

TopFormer：用于移动语义分割的令牌金字塔转换器

摘要

轻量化卷积神经网络（CNN）通常用于移动视觉任务，由于卷积神经网络具有归纳偏置，这使得它能够以更少的参数学习特征表征，从而在不同的视觉任务中发挥作用。然而，卷积神经网络主要学习局部特征，而对于学习全局表征来说，采用基于自注意力机制的视觉转换器（Vision Transformers, ViTs）成为一种选择。虽然视觉转换器在计算机视觉领域取得了重大的成功，但由于其高昂的计算代价，限制了其在移动设备上进行语义分割等密集预测任务的表现。文中针对上述问题提出了一种新的架构即移动语义分割的令牌金字塔转换器（TopFormer）。具体来说，该架构以不同尺度的 tokens 作为输入，产生尺度感知的语义特征，然后注入到相应的 tokens 中以增强特征表示。实验结果表明，该方法在 ADE20K 语义分割数据集上的性能明显优于基于 CNN 和 ViT 的网络，并在准确率和延时性之间取得了良好的平衡。在 ADE20K 数据集上，TopFormer 在基于 ARM 的移动设备上的延迟比 MobileNetV3 高 5%。复现该方法后在 ADE20K 和 COCO-Stuff 基准数据集上进行了实验，并针对复现结果提出了改进方案。

关键词：移动设备；尺度感知语义特征；语义分割；轻量化

1 引言

图像处理一直以来都是人类广泛关注的研究焦点。图像在日常生活中能够为我们提供丰富的信息，帮助我们更好地理解事物。随着数字技术的不断发展，数字图像的传播变得更加广泛。尽管人眼能够快速地识别、区分和做出决策，但对于计算机而言，这个过程却是一项复杂的任务。这是因为计算机在处理图像时涉及到低层次的基础处理技术和高层次的场景识别任务。图像语义分割作为图像处理的核心过程，在计算机领域中扮演着重要的角色。同时，这一课题一直是计算机视觉中既基础又具有代表性的任务。语义分割的目标是将待处理的对象分割成若干区域，并为每个区域赋予特定的类别标签，以确保拥有相同标签的区域具有相似的特征。作为场景理解领域的主要研究方向之一，语义分割技术在遥感、交通、医疗等领域的应用为社会创造了巨大的经济价值和社会效益。

从上世纪六七十年代开始，研究人员深入研究了语义分割领域。在深度学习兴起之前，采用的是传统图像语义分割方法。这些方法主要依赖于图像中的边缘、颜色、纹理等特征，通过基于阈值、边缘、聚类、图论等经典的分割方法，将图像分隔成不同的区域。由于计算机硬件设备的限制，最初的图像语义分割仅能处理灰度图像，直到逐渐发展到对 RGB 图像的处理阶段。随着 GPU 技术的迅猛发展，深度学习（Deep Learning, DL）[10] 为语义分割技术提供了强有力的支持。研究人员借助卷积神经网络（Convolutional Neural Network, CNN），通

过端到端的训练方式，能够推断每个像素的语义信息并实现对图像区域的有意义分类。由于 CNN 具有显著的特征学习和表达能力，逐渐成为语义分割领域的首选方法。2015 年，Long 等学者引入了全卷积神经网络 (Fully Convolutional Networks for Semantic Segmentation, FCN) [18]，这标志着图像语义分割步入了全卷积神经网络时代。随着时间推移，全卷积神经网络在深度学习领域展现出强大的潜力，并逐渐成为解决图像语义分割问题的首选方法。

深度学习技术推动了语义分割领域的显著进展，当前深度学习领域的重要关注点之一是 Transformer 模型。最初，Transformer 在自然语言处理领域 (Natural Language Processing, NLP) 取得了卓越成果，随后 Dosovitskiy 等研究者 [5] 提出了 Vision Transformer (ViT) 将其引入图像分类研究，为计算机视觉领域注入了新的活力，并在多个任务领域的实践中取得了引人瞩目的成功，其中包括应用于语义分割任务。这一创新模型在计算机视觉和深度学习社区内引起了广泛讨论和深入研究。然而，在对实时性要求较高的场景下，较大规模的 Transformer 模型可能面临推理速度较慢的问题，引发了与 Transformer 相关的一系列热点问题的探讨和研究。

实时性要求较高的场景是图像语义分割领域面临的一个重要现实。这些场景的特殊性使得图像语义分割在实际应用中必须具备快速准确的处理能力，以满足用户对系统性能和用户体验的高要求。以下是几个实时性要求较高的场景：视频监控系统、医学影像分析、自动驾驶系统 [36] [37]，这些场景中的实时性要求驱动着图像语义分割技术的不断发展和创新。轻量化 Transformer 模型设计成为解决实时性问题的关键方向之一。通过设计更为紧凑、高效的模型结构，降低模型参数量和计算复杂度，以确保在保持图像语义分割准确性的同时提高推理速度。这样的模型结构不仅适用于大规模的云端计算，更能够更好地适应资源受限的嵌入式设备或实时系统的需求。

为了使 Vision Transformers 适应各种密集预测任务，最新的 Vision Transformers 如 PVT [26]、CVT [27]、Levit [7]、MobileViT [20] 采用了分层结构，这种结构通常用于卷积神经网络 (CNN)，如 AlexNet [14]、ResNet [9]。这些视觉变形器将全局自我注意及其变体应用于高分辨率标记，由于标记数目的二次复杂性，带来了较大的计算代价。为了提高效率，最近的一些工作，如 Swin Transformer [17]，Shuffle Transformer [13]，Twin [4] 和 HR-Forform [30]，计算了局部/窗口区域内的自我注意。然而，在移动设备上，窗口分区非常耗时。此外，令牌瘦身 [25] 和移动形成器 [3] 通过减少令牌数量降低了计算量，但牺牲了它们的识别准确率。

在这些 Vision Transformers 中，MobileViT [20] 和 Mobile-Form [3] 是专门为移动设备设计的。它们都结合了 CNN 和 VITS 的优点。对于图像分类，MobileViT 在参数数量相近的情况下比 MobileNet [11] 获得更好的性能。前者比 MobileNet 获得更好的性能，失败次数更少。然而，与移动网络相比，它们在移动设备上的实际延迟并没有显示出优势，如 [20] 中所述。它提出了一个问题：是否有可能设计出移动友好的网络，在移动语义分割任务中取得比移动网络更好的性能，并且具有更低的延迟？

受 MobileViT 和 Mobile-Form 的启发，作者还利用了 CNN 和 VITS 的优势。一个基于 CNN 的模块，称为令牌金字塔模块，用于处理高分辨率图像，以快速产生局部特征金字塔。考虑到移动设备上的计算能力非常有限，这里作者使用几个堆叠的轻量级 MobileNetV2 块和快速下采样策略来构建令牌金字塔。为了获得丰富的语义和较大的接受域，采用了基于 VIT 的语义抽取器模块，并将标记作为输入。为了进一步降低计算成本，使用平均池运算符将令牌减少到极小的数字，例如，输入大小的 $1/(64 \times 64)$ 。与 VIT [6] 不同，T2T-VIT [29] 和 Levit [7]

使用嵌入层的最后输出作为输入令牌，作者将来自不同尺度（阶段）的令牌汇集成非常小的数字（分辨率），并沿着通道维度将它们连接起来。然后，新的令牌被馈送到 Transformer 块中以产生全局语义。由于 Transformer 块中的剩余连接，学习到的语义与令牌的比例相关，表示为可伸缩的全局语义。为了获得密集预测任务的强大层次特征，尺度感知的全局语义被来自不同尺度的标记通道拆分，然后将尺度感知的全局语义与相应的标记融合以增强表示。使用扩充的标记作为分割头的输入。

为了获得密集预测任务的强大层次特征，尺度感知的全局语义被来自不同尺度的标记通道拆分，然后将尺度感知的全局语义与相应的标记融合以增强表示。使用扩充的标记作为分割头的输入。为了验证该方法的有效性，作者在具有挑战性的分词数据集上进行了实验：ADE20K [34]、PASCAL CONTEXT [22] 和 COCOSTuff [1]。检查了硬件上的延迟，即基于 ARM 的现成计算核心。下面是这篇论文的主要贡献：

- TopFormer 以不同尺度的标记词作为输入，将标记词集合成很小的个数，从而以很小的计算代价获得尺度感知语义。
- 提出的语义注入模块可以将尺度感知的语义注入到相应的标记中，以构建强大的层次特征，这对于密集预测任务至关重要。
- 在基于 ARM 的移动设备上，在 ADE20K 数据集上，所提出的基本模型可以比 MobileNetV3 更好地实现 5% 的 mIoU，并且具有更低的延迟。微型版本可以在基于 ARM 的移动设备上执行实时分割，结果具有竞争力。

2 相关工作

在这一部分中，将从三个方面对最近的研究进行了综述：1) 轻量级视觉转换器，2) 高效卷积神经网络，3) 移动语义分割。

2.1 轻量化 Transformer

对于图像识别中变压器结构的使用有很多探索 [12]。ViT [6] 是第一个将纯 Transformer 应用于图像分类的工作，实现了最先进的性能。接下来，DeiT 引入了基于令牌的蒸馏，以减少训练变压器所需的数据量。T2T-ViT [29] 通过递归地将相邻标记聚合成一个标记以减少标记长度，将图像构造为标记。Swin Transformer [17] 计算每个局部窗口内的自注意力，从而导致输入标记数量的线性计算复杂性。然而，这些视觉变压器及其后续产品往往具有大量参数和繁重的计算复杂性。

为了构建轻量级的 Vision Transformer，LeViT [7] 设计了一种混合架构，该架构使用具有 stride-2 的堆叠标准卷积层来减少标记数量，然后附加改进的 Vision Transformer 来提取语义。对于分类任务，LeViT 在 CPU 上的性能明显优于 EfficientNet。MobileViT [20] 采用相同的策略，并使用 MobilenetV2 块代替标准卷积层对特征图进行下采样。Mobile-Former 采用双向桥的并行结构，充分利用了 MobileNet 和 Transformer 的优点。然而，如 [20] 中所述，MobileViT 和其他基于 ViT 的网络在移动设备上明显慢于 MobileNets [11]。对于分割任务，输入图像始终是高分辨率的。因此，基于 ViT 的网络比 MobileNet 执行得更快更具挑战性。

在本文中，目的是设计一种轻量级的 VisionTransformer 模型，它可以在分割任务中以更低的延迟超越 MobileNet。

2.2 高效的卷积神经网络

在移动和嵌入式设备上部署视觉模型的需求不断增长，鼓励了对高效卷积神经网络设计的研究。MobileNet [11] 提出了一种反向瓶颈结构，该结构堆叠了深度卷积和点卷积。IGC-Net [32] 和 ShuffleNet [19] 对通道使用通道洗牌/排列算子，使多个组卷积层实现跨组信息流。GhostNet [8] 使用更便宜的算子，即深度卷积来生成更多特征。AdderNet [2] 利用加法来进行大规模乘法运算。MobileNeXt [35] 翻转了反向残差块的结构，并提出了一个连接高维表示的构建块。EfficientNet 和 TinyNet 研究深度、宽度和分辨率的复合缩放。

2.3 移动语义分割

最准确的分割网络通常需要数十亿次的计算，这可能超出移动和嵌入式设备的计算能力。为了加速分割并降低计算成本，ICNet [33] 使用多尺度图像作为输入和级联网络以提高效率。DFANet [15] 利用轻量级主干来加速其网络，并提出跨级特征聚合来提高准确性。SwiftNet [23] 使用横向连接作为具有成本效益的解决方案来恢复预测分辨率，同时保持速度。BiSeNet [28] 引入空间路径和语义路径来减少计算量。AlignSeg [13] 和 SFNet [16] 对齐相邻级别的特征图，并使用特征金字塔框架进一步增强特征图。ESPNet [21] 通过将标准卷积分解为逐点卷积和扩张卷积的空间金字塔来节省计算量。AutoML 技术用于搜索场景解析的高效架构。NRD [31] 使用动态卷积滤波器网络动态生成神经表示。未来，随着对实时性要求的不断提高，图像语义分割技术将不断迭代和演进。轻量化 Transformer 模型的研究将在实际应用中扮演关键角色，推动图像语义分割技术更好地服务于现实场景中的需求，使得图像语义分割在实时性要求较高的领域取得更为显著的成果。

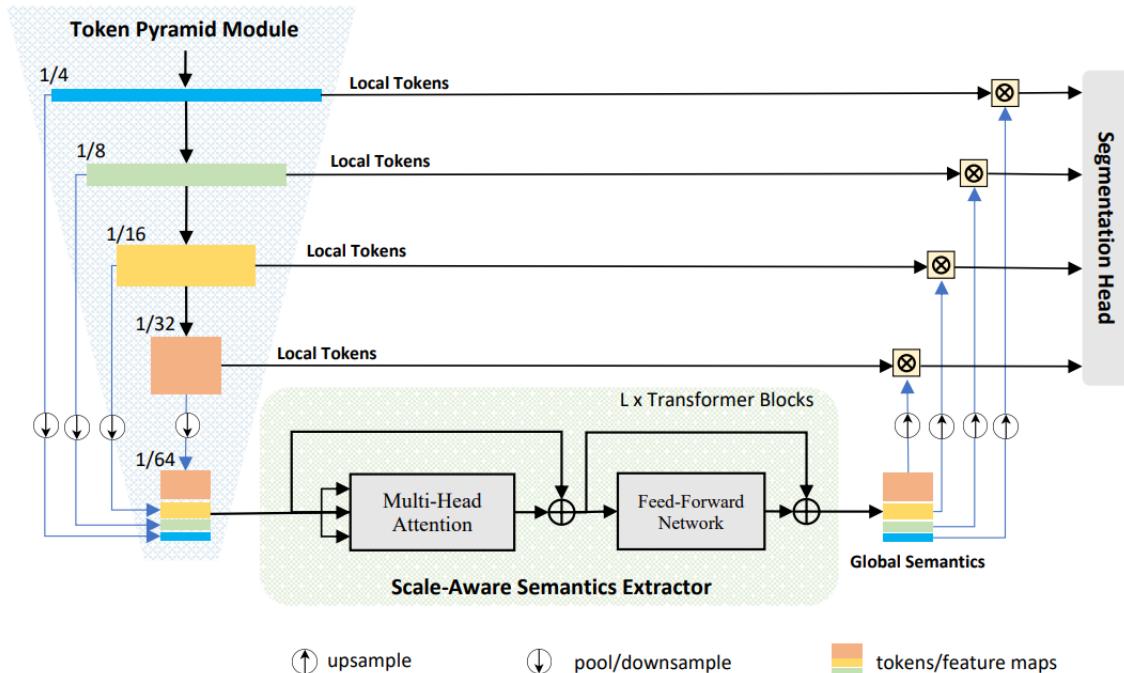


图 1. 整体网络架构

3 本文方法

3.1 本文方法概述

整个网络架构如图 2 所示。正如我们所看到的，网络由几个部分组成：Token Pyramid Module、Semantics Extractor、Semantics Injection Module 和 Segmentation Head。令牌金字塔模块将图像作为输入并生成令牌金字塔。视觉转换器被用作语义抽取器，它将令牌金字塔作为输入并产生可伸缩的语义。语义被注入到相应尺度的标记中，以通过语义注入模块来扩充表示。最后，分割头使用扩展的令牌金字塔执行分割任务。接下来，将介绍这些模块的详细信息。

3.2 Token Pyramid Module

受 MobileNet [24] 的启发，建议的令牌金字塔模块由堆叠的 MobileNet 块组成 [24]。与 MobileNet 不同，令牌金字塔模块不以获取丰富的语义和大的接受域为目标，而是使用更少的块来构建令牌金字塔。该模块的结构由一系列堆叠的 MobileNet blocks 组成。与 MobileNet 不同，该模块的目标不是获取丰富的语义信息和更大的感受野，而是通过少量的块构建 Token Pyramid。该模块将通过 MobileNetV2 处理输入图像，生成一系列的 tokens，这些 tokens 会被平均池化到目标大小，然后将这些 tokens 按通道维度拼接在一起，最后传入 L 个 Transformer blocks。MobileNet 的轻量化设计使得整个模块在保持相对低的参数量和计算复杂度的同时，能够有效地提取图像特征。

3.3 Semantics Extractor

该模块由 L 个 Transformer blocks 构成，每个 Transformer block 包含多头注意力模块 (Multi-Head Attention Module)、前馈网络 (FFN) 和残差连接。通过该模块可以获得尺度感知的语义信息。Transformer 的并行计算和注意力机制使得在相对较少的参数下实现了对复杂语义的高效提取。

3.4 Injection Module and Segmentation Head

Semantics Injection Module：该模块的内部结构如图 2(a) 所示，由于 Semantics Extractor 获得的尺度感知语义和 tokens 存在语义差距，因此引入 Semantics Injection Module 来解决这个问题。Local tokens 先经过 1×1 卷积层，然后通过 Batch Normalization (BN) 处理；全局 semantics 先进行上采样，然后进行 1×1 卷积、BN、sigmoid 处理。同时，两者经过 1×1 卷积和 BN 处理，得到三个相同大小的特征图，其中前两个进行注入操作，然后将结果相加。通过轻量的 1×1 卷积和 BN 处理，以及局部 tokens 的注入操作，该模块有效地处理了语义的一致性问题。语义注入后，来自不同尺度的扩展标记同时捕捉了丰富的空间信息和语义信息，这是语义分割的关键。此外，语义注入缓解了标记之间的语义鸿沟。该分割算法首先对低分辨率的标记点进行上采样，得到与高分辨率标记点相同的大小，然后对所有尺度的标记点进行元素求和。最后，将特征经过两层卷积得到最终的分割图。

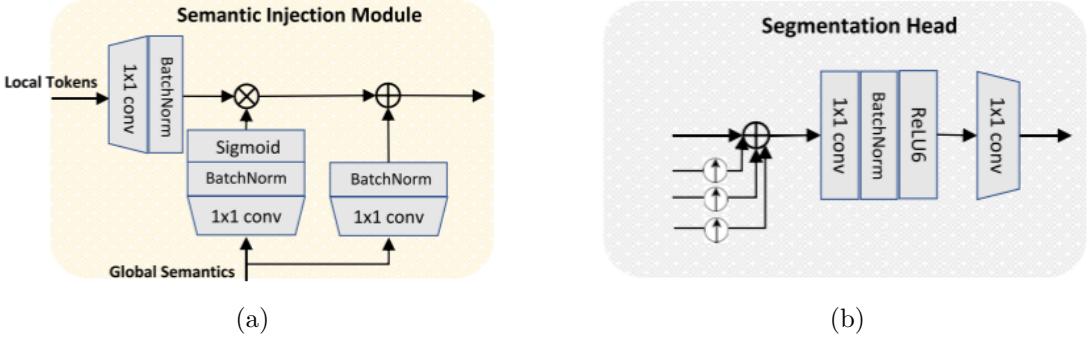


图 2. TopFormer 架构的两个模块。(a) Semantics Injection Module 模块结构 (b) Segmentation Head 模块结构

3.5 模型及其变量

为了定制各种复杂的网络，分别引入了 TopFormer-Tiny(TopFormer-T)、TopFormer-Small(TopFormer-S) 和 TopFormer-Base(TopFormer-B)。基本模型、小型模型和微型模型分别在每个多头自我注意模块中有 8、6 和 4 个头，目标通道数为 $M=256$ 、 $M=192$ 和 $M=128$ 。为了在准确率和实际时延之间实现更好的折衷，选择了最后三个尺度 T2、T3 和 T4 中的标记作为 SIM 和分割头的输入。

3.6 TopFormer 网络架构总结

整个 TopFormer 架构通过这些模块的协同工作，实现了对图像的语义分割任务。Token Pyramid Module 融合了不同尺度的 tokens，Semantics Extractor 提取尺度感知的语义信息，Semantics Injection Module 解决不同尺度语义差异的问题，而 Segmentation Head 产生最终的分割图。该架构的创新点在于通过 Token Pyramid Module 结构将 token 多尺度化，然后将不同尺度的 token 进行融合。这样做既融合了多尺度的图像又因为 H,W 减小而减小了计算量。再通过原始 token 与经过 transformer 处理后的特征图进行融合，在语义分割的任务表现出了不错的效果。

整个架构在设计上充分考虑了尺度和语义的关系，以提高语义分割的准确性。TopFormer 整体架构在设计上考虑了轻量化的思想，通过简洁而高效的模块组合，使得模型在保持良好性能的同时，具备轻量化模型所需的低计算资源和小模型体积。TopFormer 架构提供了一种有效的解决方案，使得轻量化语义分割模型能够在资源受限的环境下取得更好的性能。其对轻量化和高效性的关注为语义分割任务的移动端和嵌入式设备部署提供了实用的解决途径。

4 复现细节

4.1 与已有开源代码对比

在这一部分中，首先在几个公共数据集上进行实验。描述了语义分割任务的实现细节，并将结果与其他工作进行了比较。然后，分析不同部位的效果和效率。最后，进行改进。该论文所提的方法作者在 github 有用 pytorch 实现的 demo: <https://github.com/hustvl/TopFormer>。在本次复现中，重新使用 python 复现了整个过程。仔细研究了其各个模块的设计与实现，

这个过程包括对 Token Pyramid Module、Semantics Extractor、Semantics Injection Module 和 Segmentation Head 等关键部分的代码逻辑的仔细审查。在复现阶段，选择了 coco stuff 10k、ADE20K 数据集作为实验的基准，以保证实验结果的可比性和通用性。实验环境选用了 NVIDIA Corporation GM200，以确保计算资源的充足和性能的稳定。

4.2 数据集

本次复现使用了两个数据集，分别是原论文中使用到的 ADE20K 数据集和 coco stuff 10k 数据集。ADE20K 涵盖了场景、对象、对象部分的各种注释，在某些情况下甚至是部分的部分。有 25k 张复杂日常场景的图像，其中包含自然空间环境中的各种对象。每个图像平均有 19.5 个实例和 10.5 个对象类。COCO-Stuff 10k 是 COCO 数据集的一个子集，专注于语义分割任务。作为 Microsoft COCO 团队创建的一部分，该数据集包含丰富多样的图像，涵盖了室内外各种场景，包括人、动物、交通工具、食物等多个语义类别。每个图像都配备了详细的像素级别标注，使其成为训练和评估语义分割模型的理想选择。尽管规模相对较小，COCO-Stuff 10k 仍为研究人员和从业者提供了足够的数据量，以便于开发和评估语义分割算法。这个数据集的特点使其成为深入研究密集预测任务的有力工具，实验中，分别这两种数据集评估在 mIoU 上的效果，以及对比不同模型参数量大小。

4.3 性能指标

本次实验采用 mIoU 来评估模型的性能指标，旨在衡量模型在每个类别上的分割准确性以及整体分割性能，提供了对整体性能的综合评估。IoU 是预测区域与真实标注区域的交集与并集之比，对于多个类别，mIoU 计算如下：

$$mIoU = \frac{1}{N} \sum_{i=1}^N IoU_i \quad (1)$$

其中，N 是类别的总数， IoU_i 是第 i 个类别的 IoU。mIoU 的取值范围是 0,1，值越接近 1 表示模型的分割性能越好。

4.4 改进的 Semantics Extractor 模块

经过对 TopFormer 源代码的深入研究，对其注意力模块（即 Multi-Head Attention）和前馈神经网络（Feed-Forward Network）进行了简化。该模块将多尺度提取到的特征与经过多头注意力机制得到的特征进行相加，相加后得到的特征与经过 MLP 块得到的特征进行相加。改进灵感主要来自 EfficientViT 模块，将其应用于 Semantics Extractor 模块中，具体表现为将原有的 softmax 注意力模块替换为更为简单的 ReLU 线性注意力模块，如图 3 所示。这一改进并非简单的模块替换，而是基于对模型性能和计算效率的深入思考而得出的结论。原 Semantics Extractor 中采用的 softmax 注意力模块在计算上相对复杂，需要更多的参数和计算资源。为了进一步提高模型的计算效率和轻量化程度，选择了更为简便的 ReLU 线性注意力模块。这个替换不仅在计算复杂度上有所降低，而且减少了需要存储和处理的参数数量，为模型的轻量化提供了更为直接的途径。由于 ReLU 线性注意力模块的设计更加简单，不仅在训练阶段计算更为迅速，而且在推理时也能够更加高效地执行。这样的改进有望在一定程度上减轻模型对硬件资源的依赖，使其更适合于部署在资源受限的环境中。

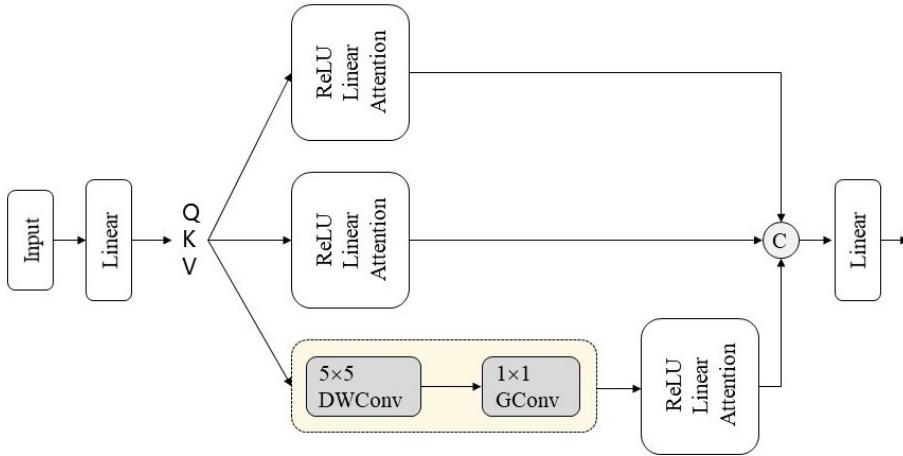


图 3. 改进后的 Semantics Extractor

5 实验结果分析

在 Topformer 的复现过程中，由于实验设备所用设备不同，且设置的 batchsize 与原论文不同，在 coco stuff 10k 实验结果显示其精度与原论文存在 0.51%-0.53% 的差异。结果表如表 1 所示和可视化效果图如图 4 所示。然而，从分割效果来看，模型在细节上并未能很好地对每个物体进行区分。例如，在大象的后背区域，由于与大象相似的墙壁颜色，模型错误地将其分类为与大象相同的类别。此外，大象的腿部也未能得到较好的分割，表现出对于小目标的识别和分割性能相对较差。这或许可以归因于 TopFormer 采用的自注意力机制更适用于处理更大范围的上下文信息，而对于小尺寸目标的精细特征提取表现相对有限。

以上问题的存在提示了 TopFormer 在应对特定场景和目标尺寸上的挑战。这促使我考虑一些改进措施，例如调整模型结构以更好地捕捉细粒度特征，或者采用其他注意力机制来提高小目标的识别性能。这样的优化可能有助于提高 TopFormer 在实际应用中的鲁棒性和性能表现。在 ADE20K 数据集上的实验结果与原论文对比如表 2 所示，其精度与原论文存在 0.4%-0.9% 的差异。将在该数据集上的训练结果进行可视化得到图 5。在改进过程中是对 Topformer 的 Tiny 模型进行改进的，在修改的过程中碰到了代码报了许多错误，好在通过不断修改修复了报错。然而出来的效果不佳，与原论文相差 8%，分析其原因，可能是修改该模块导致模型学习新的特征或失去对原始特征的捕捉能力。

表 1. 在 coco-stuff 10k 数据集上的实验结果与原论文的对比

选用模型	复现对比 mIoU
Topformer-T	论文结果 28.34
	复现结果 27.81
Topformer-B	论文结果 33.43
	复现结果 32.92



图 4. coco-stuff 10K 数据集上的分割效果

表 2. 在 ADE20K 数据集上的实验结果与原论文的对比

选用模型	复现对比 mIoU
Topformer-T	论文结果 32.8
	复现结果 32.4
Topformer-B	论文结果 37.8
	复现结果 38.7

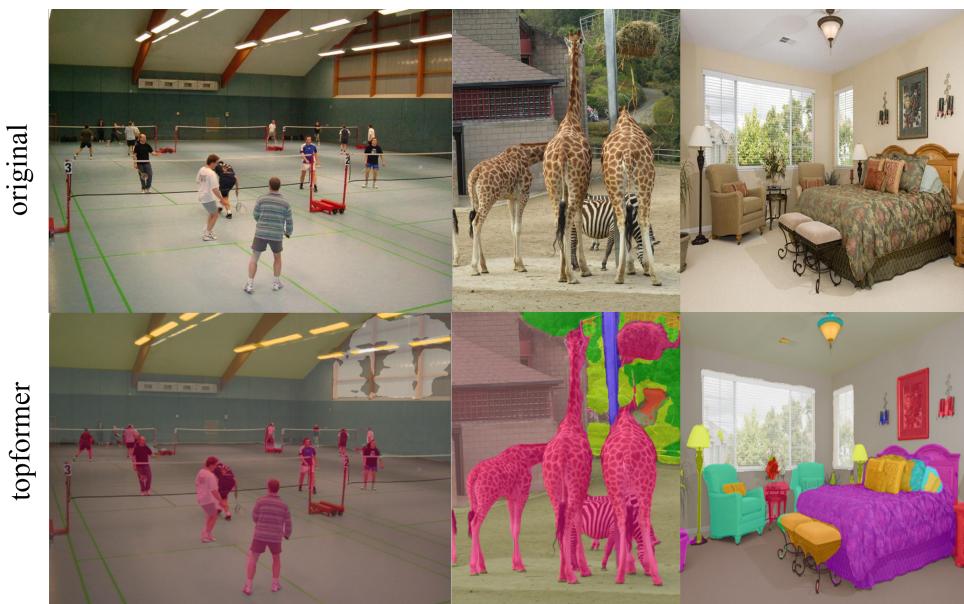


图 5. ADE20K 数据集上的分割效果

6 总结与展望

本研究聚焦于轻量化语义分割问题，深入探讨了该问题的研究背景和意义。旨在计算资源受限的情况下，寻找实现高效图像语义分割任务的有效途径。语义分割作为计算机视觉领域的重要任务，旨在将图像中的每个像素准确标记为不同的语义类别，从而实现对图像内容的细粒度解析和深入理解。提出了一种新的移动视觉任务体系结构。结合了 CNN 和 ViT 的优点，提出的 TopFormer 模型在精度和计算代价之间取得了良好的折衷。微型版的 TopFormer 可以在基于 ARM 的移动设备上产生实时推理，结果具有竞争力。实验结果证明了该方法的有效性。TopFormer 的主要局限性是在目标检测方面有很小的改进。还将在未来的工作中探索 TopFormer 在密集预测中的应用。

本文的研究对轻量化语义分割的发展和优化具有重要的意义。通过分析最新的理论框架和提出改进方案，希望能够为实现在资源受限的环境下高效进行图像语义分割任务提供有益的启示和指导。

参考文献

- [1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.
- [2] Hanting Chen, Yunhe Wang, Chunjing Xu, Boxin Shi, Chao Xu, Qi Tian, and Chang Xu. Addernet: Do we really need multiplications in deep learning? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1468–1477, 2020.
- [3] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5270–5279, 2022.
- [4] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021.
- [5] 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. arxiv 2020. *arXiv preprint arXiv:2010.11929*, 2010.
- [6] 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. *arXiv preprint arXiv:2010.11929*, 2020.

- [7] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet’s clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021.
- [8] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [10] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.
- [11] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.
- [12] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473, 2019.
- [13] Zilong Huang, Yunchao Wei, Xinggang Wang, Wenyu Liu, Thomas S Huang, and Humphrey Shi. Alignseg: Feature-aligned segmentation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):550–557, 2021.
- [14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [15] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9522–9531, 2019.
- [16] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shao-hua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 775–793. Springer, 2020.
- [17] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

- [18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [19] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.
- [20] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.
- [21] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9190–9200, 2019.
- [22] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.
- [23] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12607–12616, 2019.
- [24] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [25] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch slimming for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12165–12174, 2022.
- [26] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.
- [27] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021.
- [28] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

- [29] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021.
- [30] Y Yuan, F Rao, H Lang, W Lin, C Zhang, X Chen, and J Wang. Hrformer: High-resolution transformer for dense prediction. arxiv 2021. *arXiv preprint arXiv:2110.09408*, 19.
- [31] Bowen Zhang, Zhi Tian, Chunhua Shen, et al. Dynamic neural representational decoders for high-resolution semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17388–17399, 2021.
- [32] Ting Zhang, Guo-Jun Qi, Bin Xiao, and Jingdong Wang. Interleaved group convolutions. In *Proceedings of the IEEE international conference on computer vision*, pages 4373–4382, 2017.
- [33] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 405–420, 2018.
- [34] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.
- [35] Daquan Zhou, Qibin Hou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Rethinking bottleneck structure for efficient mobile network design. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III* 16, pages 680–697. Springer, 2020.
- [36] 徐国保, 麦锐滔, 叶昌鑫, 姚旭, and 刘^方辛. 用于自动驾驶的轻量级语义分割神经网络. *Journal of Computer Engineering & Applications*, 59(10), 2023.
- [37] 王大方, 刘磊, 曹江, 赵刚, 赵文硕, and 唐伟. 基于空洞空间池化金字塔的自动驾驶图像语义分割方法. *汽车工程*, 44(12):1818–1824, 2022.