激光与光电子学进展 先进成像

# 多尺度特征融合的道路场景语义分割

易清明，张文婷，石敏，沈佳林，骆爱文*
暨南大学信息科学技术学院，广东 广州 510632

**摘要** 针对现有语义分割网络模型难以在参数量、推理速度和精确度中取得平衡的问题，设计了一种多尺度特征信息融合的轻量级网络模型(MIFNet)。MIFNet采用编码-解码结构，在编码部分利用分离策略和非对称卷积设计了轻量型特征提取瓶颈结构，且引入空间注意力机制与Laplace边缘检测算子组成边缘-空间融合模块，将空间信息和边缘信息进行融合得到丰富的特征信息。在解码部分引入通道注意力机制恢复特征图尺寸和细节信息完成语义分割。在Cityscapes和CamVid测试集上，MIFNet仅以0.82 M的参数量分别取得了73.1%和67.7%的分割精度，同时在单个GTX 1080Ti GPU下分别获得73.68 frame/s和85.16 frame/s的推理速度，表明该方法在参数量、推理速度和精确度3个指标上得到较好平衡，实现了轻量、快速、精准的语义分割。

**关键词** 图像处理；实时语义分割；Laplace边缘检测；注意力机制；多尺度特征信息融合
**中图分类号** TP391.4 **文献标志码** A **DOI**：10.3788/LOP220914

# Semantic Segmentation for Road Scene Based on Multiscale Feature Fusion

Yi Qingming, Zhang Wenting, Shi Min, Shen Jialin, Luo Aiwen*
*College of Information Science and Technology, Jinan University, Guangzhou* 510632, *Guangdong, China*

**Abstract** A lightweight network model based on multiscale feature information fusion (MIFNet) is developed in this study owing to the imbalance among the parameter amount, inference speed, and accuracy in many existing semantic segmentation network models. The MIFNet is constructed on the encoding-decoding architecture. In the encoding part, the split strategy and asymmetric convolution are flexibly applied to design lightweight bottleneck structure for feature extraction. The spatial attention mechanism and Laplace edge detection operator are introduced to fuse spatial and edge information to obtain rich feature information. In the decoding part, a new decoder is designed by introducing a channel attention mechanism to recover the size and detail information of the feature map for a complete semantic segmentation task. The MIFNet achieves accuracies of 73.1% and 67.7% on the Cityscapes and CamVid test sets, respectively, with only approximately 0.82 M parameters. Correspondingly, it reaches up to 73.68 frame/s and 85.16 frame/s inference speed, respectively using a single GTX 1080Ti GPU. The results show that the method achieves a good balance in terms of the parameter amount, inference speed, and accuracy, yielding a lightweight, fast, and accurate semantic segmentation.

**Key words** image processing; real-time semantic segmentation; Laplace edge detection; attention mechanism; multiscale feature information fusion

## 1 引 言

图像语义分割技术在实践中得到了广泛的应用，尤其是在无人机着陆系统和无人驾驶汽车等场景中发挥着至关重要的作用。基于深度神经网络的语义分割模型具有较高的分割精度而被广泛研究[1-2]，但庞大的参数量和缓慢的推理速度阻碍了其在实际场景中的应用推广。其中，在以自动驾驶为代表的重要应用中，其对道路场景分割的准确性和时效性均提出较高要求，因此，亟须研究出一种能够在计算响应速度与分割精度之间实现较好均衡的道路场景语义分割方法。

为了解决现有语义分割模型中参数规模较大和推理速度慢等问题，许多基于“编码-解码”结构的轻量级语义分割网络[3-8]被提出。其中，编码器主要用于完成

**收稿日期**：2022-03-07；**修回日期**：2022-04-02；**录用日期**：2022-06-13；**网络首发日期**：2022-06-23
**基金项目**：国家自然科学基金(62002134)、广东省基础与应用基础研究基金(2020A1515110645)、广东省重点实验室项目(2021KSY001)、羊城创新创业领军人才支持计划(2019019)、暨南大学中央高校基本科研业务费项目(21620353)
**通信作者**：*luoaiwen@jnu.edu.cn

对图像特征的提取，解码器用于对编码后的图像进行空间细节及其相应空间维度的恢复。由于编码器与解码器之间的信息直接连接，保证语义分割网络中的目标细节信息得到更好的恢复[3]。“编码-解码”结构的主要优点是能够有效地降低网络模型的参数量。例如，ENet[6]通过使用较少的3×3标准卷积构造一个轻量级网络，参数量仅有0.36 M，网络模型的参数规模大大减少，但精确度和推理速度效果不够理想。因此一些网络在编码部分对瓶颈结构进行优化，构建轻量化骨干网络提取特征信息。如ERFNet[7]使用因式分解卷积1×3和3×1即非对称卷积替代3×3标准卷积，设计出带有空洞卷积的Non-bottleneck-1D结构。DABNet[4]在ERFNet的启发下提出了深度非对称瓶颈(DAB)模块，其仅采用了编码器主干网，舍弃了解码器，整体模型产生了0.76 M参数；加入像素级集成解码器(PAD)解码模块[5]后DABNet能以较小的计算代价有效地细化编码后的特征图，分割精度提升了1.1个百分点，这表明一个精心设计的解码器对于提高模型的性能具有很大的作用。Wang等[8]提出使用分离(split)策略的瓶颈模块SS-nbt module并将其堆叠成骨干网络，模型参数量少于1 M但精度相对DABNet有微小提升。然而，Non-bottleneck-1D和SS-nbt module使用多个非对称卷积，相比于标准卷积，非对称卷积在一定程度上损失了特征信息，导致准确率的下降；而DAB模块仅使用一次跳连接对于深层特征图的局部信息和上下文空间细节信息的融合表现不佳。因此，本文对瓶颈结构进行优化，结合非对称卷积和标准卷积实现特征信息提取，并且还进一步引入多级跳线连接(concat)补充多尺度特征信息，在保证参数量较小的同时也能够减少信息的丢失。

由于单分支网络在对分辨率较大的输入特征图进行推理时会损失掉一些信息，同时也使速度变得缓慢。针对这一问题，Fast-SCNN[9]和ContextNet[10]通过使用双分支结构分别处理低分辨率和高分辨率的图片，同时提取不同分辨率的特征信息，对网络进行压缩，实现了快速的语义分割。但是这些双分支结构的空间分支太浅，引入了较多的语义噪声，浅层中高分辨率的边缘空间信息还未提取足够多就进行信息融合，导致网络性能降低，因此需要对现有的双分支结构进行优化，以获得更丰富的特征信息。

此外，大量研究表明，注意力机制能够细化特征信息来引导网络进行更好的推理。SENet[11]和EKENet[12]通过利用注意力机制提供语义引导空间信息，引入注意力机制对通道之间的关系进行建模，让特征图进行自校正，快速恢复高分辨率的特征图。同时在解码部分，Peng等[13]和LRNNet[14]将注意力机制引入到网络的解码器中，构建轻量级解码器来减少模型参数，细化深层次特征图的语义信息和低层次特征图的边缘空间信息。因此，在受到注意力机制作用的启发下，本文在设计融合模块以及解码模块时引入注意力机制对不同层次的信息进行细化，进而提高特征信息提取能力。

综上考虑，本文设计了一个多尺度信息融合网络(MIFNet)模型。实验结果表明，MIFNet语义分割模型在参数量、推理速度和分割精度3个重要指标上取得了较好的平衡。

## 2　多尺度信息融合网络

### 2.1　总体框架

本文提出的MIFNet网络模型的整体框架如图1所示。MIFNet主要由两大部分组成：编码部分和解码部分。编码部分主要包括初始卷积模块、下采样模块、轻量级特征提取瓶颈(LFE-B)结构和边缘空间融合(ESF)模块；解码部分的核心构成为多尺度注意力融合解码器(MAFD)。

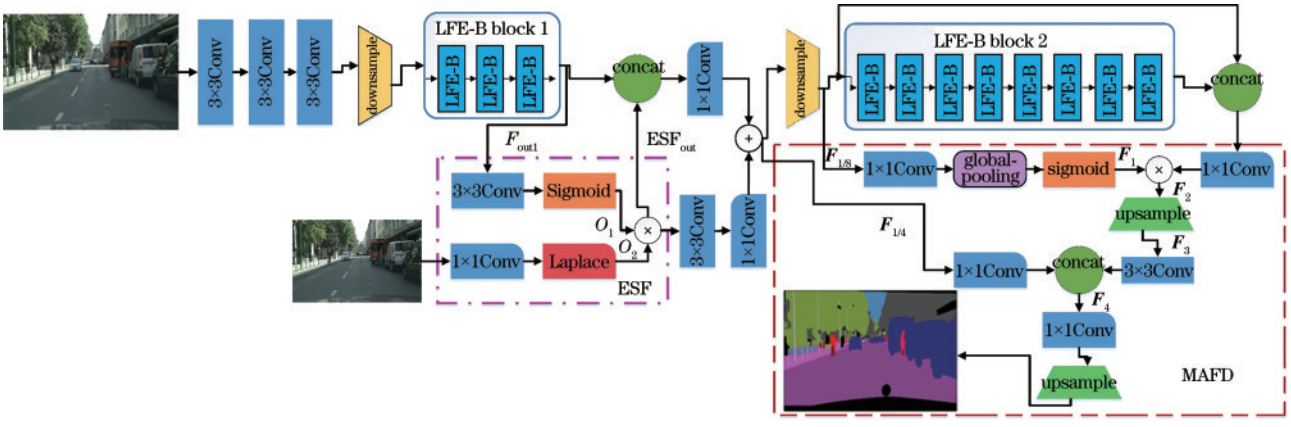



图1　MIFNet整体结构，虚线方框内是模型主要的三个模块

Fig. 1　Overall structure of MIFNet composed of three key modules marked by dashed boxes

在编码器中，初始卷积模块主要由3个3×3卷积组成，第一个3×3卷积步长设置为2，目的是改变原输入图像的大小；另外两个3×3标准卷积的步长为1，作用是对原始特征图进行信息提取。初始模块得到的输出特征图传输至下采样模块，其主要执行一个步长为2的3×3卷积和2×2的最大池化操作。下采样模块

根据式(1)和式(2)能够降低特征图的分辨率，同时在较低分辨率的情况下提高感受野，从而获得更丰富的图像特征信息。

$$F_s=\frac{(I_s+2\times P_s-k)}{s}+1, \tag{1}$$

式中：$F_s$表示特征图的尺寸；$I_s$表示输入图片的大小；$k$指卷积核的大小；$s$指步长；$P_s$指填充值的大小。当前层的感受野$r_n$可以通过前$n$-1层所有的步长$s_i$($i$=1, 2, …, $n$−1)与当前层(第$n$层)卷积核$k_n$的乘积，再与第$n$-1层的感受野$r_{n-1}$的叠加之后减去前$n$−1层所有的步长$s_i$($i$=1, 2, …, $n$−1)的乘积[15]得到，如下式所示：

$$r_n=r_{n-1}+(k_n-1)\prod_{i=1}^{n-1}s_i。 \tag{2}$$

对输入图像进行初始化和下采样处理之后，将信息传递至MIFNet模型的主干网络。MIFNet主干网络的核心组成包括：ESF特征融合模块，以及两个分别依次连接并用于提取不同深度图像特征的特征提取模块LFE-B block 1和LFE-B block 2。其中，每个特征提取模块LFE-B block分别由多个瓶颈结构LFE-B组成，作用是对降低分辨率后的特征图进行卷积操作提取特征信息，加强空间信息的交流，防止信息丢失；ESF融合模块用于对第一个特征提取模块LFE-B block 1输出的浅层空间特征信息以及由Laplace分支提取的边缘信息进行融合，通过融合图像在浅层网络中的不同特征信息，提高浅层特征的表达能力。编码器中的第二个特征提取模块LFE-B block 2经由下采样模块降低特征图分辨率之后，以较低的计算量和参数量提取小尺寸、深度的融合特征。

在解码部分，本文基于注意力机制设计了MAFD解码器，针对编码器输出的不同深度、不同尺度的融合特征，可以通过使用通道注意力机制来指导解码器进行多尺度的融合解码，恢复图像特征，从而达到图像目标对象标签准确分割的效果。

### 2.2 LFE-B结构

针对如图2(a)～2(c)所示的现有瓶颈模块所存在的特征信息丢失、信息间的融合不佳导致最终模型性能较差的问题，本文设计了一个新型的LFE-B结构，如图3所示，以解决网络退化的问题，保证一定的网络深度以提取到足够的特征信息，从而提高网络的性能。LFE-B模块通过对残差结构、非对称卷积以及split策略的灵活使用，提取多尺度上下文特征信息，能够在控制参数量的同时提升准确率。首先在输入端应用1×1卷积改变通道数，减少计算量。为了防止出现随着网络层数的增加而带来特征信息损失的问题，应用3×3标准卷积提取局部信息，接着使用split策略将输出深度降为$w$/4($w$表示输入通道数)，分别输入两个并行特征提取分支进行非对称卷积处理。其中，右分支应用1×3和3×1深度可分离卷积，能够降低模型参数并保证精度不会降低；左分支则使用带有空洞率的1×3和3×1深度可分离卷积，增加感受野的大小，使模型的深度和感受野得到平衡，进而得到更复杂的上下文空间细节信息。这两个分支通过使用1×3和3×1卷积来替代一个3×3标准卷积，在一定程度能够减少参数量，同时在左分支中加入空洞率，可以使得网络具有更大的感受野和更强的特征提取能力。紧接着使用短连接(shortcut)与局部信息拼接(concat)在一起，防止退化，保证能够很好地对网络进行优化。联合提取局部和上下文信息后再使用一个1×1卷积恢复通道数，最后引入残差结构与主干支路输出的特征信息

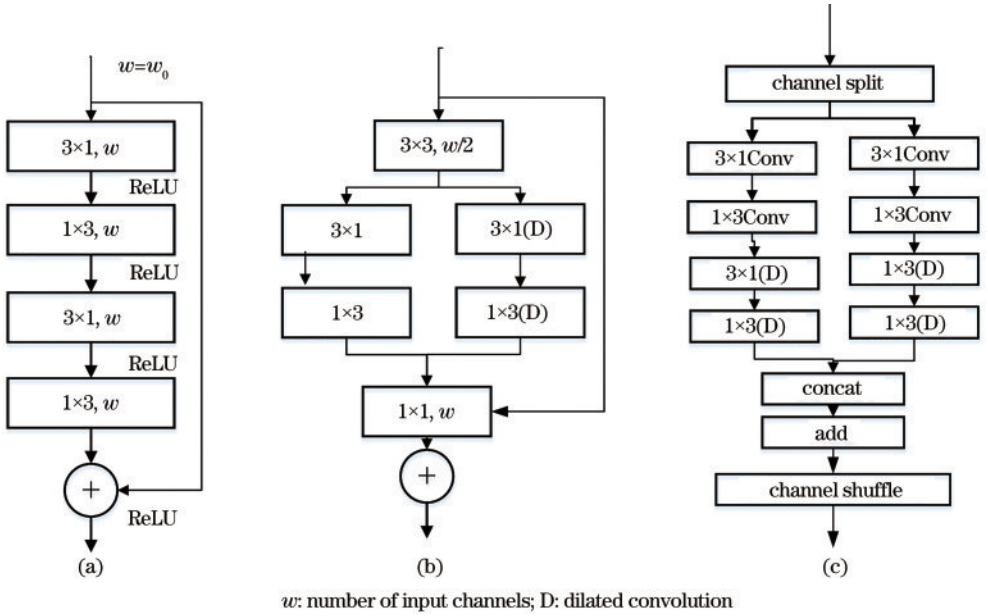
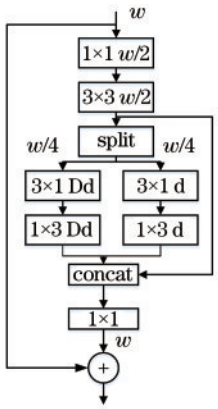



图2 不同瓶颈模块的对比。(a) Non-bottleneck-1D模块；(b) DAB模块；(c) SS-nbt模块

Fig. 2 Comparison of different bottleneck modules. (a) Non-bottleneck-1D; (b) DAB module; (c) SS-nbt module




*w*: number of input channels; Dd: depth-wise asymmetric convolution with dilated convolution; d: depth-wise asymmetric convolution

图 3　LEF-B 结构

Fig. 3　Structure of LEF-B

进行合并融合所有的通道信息。

在 MIFNet 的整体结构中，为了更好地加强空间关系和特征信息的交流，如图 1 所示，引入块间连接法，在初始卷积模块进行下采样之后连接了第一个特征提取模块 LFE-B block 1，用于对较浅层次特征信息的提取，生成含有较多空间信息的特征图通过两个支路向前传输；此外，融合后的特征图经过 1×1 卷积和第二次下采样之后连接第二特征提取模块 LFE-B block 2，用于对更深层次的特征信息进行提取。由此，可以通过在编码器中不同网络层次连接多个 LFE-B block 模块获得多个不同深度的特征图信息，从而获得丰富的空间特征信息，提高分割精度。

### 2.3　ESF 模块

在 MIFNet 编码部分中，含有 LFE-B block 1 的分支主要是对上下文空间信息进行编码，对浅层 1/2 特征图处理的分支是对边缘信息进行编码。为了有效地融合这两条路径中不同特征层次的信息，提取到足够多的语义信息，本文提出一种结合 Laplace 边缘检测算子和空间注意力机制的“边缘-空间”融合模块 ESF，一方面利用边缘检测算子去除语义噪声提取的边缘信息，另一方面利用空间注意力机制专注于上下文空间信息，通过融合两者的信息，提高网络性能。空间注意力机制定位到感兴趣的空间信息，抑制无用的信息，通过利用 3×3 卷积提取特征信息后再通过 Sigmoid 函数调节特征信息并生成空间信息权值，对该权值进行处理。如图 1 所示，其接收的一个输入是经过 LFE-B block 1 处理后的信息特征图 $\boldsymbol{F}_{\text{out1}}$，具有较多的空间信息；另外一个输入是浅层分支的 1/2 特征图经过 3×3 标准卷积信息提取处理后的输出特征图 $\boldsymbol{F}_{\text{out2}}$。特征图 $\boldsymbol{O}_1$ 由 3×3 卷积与 Sigmoid 函数运算得到，两者发挥空间注意力机制的作用专注于 $\boldsymbol{F}_{\text{out1}}$ 中的上下文空间细节信息，对特征信息进行激活。特征图 $\boldsymbol{O}_2$ 由 1×1 卷积和 Laplace 操作运算得到，传统边缘检测算子 Laplace 算子经过 Laplace 卷积形式操作提取特征图 $\boldsymbol{F}_{\text{out2}}$ 的边缘信息。特征图 $\boldsymbol{O}_1$ 和 $\boldsymbol{O}_2$ 的表达式分别为

$$\begin{cases} O_1 = \sigma\left[ f_{\text{conv3}\times 3}\left( F_{\text{out1}} \right) \right] \\ O_2 = \text{Laplace}\left[ f_{\text{conv1}\times 1}\left( F_{\text{out2}} \right) \right] \end{cases}, \tag{3}$$

式中：$\sigma$ 代表 Sigmoid 激活函数；$f_{\text{conv3}\times 3}$、$f_{\text{conv1}\times 1}$ 和 Laplace 分别表示 3×3 卷积操作、1×1 卷积操作以及 Laplace 操作。最后特征图 $\boldsymbol{O}_1$ 和 $\boldsymbol{O}_2$ 经过逐元素相乘融合通道信息，保证分割的精度。ESF 的最终输出表示为

$$F_{\text{out}} = O_1 \otimes O_2 , \tag{4}$$

式中，$\otimes$ 表示两个矩阵的逐元素级别的乘积操作。

在传统边缘检测算子中，Sobel 算子和 Prewitt 算子能够去掉一些噪声信息，但在特征图的 $x$ 和 $y$ 方向上 Laplace 算子使用一个式子就能够在 $x$ 和 $y$ 方向对边缘进行检测和提取[16]，占用较少的资源却能够提取较多的边缘信息。Laplace 算子能够先识别特征图的边缘，而不是直接将部分内容分配给特定的目标对象，效果表现较好，故将其作为融合模块中的边缘检测算子。

### 2.4　MAFD 模块

解码器是指对编码器编码的特征信息进行解码，以提供像素级上的分类。一个优秀的解码器能够使得模型提高精度的同时也能够保持较快的推理速度。为了能以较小的参数量更好地细化特征图，本文设计了一种 MAFD，如图 1 所示。

MAFD 接收的输入 $\boldsymbol{F}_{1/4}$ 是经过编码部分中 LFE-B block 1 与 ESF 模块处理后进行像素相加的特征图，输入 $\boldsymbol{F}_{1/8}$ 是经下采样模块改变特征分辨率进而控制参数量的特征图，最后一个输入 $\boldsymbol{L}_{1/8}$ 是编码部分中 LFE-B block 2 处理后的 8 倍下采样的特征图。MAFD 模块先使用 1×1 卷积使特征图 $\boldsymbol{F}_{1/8}$ 的通道数降为 64，控制计算量，然后应用全局池化（global-pooling）、Sigmoid 函数和 batch normalization 作为通道注意力机制突出重点通道特征信息，三者运算得到特征图 $\boldsymbol{F}_1$，然后与经过 1×1 卷积的 $\boldsymbol{L}_{1/8}$ 逐元素相乘融合信息得到特征图 $\boldsymbol{F}_2$。$\boldsymbol{F}_1$ 和 $\boldsymbol{F}_2$ 的表达式分别为

$$\begin{cases} F_1 = \sigma\left\{ f_{\text{gp}}\left[ f_{\text{conv1}\times 1}\left( F_{1/8} \right) \right] \right\} \\ F_2 = \left[ f_{\text{conv1}\times 1}\left( L_{1/8} \right) \right] \otimes F_1 \end{cases}, \tag{5}$$

式中，$f_{\text{gp}}$ 代表全局池化函数，主要作用是将输入特征图的分辨率降为 1，再使用 Sigmoid 函数和 batch normalization 对特征图的权值调节，生成通道注意力权值突出重点通道特征信息。通道注意力机制聚焦于不同特征通道，得到每一个特征通道的权值之后再将该权值应用到原来的通道中，学习不同通道的重要性。

对经过融合的特征图 $\boldsymbol{F}_2$ 使用双线性插值算法得到恢复 1/4 分辨率的特征图 $\boldsymbol{F}_3$，然后应用 3×3 标准卷积提取特征有效地减少信息的丢失，再与经过 1×1 卷

积处理的特征图 $\boldsymbol{F}_{1/4}$ 拼接(concat)在一起得到特征图 $\boldsymbol{F}_4$,保证网络的优化。$\boldsymbol{F}_3$ 和 $\boldsymbol{F}_4$ 的表达式分别为

$$\begin{cases} F_3 = f_{\text{upsample}}(F_2) \\ F_4 = f_{\text{concat}}\left[f_{\text{conv3}\times 3}(F_3)\right], \left[f_{\text{conv1}\times 1}(F_{1/4})\right] \end{cases}, \tag{6}$$

式中:$f_{\text{upsample}}$ 表示双线性插值上采样;$f_{\text{concat}}$ 表示拼接操作。最后再应用双线性插值恢复特征图尺寸,得到最终输出为

$$F_{\text{out}} = f_{\text{upsample}}\left[f_{\text{conv1}\times 1}(F_4)\right]。 \tag{7}$$

本文所设计的解码器加入注意力机制与多尺度的特征图进行融合,将参数量控制在可接受的范围内,加上通道注意力机制的灵活使用,保证了精度的提升。

## 3 实验验证与分析

将本文的网络模型与现有模型的参数数量、推理速度、分割精度(mIoU)进行比较,验证 MIFNet 在语义分割上的有效性。使用的数据集分别是 Cityscapes 数据集[17]和 CamVid 数据集[18]。Cityscapes 数据集是一个大型的城市街景数据集,分辨率为 1024 pixel×2048 pixel,包含 5000 幅精细注释图像和 19998 幅粗注释图像,具有 19 个语义类别。精细的注释图像分为 3 个集:297 张图像的训练集、500 张图像的验证集和 1525 张图像的测试集。另一个街景数据集 CamVid 共包含 701 张图片,分辨率为 720 pixel×960 pixel,也包括 3 个集合:367 张图像用于训练,101 张图像用于验证,233 张图像用于测试,具有 11 个道路场景语义类别。

### 3.1 实验设置

所有的模型都是在 RTX 2080Ti GPU 下 CUDA 11.2 的 PyTorch 深度学习框架上进行训练,推理的结果是在单个 GTX 1080Ti GPU 上进行评估。实验使用小批量随机梯度下降(SGD)的方法,动量设置为 0.9,权重衰减为 $10^{-4}$,使用“poly”作为学习率衰减策略,其中将初始学习率设置为 0.03,指数系数为 0.9。在训练过程中,将最大迭代次数设为 1000,同时为了增强数据,对输入图片进行随机水平翻转、平均减法和随机尺度的变换,随机尺度设置的参数为{0.75, 1.00, 1.25, 1.50, 1.75, 2.00}。在 Cityscapes 数据集中,为了与不同的语义分割网络模型进行公平比较,在训练阶段将图像的分辨率随机裁剪为 512 pixel×1024 pixel;对于 CamVid 数据集,使用分辨率为 360 pixel×480 pixel 来进行实验。需要说明的是,本文提出的新网络模型未采用预训练操作。

### 3.2 消融实验

在本节中,设计与每个模块相对应的消融实验,以验证各个核心模块的有效性。消融实验是在 Cityscapes 的验证集上进行评估,图片的输入分辨率为 512 pixel×1024 pixel。

#### 3.2.1 LFE-B 结构的性能分析

设计模块性能对比实验,评估本文 LFE-B 结构的有效性。将现有语义分割网络模型中的瓶颈结构用 LFE-B 模块替代,对比网络模型中原瓶颈结构和 LFE-B 的性能,得到的数据如表 1 所示(其中 GFLOPs 为浮点运算次数,用来衡量算法的复杂度)。这些数据都是在同一硬件平台上进行测试的,故具有一定的严谨性。

表 1 LFE-B 在不同网络结构的表现结果

Table 1 Results of different networks with LFE-B

| Network | Speed /(frame·s$^{-1}$) | | Parameters /M | | mIoU /% | | GFLOPs | |
|---|---|---|---|---|---|---|---|---|
| | Original | LFE-B | Original | LFE-B | Original | LFE-B | Original | LFE-B |
| DABNet[5] | 106.00 | 104.41 | 0.76 | 0.69 | 69.1 | 70.46 | 11.18 | 9.61 |
| ERFNet[7] | 58.57 | 41.01 | 2.07 | 0.75 | 70.0 | 71.49 | 26.86 | 9.84 |
| LEDNet[8] | 58.94 | 72.01 | 0.95 | 0.60 | 70.6 | 70.00 | 11.51 | 7.65 |
| ESNet[19] | 51.39 | 51.72 | 1.66 | 1.38 | 70.7 | 71.12 | 24.35 | 14.29 |
| MIFNet (proposed) | — | 73.68 | — | 0.82 | — | 72.50 | — | 12.03 |

从表 1 数据可以看出,替换成 LFE-B 模块的参数要比原数据低,表明了 LFE-B 模块的轻量性。除替换后的 LEDNet 的精度比原网络模型的精度要低 0.6 个百分点之外,其他三个网络的精度都有所提升。对于每秒传输帧数(FPS),替换模块后的 ERFNet 推理速度稍有变慢,但参数量得到较大的降低,精度也有所上升。整体效果来看,LFE-B 模块与现有的瓶颈结构相比在性能上得到较大的改善,同时在参数量只有 0.82 M 的情况下,能以 73.68 frame/s 的推断速度获得了 72.50 % 的 mIoU,在参数数量、推理速度以及分割精度之间得到了平衡,说明了 LFE-B 应用在语义分割网络的有效性。

#### 3.2.2 ESF 模块的性能分析

为了分析边缘信息的改善,将 Laplace 算子替换成 3×3 标准卷积,为了验证 Laplace 算子的有效性,构造了 Prewitt 算子和 Sobel 算子以及无算子部分的 ESF 模块。结果如表 2 所示,加入边缘检测算子之后的 ESF 性能有所提高,对比双方向处理边缘的边缘检测算子,Prewitt 算子的 mIoU 值为 71.79%,在 512 pixel×1024 pixel 的输入分辨率下的推理速度为 71.39 frame/s,Sobel 算子的 mIoU 值与其相比虽提高了 0.37 个百分点,但与 Laplace 算子的 mIoU 值相比,却低了 0.34 个

百分点，同时，在相同条件下，Laplace算子的推理速度比使用Sobel算子的推理速度快0.94 frame/s，因此Laplace算子具有更快的推理速度以及具有更有效的性能。同时为了更加直观地说明Laplace边缘检测算子在对边缘信息检测的敏感性，突出其有效性，对其进行了可视化处理。

图4展示了2张输入图像及其未经过Laplace算子处理的热度图以及经过处理后所对应的热度图。热度图的原理为当某些信息的温度越高时，说明对该部分的内容就越关注，从图4(b)中的方框可以看出，未使用Laplace算子处理前网络所关注的内容偏向于整体、局部等大面积的信息，而使用Laplace算子处理后，温度高的内容大多出现在特征图的边缘上，颜色比较深，关注内容偏向于边缘信息，如图4(c)中椭圆框等的一些线条显示，表明使用Laplace边缘检测算子之后的特征图边缘信息突出明显。

表2　ESF模块不同形式的表现结果

Table 2　Results of different configuration of ESF

| Configuration | Speed /(frame·s$^{-1}$) | Parameters /M | mIoU /% |
|---|---|---|---|
| None | 74.97 | 0.82 | 71.71 |
| 3×3 | 71.80 | 0.83 | 71.65 |
| Prewitt | 71.39 | 0.82 | 71.79 |
| Sobel | 72.74 | 0.82 | 72.16 |
| Laplace | 73.68 | 0.82 | 72.50 |

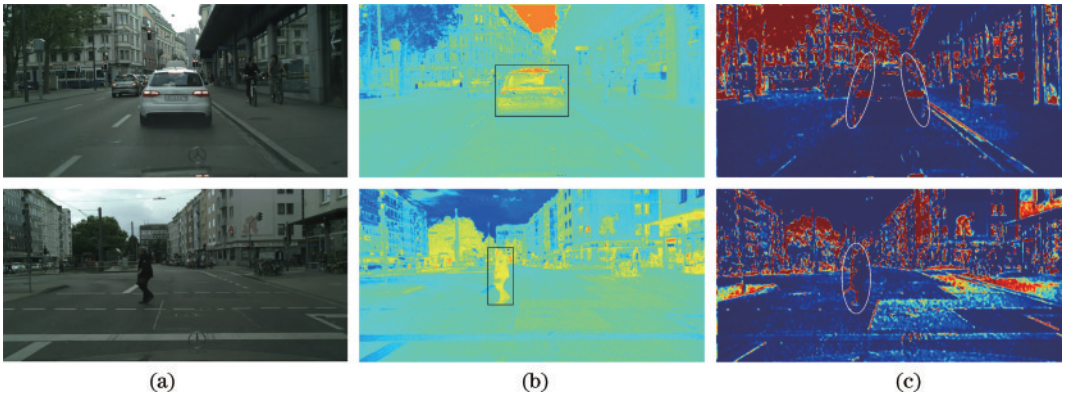



图4　融合模块中进行不同处理后的热度图。(a)输入图像；(b)未使用Laplace算子处理的热度图；(c)使用Laplace算子处理后的热度图

Fig. 4　Heat maps after different processing in fusion module. (a) Input image; (b) heat map without Laplace operator; (c) heat map with Laplace operator

### 3.2.3　MAFD模块的性能分析

在验证编码部分中的瓶颈结构以及融合模块后，分析MAFD模块对整个网络结构性能的影响。首先是在网络结构中取出MAFD，以简单的分类器作为解码器进行解码，然后将解码器换成PAD[5]、ERFD[7]以及APN[8]解码模块，得到的性能结果对比数据如表3所示。其中，带有ERFD的网络结构分割精度达到72.10%，但其整个模型比较繁重，参数数量较大；PAD解码结构推理速度以及参数量在MIFNet网络中的表现较好，但对于语义分割精度相对ERFD解码器稍差，虽然PAD解码器在推理速度和参数量表现上对比MAFD较优，但其分割精度不如MAFD；APN解码器在速度和精度上的表现均不如MAFD。因此，从总体上来看，在实时性语义分割任务中对3个性能指标的权衡方面，MAFD较其他解码模块表现更优。

表3　不同解码模块在MIFNet上的表现结果

Table 3　Results of different decoders on MIFNet

| Decoder | Speed /(frame·s$^{-1}$) | Parameters /M | mIoU /% |
|---|---|---|---|
| None | 88.20 | 0.77 | 71.10 |
| ERFD[7] | 52.91 | 1.03 | 72.10 |
| PAD[5] | 75.23 | 0.77 | 71.59 |
| APN[8] | 67.29 | 0.78 | 69.51 |
| MAFD (proposed) | 73.68 | 0.82 | 72.50 |

## 3.3　MIFNet整体网络模型的性能分析

本节基于CamVid数据集和Cityscapes数据集，将MIFNet与现有的实时性语义分割网络在参数数量、推理速度以及分割精度方面进行全面的比较，实验结果分别汇总在表4和表5中。为了进行公平的性能对比，除了特殊说明，所有网络模型的测试(test)过程均是在NVIDIA GTX 1080Ti GPU下执行，并且，基于Cityscapes测试集的实验采用512 pixel×1024 pixel图像分辨率；基于CamVid测试集的实验分别在以下两种分辨率实现：360 pixel×480 pixel和720 pixel×960 pixel。

从表4基于Cityscapes测试集和表5基于CamVid测试集的实验结果可以看出，不同的轻量化网络的性能各有优势。例如，DABNet[5]与仅使用非对称卷积的ERFNet[7]和LEDNet[8]相比精度有所提升，但与使用两次跳连接的MIFNet相比，mIoU却下降了1.9个百分点。基于Camvid数据集，DABNet[5]仍旧保持较低参数量和较快的推理速度，但在分割精度上略逊于

表 4 不同网络模型在 Cityscapes 测试集上性能的比较

Table 4 Performance comparison of different network models on Cityscapes test set

| Network | Pretrain | Speed /(frame·s$^{-1}$) | Parameters /M | mIoU (test) /% | GFLOPs |
|---|---|---|---|---|---|
| ENet[6] | No | 41.70 | 0.36 | 58.3 | 4.35 |
| ESPNet[22] | No | 146.00 | 0.36 | 60.3 | 3.50 |
| CGNet[23] | No | 44.70 | 0.50 | 65.6 | 7.00 |
| ContextNet[10] | No | 176.60 | 0.88 | 65.5 | 1.78 |
| EDANet[24] | No | 105.50 | 0.68 | 67.3 | 9.00 |
| ERFNet[7] | No | 58.57 | 2.07 | 68.0 | 26.90 |
| FastSCNN[9] | No | 198.41 | 1.10 | 62.8 | 1.76 |
| LEDNet[8] | No | 58.94 | 0.95 | 69.2 | 11.50 |
| DABNet[5] | No | 106.20 | 0.64 | 71.2 | 10.50 |
| ESNet[19] | No | 51.39 | 1.66 | 70.7 | 24.40 |
| LRNNet_C[14] | No | 71.00 | 0.68 | 72.2 | 8.58 |
| BiSeNetV1_X[20]* | ImageNet | 105.80* | 5.80 | 68.4 | 14.90 |
| BiSeNetV1_R[20]* | ImageNet | 65.50* | 49.00 | 74.7 | 55.30 |
| BiSeNetV2[21] | No | 156.00 | — | 72.6 | 21.15 |
| BiSeNetV2_L[21] | No | 47.30 | — | 75.3 | 118.51 |
| MIFNet (proposed) | No | 73.68 | 0.82 | 73.1 | 12.03 |

Note: *represents test result under NVIDIA Titan Xp GPU and resolution of 768 pixel×1536 pixel; X represents Xception39; R represents ResNet18

表 5 不同网络模型在 CamVid 测试集上性能的比较

Table 5 Performance comparison of different network models on CamVid test set

| Network | Input size /pixel | Speed /(frame·s$^{-1}$) | Parameters /M | mIoU (test) /% | GFLOPs |
|---|---|---|---|---|---|
| ENet[6] | 360×480 | 61.00 | 0.36 | 51.3 | 1.44 |
| ERFNet[7] | 360×480 | 64.30 | 2.07 | 67.1 | 8.80 |
| DABNet[5] | 360×480 | 117.00 | 0.64 | 64.6 | 3.20 |
| LEDNet[8] | 360×480 | 58.94 | 0.95 | 66.6 | 11.50 |
| EKENet[13] | 360×480 | 38.00 | 1.20 | 67.5 | — |
| ESPNet[22] | 360×480 | 132.00 | 0.36 | 55.6 | 1.10 |
| EDANet[24] | 360×480 | 163.00 | 0.68 | 66.4 | 2.90 |
| CGNet[23] | 360×480 | 112.00 | 0.50 | 65.6 | 65.60 |
| LRNNet_C[14] | 360×480 | 76.50 | 0.68 | 69.2 | — |
| BiSeNetV1_X[20]* | 720×960 | 175.00* | 49.00 | 65.6 | 8.70 |
| BiSeNetV1_R[20]* | 720×960 | 116.30* | 5.80 | 68.7 | 32.40 |
| BiSeNetV2[21] | 720×960 | 124.50 | — | 72.4 | 21.15 |
| BiSeNetV2_L[21] | 720×960 | 32.70 | — | 73.2 | 118.51 |
| MIFNet(proposed) | 720×960 | 55.02 | 0.81 | 71.1 | 15.86 |
| MIFNet(proposed) | 360×480 | 85.16 | 0.81 | 67.7 | 3.90 |

Note: *represents test result under NVIDIA Titan Xp GPU; X represents Xception39; R represents ResNet18

MIFNet。BiSeNetV1[20]利用 Xception39 或 ResNet18 作为主干网并在 ImageNet 上进行预训练之后的推理速度和分割精度均有较大提升，但由于其需要执行预训练，计算代价更高，而且其模型参数量相对其他网络模型也更加庞大(可达 49 M)；因此，BiSeNetV2[21]去除了预训练，并给出了两个不同深度与宽度的网络模型，取得了与 BiSeNetV1 相比拟的 mIoU 值，推理速度亦有所提升，但其计算所需的浮点操作数却更大，计算复杂度高，相较 BiSeNet 而言，MIFNet 更轻量。LRNNet[14]的 Model C 的整体性能与本文提出的 MIFNet 接近，但其 mIoU 值在基于 Cityscapes 测试集的实验中相对 MIFNet 下降了 0.9 个百分点，推理速度亦下降了 2.68 frame/s，基于 CamVid 测试集的实验结果其 mIoU 虽有所提高，但推理速度却慢了差不多 9 frame/s。在相同的 GTX 1080Ti GPU 环境下，ContextNet[10]推理速度相对 MIFNet 更快，但其精度在

Cityscapes测试集的实验中下降了7.6个百分点。

本文提出的MIFNet在Cityscapes测试集上以0.82 M参数量获得了73.1%的mIoU，在CamVid测试集上输入图像分辨率为720 pixel×960 pixel时获得71.1%的mIoU，在360 pixel×480 pixel的输入图像中获得67.7%的mIoU。相对其他大多数神经网络而言，MIFNet可以更好地权衡模型复杂度和分割精度，即：以更低的参数量取得更高的分割精度。在推理速度上，MIFNet属于中等水平，仍然有较大的提升空间作为未来研究工作的主要攻克方向。此外，需要说明的是，由于MAFD最后的上采样将通道数压缩至类别数（Cityscapes为19类，CamVid为11类），导致模型在Cityscapes和CamVid数据集在参数量上有细微的差别。

### 3.4 整体网络模型性能的定性对比

为了获得定性的分割结果对比分析，如图5所示，本文进一步给出了MIFNet和现有的多个轻量型语义分割网络在Cityscapes数据集上生成的可视化结果。观察图中第二行，与列出来的网络相比，MIFNet能够将细小对象中的细节信息成功分割出来，即路灯的灯杆被成功地恢复，主要得益于网络中的LFE-B模块提取到较为丰富的细节特征信息以及ESF模块边缘信息的提取，同时，从图中第一行、第三行以及第四行的结果也可以看出MIFNet对于不同类别的对象能够进行较准确的分类，比如大面积树叶和地面，以及墙面和路面的分类，进一步表明了MAFD模块对于多尺度信息融合的增强作用。与其他模型相比，MIFNet不仅能够分割出边缘细小的对象，也能够较准确地分割出较大的对象，分割效果显著，表明了MIFNet的实时性语义分割对不同尺度的对象分割效果上具有一定提升。

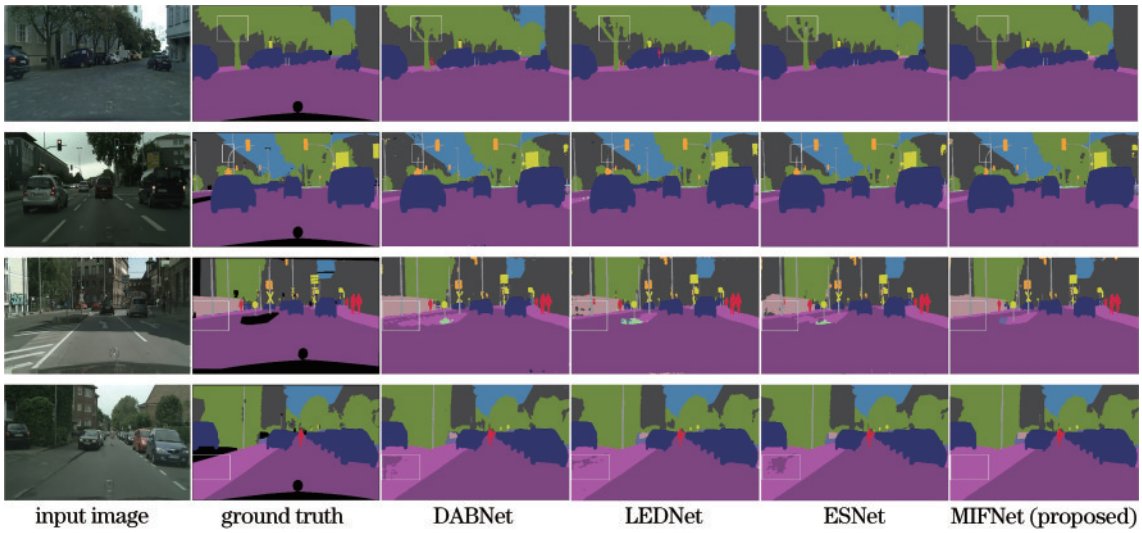



图5 在Cityscapes数据集上的语义分割结果

Fig. 5 Semantic segmentation results on Cityscapes dataset

综合比较下，MIFNet可以较好地平衡参数数量、推理速度以及分割精度三者之间的关系，表明了该模型在轻量级语义分割中具有较快和较为准确的分割性能，体现了MIFNet在轻量级语义分割方法上较佳的时效性和准确性。

## 4 结 论

设计了基于多尺度特征信息融合的道路场景实时性语义分割网络模型MIFNet。针对现有瓶颈结构出现的信息间交流不流畅的问题，设计了轻量型特征提取瓶颈结构LFE-B引入跳线连接与经过非对称卷积后的信息连接融合丰富信息，解决了网络退化的问题，同时，为了有效地融合空间信息与边缘信息，设计了Laplace边缘检测算子与空间注意力机制相结合的ESF模块来提取不同层次的信息，最后设计解码模块MAFD，进一步提取和融合多尺度的细节信息，完成特征图语义信息的提取与恢复。在Cityscapes数据集上的消融实验结果表明，LFE-B、ESF和MAFD这3个模块具有较优的性能，组合成的MIFNet与其他现有的语义分割网络相比，可以更好地平衡参数数量、推理速度以及分割精度这三者的关系，可执行轻量、实时、准确的语义分割任务，在资源受限的边缘移动终端上具有良好的应用前景。

## 参 考 文 献

[1] 陈浩，杨恺伦，胡伟健，等. 基于全景环带成像的语义视觉里程计[J]. 光学学报，2021，41(22)：2215002.
Chen H，Yang K L，Hu W J，et al. Semantic visual odometry based on panoramic annular imaging[J]. Acta Optica Sinica，2021，41(22)：2215002.

[2] 赵亮，胡杰，刘汉，等. 基于语义分割的深度学习激光点云三维目标检测[J]. 中国激光，2021，48(17)：1710004.
Zhao L，Hu J，Liu H，et al. Deep learning based on semantic segmentation for three-dimensional object

detection from point clouds[J]. Chinese Journal of Lasers, 2021, 48(17): 1710004.

[3] Takos G. A survey on deep learning methods for semantic image segmentation in real-time[EB/OL]. (2020-09-27)[2022-03-27]. https://arxiv.org/abs/2009.12942.

[4] Li G, Yun I, Kim J, et al. DABNet: depth-wise asymmetric bottleneck for real-time semantic segmentation [EB/OL]. (2019-10-01)[2022-03-27]. https://arxiv.org/abs/1907.11357.

[5] Li G, Jiang S L, Yun I, et al. Depth-wise asymmetric bottleneck with point-wise aggregation decoder for real-time semantic segmentation in urban scenes[J]. IEEE Access, 2020, 8: 27495-27506.

[6] Paszke A, Chaurasia A, Kim S, et al. ENet: a deep neural network architecture for real-time semantic segmentation[EB/OL]. (2016-06-07)[2022-03-27]. https://arxiv.org/abs/1606.02147.

[7] Romera E, Álvarez J M, Bergasa L M, et al. ERFNet: efficient residual factorized ConvNet for real-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(1): 263-272.

[8] Wang Y, Zhou Q, Liu J, et al. Lednet: a lightweight encoder-decoder network for real-time semantic segmentation[C]//2019 IEEE International Conference on Image Processing, September 22-25, 2019, Taipei, China. New York: IEEE Press, 2019: 1860-1864.

[9] Poudel R P K, Liwicki S, Cipolla R. Fast-SCNN: fast semantic segmentation network[EB/OL]. (2019-02-12) [2022-03-27]. https://arxiv.org/abs/1902.04502.

[10] Poudel R P K, Bonde U, Liwicki S, et al. ContextNet: exploring context and detail for semantic segmentation in real-time[EB/OL]. (2018-11-05) [2022-03-27]. https://arxiv.org/abs/1805.04554.

[11] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[12] Luo A, Yang F, Li X, et al. EKENet: Efficient knowledge enhanced network for real-time scene parsing [J]. Pattern Recognition, 2021, 111: 107671.

[13] Peng C L, Tian T, Chen C, et al. Bilateral attention decoder: a lightweight decoder for real-time semantic segmentation[J]. Neural Networks, 2021, 137: 188-199.

[14] Jiang W H, Xie Z Z, Li Y Y, et al. LRNNET: a light-weighted network with efficient reduced non-local operation for real-time semantic segmentation[C]//2020 IEEE International Conference on Multimedia & Expo Workshops, July 6-10, 2020, London, UK. New York: IEEE Press, 2020.

[15] Hien D H T. A guide to receptive field arithmetic for convolutional neural networks[EB/OL]. (2017-04-06) [2022-03-27]. https://blog. mlreview. com/a-guide-to-receptive-field-arithmetic-for-convolutional-neural-networks-e0f514068807.

[16] 韩利利. 基于分数阶微分的图像边缘检测算法的研究[D]. 北京: 北京印刷学院, 2021.
Han L L. Research on image edge detection algorithm based on fractional differential[D]. Beijing: Beijing Institute of Graphic Communication, 2021.

[17] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 3213-3223.

[18] Brostow G J, Fauqueur J, Cipolla R. Semantic object classes in video: a high-definition ground truth database [J]. Pattern Recognition Letters, 2009, 30(2): 88-97.

[19] Wang Y, Zhou Q, Xiong J, et al. ESNet: an efficient symmetric network for real-time semantic segmentation [M]//Lin Z C, Wang L, Yang J, et al. Pattern recognition and computer vision. Lecture notes in computer science. Cham: Springer, 2019, 11858: 41-52.

[20] Yu C Q, Wang J B, Peng C, et al. BiSeNet: bilateral segmentation network for real-time semantic segmentation[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11217: 334-349.

[21] Yu C Q, Gao C X, Wang J B, et al. BiSeNet V2: bilateral network with guided aggregation for real-time semantic segmentation[J]. International Journal of Computer Vision, 2021, 129(11): 3051-3068.

[22] Mehta S, Rastegari M, Caspi A, et al. ESPNet: efficient spatial pyramid of dilated convolutions for semantic segmentation[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11214: 561-580.

[23] Wu T Y, Tang S, Zhang R, et al. CGNet: a light-weight context guided network for semantic segmentation [J]. IEEE Transactions on Image Processing, 2021, 30: 1169-1179.

[24] Lo S Y, Hang H M, Chan S W, et al. Efficient dense modules of asymmetric convolution for real-time semantic segmentation[C]//MMAsia '19: Proceedings of the ACM Multimedia Asia, December 16-18, 2019, Beijing, China. New York: ACM Press, 2019: 1-6.