高效语义分割再探索：基于偏移学习的空间与类别特征对齐优化

张世辰[1] 李运恒[1] 吴宇寰[2] 侯淇彬[1†] 程明明[1]
[1]VCIP, CS, 南开大学 [2]IHPC, A*STAR, 新加坡
zhangshichen@mail.nankai.edu.cn
项目主页: https://github.com/HVision-NKU/OffSeg.

Abstract

语义分割是实现像素级场景理解的视觉系统中的基础任务，但在资源受限的设备上部署该任务则要求架构具备高效性。尽管现有方法通过轻量化设计实现了实时推理，我们揭示了它们的一个固有局限性：由于逐像素分类范式所导致的类别表示与图像特征之间的不对齐问题。通过实证分析，我们发现这一范式在高效场景中引入了一个极具挑战性的假设：对于同一类别，图像中像素特征在不同图像之间以及不同位置应该是相同的。为了解决这一困境，我们提出了一种耦合的双分支偏移学习（*Offset Learning*）范式，该范式显式地学习特征偏移和类别偏移，从而动态地优化类别表示和图像的空间特征。基于所提出的范式，我们构建了一个高效的语义分割网络，*OffSeg*。值得注意的是，该偏移学习范式可以无缝适配到现有方法中，而无需任何架构修改。在包括*ADE20K*、*Cityscapes*、*COCO-Stuff-164K*和*Pascal Context*在内的四个数据集上的大量实验表明，该方法在参数几乎不增加的情况下实现了稳定的性能提升。

1. Introduction

语义分割旨在为图像中的每一个像素分配类别标签，在计算机视觉应用中扮演着关键角色 [13, 15, 20, 25, 26, 28, 36, 48, 49, 63]。尽管近年来标准模型 [29, 30, 38, 39, 56, 60, 62]在分割精度上取得了显著进展，其计算复杂度与参数规模使其难以在资源受限的场景中部署。因此，研究重点逐渐转向高效语义分割模型 [18, 34, 37, 41, 46, 50]，该类方法以实时推理和参数最小化为核心目标。

传统的非高效分割框架通常依赖于高维图像特征、丰富的类别表示以及大量参数，因此在性能上普遍优于轻量化结构。这一现象符合神经网络的规模定律，即模型容量与分割精度呈正相关，直至受限于可用的计算资源。相比之下，以高效为导向的架构 [43, 46, 50, 51]面临着一个固有权衡：激进的模型压缩降低了

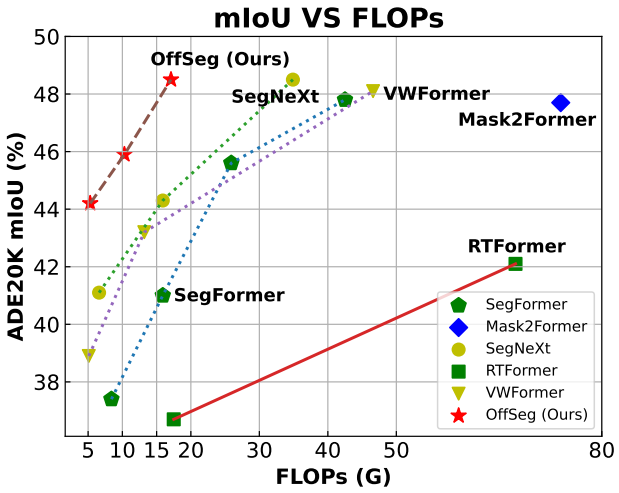

Figure 1. 与主流高效分割方法在ADE20K 数据集 [61]上的对比。图中展示了我们的方法在性能与计算量之间实现了最佳的平衡。

对类别语义与局部视觉线索的对齐能力。这种对齐障碍会导致物体边界模糊、小目标遗漏以及预测不一致等问题，并在主流的逐像素分类范式中尤为突出（见Fig. 2(a)）。尽管现有方法通过轻量化主干网络 [18, 23, 41]或空间下采样 [22, 45, 58]实现高效推理，但它们普遍忽视了在严格的参数限制下对类别与特征表示进行联合优化的根本挑战。

为揭示逐像素分类范式的根本性问题，我们采用理想类别表示（特征）挖掘方法，为每张图像提取类别专属的最优表示。统计分析结果（见Fig. 3）表明，同类别在不同图像中的最优类别表示之间相似度极低。这一发现说明，像逐像素分类方法中那样为所有图像使用固定类别表示是次优的，因为这种方式无法适应不同图像中独特的特征分布和类别细节。

基于上述观察，我们提出了偏移学习范式（Offset Learning Paradigm），这是一种新型语义分割方法，能够通过可学习的特征偏移（FOs）与类别偏移（COs）显式学习并纠正类别表示与图像特征之间的偏差。我们的重要洞察是：尽管高效模型参数有限，难以直接建模理想的类别-特征关系，但它们可以有效学习初始

†通讯作者. 邮箱: houqb@nankai.edu.cn

Table 1. 不同语义分割范式的对比。‘Fea.’ 和‘Rep.’ 分别表示特征自适应与类别表征自适应。

范式	Fea.	Rep.	交互方式	对齐方式	开销
逐像素分类（Per-Pixel Classification）	✗	✗	✗	静态单向	矩阵乘法
掩码分类（Mask Classification）	✗	✓	交叉注意力	动态但不对称	Transformer 解码器
偏移学习（Offset Learning）	✓	✓	双分离偏移	弹性双向	矩阵乘法

粗略表示与其最优表示之间的“偏移”。具体而言，我们的偏移学习范式由两个主干分支组成：类别偏移学习分支（Class Offset Learning）与特征偏移学习分支（Feature Offset Learning）。这两个分支分别学习COs与FOs，从而使图像特征与类别表示都具备灵活性。

如Fig. 2(b)所示，除了逐像素分割范式之外，还有基于掩码的分割范式 [4, 10, 11]，该方法通过交叉注意力机制使可学习查询与图像特征进行交互。这种方式使得查询能够自适应地学习图像特有的特征。然而，其存在两方面的固有限制：（1）仅调整查询而忽略图像特征本身；（2）交叉注意力机制带来显著的计算开销。如Tab. 1所总结，我们的方法相较于上述两类范式，具有两大优势：（1）图像特征与类别表示的双向自适应能力；（2）几乎可以忽略的交互开销。

基于所提出的偏移学习范式（见Fig. 2(c)），我们设计了一个简洁的语义分割网络，命名为 OffSeg，其结构仅包含一个主干网络与像素解码器。作为一种即插即用的范式，我们将其应用于SegNeXt [18]（基于CNN）、SegFormer [46]（基于Transformer）与Mask2Former [11]（掩码分类），以验证其有效性与通用性。在四个基准数据集上的大量实验结果一致验证了我们方法的高效性与有效性。不同架构与数据集上取得的性能提升进一步凸显了我们方法的鲁棒性与泛化能力。在Fig. 1中，我们展示了模型在不同规模下的性能表现。结果表明，OffSeg 在性能与计算效率之间实现了优越的平衡。

综上，我们的主要贡献可总结如下：

- 我们通过统计分析与理想类别表示（特征）挖掘，揭示了逐像素分割范式的核心局限性，即静态图像特征与类别表示之间存在内在不对齐。
- 我们提出了一种参数高效的偏移学习范式，采用双分支结构，在几乎无额外计算开销的前提下联合自适应图像特征与类别表示。
- 大量实验验证了我们提出的 OffSeg 方法在性能上的优势，并在CNN、Transformer 与掩码分类等不同范式下均表现出优异的适应性与有效性。

2. Related work

2.1. 传统语义分割方法

语义分割在以精度为优先的大规模模型推动下取得了显著进展。开创性工作如全卷积网络（FCN）[31]通过将全连接层替换为卷积操作，实现了密集的逐像素预测，为后续研究奠定了基础。在这一范式

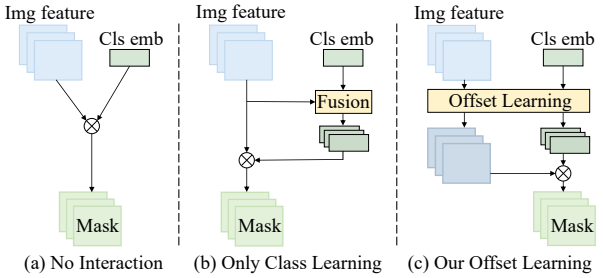

Figure 2. 不同语义分割范式的可视化对比。从左至右依次展示逐像素分类、掩码分类以及我们提出的偏移学习范式。

确立之后，后续基于CNN 的研究 [1, 17, 27, 36, 42, 52, 54, 57, 59]从多个方向对FCN 进行了改进。例如，U-Net [36] 通过对称的编码器-解码器结构与跳跃连接进一步增强了特征的定位能力。从上下文聚合的角度来看，DeepLab 系列 [6–9]引入了扩张空间金字塔池化（ASPP），以捕获多尺度上下文信息；而PSPNet [59]提出的金字塔池化模块可在不同子区域间聚合全局上下文。随着注意力机制 [14, 40]的发展，基于Transformer 的方法 [21, 35, 38, 55, 60]也取得了显著成果。例如，SERE [60]将语义分割重新定义为序列到序列的预测任务，利用全局自注意力建模整图上下文信息。不同于逐像素分类范式，MaskFormer 系列 [10, 11]引入了掩码分类范式，其中可学习的查询通过Transformer 解码器与图像特征进行交互。

2.2. 高效语义分割

尽管传统模型在分割精度方面表现优异，但其庞大的计算开销限制了在实时场景中的应用，这促使研究者发展高效语义分割方法 [4, 16, 18, 37, 41, 43–46, 50, 51, 58]。从主干网络的角度，SegFormer [46]提出了一种轻量化且具有层次结构的Transformer 编码器；而SegNeXt [18] 则通过多尺度卷积构建高效主干，仅依赖卷积注意力机制来增强表达能力。LRFormer [45] 提出了在极低分辨率空间中计算的线性注意力Transformer，进一步提升效率。在解码器设计方面，FeedFormer [37] 使用Transformer 架构，以图像特征作为查询提取结构信息；VWFormer [47] 通过跨尺度窗口交互增强多尺度表示；CGRSeg [34] 则采用金字塔上下文引导的空间特征重建机制，从水平与垂直两个维度增强前景物体表征能力。

对于高效语义分割模型而言，由于掩码分类范式依赖计算成本极高的Transformer 解码器进行特征交互，因此这些方法普遍采用逐像素分类范式。虽然逐像

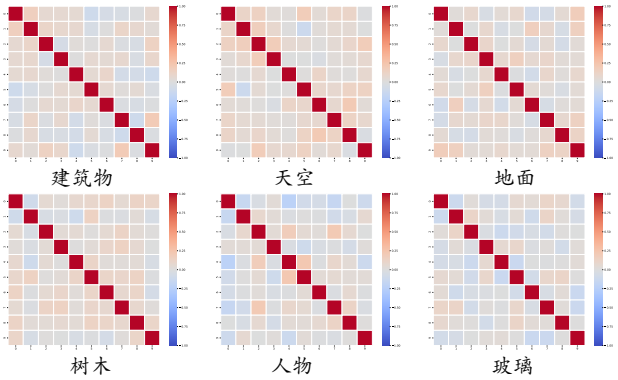

Figure 3. 理想类别表示之间相似性的热力图可视化。我们可以观察到，同一类别在不同图像中的理想类别表示之间的相关性非常低。

素分类几乎不引入额外计算开销，但其固有的问题在于图像特征与类别表示之间存在错位。在轻量化场景下，这一困境尤为明显（见Fig. 3），这也驱动我们探索一种专门面向高效语义分割任务的新型分割范式。

3. Method

3.1. 重新审视逐像素分类

逐像素分类作为传统语义分割的基础方法，通过将每个像素的特征向量与预定义的类别原型进行比较，独立地为每个像素分配标签。传统的逐像素分类方法通过一个1×1卷积将像素嵌入$\mathbf{E} \in \mathbb{R}^{HW \times C}$映射到类别得分$\mathbf{P}$：

$$\mathbf{P}_{i,j} = \mathbf{W}_c \cdot {\mathbf{E}^\top}_{i,j}, \tag{1}$$

其中$\mathbf{W} \in \mathbb{R}^{K \times C}$为$K$个类别的可学习参数，$c$为类别索引。该范式将每个像素视为独立个体，忽略了上下文之间的关联性。

尽管这一范式被广泛采用，但在高效分割场景中仍面临两个关键问题。首先，逐像素分类依赖于固定的类别表示进行像素分类。其次，该方法默认假设网络能够为同一类别在不同图像中学习到相同的特征。然而，§3.2 中的分析表明，这一假设在本质上是无法实现的。这种固定类别表示与多变图像特征之间的错配，强烈说明了在现代分割框架中引入自适应机制的必要性。

3.2. 理想类别表示（特征）挖掘

为了从理论上验证自适应类别表示与图像特征的必要性，我们基于真实标注掩码，反向推导每张图像的最优类别原型。给定一张输入图像，其真实掩码为$\mathbf{M} \in \mathbb{R}^{K \times HW}$，深度特征为$\mathbf{E} \in \mathbb{R}^{HW \times C}$，其理想类别原型$\mathbf{W}^* \in \mathbb{R}^{K \times C}$应满足：$\mathbf{M} = \mathbf{W}^* \cdot \mathbf{E}^\top$，其中$\mathbf{W}^*$的每一行代表某一类别的最优原型。

解该线性系统可得：

$$\mathbf{W}^* = \mathbf{M} \cdot \left(\mathbf{E}^\top\right)^\dagger, \tag{2}$$

其中$(\cdot)^\dagger$表示Moore–Penrose 伪逆。

通过$\mathbf{M}_{\text{pred}} = \mathbf{W}^* \cdot \mathbf{E}^\top$重新计算掩码，我们在ADE20K 数据集上实现了约95% 的mIoU，验证了$\mathbf{W}^*$的理论有效性。基于上述数学推导，我们采用SegFormer [46]（一个仅含4.3M 参数的高效网络）进行相似性分析。我们随机选取了六个类别（即建筑物、天空、地面、树木、人物、玻璃），并为每个类别选取10 张图像计算$\mathbf{W}^*$。如Fig. 3 所示，我们通过热力图展示其两两之间的相似性。令人惊讶的是，这些理想类别表示之间的相似性远低于我们的常识预期。热力图中低相关性的模式揭示了：即使是同一类别，其最优表示在不同图像中也存在显著差异。这一现象源于高效模型中的根本矛盾：激进的特征压缩加剧了类内特征差异，迫使$\mathbf{W}^*$发生显著偏移以适应扭曲的特征分布。上述分析揭示出两个关键结论：

- 固定原型失效：在高效模型中，固定的类别表示无法适应不同图像中高度变化的特征。
- 固定特征失效：关系$\mathbf{W}^* \propto f(\mathbf{E})$表明，扭曲的特征也会破坏原型的稳定性，呈现出一种需要联合纠正的恶性循环。

3.3. 偏移学习范式

我们的偏移学习范式将语义分割重新定义为一种双分支解耦对齐过程：

$$\mathbf{M} = (\mathbf{W} + \Delta\mathbf{W}) \cdot (\mathbf{E} + \Delta\mathbf{E})^\top, \tag{3}$$

该公式将我们的方法与传统的逐像素分类方法 [7, 18, 19, 59]以及基于掩码的方法 [10, 11]区分开来。其核心创新在于类别偏移学习和特征偏移学习两个分支，这两个分支通过解耦的注意力机制协同优化类别原型与空间特征，如Fig. 4 所示。

具体而言，给定图像特征$\mathbf{E} \in \mathbb{R}^{HW \times C}$和类别嵌入$\mathbf{W} \in \mathbb{R}^{K \times C}$，我们计算耦合注意力矩阵$\mathbf{A}_c$：

$$\mathbf{A}_c = \mathbf{W} \cdot \mathbf{E}^\top, \tag{4}$$

其中$\mathbf{A}_c \in \mathbb{R}^{K \times HW}$。该矩阵编码了类别与空间位置之间的相关性，并被用于后续的类别偏移学习分支与特征偏移学习分支。

类别偏移学习 根据空间上下文动态调整类别表示，从而缓解固定类别嵌入带来的刚性问题。首先，我们在空间维度上应用softmax 归一化Softmax_S，生成类别注意力权重：

$$\mathbf{A}_{\text{cls}} = \text{Softmax}_S(\mathbf{A}_c) \in \mathbb{R}^{K \times HW}, \tag{5}$$

其中每一行$\mathbf{a}_k \in \mathbf{A}_{\text{cls}}$表示第$k$类在空间位置上的注意力分布。随后，我们使用类别注意力对空间特征进行加权聚合：

$$\mathbf{F}_{\text{cls}} = \mathbf{A}_{\text{cls}} \cdot \mathbf{E} \in \mathbb{R}^{K \times C}, \tag{6}$$

其中$\mathbf{F}_{\text{cls}}$包含了编码全局空间分布的类别特定原型。最后，我们通过一个MLP 生成类别偏移：

$$\Delta\mathbf{W} = \text{MLP}(\mathbf{F}_{\text{cls}}) \in \mathbb{R}^{K \times C}, \tag{7}$$

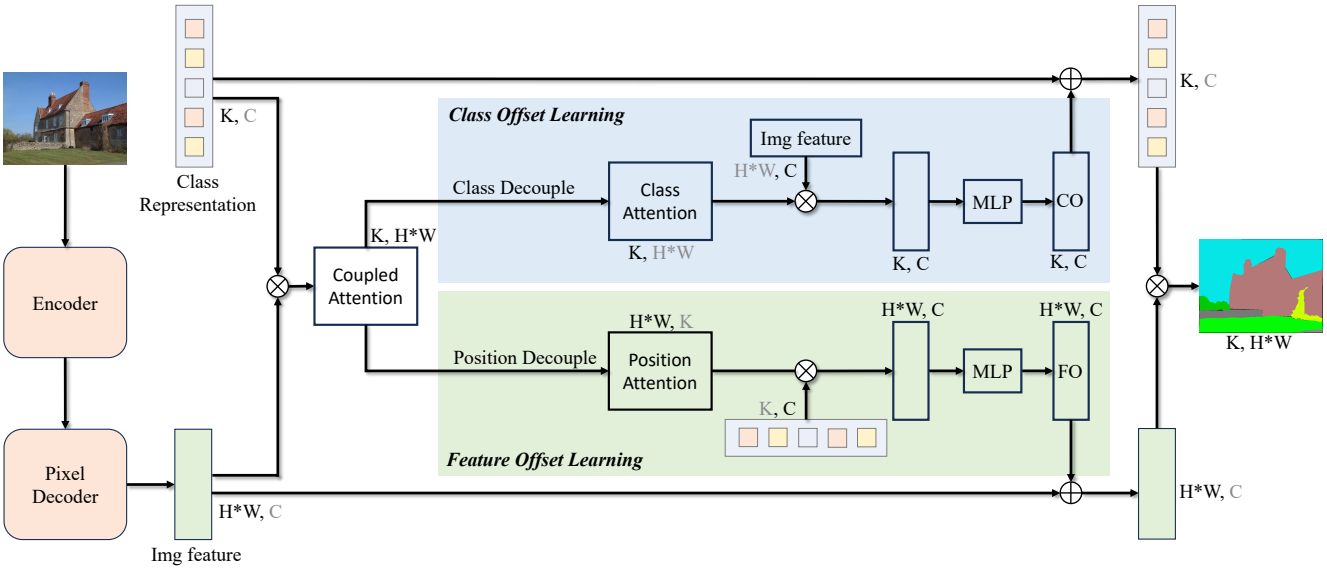

Figure 4. 所提出方法 OffSeg的整体框架。对于输入图像，我们首先使用编码器提取多尺度特征，然后通过像素解码器生成图像特征。偏移学习范式包含两个分支：类别偏移学习分支和特征偏移学习分支。通过学习得到的类别偏移（CO）和特征偏移（FO），我们引导初始特征对齐至更合理的空间。矩阵乘法的维度在图中以灰色标注。

并将其加至原始表示中进行调整：

$$\mathbf{W}_{\text{adj}} = \mathbf{W} + \Delta\mathbf{W}. \tag{8}$$

该分支学习图像特定的偏移量，以对齐类别嵌入与对应图像特征，从而缩小它们之间的表示差距。

特征偏移学习 通过注入类别感知语义来优化图像特征，旨在缓解逐像素分类中的局部歧义问题。如Fig. 4 所示，特征偏移学习分支与类别偏移学习分支形成对偶结构。

具体而言，与类别偏移学习类似，我们首先在类别维度上应用softmax Softmax_K，并转置以实现空间对齐：

$$\mathbf{A}_{\text{pos}} = (\text{Softmax}_K(\mathbf{A}_c))^T \in \mathbb{R}^{HW\times K}, \tag{9}$$

其中每一行$\mathbf{a}_i \in \mathbf{A}_{\text{pos}}$ 表示第i 个位置在所有类别上的概率分布。然后，我们通过下式将类别语义注入至空间位置：

$$\mathbf{F}_{\text{pos}} = \mathbf{A}_{\text{pos}} \cdot \mathbf{W} \in \mathbb{R}^{HW\times C}, \tag{10}$$

其中$\mathbf{F}_{\text{pos}}$ 表示来自所有类别的、用于空间引导的语义信息。接着，我们使用MLP 生成特征偏移：

$$\Delta\mathbf{E} = \text{MLP}(\mathbf{F}_{\text{pos}}) \in \mathbb{R}^{HW\times C}, \tag{11}$$

并利用偏移量引导原始特征修正为：

$$\mathbf{E}_{\text{adj}} = \mathbf{E} + \Delta\mathbf{E}. \tag{12}$$

最终的分割掩码通过双向弹性对齐生成：

$$\mathbf{M} = \mathbf{W}_{\text{adj}} \cdot \mathbf{E}_{\text{adj}}^T, \tag{13}$$

这实际上是Eqn. (3) 的另一种形式。

我们在 Tab. 1 中总结了我们的偏移学习范式与其他语义分割方法的主要区别。与逐像素分类方法（如SegFormer [46]、SegNeXt [18]）依赖固定特征与静态类别嵌入之间的静态对齐不同；与掩码分类方法（如MaskFormer [10]、Mask2Former [11]）通过重量级Transformer 解码器动态优化类别查询不同；我们的方法独特地引入了具有极少可学习参数的双向偏移学习机制。该方法实现了对称适配：类别表示可以通过类别特定的空间原型进行调整，而图像特征则可通过位置感知的语义信息进行细化。通过将类别维和位置维的交互解耦为两个独立路径，我们的方法实现了弹性的特征-类别对齐，使得两种模态能够共同演化，以捕捉实例特有的几何结构与上下文语义。这与现有范式中单向或硬编码的对齐策略形成了鲜明对比。

3.4. 整体架构

为了验证我们提出的偏移学习范式在效率与效果上的优势，我们设计了一个标准的语义分割模型，整合以下高效组件，且不进行结构修改。

在主干网络方面，我们采用EfficientFormerV2 [24]，这是一种混合架构，通过细粒度联合搜索策略在参数效率与性能之间取得了良好平衡。在多尺度特征融合方面，我们选择FreqFusion [5]，该模块基于频率感知算子对两个尺度的特征进行融合。值得注意的是，当与我们的偏移学习范式结合时，整个模型仅引入了约0.1-0.2M 的可学习参数，几乎可以忽略不计。

Table 2. 在ADE20K、Cityscapes 和COCO-Stuff 数据集上，主流方法的性能对比。FLOPs（G）在ADE20K 和COCO-Stuff 数据集上以512×512 输入尺寸计算，在Cityscapes 上以2048×1024 输入尺寸计算。

方法	参数量(M)	ADE20K		Cityscapes		COCO-Stuff	
		FLOPs (G)	mIoU	FLOPs (G)	mIoU	FLOPs (G)	mIoU
SegFormer-B0 [46]	3.8	8.4	37.4	125.5	76.2	8.4	35.6
RTFormer-Slim [43]	4.8	17.5	36.7	-	76.3	-	-
FeedFormer-B0 [37]	4.5	7.8	39.2	107.4	77.9	-	-
Seaformer-L [41]	14.0	6.5	42.7	-	-	-	-
VWFormer-B0 [47]	3.7	5.1	38.9	-	77.2	5.1	36.2
CGRSeg-T [34]	9.4	4.0	43.6	-	-	4.0	42.2
EDAFormer-T [53]	4.9	5.6	42.3	151.7	78.7	5.6	40.3
OffSeg-T	6.2	5.3	44.2	44.8	78.9	5.3	41.9
SegFormer-B1 [46]	13.7	15.9	42.2	243.7	78.5	15.9	40.2
SegNeXt-S [18]	13.9	15.9	44.3	124.6	81.3	15.9	42.2
RTFormer-Base [43]	16.8	67.4	42.1	-	79.3	26.6	35.3
VWFormer-B1 [47]	13.7	13.2	43.2	-	79.0	-	41.5
PEM-STDC1 [4]	17.0	16.0	39.6	-	-	-	-
OffSeg-B	13.0	10.3	45.9	86.5	80.5	10.3	44.3
SenFormer [2]	59.0	179.0	46.0	-	-	-	-
SegFormer-B2 [46]	27.5	25.9	45.6	717.1	81.0	26.0	44.6
MaskFormer [10]	42.0	55.0	46.7	-	-	-	-
Mask2Former [11]	47.0	74.0	47.7	-	-	-	-
FeedFormer-B2 [37]	29.1	42.7	48.0	522.7	81.5	-	-
PEM-STDC2 [4]	21.0	19.3	45.0	-	-	-	-
OffSeg-L	26.4	17.1	48.5	143.4	81.6	17.1	46.0

4. Experiments

4.1. 实验设置

数据集： 我们在四个广泛使用的语义分割基准数据集上评估了我们的方法：ADE20K [61]、Cityscapes [13]、COCO-Stuff [3]和Pascal Context [33]。ADE20K [61] 是一个包含150个物体/背景类别的场景解析数据集，包含20K/2K/3K张训练/验证/测试图像，涵盖了多样的室内外场景与复杂遮挡。Cityscapes [13] 专注于城市驾驶场景，提供了5,000 张高分辨率图像（2048×1024），包含19个语义类别。COCO-Stuff [3] 包含118K 张训练图像和5K 张验证图像，涵盖171 个类别（80 个物体类+91 个背景类），其长尾分布对模型泛化能力提出挑战。PASCAL Context [33] 包含59 个语义前景类别，共有4,996 张训练图像和5,104 张验证图像。

实现细节： 我们的实现基于MMSegmentation [12]和PyTorch。参考已有工作 [10, 11, 18, 34, 46]，我们对所有模型使用AdamW [32] 优化器，结合poly 学习率衰减策略，并在训练初期进行1500次线性warmup，而未对其它设置进行额外调参。ADE20K/COCO-Stuff/Pascal Context 的batch size设置为16，Cityscapes 为8。训练时，图像尺寸裁剪为：ADE20K 和COCO-Stuff 为512×512，Pascal Context 为480×480，Cityscapes 为1024×1024。我们采用标准的数据增强策略，在ADE20K 和Cityscapes 上训练160k 次迭代，在COCO-Stuff 和Pascal Context 上训练80k 次。推理阶段，所有数据集均采用单尺度测试。所有实验均在8 张NVIDIA RTX 3090 GPU 上进行。

4.2. 主要结果

我们在ADE20K、Cityscapes 和COCO-Stuff 三个标准语义分割数据集上评估了我们的方法，详见 Tab. 2。在ADE20K 上，我们提出的OffSeg-T 取得了44.2的mIoU，超过EDAFormer-T 1.9 点，同时计算量减少了24%。OffSeg-B 展现出良好的精度-效率权衡：在ADE20K 上达到45.9 mIoU（10.3G FLOPs），超过SegNeXt-S（+1.6）和PEM-STDC1（+6.3），FLOPs比SegNeXt-S 低35%。在大模型尺度下，OffSeg-L在ADE20K 上取得48.5 mIoU，仅需17.1G FLOPs，超过Mask2Former（+0.8），计算量仅为其四分之一。

在Cityscapes 上，OffSeg-L 的性能优于FeedFormer-B2，计算成本却仅为其四分之一。在COCO-Stuff上，OffSeg-B 相较RTFormer-Base 提升9.0 mIoU，计算量却不到一半。

这些实验结果表明，我们提出的双解耦偏移学习范式能够有效缓解类别表示与图像特征之间的错配问题，尤其在类别密集和挑战性较高的数据集如ADE20K 和COCO-Stuff 上表现尤为显著。

4.3. 泛化能力

为验证我们偏移学习范式的广泛适用性，我们将其集成到三种代表性模型中：SegNeXt [18]（基

Table 3. SegNeXt [18] 与加入偏移学习后的SegNeXt 在ADE20K、Cityscapes、Pascal Context 和COCO-Stuff 上的性能对比。FLOPs（G）的计算分辨率为：Cityscapes 为2048×1024，其它数据集为512×512。

方法	偏移学习	参数量(M)	ADE20K		Cityscapes		Pascal Context		COCO-Stuff	
			FLOPs (G)	mIoU	FLOPs (G)	mIoU	FLOPs (G)	mIoU	FLOPs (G)	mIoU
SegNeXt-T		4.3	6.6	41.1	50.5	79.8	6.6	51.2	6.6	38.7
SegNeXt-T	✓	4.4	7.2	43.0$_{(+1.9)}$	53.1	80.0$_{(+0.2)}$	6.8	53.2$_{(+2.0)}$	7.3	40.0$_{(+1.3)}$
SegNeXt-S		13.9	15.9	44.3	124.6	81.3	15.9	54.2	15.9	42.2
SegNeXt-S	✓	14.1	16.5	45.6$_{(+1.3)}$	127.2	81.7$_{(+0.4)}$	16.1	55.9$_{(+1.7)}$	16.6	43.5$_{(+1.3)}$
SegNeXt-B		27.6	34.9	48.5	275.7	82.6	34.9	57.0	34.9	45.8
SegNeXt-B	✓	28.2	34.8	49.4$_{(+0.9)}$	269.6	82.8$_{(+0.2)}$	34.1	58.0$_{(+1.0)}$	35.0	45.8$_{(+0.0)}$

Table 4. SegFormer [46] 与加入偏移学习后的SegFormer 在ADE20K 和COCO-Stuff 数据集上的性能比较。FLOPs（G）均在512×512 输入大小下计算。

方法	偏移学习	参数量	ADE20K		COCO-Stuff	
			FLOPs	mIoU	FLOPs	mIoU
SegFormer-B0		3.8M	8.4	37.4	8.6	35.6
SegFormer-B0	✓	3.9M	8.8	40.1$_{(+2.7)}$	8.9	38.3$_{(+2.7)}$
SegFormer-B1		13.7M	15.9	41.0	16.1	40.2
SegFormer-B1	✓	13.9M	16.3	43.7$_{(+2.7)}$	16.4	41.9$_{(+1.7)}$
SegFormer-B2		24.8M	25.9	45.6	26.0	44.6
SegFormer-B2	✓	24.9M	26.1	47.3$_{(+1.7)}$	26.2	45.2$_{(+0.6)}$
SegFormer-B3		44.6M	42.5	47.8	42.6	45.5
SegFormer-B3	✓	44.8M	42.8	49.5$_{(+1.7)}$	42.9	46.3$_{(+0.8)}$
SegFormer-B4		61.4M	59.2	48.5	59.3	46.5
SegFormer-B4	✓	61.6M	59.5	50.1$_{(+1.6)}$	59.6	47.0$_{(+0.5)}$
SegFormer-B5		82.0M	75.2	49.1	75.3	46.7
SegFormer-B5	✓	82.2M	75.5	50.6$_{(+1.5)}$	75.5	47.2$_{(+0.5)}$

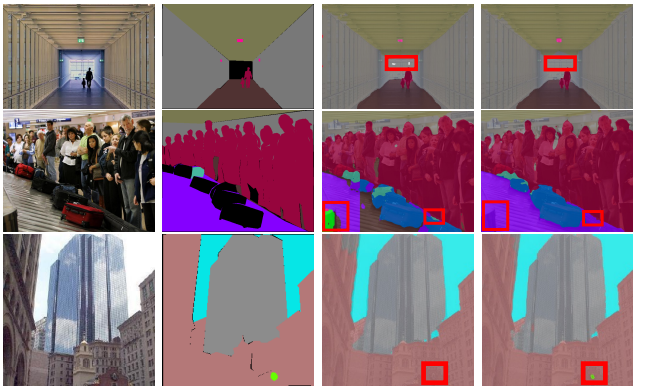

Figure 5. 在SegNeXt 中应用偏移学习范式的可视化效果。相较于原始的SegNeXt-T，加入偏移学习后模型能更准确地完成分割，特别是在小物体（如第三张图中的时钟）上表现更好。

于CNN）、SegFormer [46]（基于Transformer）以及Mask2Former [11]（基于掩码分类）。对于逐像素分类框架（SegNeXt 和SegFormer），我们仅将最终的1×1 卷积层替换为我们提出的偏移学习模块。对于掩码分类框架（Mask2Former），我们仅使用偏移学习对掩码嵌入与像素嵌入进行对齐，其余部分保持不变。

带有偏移学习的**SegNeXt:** 为评估我们方法的鲁棒性，我们在四个数据集上进行了实验。如表 3 所示，将偏移学习范式集成到SegNeXt 中，在几乎不增加参数量的前提下，带来了稳定的性能提升。总体而言，我们的方法在Tiny、Small 和Base 模型规模上分别平均提升了1.4、1.2 和0.5 的mIoU，仅引入0.1-0.2M 的额外参数。这些结果表明，我们的方法在提升分割性能方面具备良好的效果与效率。我们还在SegNeXt-T 模型中使用多尺度集成策略评估了偏移学习，其在ADE20K、Cityscapes、Pascal Context 和COCO-Stuff 上分别取得了43.2、81.5、54.5 和40.5 的mIoU，进一步提升了分割性能。

为进一步展示我们方法的优势，我们在Fig. 5 中展示了基于SegNeXt-T 模型的分割可视化结果。可视化结果显示，我们的方法能更准确地完成分割，尤其在背景区域和小物体（如第三张图中的时钟）识别方面表现显著。这从定性角度验证了我们方法在图像特征与类别表示对齐方面的有效性。

此外，随着模型规模的增大，性能增益从1.4 降至0.5 的趋势表明，大模型本身具备更强的特征-表示对齐能力。这一实验观察直接验证了我们关于“模型规模越大，其类别对齐能力越强”的假设。同时也表明，我们的方法正如分析所预测的那样，能高效解决高效模型中存在的类别错配问题。

带有偏移学习的**SegFormer:** 为系统评估偏移学习范式在不同模型容量下的适配性，我们在六种SegFormer 架构（B0-B5）上进行了全面实验，并在ADE20K 和COCO-Stuff 数据集上报告结果。为确保公平性，我们采用mmsegmentation [12] 提供的结果，其FLOPs 数值略低于原论文报告。从B0 到B5，在两个数据集上的平均mIoU 提升分别为2.7、2.2、1.2、1.3、1.1 和1.0。随着模型规模的增大，性能提升逐渐减弱，这一趋势与SegNeXt 实验中的结论一致，进一步验证了我们关于“高效模型中存在类别错配问题”的假设。也再次证明了我们提出的偏移学习范式在解决传统逐像素分类

Table 5. Mask2Former [11] 与其偏移学习版本在ADE20K 数据集上的性能对比。

方法	偏移学习	参数量(M)	mIoU
Mask2Former-Tiny		47.4	47.7
Mask2Former-Tiny	✓	47.6	50.3$_{(+2.6)}$

中类别对齐问题上的有效性。

带有偏移学习的**Mask2Former:** 我们在Mask2Former-Tiny 模型上开展实验，尽管其为Tiny 版本，其参数量已达47.4M。如表 5 所示，将偏移学习集成到掩码分类框架中，在仅增加0.2M 参数的情况下，mIoU 提升了2.6。正如Tab. 1 中的对比所示，掩码分类方法仅调整类别表示，而我们的偏移学习同时对图像特征和类别表示进行对齐。这表明，我们的方法在实现图像特征（逐像素嵌入）与类别表示（掩码嵌入）高效协同对齐方面具有显著优势。

4.4. 消融实验

核心组件的消融分析**:** 为系统验证 OffSeg 框架中各个组件的有效性，我们在ADE20K 数据集上进行了消融实验，具体结果见Tab. 6。基线模型（第一行）使用简单的卷积像素解码器，未引入FreqFusion [5]，其mIoU为40.7。仅引入FreqFusion 模块后，mIoU 提升了1.3。在引入FreqFusion 的基础上，加入类别偏移学习与特征偏移学习，分别带来0.9 和1.5 的mIoU 提升。这表明两个分支都能独立提升模型性能，其中可自适应的图像特征对性能的提升作用大于可调整的类别表示。同时引入两个分支后取得了最佳性能，展示了它们在图像特征与类别表示对齐方面的协同增益效果。

通道数影响的消融分析**:** 如 Tab. 7 所示，我们以仅包含FreqFusion 的基线模型作为基础，验证通道数对逐像素分类范式的影响。当通道数从64 扩展到1024 时，结果表明，尽管增加特征维度可以提升模型表达能力，但当通道数达到2048 时，性能反而有所下降。这说明在逐像素分类范式中，增加图像特征通道数与类别表示维度确实可以提升表达能力，因为高维向量可表示更高维的空间。

然而，简单地通过增加通道数来提升模型性能会带来显著的计算开销，并存在上限。例如，在1024 通道下，模型可达43.5 mIoU，但FLOPs 达到11.1G；继续增加通道数并未带来进一步性能提升。相比之下，我们的OffSeg-T 模型采用偏移学习范式，在仅使用一半计算资源的情况下即可达到44.2 mIoU。该对比进一步验证了我们所提出方法在实现高精度语义分割的同时具备良好的计算效率。

5. Conclusions

本文系统分析了逐像素分类范式的局限性，特别是图像特征与类别表示之间的对齐失配问题。为了解决这一问题，我们提出了偏移学习范式，通过引入特征偏移学习与类别偏移学习两个分支，显式学习图像特征与其对应类别表示之间所需的偏移量，从而实现有效对齐。基于该范式，我们设计了高效的语义分割网络，OffSeg，并提供三种不同的模型规模以适应不同应用需求。作为一种通用的分割范式，我们还将偏移学习范式集成到三种具有代表性的语义分割方法中，包括SegFormer、SegNeXt 和Mask2Former，且仅引入了极小的参数量。在四个广泛使用的数据集上的大量实验结果充分验证了我们所提出的偏移学习范式在准确性与效率方面的优势。

Table 6. 对 OffSeg 中不同组件的消融实验。FF.、CO. 和FO. 分别表示FreqFusion、类别偏移学习和特征偏移学习。

FF.	CO.	FO.	参数量(M)	FLOPs (G)	mIoU
			5.9	5.1	40.7
✓			6.1	6.0	42.0
✓	✓		6.2	5.3	42.9
✓		✓	6.2	5.3	43.5
✓	✓	✓	6.2	5.3	44.2

Table 7. 通道数对图像特征与类别表示的影响消融实验。所有模型均为逐像素分类范式，不包含偏移学习。

通道数	64	128	256	512	768	1024	2048
参数量(M)	6.0	6.0	6.1	6.2	6.3	6.4	6.8
FLOPs (G)	4.7	5.1	6.0	7.7	9.4	11.1	17.9
mIoU (%)	41.2	41.7	42.0	42.8	42.9	43.5	43.0

Acknowledgment

本研究部分得到了以下项目的资助支持：国家自然科学基金（编号：62495061，62276145），国家重点研发计划（编号：2024YFE0100700），天津市科技支撑计划项目（编号：23JCZDJC01050），以及新加坡A*STAR 青年科学家发展基金（编号：C233312006）。本研究还得到了南开大学超级计算中心的部分支持。

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12):2481–2495, 2017. 2

[2] Walid Bousselham, Guillaume Thibault, Lucas Pagano, Archana Machireddy, Joe Gray, Young Hwan Chang, and Xubo Song. Efficient self-ensemble for semantic segmentation. In *BMVC*, 2022. 5

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *IEEE CVPR*, pages 1209–1218, 2018. 5

[4] Niccolo Cavagnero, Gabriele Rosi, Claudia Cuttano, Francesca Pistilli, Marco Ciccone, Giuseppe Averta, and Fabio Cermelli. Pem: Prototype-based efficient maskformer

for image segmentation. In *IEEE CVPR*, pages 15804–15813, 2024. 2, 5

[5] Linwei Chen, Ying Fu, Lin Gu, Chenggang Yan, Tatsuya Harada, and Gao Huang. Frequency-aware feature fusion for dense image prediction. *IEEE TPAMI*, 2024. 4, 7

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 2

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2017. 3

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 2

[10] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 34:17864–17875, 2021. 2, 3, 4, 5

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE CVPR*, pages 1290–1299, 2022. 2, 3, 4, 5, 6, 7

[12] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 5, 6

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE CVPR*, pages 3213–3223, 2016. 1, 5

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

[15] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *Visual Intelligence*, 1(1):16, 2023. 1

[16] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *IEEE CVPR*, pages 9716–9725, 2021. 2

[17] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE CVPR*, pages 3146–3154, 2019. 2

[18] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *NeurIPS*, 35:1140–1156, 2022. 1, 2, 3, 4, 5, 6

[19] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *IEEE CVPR*, pages 4003–4012, 2020. 3

[20] Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications, 2024. 1

[21] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *IEEE CVPR*, pages 7287–7296, 2022. 2

[22] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *IEEE CVPR*, pages 9522–9531, 2019. 1

[23] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *NeurIPS*, 35:12934–12949, 2022. 1

[24] Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. In *IEEE ICCV*, pages 16889–16900, 2023. 4

[25] Yunheng Li, Yuxuan Li, Quansheng Zeng, Wenhai Wang, Qibin Hou, and Ming-Ming Cheng. Unbiased region-language alignment for open-vocabulary dense prediction. *arXiv preprint arXiv:2412.06244*, 2024. 1

[26] Yunheng Li, Zhong-Yu Li, Quan-Sheng Zeng, Qibin Hou, and Ming-Ming Cheng. Cascade-CLIP: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28243–28258. PMLR, 2024. 1

[27] Dong Liang, Yue Sun, Yun Du, Songcan Chen, and Sheng-Jun Huang. Relative difficulty distillation for semantic segmentation. *Science China Information Sciences*, 67(9): 192105, 2024. 2

[28] Zheng Lin, Zhao Zhang, Zi-Yue Zhu, Deng-Ping Fan, and Xia-Lei Liu. Sequential interactive image segmentation. *Computational Visual Media*, 9(4):753–765, 2023. 1

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE ICCV*, pages 10012–10022, 2021. 1

[30] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *IEEE CVPR*, pages 12009–12019, 2022. 1

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE CVPR*, pages 3431–3440, 2015. 2

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[33] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE CVPR*, pages 891–898, 2014. 5

[34] Zhenliang Ni, Xinghao Chen, Yingjie Zhai, Yehui Tang, and Yunhe Wang. Context-guided spatial feature reconstruction

for efficient semantic segmentation. In *ECCV*, pages 239–255. Springer, 2024. 1, 2, 5

[35] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE ICCV*, pages 12179–12188, 2021. 2

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1, 2

[37] Jae-hun Shim, Hyunwoo Yu, Kyeongbo Kong, and Suk-Ju Kang. Feedformer: Revisiting transformer decoder for efficient semantic segmentation. In *AAAI*, pages 2263–2271, 2023. 1, 2, 5

[38] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *IEEE ICCV*, pages 7262–7272, 2021. 1, 2

[39] Changki Sung, Wanhee Kim, Jungho An, Wooju Lee, Hyungtae Lim, and Hyun Myung. Contextrast: Contextual contrastive learning for semantic segmentation. In *IEEE CVPR*, pages 3732–3742, 2024. 1

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2

[41] Qiang Wan, Zilong Huang, Jiachen Lu, YU Gang, and Li Zhang. Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation. In *The eleventh international conference on learning representations*, 2023. 1, 2, 5

[42] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 43(10): 3349–3364, 2020. 2

[43] Jian Wang, Chenhui Gou, Qiman Wu, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Rtformer: Efficient design for real-time semantic segmentation with transformer. *NeurIPS*, 35:7423–7436, 2022. 1, 2, 5

[44] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. P2T: Pyramid pooling transformer for scene understanding. *IEEE TPAMI*, 45(11):12760–12771, 2023.

[45] Yu-Huan Wu, Shi-Chen Zhang, Yun Liu, Le Zhang, Xin Zhan, Daquan Zhou, Jiashi Feng, Ming-Ming Cheng, and Liangli Zhen. Low-resolution self-attention for semantic segmentation. *IEEE TPAMI*, 2025. 1, 2

[46] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34:12077–12090, 2021. 1, 2, 3, 4, 5, 6

[47] Haotian Yan, Ming Wu, and Chuang Zhang. Multi-scale representations by varying window attention for semantic segmentation. In *ICLR*, 2024. 2, 5

[48] Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. In *ICLR*, 2024. 1

[49] Bo-Wen Yin, Jiao-Long Cao, Ming-Ming Cheng, and Qibin Hou. Dformerv2: Geometry self-attention for rgbd semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19345–19355, 2025. 1

[50] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 325–341, 2018. 1, 2

[51] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *IJCV*, 129:3051–3068, 2021. 1, 2

[52] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2

[53] Hyunwoo Yu, Yubin Cho, Beoungwoo Kang, Seunghun Moon, Kyeongbo Kong, and Suk-Ju Kang. Embedding-free transformer with inference spatial reduction for efficient semantic segmentation. In *ECCV*, pages 92–110. Springer, 2024. 5

[54] Yuhui Yuan, Xilin Chen, and Jingdong 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. 2

[55] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *NeurIPS*, 34: 7281–7293, 2021. 2

[56] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, et al. Segvit: Semantic segmentation with plain vision transformers. *NeurIPS*, 35:4971–4982, 2022. 1

[57] Dong Zhang, Liyan Zhang, and Jinhui Tang. Augmented fcn: rethinking context modeling for semantic segmentation. *Science China Information Sciences*, 66(4):142105, 2023. 2

[58] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation. In *IEEE CVPR*, pages 12083–12093, 2022. 1, 2

[59] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE CVPR*, pages 2881–2890, 2017. 2, 3

[60] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE CVPR*, pages 6881–6890, 2021. 1, 2

[61] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE CVPR*, pages 633–641, 2017. 1, 5

[62] Tianfei Zhou and Wenguan Wang. Cross-image pixel contrasting for semantic segmentation. *IEEE TPAMI*, 2024. 1

[63] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 1