Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(6), 1476-1488
Published Online November 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.146138

Hans汉斯

ESINet

——基于空间信息增强的实时语义分割网络

梅金庄，王　超

北京信息科技大学计算机学院，北京

收稿日期：2025年4月5日；录用日期：2025年11月7日；发布日期：2025年11月17日

摘　要

为提高语义分割实时性，现有的方法细节分支往往设计简单，对空间上下文信息提取极不充分，另外传统卷积操作伴随大量的冗余特征计算对推理速度形成瓶颈。本文提出实时语义分割模型ESINet，引入改进的轻量化空间优化卷积模块(SpaceOptiConv)，采用少量生成特征图并与本征特征进行拼接，减少冗余特征图的计算，降低计算复杂度和参数量；在混合注意力的基础上引入多尺度特征提取，设计了一种多尺度混合注意力机制(HMA)，高效捕捉不同尺度的特征信息；提出一种复合损失函数，交并比损失(IoULoss)优化分割区域的整体重叠度，在线难例挖掘交叉熵损失(OhemCELoss)聚焦类别不平衡、小目标和复杂边界增强局部分类的准确性。对于2048 × 1024的输入，ESINet在Cityscapes测试集上实现81.6%的mIoU，NVIDIA TITAN RTX上的速度为144.5 FPS。相较于基线模型mIoU和分割速度分别取得5.6%和16.4 fps的提高。

关键词

实时语义分割，注意力机制，多尺度特征提取

ESINet

—Real-Time Semantic Segmentation Network Based on Enhanced Spatial Information

Jinzhuang Mei, Chao Wang

Computer School of Beijing Information Science & Technology University, Beijing

Received: April 5, 2025; accepted: November 7, 2025; published: November 17, 2025

Abstract

To improve the real-time performance of semantic segmentation, existing methods often design

文章引用: 梅金庄, 王超. ESINet——基于空间信息增强的实时语义分割网络[J]. 人工智能与机器人研究, 2025, 14(6): 1476-1488. DOI: 10.12677/airr.2025.146138

detail branches that are relatively simple and fail to extract spatial context information effectively. In addition, traditional convolution operations involve significant redundant feature computations, which become bottlenecks for inference speed. This paper proposes a real-time semantic segmentation model, ESINet, which introduces an improved lightweight spatial optimization convolution module (SpaceOptiConv). The module generates a few feature maps and concatenates them with intrinsic features, reducing redundant feature map computations, thus lowering computational complexity and the number of parameters. Based on a hybrid attention mechanism, a multi-scale feature extraction approach is introduced, and a multi-scale hybrid attention mechanism (HMA) is designed to efficiently capture feature information at different scales. Moreover, a composite loss function is proposed, including Intersection over Union loss (IoULoss) to optimize the overall overlap of segmented regions, and Online Hard Example Mining Cross-Entropy Loss (OhemCELoss) to focus on class imbalance, small targets, and complex boundaries, enhancing the accuracy of local classification. With an input size of 2048 × 1024, ESINet achieves 81.6% mIoU on the Cityscapes test set, with a speed of 144.5 FPS on an NVIDIA TITAN RTX. Compared to the baseline model, ESINet achieves improvements of 5.6% in mIoU and 16.4 FPS in segmentation speed.

Keywords

Real-Time Semantic Segmentation, Attention Mechanism, Multi-Scale Feature Extraction

1. 引言

语义分割[1]作为计算机视觉领域的重要任务之一，旨在为图像中的每一个像素赋予语义标签，在自动驾驶、医学图像分析和缺陷检测等领域至关重要。近年来，为了更好地捕获空间语义信息诞生诸多架构，包括多尺度融合(HRNet [2])、编码器–解码器架构(SegNet [3])以及 TransFormer 架构(SegVit [4])等。在实际应用中，尤其是在资源受限的环境下，如移动终端和自动驾驶中模型的推理速度和分割精度仍有待提高。

相较于传统方法，为了特定应用场景的需求，实时语义分割方法通过权衡推理速度与分割精度，提出了多种网络架构。Enet [5]采用降采样策略降低计算成本，BiSeNet [6] [7]系列采用较深的语义分支提取深层语义，浅层细节分支提取空间细节，但细节分支过于简单，未能充分提取特征；为此，STDCSeg [8]采用短期密集级联的单流网络，去除冗余分支，但过多的跳连接同样需要大量计算资源；DDRNet [9]则在早期阶段共享分支，中期通过深度聚合金字塔池化模块级联双分辨率分支；PIDNet [10]则增加边缘引导分支，配合细节分支和语义分支，通过边界注意力机制指导分支融合。以上方法，均涉及大量卷积操作，在大分辨率的细节特征提取过程中存在大量的冗余计算，“幽灵”卷积(Ghost Convolutions) [11]在特征提取时仅生成部分特征图与本征特征拼接，能够有效降低冗余计算，主要区别如图 1 所示。

注意力机制的引入为改善语义分割性能提供了新思路。通过动态聚焦于特征图中的重要区域，使深度学习模型能够捕获长距离的依赖关系，有效地克服了传统 CNN 在全局信息建模方面的限制。PP-LiteSeg [12]引入的注意力融合模块(UAFM)在解码阶段融合了通道和空间注意力。为进一步降低计算复杂度，RTFormer [13]提出了具有线性复杂度的 GPU 友好注意力机制(GFA)。SCTNet [14]在 GFA 的基础上，提出卷积注意力机制，且仅在训练阶段使用的 Transformer 分支监督 CNN 分支的特征学习。CBAM [15]、

CPCA [16]采用双注意力设计，聚合了通道和空间层面信息，但均衡的算子将忽略各通道的个性。本文引入多尺度，设计轻量级混合注意力模块增强特征表达。

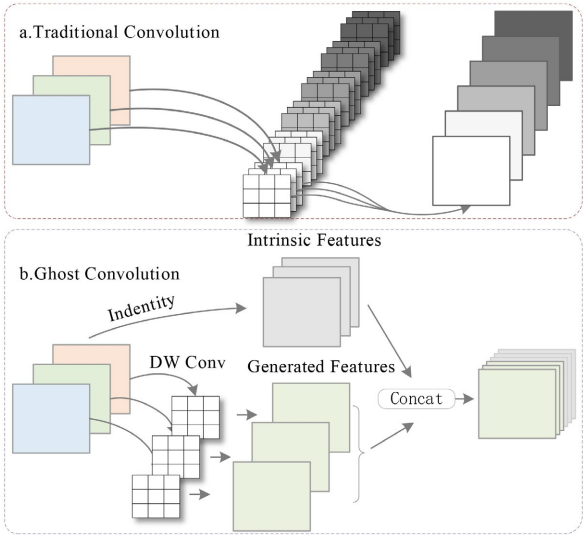

Figure 1. Comparison between Ghost convolution and standard convolution
图 1. Ghost 卷积与传统卷积的对比图

本文针对上述挑战，提出一种基于空间信息增强的双分支实时语义分割网络(ESINet)，通过引入轻量级的混合注意力模块和空间优化卷积模块，在模型性能和实时性方面取得了较好的提升，具体情况如图2所示。具体而言，本研究的主要贡献包括：

1) 提出了空间优化卷积模块(SpaceOptiConv)，通过线性计算获得部分生成特征图并与本征特征进行拼接，显著降低计算复杂度和参数量。

2) 设计轻量级混合注意力模块(HMA)，通过通道注意力和改进的多尺度空间注意力的结合，增强特征图的空间细节表达能力。

3) 构建复合损失函数方法，利用 IoULoss 的全局优化能力和 OhemCELoss 在局部细节的挖掘能力，提升难样本的学习能力和分割边缘的精度。

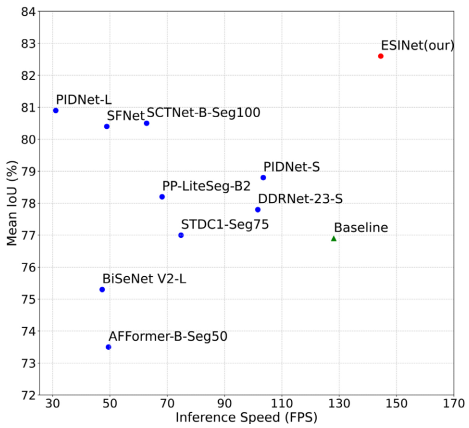

Figure 2. Comparison of speed and accuracy with existing models on the Cityscapes dataset
图 2. Cityscapes 数据集上与现有模型速度和精度的对比图

2. 相关工作

2.1. 通用语义分割

全卷积网络(Fully Convolutional Network, FCN) [17]提出以来，语义分割领域取得显著的进展。DeepLab [20]系列采用了具有不同膨胀率的空洞卷积来扩大感受野。PSPNet [18]通过金字塔池化模块以捕获多层次的语义特征。SegNet [3]通过对称的编码器–解码器结构，实现编码器为解码器提供池化索引。HRNet [2]利用多分辨率架构和跳连接实现多尺度特征融合。然而，这些方法通常依赖深层网络和复杂连接策略，导致高计算成本，难以应用于实时场景。

2.2. 实时语义分割

为了满足实时性需求，研究者提出了多种高效的网络架构。Enet [5]使用小分辨率的图像输入；BiSeNet [6]通过分支设计实现语义与细节特征的分离；DDRNet [9]通过早期共享分支降低，中期采用跳连接级联不同分辨率分支；PIDNet [10]增加边缘引导分支配合细节分支和语义分支。部分单分支架构采用监督策略，STDCSeg [8]通过采用短时稠密连接网络，实现高分辨率特征对低分辨率特征的监督；SCTNet [14]则设计了单独的监督分支。此外，轻量化的 Transformer 架构开始受到青睐，例如 RTFormer [13]、SegVit [4]等。然而，跳连接、增加分支以及堆叠 Transformer 层的往往会带来更高的计算量。

2.3. 注意力机制

注意力机制已成为提升语义分割模型性能的重要技术手段。其中，外部注意力(EA) [19]通过引入两个轻量级参数，仅使用两个线性层和归一化层，将自注意力的计算复杂度从 O(n2)降低到 O(n)。GPU 友好注意力机制(GFA) [13]模块通过优化矩阵乘法操作，显著提升了其在 GPU 上的并行处理效率。此外，注意力融合模块(UAFM) [12]、CBAM [15]和 CPCA [16]等方法，巧妙地结合通道信息与空间信息，设计出高效的注意力机制，有效增强了特征表达能力。

3. 方法

本文设计了双分支架构网络 ESINet，主要基于空间特征增强和注意力机制实现，目的在于提高双分支架构网络在分割速度和分割精度，并优化其在小目标检测和边缘细节上的不足。整体网络架构采用语义分支和细节分支，融合层聚合双分支特征。本节，首先介绍整体网络架构，然后探讨 SpaceOptiConv 模块和 HAM 模块，最后讨论复合损失函数的设计。

3.1. ESINet 整体网络架构

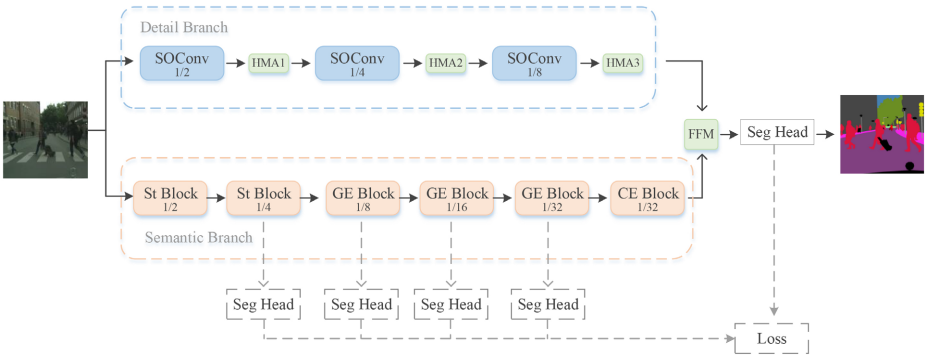

Figure 3. Overall network architecture of ESINet
图 3. ESINet 整体网络架构图

ESINet 采用双分支实时语义分割网络结构，通过增强空间信息的提取能力，结合高分辨率细节特征与低分辨率语义特征，提高图像分割的精度和推理速度。整体框架如图 3 所示，由细节分支、语义分支和特征融合三部分组成。其中，语义分支保持原结构不变，逐步下采样至 1/32。细节分支，由 SpaceOptiConv 模块构成，逐步将特征图的分辨率下采样到 1/2、1/4 和 1/8，每次下采样后的特征图传入 HMA 模块进一步增强空间细节描述。高分辨率细节特征和低分辨率语义特征，通过特征融合模块进行加权融合。融合后的特征图通过分割头处理，生成最终的分割结果。训练过程中，采用 5 个分割头参与损失函数计算。

3.2. 空间优化卷积模块

传统卷积操作虽然在提取特征方面表现出色，但特征图的维度往往在网络中逐渐增大，而许多高维特征图的生成可能包含了大量冗余信息，这些信息对最终输出没有显著贡献，直观表现为特征图的高度相似，直接造成了大量冗余计算。SpaceOptiConv 模块，通过分组卷积进行原始特征的提取，再采用深度可分离卷积生成少量生成特征图，并与原始特征进行通道拼接，而不是像传统卷积那样生成大量高维特征图，从而避免了冗余计算。同时，由于本征特征图监督了下一阶段的计算，进一步提升了空间特征的表达能力，具体模块设计如图 4 所示。

3.2.1. SpaceOptiConv 模块设计

为充分提取空间信息的同时减少模型的计算量，采用三个阶段的设计。第一阶段，采用分组卷积提取本征特征，从输入特征图 $X \in R^{c\times h\times w}$，通过轻量化的分组卷积生成 m 个本征特征图 $Y_0 \in R^{m\times h\times w}$，其过程可以表述为：

$$Y_0 = \text{GroupConv}_{s=1,g=C/2}(X) \tag{1}$$

其中，s 表示步幅，g 表示分组数，c 为输入通道数；第二阶段采用深度可分离卷积生成 m 个生成特征图，再与本征特征图拼接产生充足的特征图 $Y_0 \in R^{m\times h\times w}$，其过程可以表述为：

$$Y_1 = \text{Concat}\left(\text{DWConv}_{s=1}(Y_0),\ Y\right)_0 \tag{2}$$

第三阶段，采用分组卷积对拼接后的特征图进行下采样。输出特征图 $Y \in R^{n\times h'\times w'}$，其过程可以表述为：

$$Y = \text{GroupConv}_{s=2,g=m/2}(Y_1) \tag{3}$$

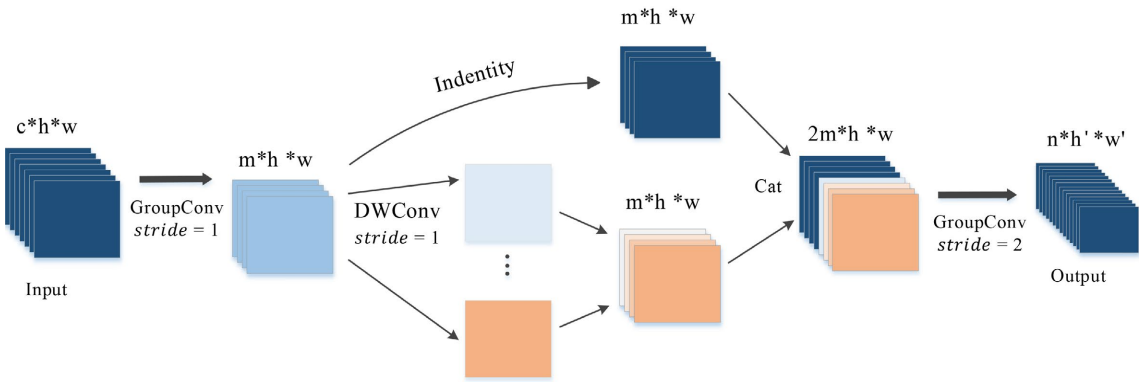

Figure 4. Structural diagram of the SpaceOptiConv module
图 4. SpaceOptiConv 模块结构图

3.2.2. SpaceOptiConv 模块复杂度分析

由于提取网络仅采用了低运算量的分组卷积和深度可分离卷积，计算量取得显著降低。与 Baseline

下采样模块对比(如表 1 所示)，SpaceOptiConv 模块的加速比为 r_s；参数比 r_c，如表 1 所示：

$$r_s = \frac{\left(k^2 \times c \times \frac{chw}{8}\right) + \left(k^2 \times \frac{c}{2} \times \frac{chw}{8}\right) + 1 \times 1 \times \frac{c}{2} \times \frac{hw}{16}}{\left(k^2 \times c \times mhw\right) + \left(k^2 \times mhw\right) + \left(k^2 \times 2m \times \frac{hw}{4} \times n\right)} \approx \frac{c}{m} \tag{4}$$

其中 $k=3$， m 表示分组卷积所产生的滤波器数量，并且 $m \ll c$。同样，可以计算为：

$$r_c = \frac{(k^2 \times c \times n) + \left(k^2 \times n \times n\right) + \left(1 \times 1 \times n \times n\right)}{\left(k^2 \times c \times m\right) + \left(k^2 \times m\right) + \left(k^2 \times 2m \times n\right)} \approx \frac{n}{m} \tag{5}$$

Table 1. Complexity analysis of SpaceOptiConv module
表 1. SpaceOptiConv 模块复杂度分析

	Baseline	SpaceOptiConv
S_1	3 × 3 Conv	3×3 GroupConv
S_2	3 × 3 Conv	3×3 DWConv
S_3	1 × 1 Conv	3×3 GroupConv
r_s	1	≈ m/c
r_c	1	≈ m/n

3.3. 多尺度混合注意力(HMA)

现有的混合注意力方法结合了通道注意力和空间注意力，但在通道之间表现出均匀的空间注意力权重分布，导致了大量的噪声。然而，不同通道在特征表示中具有不同的重要性。为捕捉这些差异，在空间注意力中采用多尺度特征提取，从而关注不同尺寸的空间细节。基于此思想本文提出 HMA 注意力机制。

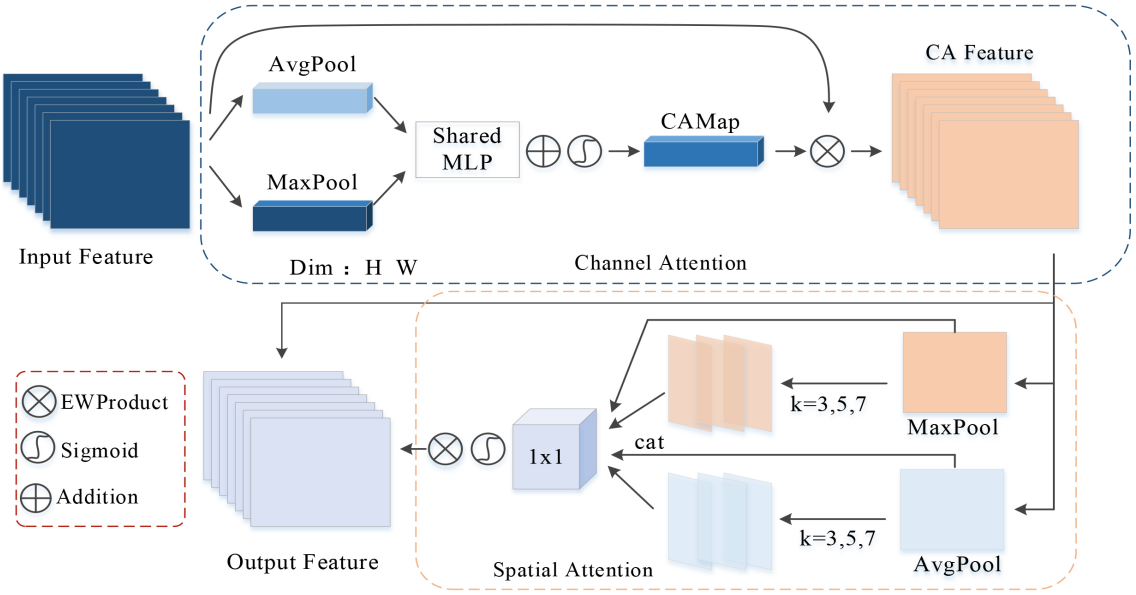

Figure 5. Multiscale Hybrid Attention (HMA) module
图 5. 多尺度混合注意力模块

3.3.1. 多尺度混合注意力架构

多尺度混合注意力模块依次执行通道注意力和空间注意力。如图 5 所示，输入特征图 $F_{in} \in R^{c\times h\times w}$，先进行通道注意力计算产生 1D 特征图 $M_c \in R^{c\times 1\times 1}$，再与本征特征逐元素相乘输出 $F_c \in R^{c\times h\times w}$，然后进行空间注意力计算产生空间注意力图 $M_s \in R^{m\times h\times w}$，再与输入 F_c 逐元素相乘，最终输出特征图 $F_{out} \in R^{c\times h\times w}$。整个过程总结为：

$$F_c = M_c\left(F_{in}\right)\otimes F_{in} \tag{6}$$

$$F_{out} = M_s\left(F_c\right)\otimes F_c \tag{7}$$

3.3.2. 通道注意力

在空间维度(H, W)使用全局平均池化和全局最大池化聚合空间信息，生成两个不同的空间上下文描述符；$F_{c\_avg}, F_{c\_max} \in R^{1\times 1\times c}$。再将描述符传入共享参数的感知机网络，并进行相加，生成通道注意力图 $M_c \in R^{c\times 1\times 1}$。最终注意力权重与原始特征图按元素相乘。为保持高效的推理速度，共享网络只有一个大小为 $R^{c/r\times 1\times 1}$ 的隐层。这一过程可以表述为：

$$F_{c\_avg} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}F_{in}\left(i,j\right) \tag{8}$$

$$F_{c\_max} = max_{i=1}^{H}max_{j=1}^{W}F\left(i,j\right) \tag{9}$$

$$\begin{aligned} M_c &= \sigma\left(\mathrm{MLP}\left(F_{c\_avg}\right)+\mathrm{MLP}\left(F_{c\_max}\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F_{c\_avg}\right)\right)+W_1\left(W_0\left(F_{c\_max}\right)\right)\right) \end{aligned} \tag{10}$$

其中，σ 表示 sigmoid 激活函数，MLP 共享参数 $W_0 \in R^{c/r\times c}, W_1 \in R^{c\times c/r}$，$r$ 表示隐层缩减比。

3.3.3. 空间注意力

空间注意力目的在于提取空间映射关系。通道注意力加权后的特征图 $F_c \in R^{c\times h\times w}$，对通道维度(C)分别进行全局平均池化和最大池化操作。采用多尺度卷积核捕捉注意力图，然后使用 1×1 卷积进行通道混合，生成空间注意力图 $M_s \in R^{m\times h\times w}$，该过程表述为：

$$F_{s\_avg} = AvgPool\left(x\right) = \frac{1}{c}\sum_{i=1}^{h}\sum_{j=1}^{w}x_{ij} \tag{11}$$

$$F_{s\_max} = MaxPool\left(x\right) = max_{i=1,\,j=1}^{h,w}x_{ij} \tag{12}$$

$$F_{concat} = Concat\left(F_{s\_max}^{k}; F_{s\_avg}^{k}\right) \tag{13}$$

$$M_s\left(F_c\right) = \sigma\left(Conv_{1\times 1}\left(F_{concat}\right)\right) \tag{14}$$

其中，σ 表示 sigmoid 函数，F^k 表示卷积核为 k 的多尺度卷积操作，$k = \{3,5,7\}$。

3.4. 损失函数设计

设计复合损失函数用于训练网络，该损失函数由主分支和多个辅助分支共同组成。辅助分支使用 OhemCELoss，增强难样本的学习能力，主分支使用 IoULoss，对分割边缘学习优化。OhemCELoss 是一种改进的交叉熵损失函数，旨在通过关注难以分类的样本来优化模型。首先，定义交叉熵损失函数：

$$\mathcal{L}_{CE}\left(p_i, y_i\right) = -\sum_{c=1}^{C}y_{i,c}\log\left(p_{i,c}\right) \tag{15}$$

其中预测输出样本为 p_i、y_i 分别表示第 i 个样本的预测概率分布和真实标签。然后筛选出难样本集合，

难样本的定义为 $\mathcal{L}_{CE}\left(p_i, y_i\right) > \tau$ 的样本，τ 表示难样本设定的阈值(实验中 $\tau = 0.7$)。再计算难样本集合的平均交叉熵损失：

$$\mathcal{L}_{hard} = \left\{\mathcal{L}_{CE}\left(p_i, y_i\right) | \mathcal{L}_{CE}\left(p_i, y_i\right) > \tau\right\} \tag{16}$$

$$\mathcal{L}_{OhemCE} = \frac{1}{\left|\mathcal{L}_{hard}\right|} \sum_{i \in \mathcal{L}_{hard}} \mathcal{L}_{CE}\left(p_i, y_i\right) \tag{17}$$

IoULoss 用于评估预测分割区域与真实分割区域的重叠程度。其计算过程如下：

$$\mathrm{IoU} = \frac{|P \cap Y|}{|P \cup Y|} = \frac{\sum_{i=1}^{j}\left(p_i \cdot y_i\right)}{\sum_{i=1}^{j}\left(p_i + y_i - p_i \cdot y_i\right)} \tag{18}$$

$$\mathcal{L}_{iou} = 1 - \mathrm{IoU} \tag{19}$$

其中 P 、Y 分别表示预测分割区域和真实分割区域。最终复合损失函数：

$$l_{pre} = \mathcal{L}_{iou}\left(p_{pre}, y\right) \tag{20}$$

$$l_{aux} = \sum_{i=1}^{k} \mathcal{L}_{OhemCE}\left(p_{aux,i}, y\right) \tag{21}$$

$$l_{total} = l_{pre} + l_{aux} \tag{22}$$

其中，k 表示辅助分割头的数量。

4. 实验

4.1. 数据集

Cityscapes 是一个广泛认可的城市驾驶场景的语义分割数据集。该数据集包含 5,000 张精细标注的图像，划分为训练集(2975 张)、验证集(500 张)和测试集(1525 张)。为确保与先前研究的一致性和结果的可比性，专注于 19 个类别的语义分割。Cityscapes 数据集的图像分辨率为 2048 × 1024 像素。

COCO-Stuff10K 数据集包含 10,000 张图像，提供详细的像素级标注，适用于语义分割和目标检测等任务。该数据集涵盖 80 个实体类别(thing)和 91 个背景类别(stuff)，总共 171 个类别。图像分辨率通常为 640 × 480 像素。另外增加背景元素的标注，增强了场景复杂性，有助于提高模型的泛化性能。

ADE20K 数据集是一个用于场景解析的大规模数据集，包含超过 25,000 张图像。这些图像被划分为训练集(20,210 张)、验证集(2000 张)和测试集(3000 张)。数据集涵盖 150 个类别，包括离散对象和背景区域。每张图像都经过详细的像素级标注，适用于验证语义分割精确度和泛化性能。

4.2. 训练

模型训练采用从零开始的方法，初始化采用“kaiming normal”方法。训练算法为随机梯度下降(SGD)，并设置动量为 0.9。不同数据集的迭代次数分别为 Cityscapes 为 150k，COCO-Stuff 和 ADE20K 为 180 k。Cityscapes 权重衰减为 0.0005，其余数据集为 0.0001，且权重衰减仅适用于卷积层参数。学习率初始值设为 5×10^{-2}，并采用“poly”策略调整，每次迭代后学习率乘以 $\left(1 - iter / iter_{max}^{0.9}\right)$。数据增强技术包括随机水平翻转、随机缩放和随机裁剪。

4.3. 推理

模型推理过程中，Cityscapes 减至 1024 × 512 分辨率进行推理，推理结果扩展回原始尺寸 2048 × 1024，

调整所需时间包含在推理时间的测量中。推理测试在单个 GPU 上完成的，采用了平均交并比(mIoU)作为评估标准。

实验基于 PyTorch2.0.1，推理时间的测量 NVIDIA TITAN RTX 上进行配备了 CUDA11.7、CUDNN8.0。

4.4. 消融研究

为验证 SpaceOptiConv 模块、HMA 机制和复合损失函数设计的有效性，在 Cityscapes 数据集上进行了消融实验，结果如表 2 所示。

在基线模型(Baseline)中，保持细节分支的相关配置，并逐步在空间分支中引入 SpaceOptiConv 模块作为主干网络，同时在细节分支中进一步引入注意力机制。与 Baseline 模型相比，当空间分支采用 SpaceOptiConv 模块时，mIoU 和 FPS 分别提升了 1.5%和 29.7 fps。最终，通过结合 SpaceOptiConv 模块、HMA 机制和复合损失函数，mIoU 和 FPS 分别取得了 5.6%和 16.4 fps 的提升。

Table 2. Effectiveness study of each module
表 2. 模块有效性研究

模块	HMA	SpaceOptiConv	IoULoss	mIoU (%)	#FPS
Baseline				76.9	128.1
Baseline	✓			77.6	107.7
Baseline-Detail Branch		✓		78.4	157.8
Baseline-Detail Branch	✓	✓		78.9	144.5
Baseline-Loss			✓	77.8	-
ESINet (our)	✓	✓	✓	82.5	144.5

其中“–”表示去除相应模块，“✓”表示添加相应模块。

4.4.1. 多尺度混合注意力机制有效性研究

为进一步验证轻量化混合注意力模块(HMA)对性能的影响。在 Baseline 和 Baseline + SpaceOptiConv 模型的基础上逐步添加 HMA1、HMA2、HAM3 以增强网络的特征捕获特征，最终 mIoU 分别提高了 0.5%和 2%，具体结果如表 2 和表 3 所示。

Table 3. Ablation study of HMA based on the baseline model
表 3. HMA 在 Baseline 模型基础上的消融研究

空间分支			mIoU (%)
HMA1	HMA2	HMA3	
			76.9
✓			77.3
✓	✓		77.4
✓	✓	✓	77.6

4.4.2. 复合损失函数有效性研究

为了增强训练，本文提出的方法引入 5 个分割头参与损失计算，其中 $l_0 \sim l_3$ 采用 OhemCELoss，主干采用 IoULoss。在其他配置不变的情况下，逐步添加分割头(IoULoss, $l_0 \sim l_3$)参与损失计算。最终 mIoU 取得了 0.9%的提升，说明复合损失函数的设计对于分割精度的提升有明显的效果，具体结果如表 4 所示。

Table 4. Ablation study of composite loss function based on the baseline model
表 4. 复合损失函数在 Baseline 模型上的消融研究

OhemCELoss				IoULoss	mIoU (%)
l_0	l_1	l_2	l_3		
				✓	76.9
			✓	✓	76.8
	✓	✓	✓	✓	77.4
✓	✓	✓	✓	✓	77.8

5. 结果和比较

在本章中，我们详细展示了 ESINet 在多个主流语义分割数据集上的性能表现，重点分析其与现有最优方法在精度和速度上的比较结果。

5.1. Cityscapes 的结果比较

ESINet 的方法在 Cityscapes 测试集上在 Torch 实现中实现了最佳的速度-精度权衡。评估时采用的分辨率为 2048 × 1024。在相同的环境下，测试集(Test)上取得了 Troch 推理速度和 mIoU 分别为 144.5 fps 和 81.6%，相较于基线模型精度测试集取得了 4.9%的提升，同时推理速度提升了 16.4 fps，详细结果如表 5 所示，标有*的模型为在本地平台上进行部署测试得到的推理速度。

Table 5. Accuracy and speed on Cityscapes
表 5. Cityscapes 上的精度与速度

模型	mIoU		#FPS	分辨率	GPU
	Val	Test			
Baseline(BisenetV2)*	**76.9**	**76.7**	**128.1**	1024 × 512	RTX TITAN
STDC1-Seg75	77.0	76.8	74.8	1536 × 768	RTX 3090
AFFormer-B-Seg50	73.5	-	49.5	1024 × 512	RTX 3090
PP-LiteSeg-B2	78.2	77.5	68.2	1536 × 768	RTX 3090
RTFormer-S	76.3	-	110.0	512 × 2048	RTX 2080Ti
DDRNet-23-S	77.8	77.4	101.6	2048 × 1024	GTX 2080Ti
SCTNet-B-Seg100	80.5	-	62.8	2048 × 1024	RTX 3090
SFNet-R18	-	80.4	48.9	1024 × 2048	RTX 3090
BisenetV3	78.8	79.0	93.8	768 × 1024	GTX 1080Ti
PIDNet-S*	78.8	78.6	103.5	2048 × 1024	RTX TITAN
PIDNet-L*	80.9	80.1	31.1	2048 × 1024	RTX TITAN
ESINet(our)	**82.5**	**81.6**	**144.5**	1024 × 512	RTX TITAN

5.2. COCO-Stuff-10K 的结果比较

COCO-Stuff-10K 上的相应结果如表 3 所示。ESINet 以较高正确率，在实时语义分割方法中 COCO-Stuff-10K 上的最高推理速度。在输入大小为 640 × 640 的情况下，ESINet 在 169.8FPS 实现了 34.9%的 mIoU，比基线模型提高 3.2%，详细结果如表 6 所示。

Table 6. Comparison with other models on the COCO-Stuff-10K dataset
表 6. 在 COCO-Stuff-10K 数据集上的模型对比

模型	mIoU (%)↑	#FPS↑
Baseline *	31.7	138.1
SeaFormer-B	34.1	41.9

续表

AFFormer-B	35.1	46.5
RTFormer-B	35.3	90.9
SCTNet-B	35.9	141.5
ESINet (our)	34.9	156.3

5.3. ADE20K 的结果

在 ADE20K 数据集，ESINet 在基线模型的基础上取得 2.4%提升，并在众多模型中实现了最快的分割速度。考虑到 ADE20K 中的图像数量和大量的语义类别，这结果也进一步证明了 ESINet 具备良好的泛化能力，详细结果如表 7 所示。

Table 7. Comparison with state-of-the-art models on the ADE20K dataset
表 7. 在 ADE20K 数据集上与其他最先进模型的对比

模型	mIoU (%)↑	#FPS↑
Baseline*	32.5	158.1
SeaFormer-B	34.1	44.5
AFFormer-B	41.8	49.6
RTFormer-B	42.1	93.4
SCTNet-B	43.0	145.1
ESINet(our)	34.9	169.8

5.4. 结果可视化

图 6 展示了本文所提出的 ESINet 在不同场景下的语义分割效果。图中从左至右依次为原始图像、BiseNet*分割结果、本文方法分割结果及真实标签。可以看出，ESINet 在小目标识别、边界处理和复杂场景分割等方面均优于基线模型。第一行为细节结构(如围栏)分割效果；第二行为人群目标的识别结果；第三与第四行体现模型对空间上下文的建模能力；最后一行展示其在复杂交通场景中的分割表现。

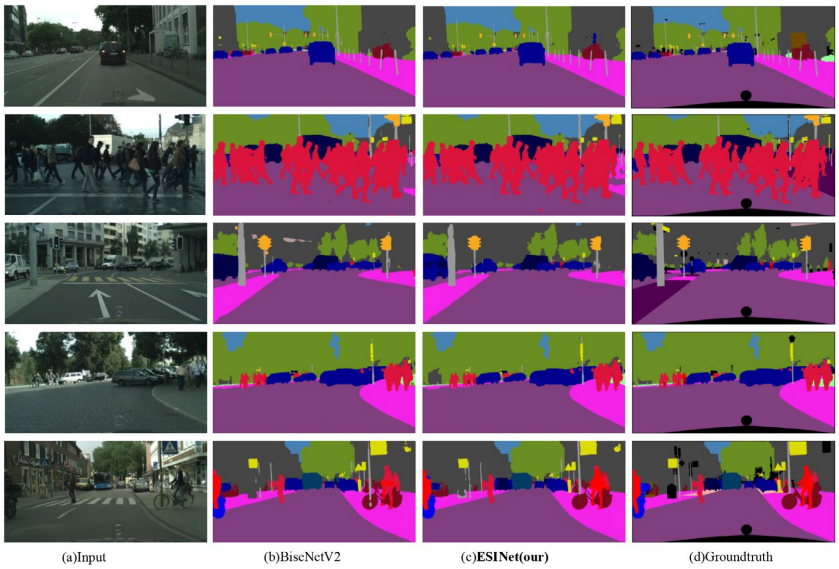

Figure 6. Visualization examples on the Cityscapes dataset
图 6. Cityscapes 数据集的可视化示例图

6. 总结

本文提出了一种双分支实时语义分割模型——ESINet，旨在解决现有语义分割方法在准确性和实时性方面的不足。为提高模型的分割精度和推理速度，ESINet 引入了改进的空间优化卷积模块(SpaceOptiConv)，通过少量生成特征图并与本征特征拼接，减少了冗余计算，降低了计算复杂度和参数量。同时，模型基于混合注意力机制设计了多尺度特征提取模块(HMA)，有效捕捉不同尺度的特征信息，增强了空间细节表达能力。此外，ESINet 还提出了一种复合损失函数，结合了交并比损失(IoULoss)和在线难例挖掘交叉熵损失(OhemCELoss)，提升了模型对难样本的学习能力，优化了分割边界的准确性。在 Cityscapes 数据集上的实验结果表明，分割精度和实时性均得到了显著提升。

参考文献

[1] Cheng, B., Schwing, A. and Kirillov, A. (2021) Per-Pixel Classification Is Not All You Need for Semantic Segmentation. *Advances in Neural Information Processing Systems*, **34**, 17864-17875.

[2] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., *et al.* (2021) Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 3349-3364. https://doi.org/10.1109/tpami.2020.2983686

[3] Badrinarayanan, V., Kendall, A. and Cipolla, R. (2017) SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 2481-2495. https://doi.org/10.1109/tpami.2016.2644615

[4] Zhang, B., Tian, Z., Tang, Q., *et al.* (2022) Segvit: Semantic Segmentation with Plain Vision Transformers. *Advances in Neural Information Processing Systems*, **35**, 4971-4982.

[5] Paszke, A., Chaurasia, A., Kim, S., *et al.* (2016) Enet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. arXiv: 1606.02147.

[6] Tsai, T. and Tseng, Y. (2023) BiSeNet V3: Bilateral Segmentation Network with Coordinate Attention for Real-Time Semantic Segmentation. *Neurocomputing*, **532**, 33-42. https://doi.org/10.1016/j.neucom.2023.02.025

[7] Yu, C., Gao, C., Wang, J., Yu, G., Shen, C. and Sang, N. (2021) BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *International Journal of Computer Vision*, **129**, 3051-3068. https://doi.org/10.1007/s11263-021-01515-2

[8] Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., *et al.* (2021) Rethinking BiSeNet for Real-Time Semantic Segmentation. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 9711-9720. https://doi.org/10.1109/cvpr46437.2021.00959

[9] Pan, H., Hong, Y., Sun, W. and Jia, Y. (2023) Deep Dual-Resolution Networks for Real-Time and Accurate Semantic Segmentation of Traffic Scenes. *IEEE Transactions on Intelligent Transportation Systems*, **24**, 3448-3460. https://doi.org/10.1109/tits.2022.3228042

[10] Xu, J., Xiong, Z. and Bhattacharyya, S.P. (2023) PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 19529-19539. https://doi.org/10.1109/cvpr52729.2023.01871

[11] Tang, Y., Han, K., Guo, J., *et al.* (2022) GhostNetv2: Enhance Cheap Operation with Long-Range Attention. *Advances in Neural Information Processing Systems*, **35**, 9969-9982.

[12] Peng, B., Liu, Y., Zhu, X., Ikeda, S. and Tsunoda, S. (2022) Femoral Segmentation of MRI Images Using PP-LiteSeg. 2022 *IEEE-EMBS International Conference on Biomedical and Health Informatics* (*BHI*), Ioannina, 27-30 September 2022, 1-4. https://doi.org/10.1109/bhi56158.2022.9926879

[13] Lu, C., de Geus, D. and Dubbelman, G. (2023) Content-Aware Token Sharing for Efficient Semantic Segmentation with Vision Transformers. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 23631-23640. https://doi.org/10.1109/cvpr52729.2023.02263

[14] Xu, Z., Wu, D., Yu, C., Chu, X., Sang, N. and Gao, C. (2024) SCTNet: Single-Branch CNN with Transformer Semantic Information for Real-Time Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 6378-6386. https://doi.org/10.1609/aaai.v38i6.28457

[15] Woo, S., Park, J., Lee, J. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, Springer, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[16] Huang, H., Chen, Z., Zou, Y., Lu, M., Chen, C., Song, Y., *et al.* (2024) Channel Prior Convolutional Attention for

Medical Image Segmentation. *Computers in Biology and Medicine*, **178**, 108784. https://doi.org/10.1016/j.compbiomed.2024.108784

[17] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/cvpr.2015.7298965

[18] Yan, L., Liu, D., Xiang, Q., Luo, Y., Wang, T., Wu, D., *et al.* (2021) PSP Net-Based Automatic Segmentation Network Model for Prostate Magnetic Resonance Imaging. *Computer Methods and Programs in Biomedicine*, **207**, Article ID: 106211. https://doi.org/10.1016/j.cmpb.2021.106211

[19] Chua, L., Jimenez-Diaz, J., Lewthwaite, R., Kim, T. and Wulf, G. (2021) Superiority of External Attentional Focus for Motor Performance and Learning: Systematic Reviews and Meta-Analyses. *Psychological Bulletin*, **147**, 618-645. https://doi.org/10.1037/bul0000335

[20] Chen, L.C. (2017) Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv: 1706.05587.