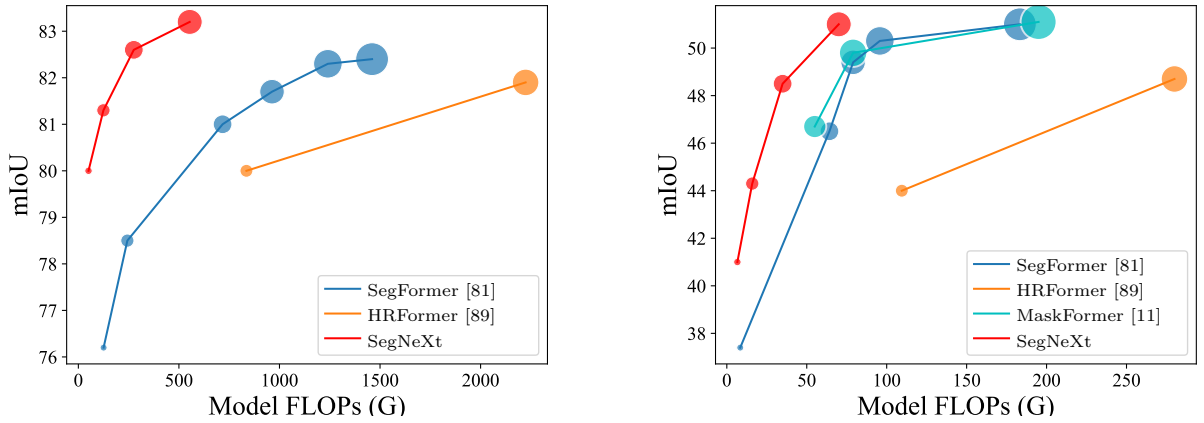


图 1. 在 Cityscapes（左）和 ADE20K（右）验证集上的性能-计算曲线。计算 FLOPs 时，Cityscapes 的输入尺寸为 $2,048 \times 1,024$，ADE20K 为 512×512。圆圈的大小表示参数的数量。更大的圆圈意味着更多的参数。我们可以看到，我们的 SegNeXt 实现了分割性能和计算复杂性之间的最佳权衡。计算复杂度之间的最佳平衡。

交互。与图像分类任务大多识别单一物体不同，语义分割是一个密集的预测任务，因此需要处理单一图像中不同大小的物体。(iii) 空间注意力。空间注意力允许模型通过对语义区域内的区域进行优先排序来进行分割。(iv) 低计算复杂度。这在处理遥感和城市场景的高分辨率图像时尤其关键。

考虑到上述分析，在本文中，我们重新思考了卷积注意力的设计，并提出了一个高效而有效的用于语义分割的编码器-解码器架构。与之前的基于 transformer 的模型不同，我们的方法是将解码器中的卷积作为特征细化器，倒置了 transformer-卷积的编码器-解码器架构。具体来说，对于我们编码器中的每个块，我们翻新了传统卷积块的设计，并利用多尺度的卷积特征，通过一个简单的元素 element-wise 来唤起空间注意 [25]。我们发现这样一种简单的建立空间注意力的方法比标准卷积和空间信息编码中的自我注意力都要有效。对于解码器，我们从不同的阶段收集多层次的特征，并使用 Hamburger [21] 来进一步提取全局上下文信息。在这种情况下，我们的方法可以获得从局部到全局的多尺度背景，实现空间和通道维度的适应性，并从低级到高级的信息聚合。

我们的网络，称为 SegNeXt，除了解码器部分，主要由卷积运算组成，其中包含一个基于分解的 Hamburger 模块 [21]（Ham）用于全局信息提取。这使得我们的 SegNeXt 比以前严重依赖 transformer 的分割方法更有效率。如图 1所示，SegNeXt 明显优于最近的基于 transformer 的方法。特别是我们的 SegNeXt-S 在处理 Cityscapes 数据集中的高分辨率城市场景时，只用了大约 1/6 的计算成本（124.6G vs. 717.1G）和 1/2 的参数（13.9M vs. 27.6M）就超过了 SegFormer-B2（81.3% vs. 81.0%）

我们的贡献可以概括为以下几点:

- 我们确定了一个好的语义分割模型应该拥有的特征，并提出了一个新颖的定制网络架构，称为 SegNeXt，它通过多尺度卷积特征唤起空间注意。
- 我们表明具有简单而廉价的卷积的编码器仍然可以比 vision transformer 表现得更好，特别是在处理物体细节时，而它需要的计算成本要少得多。
- 我们的方法将最先进的语义分割方法在各种分割基准上的性能提高了一大截，包括 ADE20K、Cityscapes, COCO-Stuff, Pascal VOC, Pascal Context 和 iSAID 数据集。

2. 相关工作

2.1. 语义分割

语义分割是一项基本的计算机视觉任务。自 FCN [54] 提出以来，卷积神经网络（CNN） [1, 65, 87, 95, 19, 88, 72, 20, 45] 已经取得了巨大的成功，并成为语义分割的一个流行架构。最近，基于 transformer 的方法 [97, 81, 89, 66, 64, 44, 11, 10] 显示出巨大的潜力，并超过了基于 CNN 的方法。

在深度学习时代，分割模型的架构可以大致分为两部分：编码器和解码器。对于编码器，研究人员通常采用流行的分类网络 (例如 ResNet [28]、ResNeXt [82] 和 DenseNet [33])，而不是定制的架构。然而，语义分割是一种密集预测任务，它与图像分类不同。分类的改进可能不会出现在具有挑战性的分割任务中 [29]。因此，出现了一些定制的编码器，包括 Res2Net [20], HRNet [72], SETR [97], SegFormer [81], HRFormer [89], MPViT [39], DPT [64] 等。对于解码器来说，它经常与编码器配合使用，以达到更好的效果。针对不同的目标，有不同类型的解码器，包括实现多尺度的感受野 [95, 8, 79]，收集多尺度语义 [65, 81, 9]，扩大感受野 [6, 5, 63]，加强边缘特征 [96, 2, 16, 42, 91]，以及捕捉到全局上下文信息 [19, 35, 90, 46, 23, 27, 92]。

在本文中，我们总结了那些为语义分割而设计的成功模型的特点，并提出了一个基于 CNN 的模型，名为 SegNeXt。与我们的论文最相关的工作是 [63]，它将 $k \times k$ 卷积分解为一对 $k \times 1$ 和 $1 \times k$ 卷积。虽然这项工作表明大卷积核在语义分割中很重要，但它忽略了多尺度感受野的重要性，也没有考虑如何利用大卷积核提取的这些多尺度特征以注意力的形式进行分割。

2.2. 多尺度网络

设计多尺度网络是计算机视觉的热门方向之一。对于分割模型，多尺度出现在编码器 [72, 20, 68] 和解码器 [95, 87, 7] 两部分。GoogleNet [68] 是与我们的方法最相关的多尺度架构之一，它使用多分支结构来实现多尺度特征提取。另一项与我们的方法相关的工作是 HRNet [72]。在深层阶段，HRNet 也保留了高分辨率的特征，与低分辨率的特征聚合在一起，以实现多尺度特征提取。

与以往的方法不同，SegNeXt 除了在编码器中捕获多尺度特征外，还引入了一个有效的注意力机制，并采用了更便宜和更大核的卷积。这些使我们的模型能够达到比上述分割方法更高的性能。

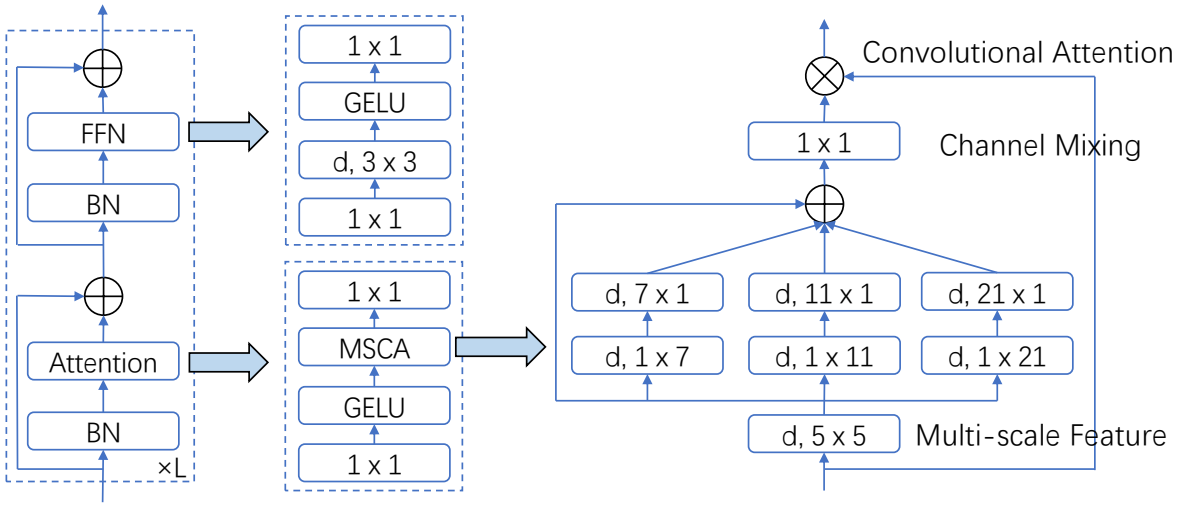
2.3. 注意力机制

注意机制是一种自适应的选择过程，其目的是使网络集中于重要的部分。一般来说，它在语义分割中可以分为两类 [26]，包括通道注意力和空间注意力。不同类型的注意力起着不同的作用。例如，空间注意力主要关心重要的空间区域 [17, 14, 58, 52, 22]. 不同的是，使用通道注意力的目的是使网络有选择地注意到那些重要的对象。到那些重要的物体，这在以前的工作中已经被证明是很重要的 [31, 4, 73]。说到最近流行的 vision transformer [17, 52, 83, 75, 74, 51, 81, 34, 50, 89]，他们通常忽略了通道维度的适应性。

视觉注意力网络（VAN） [25] 是与 SegNeXt 最相关的工作，它也提出利用大核注意力（LKA）机制来建立通道注意力和空间注意力。虽然 VAN 在图像分类中取得了很好的表现，但它在网络设计过程中忽略了多尺度特征聚合的作用，而这对于类似分割的任务是至关重要的。

3. 方法

在这一节中，我们将详细描述提出的 SegNeXt 的结构。基本上，我们采用了一个编码器-解码器的架构，这与以前的大多数工作一样，简单易行。



(a) A stage of MSCAN (b) MSCA

图 2. 拟议的 MSCA 和 MSCAN 的说明。这里，$d, k_1 \times k_2$ 意味着使用 $k_1 \times k_2$ 的核大小的深度卷积 (d)。我们使用卷积提取多尺度特征，然后利用它们作为注意力权重来重新权衡 MSCA 的输入。

3.1. 卷积编码器

我们的编码器采用了金字塔结构，这与之前的大多数工作 [81, 6, 19] 一致。对于我们编码器中的构件，我们采用了与 ViT [17, 81] 类似的结构，但不同的是，我们没有使用自我注意机制，而是设计了一个新颖的多尺度卷积注意（MSCA）模块。如图 2(a) 所示，MSCA 包含三个部分：一个深度卷积来聚集局部信息，多分支深度卷积来捕捉多尺度上下文信息，以及一个 1×1 卷积来模拟不同通道之间的关系。1×1 卷积的输出被直接用作注意力权重，以重新权衡 MSCA 的输入。在数学上，我们的 MSCA 可以写成：

$$\text{Att} = \text{Conv}_{1\times1}(\sum_{i=0}^{3} \text{Scale}_i(\text{DW-Conv}(F))), \tag{1}$$

$$\text{Out} = \text{Att} \otimes F. \tag{2}$$

其中 F 代表输入特征。Att 和 Out 分别为注意图和输出。$\otimes$ 是逐元素的矩阵乘法运算。DW-Conv 表示深度卷积，Scale_i，$i \in \{0, 1, 2, 3\}$，表示 图 2（b）中的第 i 个分支。Scale_0 是身份连接。按照 [63]，在每个分支中，我们使用两个深度的带状卷积来近似于大卷积核的标准 depth-wise 的卷积。这里，每个分支的核大小分别被设定为 7、11 和 21。我们选择 depth-wise 条状卷积的原因有两个方面。一方面，带状卷积是轻量级的。为了模仿核大小为 7×7 的标准二维卷积，我们只需要一对 7×1 和 1×7 的卷积。另一方面，在分割场景中存在一些条状物体，如人和电线杆。因此，条状卷积可以作为网格卷积的补充，有助于提取条状特征 [63, 30]。

将一连串的构件堆叠在一起，就得到了拟议的卷积编码器，命名为 MSCAN。对于 MSCAN，我们采用一个普通的分层结构，它包含四个阶段 $\frac{H}{4} \times \frac{W}{4}$，$\frac{H}{8} \times \frac{W}{8}$，$\frac{H}{16} \times \frac{W}{16}$ 和 $\frac{H}{32} \times \frac{W}{32}$，其空间分辨率递减。这里，$H$ 和 W 分别是输入图像的高度和宽度。每个阶段都包含一个下采样块和一个堆栈的构建块，如上所述。下采样块有一个具有步长为 2 和核大小 3×3 的卷积，然后是一个批处理归一化层 [36]。请注意，在 MSCAN 的每个构建块中，我们使用批量归一化而不是层归一化，因为我们发现批量归一化对分割性能的增益更大。

我们设计了四个不同规模的编码器模型，分别命名为 MSCAN-T、MSCAN-S、MSCAN-B 和 MSCAN-L。相应的整体分割模型分别称为 SegNeXt-T, SegNeXt-S, SegNeXt-B, SegNeXt-L。详细的网络设置显示在表 2.

表 2. 拟议中的 SegNeXt 的不同尺寸的详细设置。在这个表中，‘e.r.’ 代表前馈网络的扩展率. ‘C’ 和 ‘L’ 分别是通道和 block 的数量 ‘Decoder dimension’ 表示解码器中的 MLP 尺寸。‘Parameters’ 是根据 ADE20K 数据集 [99] 计算的。由于不同的数据集中类别的数量不同，参数的数量也不同。数据集中类别的数量不同，参数的数量可能略有变化。

stage	output size	e.r.	SegNeXt-T	SegNeXt-S	SegNeXt-B	SegNeXt-L
1	$\frac{H}{4} \times \frac{W}{4} \times C$	8	$C = 32,\ L = 3$	$C = 64,\ L = 2$	$C = 64,\ L = 3$	$C = 64,\ L = 3$
2	$\frac{H}{8} \times \frac{W}{8} \times C$	8	$C = 64,\ L = 3$	$C = 128,\ L = 2$	$C = 128,\ L = 3$	$C = 128\ ,\ L = 5$
3	$\frac{H}{16} \times \frac{W}{16} \times C$	4	$C = 160,\ L = 5$	$C = 320, L = 4$	$C = 320, L = 12$	$C = 320, L = 27$
4	$\frac{H}{32} \times \frac{W}{32} \times C$	4	$C = 256, L = 2$	$C = 512, L = 2$	$C = 512, L = 3$	$C = 512, L = 3$
Decoder dimension			256	256	512	1,024
Parameters (M)			4.3	13.9	27.6	48.9

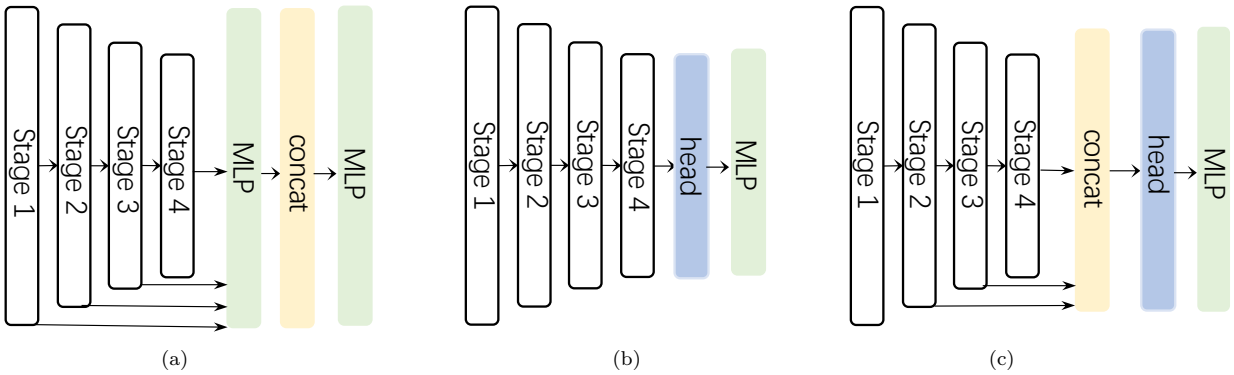


图 3. Three different decoder designs.

3.2. 解码器

在分割模型 [81, 97, 6] 中，编码器大多是在 ImageNet 数据集上预训练的。为了捕捉高层次的语义，通常需要一个解码器，它被应用在编码器之后。在这项工作中，我们研究了三种简单的解码器结构，如图 3所示。第一种，在 SegFormer [81] 中采用，是一种纯粹的基于 MLP 的结构。第二种是主要采用基于 CNN 的模型。在这种结构中，编码器的输出被直接用作重型解码器头的输入，如 ASPP [6]、PSP [95] 和 DANet [19]。最后一种是我们的 SegNeXt 中采用的结构。我们将后三个阶段的特征集合起来，并使用一个轻量级的 Hamburger [21] 来进一步对全局环境进行建模。结合我们强大的卷积编码器，我们发现使用一个轻量级的解码器可以提高性能与计算效率。

值得注意的是，与 SegFormer 不同的是，SegFormer 的解码器聚集了第一阶段到第四阶段的特征，而我们的解码器只接收最后三个阶段的特征。这是因为我们的 SegNeXt 是基于卷积的。第一阶段的特征包含了太多的低层次信息，会影响性能。此外，对第一阶段的操作带来了沉重的计算开销。在实验部分，我们将展示我们的卷积 SegNeXt 比最近最先进的基于 transformer 的 SegFormer [81] 和 HRFormer [89] 表现得更好。

表 3. 在 ImageNet 验证集上与最先进的方法进行比较。'Acc.' 表示 Top-1 准确率.

Method	Params. (M)	Acc. (%)
MiT-B0 [81]	3.7	70.5
VAN-Tiny [25]	4.1	75.4
MSCAN-T	4.2	75.9
MiT-B1 [81]	14.0	78.7
VAN-Small [25]	13.9	81.1
MSCAN-S	14.0	81.2
MiT-B2 [81]	25.4	81.6
Swin-T [52]	28.3	81.3
ConvNeXt-T [53]	28.6	82.1
VAN-Base [25]	26.6	82.8
MSCAN-B	26.8	83.0
MiT-B3 [28]	45.2	83.1
Swin-S [52]	49.6	83.0
ConvNeXt-S [52]	50.1	83.1
VAN-Large [25]	44.8	83.9
MSCAN-L	45.2	83.9

表 4. 在遥感数据集 iSAID 上与最先进的方法比较。默认采用的是单一尺度（SS）测试。我们的 SegNeXt-T 已经取得了卓越的性能。

Method	Backbone	mIoU (%)
DenseASPP [84]	ResNet50	57.3
PSPNet [95]	ResNet50	60.3
SemanticFPN [37]	ResNet50	62.1
RefineNet [48]	ResNet50	60.2
HRNet [72]	HRNetW-18	61.5
GSCNN [69]	ResNet50	63.4
SFNet [43]	ResNet50	64.3
RANet [60]	ResNet50	62.1
PointRend [38]	ResNet50	62.8
FarSeg [98]	ResNet50	63.7
UperNet [80]	Swin-T	64.6
PointFlow [41]	ResNet50	66.9
SegNeXt-T	MSCAN-T	68.3
SegNeXt-S	MSCAN-S	68.8
SegNeXt-B	MSCAN-B	69.9
SegNeXt-L	MSCAN-L	70.3

4. 实验

数据集: 我们在七个流行的数据集上评估了我们的方法，包括 ImageNet-1K [15]、ADE20K [99]、Cityscapes [13]、Pascal VOC [18]、Pascal Context [59]、COCO Stuff [3]，以及 iSAID [77]。ImageNet [15] 是最著名的图像分类数据集，它包含 1000 个类别。与大多数分割方法类似，我们用它来预训练我们的 MSCAN 编码器。ADE20K [99] 是一个具有挑战性的数据集，包含 150 个语义类别。它由 20,210/2,000/3,352 张图像组成的训练、验证和测试集。Cityscapes [13] 主要关注城市场景，包含 5,000 张高分辨率图像，有 19 个类别。其中有 2,975/500/1,525 张图像，分别用于训练、验证和测试。Pascal VOC [18] 涉及 20 个前景类和一个背景类。扩增后，它有 10, 582/1, 449/1, 456 个图像，分别用于训练、验证和测试。Pascal Context [59] 包含 59 个前景类和一个背景类。类和一个背景类。训练集和验证集分别包含 4,996 和 5,104 张图像。COCO-Stuff [3] 也是一个具有挑战性的数据集，它包含 172 个语义类别和总共 164k 张图像。它包括 15 个前景类和一个背景类。iSAID [77] 是一个大规模的航空图像分割基准，它包括 15 个前景类和一个背景类。其训练、验证和测试集分别涉及 1,411/458/937 张图像。

实验细节: 我们通过使用 Jittor [32] 和 [62] 进行实验。我们的实现是基于 timm（Apache-2.0） [78] 和 mmsegmentation（Apache-2.0） [12] 库，分别用于分类和分割。我们的分割模型的所有编码器都在 ImageNet-1K 数据集 [15] 上进行了预训练。我们分别采用 Top-1 准确率和平均交集大于联盟（mIoU）作为分类和分割的评价指标。所有模型都是在一个有 8 个 RTX 3090 GPU 的节点上训练的。

对于 ImageNet 预训练，我们的数据增强方法和训练设置与 DeiT [71] 相同。对于分割实验，我们采用了一些常见的数据增强方法，包括随机水平翻转、随机缩放（从 0.5 到 2）和随机剪裁。Cityscapes 数据集的批量大小被设置为 8，其他所有数据集的批量大小为 16。AdamW [55] 被用来训练我们的模型。我们将初始学习率设置为 0.00006，并采用多学习率衰减策略。我们对 ADE20K、Cityscapes 和 iSAID 数据集进行了 16 万次迭代训练，对 COCO-Stuff、Pascal VOC 和 Pascal Context 数据集进行了 8 万次迭代。在测试过程中，我们同时使用单尺度（SS）和多尺度（MS）的翻转测试策略进行公平的比较。更多的细节可以在我们的补充

表 5. 解码器中不同注意力机制的表现。SegNeXt-B w/ Ham 是指 MSCAN-B 编码器加上 Ham 解码器。FLOPs 是用 512×512 的输入尺寸计算的。

Architecture	Params. (M)	GFLOPs	mIoU (SS)	mIoU (MS)
SegNeXt-B w/ CC [35]	27.8	35.7	47.3	48.6
SegNeXt-B w/ EMA [46]	27.4	32.3	48.0	49.1
SegNeXt-B w/ NL [76]	27.6	40.9	48.6	50.0
SegNeXt-B w/ Ham [21]	27.6	34.9	48.5	49.9

材料中找到。

4.1. 编码器在 **ImageNet** 上的表现

ImageNet 预训练是训练分割模型的一个常见策略 [95, 7, 81, 89, 6]。在这里，我们将我们的 MSCAN 与最近流行的几个基于 CNN 和基于 transformer 的分类模型的性能进行比较。如表 3所示，我们的 MSCAN 取得了比最近最先进的基于 CNN 的方法 ConvNeXt [53] 更好的结果，并且超过了流行的基于 transformer 的方法，如 Swin Transformer [52] 和 MiT，SegFormer [81] 的编码器。

表 6. 关于 MSCA 设计的消融研究。Top-1 表示 ImageNet 数据集上的 Top-1 准确性，，mIoU 表示 ADE20K 数据集上的 mIoU。Br: 分支。

7 × 7 Br	11 × 11 Br	21 × 21 Br	1 × 1 Conv	Attention	Top-1	mIoU
✓	✗	✗	✓	✓	74.7	39.6
✗	✓	✗	✓	✓	75.2	39.7
✗	✗	✓	✓	✓	75.3	40.0
✓	✓	✓	✗	✓	74.8	39.1
✓	✓	✓	✓	✗	75.5	40.5
✓	✓	✓	✓	✓	75.9	41.1

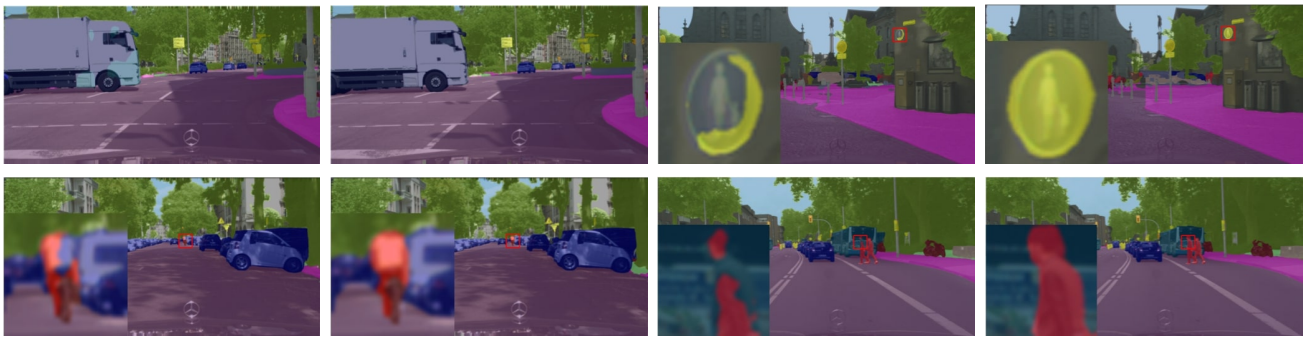
4.2. 消融实验

对 MSCA 设计进行消融： 我们在 ImageNet 和 ADE20K 数据集上进行 MSCA 设计的消融研究。K × K 分支包含 1 × K 的深度卷积和 K × 1 的深度卷积。1 × 1 conv 指的是通道混合操作。Attention 是指 element-wise 的乘积，这使得网络获得自适应能力。结果显示在表 6。我们可以发现，每个部分都对最终的性能做出了贡献。

解码器的全局上下文信息： 解码器在整合来自多尺度特征的全局背景方面发挥着重要的作用。在这里，我们研究了不同的全局背景模块对解码器的影响。正如大多数以前的工作 [76, 19] 所示，基于注意力的解码器对 CNN 的性能比金字塔结构 [95, 6] 更好，因此我们只展示使用基于注意力的解码器的结果。具体来说，我们展示了 4 种不同类型的基于注意力的解码器的结果，包括具有 $\mathcal{O}(n^2)$ 复杂度的非局部（NL）注意力 [76] 和 CCNet [35]，具有 $\mathcal{O}(n)$ 复杂度的 EMANet [46] 和 HamNet [21] 。如表 5所示，Ham 在复杂性和性能之间实现了最佳的权衡。因此，我们在解码器中使用 Hamburger [21]。

表 7. 不同解码器结构的性能。SegNeXt-T (a) 是指图 图 3 (a) 被用于解码器。FLOPs 是用 512×512 的输入尺寸计算的。SegNeXt-T (c) w/ stage 1 意味着阶段 1 的输出也被送入解码器。

Architecture	Params. (M)	GFLOPs	mIoU (SS)	mIoU (MS)
SegNeXt-T (a)	4.4	10.0	40.3	41.1
SegNeXt-T (b)	4.2	4.9	30.9	40.6
SegNeXt-T (c)	4.3	6.6	41.1	42.2
SegNeXt-T (c) w/ stage 1	4.3	12.1	40.7	42.2



SegFormer-B2 SegNeXt-B SegFormer-B2 SegNeXt-B

图 4. SegNeXt-B 和 SegFormer-B2 在 Cityscapes 数据集上的定性比较。更多可视化结果可以在我们的补充材料中找到。

解码器结构: 与图像分类不同，分割模型需要高分辨率的输出。我们为分割设计了三种不同的解码器，所有这些都已在图 3中显示。相应的结果列于表 7。我们可以看到，SegNeXt(c) 取得了最好的性能，而且计算成本也很低。

我们 MSCA 的重要性: 在这里，我们进行实验来证明 MSCA 对于分割的重要性。作为比较，我们跟随 VAN [25] ，用一个大核的单一卷积来代替我们的 MSCA 中的多个分支。如表 3所示，我们可以观察到，虽然两个编码器在 ImageNet 分类中的性能接近，但 SegNeXt w/MSCA 产生的结果比设置 w/o MSCA 要好得多。这表明多尺度特征的聚合对编码器的语义分割至关重要。

4.3. 与最先进的方法比较

在本小节中，我们将我们的方法与最先进的基于 CNN 的方法，如 HRNet [72]、ResNeSt [93] 和 EfficientNet [70]，以及基于 transformer 的方法，如 Swin Transformer [52]、SegFormer [81]、HRFormer [89]、MaskFormer [11] 和 Mask2Former [10] 进行比较。

性能-计算的权衡: ADE20K 和 Cityscapes 是语义分割中两个广泛使用的基准。如图 1所示，我们绘制了不同方法在 Cityscape 和 ADE20K 验证集上的性能-计算曲线。显然，与其他最先进的方法相比，我们的方法在性能和计算量之间实现了最佳的折衷。如 SegFormer [81], HRFormer [89], 和 MaskFormer [11]。

与最先进的 transformer 比较: 我们在 ADE20K、Cityscapes、COCO-Stuff 和 Pascal Context 基准上将 SegNeXt 与最先进的 transformer 模型进行了比较。如 表 8所示，SegNeXt-L 超过了 Mask2Former。在 ADE20K 数据集上，SegNeXt-L 在参数和计算成本相似的情况下，以 3.3 mIoU (51.0 v. 47.7) 的优势超过了以 Swin-T

表 8. 在 ADE20K、Cityscapes 和 COCO-Stuff 基准上与最先进的方法比较。ADE20K 和 COCO-Stuff 的 FLOPs 数（G）是按 512×512 的输入尺寸计算的，Cityscapes 是按 2,048×1,024 计算的。† 表示在 ImageNet-22K 上预训练的模型。

Model	Params	ADE20K			Cityscapes			COCO-Stuff		
	(M)	GFLOPs	mIoU (SS/MS)		GFLOPs	mIoU (SS/MS)		GFLOPs	mIoU (SS/MS)	
Segformer-B0 [81]	3.8	8.4	37.4	38.0	125.5	76.2	78.1	8.4	35.6	-
SegNeXt-T	4.3	6.6	41.1	42.2	50.5	79.8	81.4	6.6	38.7	39.1
Segformer-B1 [81]	13.7	15.9	42.2	43.1	243.7	78.5	80.0	15.9	40.2	-
HRFormer-S [89]	13.5	109.5	44.0	45.1	835.7	80.0	81.0	109.5	37.9	38.9
SegNeXt-S	13.9	15.9	44.3	45.8	124.6	81.3	82.7	15.9	42.2	42.8
Segformer-B2 [81]	27.5	62.4	46.5	47.5	717.1	81.0	82.2	62.4	44.6	-
MaskFormer [11]	42	55	46.7	48.8	-	-	-	-	-	-
SegNeXt-B	27.6	34.9	48.5	49.9	275.7	82.6	83.8	34.9	45.8	46.3
SETR-MLA†[97]	310.6	-	48.6	50.1	-	79.3	82.2	-	-	-
DPT-Hybrid [64]	124.0	307.9	-	49.0	-	-	-	-	-	-
Segformer-B3 [81]	47.3	79.0	49.4	50.0	962.9	81.7	83.3	79.0	45.5	-
Mask2Former [10]	47	74	47.7	49.6	-	-	-	-	-	-
HRFormer-B [89]	56.2	280.0	48.7	50.0	2223.8	81.9	82.6	280.0	42.4	43.3
MaskFormer [11]	63	79	49.8	51.0	-	-	-	-	-	-
SegNeXt-L	48.9	70.0	51.0	52.1	577.5	83.2	83.9	70.0	46.5	47.2

为主干的 Mask2Former。此外，与 SegFormer-B2 相比，SegNeXt-B 在 ADE20K 数据集上只用了 56% 的计算量就获得了 2.0 mIoU 的改进（48.5 v.s. 46.5）。特别是，由于 SegFormer [81] 中的自我关注是二次复杂的，而我们的方法使用卷积，这使得我们的方法在处理 Cityscapes 数据集的高分辨率图像时表现非常好。例如，SegNeXt-B 比 SegFormer-B2 增加了 1.6 mIoU（81.0 v. 82.6），但使用的计算量减少了 40%。在图 4中，我们还展示了与 SegFormer 的定性比较。我们可以看到，由于提出了 MSCA，我们的方法在处理物体细节时识别度很高。

与最先进的 CNNs 比较: 如 表 4, 表 9, and 表 11所示，我们在 Pascal VOC 2012、Pascal Context 和 iSAID 数据集上将我们的 SegNeXt 与最先进的 CNN 如 ResNeSt-269 [93]、EfficientNet-L2 [100] 和 HRNetW48 [72] 进行比较。SegNeXt-L 的性能优于流行的 HRNet（OCR） [72, 88] 模型（60.3 v.56.3），它使用的参数和计算量甚至更少，是为分割任务精心设计的。此外，SegNeXt-L 在 Pascal VOC 2012 测试排行榜上的表现甚至优于 EfficientNet-L2（NAS-FPN），后者在额外的 3 亿张不可用图像上进行了预训练。值得注意的是，EfficientNet-L2（NAS-FPN）有 485M 的参数，而 SegNeXt-L 只有 48.7M 的参数。

与实时方法的比较。 除了最先进的性能外，我们的方法也适合于实时部署。即使没有任何特定的软件或硬件加速，SegNeXt-T 在处理 768×1,536 尺寸的图像时，使用单个 3090 RTX GPU 实现了每秒 25 帧（FPS）。如 表 10所示，我们的方法为 Cityscapes 测试集的实时分割创造了新的最先进的结果。

表 9. 在 Pascal VOC 数据集上与最先进的方法比较。* 表示 COCO [49] 预训练。† 表示 JFT-300M [67] 预训练。$ 利用额外的 3 亿张无标签图像进行预训练。

Method	Backbone	mIoU
DANet [19]	ResNet101	82.6
OCRNet [88]	HRNetV2-W48	84.5
HamNet [21]	ResNet101	85.9
EncNet* [92]	ResNet101	85.9
EMANet* [46]	ResNet101	87.7
DeepLabV3+* [9]	Xception-71	87.8
DeepLabV3+† [9]	Xception-JFT	89.0
NAS-FPN$ [100]	EfficientNet-L2	90.5
SegNeXt-T	MSCAN-T	82.7
SegNeXt-S	MSCAN-S	85.3
SegNeXt-B	MSCAN-B	87.5
SegNeXt-L*	MSCAN-L	90.6

表 10. 与最先进的实时方法在 Cityscapes 测试数据集上的比较。我们用单个 RTX-3090 GPU 和 AMD EPYC 7543 32 核处理器 CPU 测试我们的方法。在不使用任何优化的情况下，SegNeXt-T 可以达到每秒 25 帧（FPS），这符合实时应用的要求。

Method	Input size	mIoU
ESPNet [56]	512×1,024	60.3
ESPNetv2 [57]	512×1,024	66.2
ICNet [94]	1,024 × 2,048	69.5
DFANet [40]	1,024 × 1,024	71.3
BiSeNet [86]	768 × 1,536	74.6
BiSeNetv2 [85]	512 × 1,024	75.3
DF2-Seg [47]	1,024 × 2,048	74.8
SwiftNet [61]	1,024 × 2,048	75.5
SFNet [43]	1,024 × 2,048	77.8
SegNeXt-T	768 × 1,536	78.0

表 11. 在 Pascal Context 基准上的比较。FLOPs 的数量是以 512×512 的输入尺寸计算的。* 表示 ImageNet-22K 预训练。† 表示 ADE20K 预训练。

Method	Backbone	Params.(M)	GFLOPs	mIoU (SS/MS)	
PSPNet [95]	ResNet101	-	-	-	47.8
DANet [19]	ResNet101	69.1	277.7	-	52.6
EMANet [46]	ResNet101	61.1	246.1	-	53.1
HamNet [21]	ResNet101	69.1	277.9	-	55.2
HRNet(OCR) [72]	HRNetW48	74.5	-	-	56.2
DeepLabV3+ [9]	ResNeSt-269	-	-	-	58.9
SETR-PUP* [97]	ViT-Large	317.8	-	54.4	55.3
SETR-MLA* [97]	ViT-Large	309.5	-	54.9	55.8
HRFormer-B [89]	HRFormer-B	56.2	280.0	57.6	58.5
DPT-Hybrid† [64]	ViT-Hybrid	124.0	-	-	60.5
SegNeXt-T	MSCAN-T	4.2	6.6	51.2	53.3
SegNeXt-S	MSCAN-S	13.9	15.9	54.2	56.1
SegNeXt-B	MSCAN-B	27.6	34.9	57.0	59.0
SegNeXt-L	MSCAN-L	48.8	70.0	58.7	60.3
SegNeXt-L†	MSCAN-L	48.8	70.0	59.2	60.9

5. 总结与讨论

在本文中，我们分析了以前成功的分割模型，并找到它们所拥有的良好特性。基于这些发现，我们提出了一个定制的卷积注意模块 MSCA 和一个 CNN 式网络 SegNeXt。实验结果表明，SegNeXt 在相当大的程度上超过了目前最先进的基于 transformer 的方法。

最近，基于 transformer 的模型在各种分割排行榜上占主导地位。相反，本文显示，当使用适当的设计时，基于 CNN 的方法仍然可以比基于 transformer 的方法表现得更好。我们希望本文能够鼓励研究人员进一步研究 CNNs 的潜力。

我们的模型也有其局限性，例如，将这种方法扩展到具有 100M 以上参数的大规模模型以及在其他视觉或 NLP 任务上的表现。这些将在我们未来的工作中解决。

参考文献

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 39(12):2481–2495, 2017.

[2] G. Bertasius, J. Shi, and L. Torresani. Semantic segmentation with boundary neural fields. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3602–3610, 2016.

[3] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1209–1218, 2018.

[4] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5659–5667, 2017.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062, 2014.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell., 40(4):834–848, 2017.

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017.

[8] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3640–3649, 2016.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Eur. Conf. Comput. Vis., pages 801–818, 2018.

[10] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation, 2021.

[11] B. Cheng, A. G. Schwing, and A. Kirillov. Per-pixel classification is not all you need for semantic segmentation. In NeurIPS, 2021.

[12] M. Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation(Apache-2.0), 2020.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3213–3223, 2016.

[14] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In Int. Conf. Comput. Vis., pages 764–773, 2017.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE Conf. Comput. Vis. Pattern Recog., pages 248–255. Ieee, 2009.

[16] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang. Boundary-aware feature propagation for scene segmentation. In Int. Conf. Comput. Vis., pages 6819–6829, 2019.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Int. Conf. Learn. Represent., 2020.

[18] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis., 88(2):303–338, 2010.

[19] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3146–3154, 2019.

[20] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr. Res2net: A new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell., 43(2):652–662, 2021.

[21] Z. Geng, M.-H. Guo, H. Chen, X. Li, K. Wei, and Z. Lin. Is attention better than matrix decomposition? In Int. Conf. Learn. Represent., 2021.

[22] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu. Pct: Point cloud transformer. Computational Visual Media, 7(2):187–199, 2021.

[23] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu. Beyond self-attention: External attention using two linear layers for visual tasks. arXiv preprint arXiv:2105.02358, 2021.

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

[25] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu. Visual attention network. arXiv preprint arXiv:2202.09741, 2022.

[26] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu. Attention mechanisms in computer vision: A survey. arXiv preprint arXiv:2111.07624, 2021.

[27] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao. Adaptive pyramid context network for semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 7519–7528, 2019.

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE Conf. Comput. Vis. Pattern Recog., pages 770–778, 2016.

[29] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li. Bag of tricks for image classification with convolutional neural networks. In IEEE Conf. Comput. Vis. Pattern Recog., pages 558–567, 2019.

[30] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4003–4012, 2020.

[31] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In IEEE Conf. Comput. Vis. Pattern Recog., pages 7132–7141, 2018.

[32] S.-M. Hu, D. Liang, G.-Y. Yang, G.-W. Yang, and W.-Y. Zhou. Jittor: a novel deep learning framework with meta-operators and unified graph execution. Science China Information Sciences, 63(12):1–21, 2020.

[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4700–4708, 2017.

[34] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li. Flowformer: A transformer architecture for optical flow. arXiv preprint arXiv:2203.16194, 2022.

[35] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. Ccnet: Criss-cross attention for semantic segmentation. In Int. Conf. Comput. Vis., pages 603–612, 2019.

[36] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Int. Conf. Mach. Learn., pages 448–456. PMLR, 2015.

[37] A. Kirillov, R. Girshick, K. He, and P. Dollár. Panoptic feature pyramid networks. In IEEE Conf. Comput. Vis. Pattern Recog., pages 6399–6408, 2019.

[38] A. Kirillov, Y. Wu, K. He, and R. Girshick. Pointrend: Image segmentation as rendering. In IEEE Conf. Comput. Vis. Pattern Recog., pages 9799–9808, 2020.

[39] Y. Lee, J. Kim, J. Willette, and S. J. Hwang. Mpvit: Multi-path vision transformer for dense prediction. In IEEE Conf. Comput. Vis. Pattern Recog., 2022.

[40] H. Li, P. Xiong, H. Fan, and J. Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 9522–9531, 2019.

[41] X. Li, H. He, X. Li, D. Li, G. Cheng, J. Shi, L. Weng, Y. Tong, and Z. Lin. Pointflow: Flowing semantics through points for aerial image segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4217–4226, 2021.

[42] X. Li, X. Li, L. Zhang, G. Cheng, J. Shi, Z. Lin, S. Tan, and Y. Tong. Improving semantic segmentation via decoupled body and edge supervision. In European Conference on Computer Vision, pages 435–452. Springer, 2020.

[43] X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, S. Tan, and Y. Tong. Semantic flow for fast and accurate scene parsing. In European Conference on Computer Vision, pages 775–793. Springer, 2020.

[44] X. Li, W. Zhang, J. Pang, K. Chen, G. Cheng, Y. Tong, and C. C. Loy. Video k-net: A simple, strong, and unified baseline for video segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18847–18857, 2022.

[45] X. Li, H. Zhao, L. Han, Y. Tong, S. Tan, and K. Yang. Gated fully fusion for semantic segmentation. In Proceedings of the AAAI conference on artificial intelligence, pages 11418–11425, 2020.

[46] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu. Expectation-maximization attention networks for semantic segmentation. In Int. Conf. Comput. Vis., pages 9167–9176, 2019.

[47] X. Li, Y. Zhou, Z. Pan, and J. Feng. Partial order pruning: for best speed/accuracy trade-off in neural architecture search. In IEEE Conf. Comput. Vis. Pattern Recog., pages 9145–9153, 2019.

[48] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1925–1934, 2017.

[49] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In Eur. Conf. Comput. Vis., pages 740–755. Springer, 2014.

[50] R. Liu, H. Deng, Y. Huang, X. Shi, L. Lu, W. Sun, X. Wang, J. Dai, and H. Li. Decoupled spatial-temporal transformer for video inpainting. arXiv preprint arXiv:2104.06637, 2021.

[51] R. Liu, H. Deng, Y. Huang, X. Shi, L. Lu, W. Sun, X. Wang, J. Dai, and H. Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In Int. Conf. Comput. Vis., pages 14040–14049, 2021.

[52] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Int. Conf. Comput. Vis., 2021.

[53] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s, 2022.

[54] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3431–3440, 2015.

[55] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[56] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In Proceedings of the european conference on computer vision (ECCV), pages 552–568, 2018.

[57] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In IEEE Conf. Comput. Vis. Pattern Recog., pages 9190–9200, 2019.

[58] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In Adv. Neural Inform. Process. Syst., pages 2204–2212, 2014.

[59] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In IEEE Conf. Comput. Vis. Pattern Recog., pages 891–898, 2014.

[60] L. Mou, Y. Hua, and X. X. Zhu. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In IEEE Conf. Comput. Vis. Pattern Recog., pages 12416–12425, 2019.

[61] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In IEEE Conf. Comput. Vis. Pattern Recog., pages 12607–12616, 2019.

[62] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 2019.

[63] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4353–4361, 2017.

[64] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In Int. Conf. Comput. Vis., pages 12179–12188, 2021.

[65] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

[66] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. In Int. Conf. Comput. Vis., pages 7262–7272, 2021.

[67] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision, pages 843–852, 2017.

[68] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1–9, 2015.

[69] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In Int. Conf. Comput. Vis., pages 5229–5238, 2019.

[70] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, pages 6105–6114. PMLR, 2019.

[71] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In Int. Conf. Mach. Learn., pages 10347–10357. PMLR, 2021.

[72] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell., 2020.

[73] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu. Eca-net: Efficient channel attention for deep convolutional neural networks, 2020.

[74] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pvtv2: Improved baselines with pyramid vision transformer. arXiv preprint arXiv:2106.13797, 2021.

[75] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Int. Conf. Comput. Vis., 2021.

[76] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In IEEE Conf. Comput. Vis. Pattern Recog., pages 7794–7803, 2018.

[77] S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 28–37, 2019.

[78] R. Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models(Apache-2.0), 2019.

[79] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In European Conference on Computer Vision, pages 648–663. Springer, 2016.

[80] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In Eur. Conf. Comput. Vis., pages 418–434, 2018.

[81] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inform. Process. Syst., 34, 2021.

[82] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1492–1500, 2017.

[83] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao. Focal self-attention for local-global interactions in vision transformers, 2021.

[84] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. Denseaspp for semantic segmentation in street scenes. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3684–3692, 2018.

[85] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. Int. J. Comput. Vis., 129(11):3051–3068, 2021.

[86] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Eur. Conf. Comput. Vis., pages 325–341, 2018.

[87] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015.

[88] Y. Yuan, X. Chen, and J. Wang. Object-contextual representations for semantic segmentation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, pages 173–190. Springer, 2020.

[89] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang. Hrformer: High-resolution vision transformer for dense predict. Adv. Neural Inform. Process. Syst., 34, 2021.

[90] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang. Ocnet: Object context network for scene parsing. arXiv preprint arXiv:1809.00916, 2018.

[91] Y. Yuan, J. Xie, X. Chen, and J. Wang. Segfix: Model-agnostic boundary refinement for segmentation. In European Conference on Computer Vision, pages 489–506. Springer, 2020.

[92] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 7151–7160, 2018.

[93] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, et al. Resnest: Split-attention networks. arXiv preprint arXiv:2004.08955, 2020.

[94] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. Icnet for real-time semantic segmentation on high-resolution images. In Eur. Conf. Comput. Vis., pages 405–420, 2018.

[95] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2881–2890, 2017.

[96] M. Zhen, J. Wang, L. Zhou, S. Li, T. Shen, J. Shang, T. Fang, and L. Quan. Joint semantic segmentation and boundary detection using iterative pyramid contexts. In IEEE Conf. Comput. Vis. Pattern Recog., pages 13666–13675, 2020.

[97] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In IEEE Conf. Comput. Vis. Pattern Recog., pages 6881–6890, 2021.

[98] Z. Zheng, Y. Zhong, J. Wang, and A. Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4096–4105, 2020.

[99] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In IEEE Conf. Comput. Vis. Pattern Recog., pages 633–641, 2017.

[100] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le. Rethinking pre-training and self-training. Adv. Neural Inform. Process. Syst., 33:3833–3845, 2020.