中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2025)07-2437-14

论文引用格式:Hou Z Q, Qu M J, Li J G, Ma S G, Wang Y C and Yang X B. 2025. Lightweight CNN-Transformer combined network for real-time semantic segmentation. Journal of Image and Graphics, 30(7):2437-2450(侯志强, 屈敏杰, 李俊歌, 马素刚, 王昀琛, 杨小宝. 2025. 轻量级CNN-Transformer相结合的实时语义分割网络. 中国图象图形学报, 30(7):2437-2450)[DOI:10.11834/jig.240527]

轻量级CNN-Transformer相结合的实时语义分割网络

侯志强,屈敏杰*,李俊歌,马素刚,王昀琛,杨小宝

1. 西安邮电大学计算机学院,西安　710121; 2. 陕西省网络数据分析与智能处理重点实验室,西安　710121

摘　要:**目的**　在语义分割领域,基于CNN的分割模型虽然在获取局部信息方面表现优越,但往往无法有效提取全局语义信息;相比之下,Transformer在提取全局语义信息方面明显优于CNN,为了更好地利用局部信息和全局语义信息,提出一种轻量级CNN-Transformer相结合的实时语义分割网络(lightweight CNN-Transformer combined network, LCTNet),以适应城市街景下的自动驾驶场景分割。**方法**　首先,设计了轻量级的小波增强卷积块(wavelet-enhanced convolution block, WECB)用于下采样,该模块将传统卷积和小波变换相结合,加强了对边缘纹理细节的提取能力;其次,为提取多尺度上下文信息,设计了一个轻量级的多尺度融合模块(multi-scale fusion module, MFM),该模块能高效地提取并融合多尺度上下文信息;最后,为了捕捉长距离依赖关系和全局上下文信息,设计了轻量级的Transformer(lightweight Transformer, LWT),通过自注意力机制直接关注输入序列的任意位置,捕捉长距离依赖关系,从而改善城市街景分割效果。**结果**　所提算法在Cityscapes和CamVid数据集上进行实验,平均交并比(mean intersection over union,mIoU)分别达到75.9%和69.7%,速度分别为63帧/s和84帧/s,参数量仅有1.06 M。**结论**　所提算法在Cityscapes和CamVid数据集上与近年优秀的方法进行了比较,LCTNet有效地提高了分割性能,同时在分割精度、推理速度和参数量方面取得了较好的平衡。

关键词:轻量级;实时语义分割;小波变换;Transformer;多尺度上下文信息

Lightweight CNN-Transformer combined network for real-time semantic segmentation

Hou Zhiqiang, Qu Minjie*, Li Junge, Ma Sugang, Wang Yunchen, Yang Xiaobao

1. School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an　710121, China;

2. Key Laboratory of Network Data Analysis and Intelligent Processing of Shaanxi Province, Xi'an　710121, China

Abstract: **Objective** Convolutional neural networks (CNNs) have been widely used in the field of semantic segmentation due to their excellent ability to handle local information in images. However, they frequently struggle when capturing global semantic information, limiting their performance in complex scenes with long-range dependencies. By contrast, Transformer models have demonstrated superior performance in extracting global semantic information. The self-attention mechanism in Transformers allows them to capture information across an entire image. This capability is particularly impor-

收稿日期:2024-09-18;**修回日期**:2024-11-26;**预印本日期**:2024-12-03

*通信作者:屈敏杰　qmj_mu@163.com

基金项目:国家自然科学基金项目(62072370);陕西省自然科学基金项目(2023-JC-YB-598)

Supported by: National Natural Science Foundation of China (62072370); Natural Science Foundation of Shaanxi Province, China (2023-JC-YB-598)

tant in the context of complex urban street scene segmentation. **Method** To better leverage local information and global semantic information, this study proposes a lightweight CNN-Transformer combined network for real-time semantic segmentation, referred to as LCTNet. By combining the advantages of traditional CNNs and Transformers, the objective is to maintain high segmentation accuracy while achieving model lightweighting and real-time processing capabilities. First, although highly effective in extracting local features, traditional convolution operations frequently lose some key details when dealing with complex edges and fine textures, especially during down-sampling. To enhance the model's ability to extract edge and texture details, this study designs a lightweight wavelet-enhanced convolution block. This module innovatively combines traditional convolution operations with wavelet transform to enhance the extraction of detailed texture information. Second, the variation in object scale is highly significant in complex urban street scenes, such as small cars in the distance and pedestrians nearby, which vary considerably in their proportion within an image. To address this challenge, this study designs a lightweight multi-scale fusion module (MFM) for efficiently extracting and integrating multi-scale contextual information. By using pooling kernels with varying sizes, MFM can extract feature information at different scales, ensuring the capture of small-scale details and large-scale global features. During the fusion stage, MFM uses a dense connection strategy to efficiently integrate the features extracted from different-scale pooling. By directly connecting each scale's feature layer with others, all scale information is fully shared and interacted with. This dense connection method effectively enhances feature representation capability, enabling the model to comprehensively understand a scene across multiple scales and improve the recognition and segmentation accuracy of objects of different scales. Through this design, MFM not only improves the model's segmentation performance in complex urban street scenes but also maintains lightweight characteristics, achieving efficient multi-scale information fusion while maintaining low computational overhead. Finally, Transformer models excel in terms of capturing long-range dependencies and global contextual information, particularly when dealing with complex scenes, significantly improving model performance. To fully utilize the advantages of Transformers, this study designs a lightweight Transformer module (LWT). This module focuses on capturing long-range dependencies at any position in the input sequence, ensuring that the model can understand the semantic information of a scene on a global scale. LWT receives features from different layers of the network as input, encompassing local and global semantic information. By integrating multilayer features, LWT can comprehensively capture the multilevel semantic information of an image. This multilayer feature input allows the model to understand not only detailed information but also the overall semantic structure. Such ability is crucial for complex urban street scene segmentation. **Result** The proposed algorithm was tested on two standard datasets that are widely used for urban street scene segmentation, namely, Cityscapes and CamVid. The experimental results showed that the mean intersection over union reached 75.9% on the Cityscapes dataset and 69.7% on the CamVid dataset, with inference speeds of 63 frame/s and 84 frame/s, respectively. The model's parameter count was only 1.06 M. **Conclusion** The proposed algorithm was compared with several excellent methods from recent years on the Cityscapes and CamVid datasets. The results indicate that the proposed algorithm achieves lightweight and real-time processing while maintaining high segmentation accuracy. Compared with traditional CNN models and some of the latest segmentation algorithms, the proposed algorithm exhibits advantages in terms of parameter count and computational complexity, achieving a good balance between accuracy and speed.

Key words: lightweight; real-time semantic segmentation; wavelet transform; Transformer; multi-scale contextual information

0 引 言

在计算机视觉领域中,语义分割是一个重要的研究方向,目的是将图像中的每个像素分配给一个特定的类别,从而实现对图像内容的细粒度理解。它具有重要的应用价值,例如自动驾驶(Siam等,2017)、医疗影像分析(曹伟杰 等,2024)和场景理解(Hofmarcher等,2019)等。

随着全卷积网络(fully convolution network,FCN)(Long等,2015)的出现,语义分割实现了端到端的学习。但由于传统卷积层感受野受限,因此其分割精度也受到限制。为此,DeepLabv2(Chen等,2018)提出空洞空间金字塔池化(atrous spatial pyra-

mid pooling，ASPP），通过在不同尺度上并行应用空洞卷积，捕捉更多上下文信息，大大提高了分割精度。PSPNet（pyramid scene parsing network）（Zhao等，2017）引入了空间金字塔池化（spatial pyramid pooling，SPP），在不同尺度上对特征进行池化，从而捕捉多尺度语义信息。

但是，上述算法的复杂度以及内存需求较高，难以在移动设备上应用。于是，一些轻量化的模型相继提出以实现在移动设备上的应用。ENet（efficient neural network）（Paszke等，2016）在编码器解码器结构中，使用深度可分离卷积和瓶颈模块以减少计算复杂度和参数量，显著优化了计算和内存消耗。ESPNet（Mehta等，2018）设计了ESP（efficient spatial pyramid）模块，通过将传统卷积分解为点卷积和空洞卷积，从而减少计算量并扩大感受野。ICNet（image cascade network）（Zhao等，2018）结合了轻量级的骨干网络和级联结构，通过下采样和逐步恢复分辨率的方式减少计算量，同时保持较好的分割效果。LEANet（lightweight and efficient asymmetric network）（Zhang等，2022a）提出一种轻量且高效的不对称网络，在解码器中设计了基于分离和混洗操作的深度不对称瓶颈模块，用于共同提取局部和上下文信息。LCNet（lightweight context-aware network）（Shi等，2024）使用三分支上下文聚合（three-branch context aggregation，TCA）模块扩展了特征感受野，捕捉多尺度的上下文信息，并引入一种部分通道变换（partial-channel transformation，PCT）策略，以最小化基本单元的计算延迟和硬件需求。为了提高对象边界的分割效果，BSCNet（boundary-guide semantic context network）（Zhou等，2024）提出一种具有多尺度上下文的边界引导双分辨率轻量级网络，利用一个级轻量级金字塔池化模块，在低分辨率分支获取多尺度上下文信息。ELANet（effective lightweight attention-guided network）（Yi等，2023）结合了空洞卷积和深度卷积，设计了两种轻量级的上下文引导模块，用于学习局部信息和上下文信息。

上述算法都采用了传统卷积神经网络（convolution neural network，CNN），该网络能充分地捕获局部细节信息，但难以捕捉长距离依赖关系和全局上下文信息。相较于CNN结构，Transformer通过自注意力机制（self-attention），可以直接关注输入序列中的任意位置，从而捕捉长距离依赖关系。近年来，Transformer在语义分割领域的应用也越来越广泛。Segmenter（Strudel等，2021）基于视觉Transformer（vision Transformer，ViT）将其扩展到语义分割中，在整个网络中对全局上下文信息进行建模，获得了良好的分割性能。SegFormer（Xie等，2021）将Transformer的自注意力机制与轻量级的卷积网络相结合，采用高效的编码器结构，并结合了多尺度特征提取的解码器部分，使得整体分割性能表现更加优异。SSFormer（Shi等，2022）考虑到SwinTransformer的分层设计，提出一种解码器来聚合不同层的信息，从而获得局部和全局的关注。COVID-TransNet（樊圣澜等，2023）充分利用Transformer在捕获全局上下文信息方面的优势，提出一种基于Transformer的肺部CT（computed tomography）图像分割网络，用于COVID-19患者肺部CT图像分割。但此类模型参数量大且计算复杂度较高，在移动设备上实现起来较为困难。

为了更好地利用局部信息和全局语义信息，同时实现在移动设备上的应用，本文提出一种轻量级CNN-Transformer相结合的实时语义分割网络（lightweight cnn-transformer combined network，LCTNet），主要工作总结如下：1）在编码器中的特征提取部分，设计了小波增强卷积块（wavelet-enhanced convolution block，WECB），该模块利用Haar小波结合传统卷积，帮助捕捉图像中的边缘和纹理特征，增强了编码器的特征表达。2）为了提取并融合多尺度上下文信息，设计了轻量型的多尺度融合模块（multi-scale fusion module，MFM），该模块能高效地获取并融合多尺度上下文信息，通过提高计算效率，显著减少了计算量，对模型的推理时间影响较小。3）在解码器的末端，设计了轻量级的Transformer（lightweight Transformer，LWT），利用自注意力机制捕获输入序列的长距离依赖关系和全局语义信息，从而弥补CNN在处理长距离依赖关系时的不足，有效提升了模型的分割性能。4）在Cityscapes和CamVid数据集上对所提算法进行了测试，实验结果表明，LCTNet在分割性能、推理速度以及参数量上实现了良好的平衡。

1 本文算法

轻量级CNN-Transformer相结合的实时语义分

割网络(LCTNet)的整体架构如图1所示。

LCTNet由编码器解码器构成。首先,输入图像经过卷积块进行下采样,生成特征$\boldsymbol{X}_{1/2}$。将特征$\boldsymbol{X}_{1/2}$作为输入,使用WECB进行下采样,然后采用不带有小波变换的WECB进行特征融合,得到特征$\boldsymbol{X}_{1/4}$,同理,再得到特征$\boldsymbol{X}_{1/8}$,特征$\boldsymbol{X}_{1/2}$、$\boldsymbol{X}_{1/4}$、$\boldsymbol{X}_{1/8}$大小分别为原始特征的1/2、1/4和1/8。其次,为了更高效地提取多尺度信息,帮助模型更好地识别和分割图像中的不同大小和不同上下文的对象,设计了多尺度信息融合模块(MFM),进一步增强模型对于多尺度对象的分割能力,前一层的特征经过MFM处理后,输出为$\boldsymbol{M}_{1/8}$,其大小为原图的1/8。最后,在编码器的末端,利用轻量级的Transformer(LWT)获取长距离依赖关系,输出特征$\boldsymbol{L}_{1/16}$大小为原图的1/16,该特征为解码器提供了丰富的语义信息,使得模型更好地理解整个图像的内容和结构,从而有效地提升分割效果。

在解码器中,将编码器末端的特征上采样,逐步与网络浅层特征融合,恢复细节信息。首先,将特征$\boldsymbol{L}_{1/16}$上采样,与MFM的输出融合,得到特征$\boldsymbol{C}_{1/8}$,输出特征$\boldsymbol{C}_{1/4}$。然后,与编码器特征$\boldsymbol{X}_{1/2}$进行融合,恢复细节信息,改善小目标分割效果。特征$\boldsymbol{C}_{1/8}$、$\boldsymbol{C}_{1/4}$和$\boldsymbol{C}_{1/2}$大小分别为原始特征的1/8、1/4和1/2。

下面详细介绍LCTNet的3个模块:小波增强卷积块(WECB)、多尺度信息融合模块(MFM)以及轻量级的Transformer(LWT)。

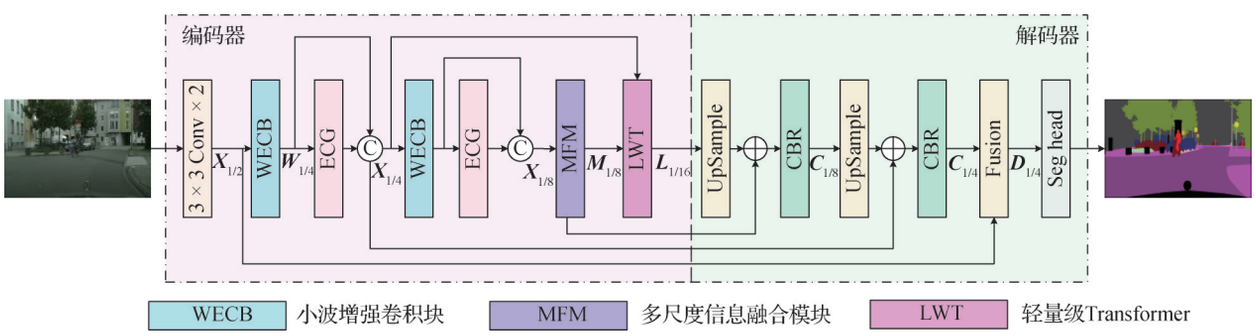


图1 LCTNet整体架构示意图

Fig. 1 Schematic diagram of the overall architecture of LCTNet

1.1 小波增强卷积块(WECB)

传统卷积能够提取图像中的局部信息,但随着多层卷积的堆叠,利用传统卷积进行下采样会损失细节信息,如边缘和纹理信息。即使在解码器部分使用浅层信息,也无法恢复已经损失的边缘和纹理信息。在语义分割领域,对象边缘的分割对于分割效果的提升有着重要的影响。小波变换能够在时间和频率两个域中进行局部信息的提取,同时有效地去除噪声,在图像处理和边缘检测中具有显著优势。因此,设计了小波增强卷积块(WECB),同时利用传统卷积和Haar小波变换各自的优势完成对图像的下采样,还使用空洞卷积扩大传统卷积的感受野,既提取了图像中的局部信息,又利用了上下文信息,帮助模型更好地识别和分割对象,WECB结构如图2所示。

首先,输入特征$\boldsymbol{X}_{1/2} \in \mathbf{R}^{C_0 \times H_0 \times W_0}$输入到卷积和小波变换下采样块中,分别进行下采样。在左分支中使用步长为2的3×3卷积,将特征$\boldsymbol{X}_{1/2}$的大小缩小为一半,得到$\boldsymbol{L} \in \mathbf{R}^{C_0 \times (H_0/2) \times (W_0/2)}$。同时,为了更好地提取边缘纹理信息,减少传统卷积下采样带来的信息损失,在右分支中,特征图进入小波变换下采样块(wavelet-enhanced convolution block, WTDB)模块中,利用Haar小波分解为1个低频分量$\boldsymbol{y}_l \in \mathbf{R}^{C_0 \times (H_0/2) \times (W_0/2)}$和3个高频分量$\boldsymbol{y}_{hl} \in \mathbf{R}^{C_0 \times (H_0/2) \times (W_0/2)}$、$\boldsymbol{y}_{lh} \in \mathbf{R}^{C_0 \times (H_0/2) \times (W_0/2)}$以及$\boldsymbol{y}_{hh} \in \mathbf{R}^{C_0 \times (H_0/2) \times (W_0/2)}$,分别表示水平方向、垂直方向和对角线方向的高频特征。上述过程可表示为

$$\begin{cases} \boldsymbol{L} = Conv_{3\times3}(\boldsymbol{X}_{1/2}) \\ (\boldsymbol{y}_l, [\boldsymbol{y}_{hl}, \boldsymbol{y}_{lh}, \boldsymbol{y}_{hh}]) = Haar(\boldsymbol{X}_{1/2}) \end{cases} \tag{1}$$

式中,$Conv_{3\times3}$代表3×3卷积层,$Haar$指哈尔小波变换操作。

随后,低频和高频分量在通道维度上进行拼接,得到包含多尺度上下文信息的特征$\boldsymbol{C} \in \mathbf{R}^{4C_0 \times (H_0/2) \times (W_0/2)}$。拼接后的特征图经过1×1卷积、批

归一化和ReLU(rectified linear unit)激活函数的序列操作，进一步增强和提取特征，得到特征 $\boldsymbol{R}\in\mathbf{R}^{C_0\times(H_0/2)\times(W_0/2)}$，上述操作可表示为

$$\begin{cases}\boldsymbol{C}=C(\boldsymbol{y}_l,\boldsymbol{y}_{hl},\boldsymbol{y}_{lh},\boldsymbol{y}_{hh})\\ \boldsymbol{R}=ReLU(BN(Conv_{1\times1}(\boldsymbol{C})))\end{cases}\tag{2}$$

式中，C 指拼接操作，$ReLU$ 为激活函数，BN(batch normalization)指批量归一化，$Conv_{1\times1}$ 表示 1 × 1 卷积层。

将两个分支得到的特征融合后，输入到非对称卷积中进行特征提取，其中左分支为深度可分离非对称卷积，为了扩大感受野提取上下文信息，右分支为带有空洞率的非对称卷积，分别得到特征 $\boldsymbol{L}_1$ 和 $\boldsymbol{R}_1$，非对称卷积能更好地关注到和垂直方向的不同特征，并且在不额外增加计算成本的情况下，提升模型的分割性能，上述过程可表示为

$$\begin{cases}\boldsymbol{L}_1=DWConv_{1\times3}(DWConv_{3\times1}(Conv_{1\times1}(\boldsymbol{L}\oplus\boldsymbol{R})))\\ \boldsymbol{R}_1=DWConv^r_{1\times3}(DWConv^r_{3\times1}(Conv_{1\times1}(\boldsymbol{L}\oplus\boldsymbol{R})))\end{cases}\tag{3}$$

式中，$DWConv_{1\times3}$ 和 $DWConv_{3\times1}$ 指非对称卷积，$DWConv^r_{1\times3}$ 和 $DWConv^r_{3\times1}$ 表示带空洞率的非对称卷积，$\oplus$ 表示逐元素相加。

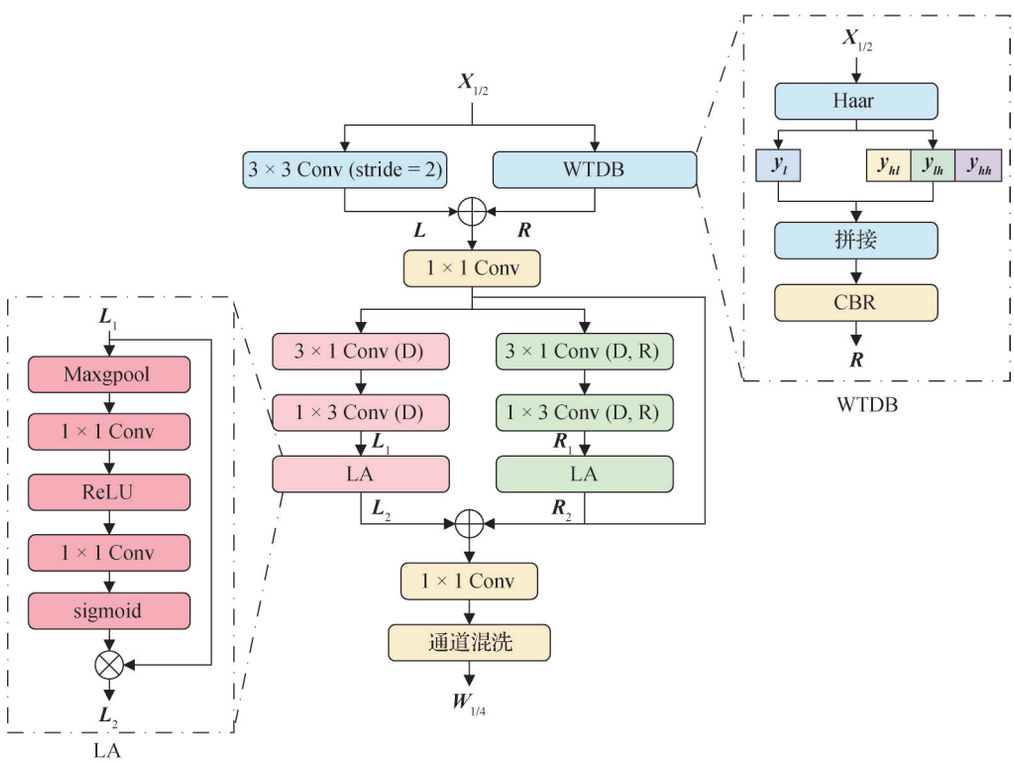


图2　小波增强卷积块

Fig. 2　Wavelet-enhanced convolution block (WECB)

另外，在每个分支的末端，设计了局部注意力(local attention，LA)来增强每个通道的重要特征。LA主要由最大池化和一系列卷积和激活函数组成，并使用sigmoid函数计算通道维度的权重并与原特征相乘，$\boldsymbol{L}_2\in\mathbf{R}^{C_0\times(H_0/2)\times(W_0/2)}$ 计算过程为

$$\begin{cases}\boldsymbol{L}_1'=Conv_{1\times1}(Maxpool(\boldsymbol{L}_1))\\ \boldsymbol{L}_2=sigmoid(Conv_{1\times1}(ReLU(\boldsymbol{L}_1')))\end{cases}\tag{4}$$

式中，$Maxpool$ 为最大池化操作。

最后，通过 1 × 1 卷积和通道混洗操作融合特征 $\boldsymbol{W}_{1/4}\in\mathbf{R}^{C_0\times(H_0/2)\times(W_0/2)}$，生成最终的输出特征图，上述流程可表示为

$$\boldsymbol{W}_{1/4}=ChannelShuffle(Conv(\boldsymbol{L}_2\oplus\boldsymbol{R}_2))\tag{5}$$

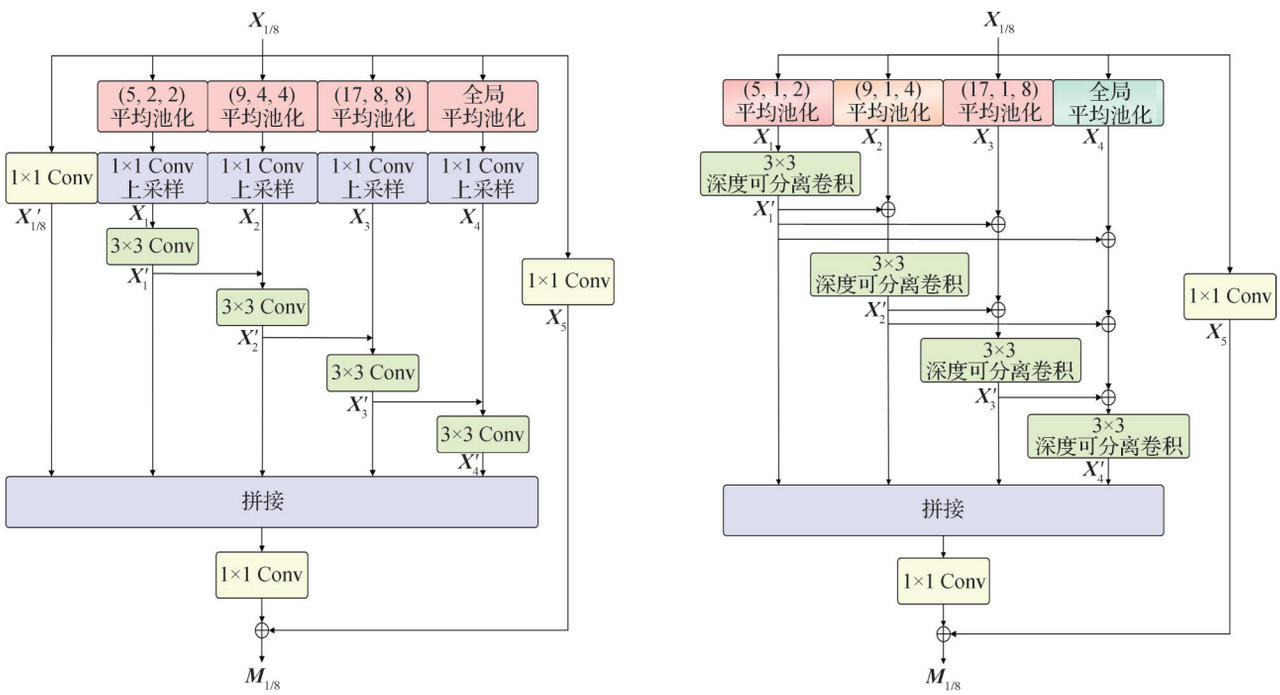
式中，$ChannelShuffle$ 为通道混洗操作。

这种结合了小波变换的下采样方法提升了模型对边缘纹理特征的捕捉能力，同时使用非对称卷积和空洞卷积使模型能够灵活地捕捉水平和垂直方向的特征，有效地提升了编码器的特征提取能力。

1.2　多尺度信息融合模块(MFM)

多尺度上下文信息在增强模型对不同尺度物体和场景的理解能力上至关重要，有助于提高语义分割的精度和鲁棒性。近年来，一些语义分割模型采用了金字塔网络结构，使模型能够在不同层次上获

取丰富的上下文信息。例如，DeepLabv2（Chen等，2018）引入了空洞空间金字塔池化（ASPP），通过在多个尺度上并行使用空洞卷积增强上下文信息的捕捉能力。PSPNet（Zhao等，2017）采用空间金字塔池化（SPP）在不同尺度上进行特征池化，以有效捕捉多尺度的语义信息。DDRNet（deep dual-resolution networks）（Hong等，2021）算法中的深度聚合金字塔池化模块（deep aggregation pyramid pooling module，DAPPM）通过并行的多尺度池化操作提取上下文信息，如图3（a）所示。然而，DAPPM模块中较大的池化步长和上采样操作可能会导致特征信息的丢失，并引入冗余信息。因此，为了更加高效地提取多尺度信息，提出一种新的多尺度上下文融合模块（MFM），如图3（b）所示。



(a) 深度聚合金字塔池化模块　　(b) 多尺度融合模块

图3　深度聚合金字塔池化模块和多尺度融合模块

Fig. 3　Deep aggregation pyramid pooling module and multi-scale fusion module (MFM)

((a) deep aggregation pyramid pooling module; (b) multi-scale fusion module)

如图3（a）所示，DAPPM模块的3个平均池化核大小分别为（5，9，17），步长为（2，4，8），padding为（2，4，8），输出特征图的大小分别为8×8、4×4、2×2、1×1大小，再将多尺度特征上采样到16×16大小。为了减少多尺度特征信息的损失，将MFM模块的池化操作步长均设置为1，3个平均池化核大小分别为（5，9，17），同时通过自适应全局平均池化获取全局上下文信息。输入特征$\boldsymbol{X}_{1/8} \in \mathbf{R}^{C_1 \times H_1 \times W_1}$经过并行的平均池化和全局平均池化操作后，直接得到输出特征$\boldsymbol{X}_1 \in \mathbf{R}^{C_2 \times H_1 \times W_1}$、$\boldsymbol{X}_2 \in \mathbf{R}^{C_2 \times H_1 \times W_1}$、$\boldsymbol{X}_3 \in \mathbf{R}^{C_2 \times H_1 \times W_1}$和$\boldsymbol{X}_4 \in \mathbf{R}^{C_2 \times H_1 \times W_1}$，它们的输出大小均为16×16。相较于DAPPM，该方法能够保证在提取到多尺度信息的同时减少输入图像信息的损失，也避免了不同尺度的特征图进行双线性差值上采样操作时引入冗余信息。上述过程可表示为

$$\begin{cases} \boldsymbol{X}_{1,2,3} = Avgpool(\boldsymbol{X}_{1/8}) \\ \boldsymbol{X}_4 = GAP(\boldsymbol{X}_{1/8}) \end{cases} \tag{6}$$

式中，*AvgPool*指平均池化操作，*GAP*指全局平均池化（global average pooling，GAP）操作。

从输入特征中获取到多尺度上下文信息后，DAPPM模块将多尺度特征从左至右依次相加融合。为了更好地利用多尺度信息，同时尽量不增加模型的参数量，在MFM中，设计了新的多尺度融合方式。如图3（b）所示，从第1个分支开始，将多尺度特征输入到深度可分离卷积中整理特征，再依次将特征与不同分支的特征相加融合，其余分支进行相同的操作。特征$\boldsymbol{X}'_1 \in \mathbf{R}^{C_2 \times H_1 \times W_1}$、$\boldsymbol{X}'_2 \in \mathbf{R}^{C_2 \times H_1 \times W_1}$、$\boldsymbol{X}'_3 \in \mathbf{R}^{C_2 \times H_1 \times W_1}$和$\boldsymbol{X}'_4 \in \mathbf{R}^{C_2 \times H_1 \times W_1}$实现过程可表示为

$$\begin{cases} \boldsymbol{X}_1' = DWConv_{3\times3}(\boldsymbol{X}_1) \\ \boldsymbol{X}_2' = DWConv_{3\times3}(\boldsymbol{X}_1' \oplus \boldsymbol{X}_2) \\ \boldsymbol{X}_3' = DWConv_{3\times3}(\boldsymbol{X}_1' \oplus \boldsymbol{X}_2' \oplus \boldsymbol{X}_3) \\ \boldsymbol{X}_4' = DWConv_{3\times3}(\boldsymbol{X}_1' \oplus \boldsymbol{X}_2' \oplus \boldsymbol{X}_3' \oplus \boldsymbol{X}_4) \end{cases} \tag{7}$$

式中，$DWConv_{3\times3}$指3×3的深度可分离卷积。

最后将多尺度特征拼接起来，利用1×1卷积调整通道并与特征$\boldsymbol{X}_5 \in \mathbf{R}^{C_3\times H_3\times W_3}$相加融合得到最终输出结果$\boldsymbol{M}_{1/8} \in \mathbf{R}^{C_1\times H_1\times W_1}$，实现过程可表示为

$$\boldsymbol{M}_{1/8} = Conv_{1\times1}(C(\boldsymbol{X}_1', \boldsymbol{X}_2', \boldsymbol{X}_3', \boldsymbol{X}_4')) \oplus \boldsymbol{X}_5 \tag{8}$$

式中，C代表拼接操作，⊕指逐元素相加操作。

1.2 轻量级的Transformer（LWT）

在语义分割领域，传统卷积神经网络（CNN）感受野受限，难以捕捉长距离依赖关系，而Transformer利用自注意力机制，在获取长距离依赖关系方面展现出强大的能力。但是，由于标准Transformer庞大的计算量和参数量，难以实现在移动设备上的应用。为了提高模型的分割性能同时实现更广泛的应用场景，设计了轻量级的Transformer，以提高模型的分割性能，同时平衡计算效率。Transformer的操作流程如图4所示，包括下采样卷积层、轻量级的双输入多头注意力（lightweight dual-input multi-head attention，LDMA）、归一化层和多层感知机（multilayer perceptron，MLP）。为了利用网络浅层细节信息，分别将浅层特征$\boldsymbol{X}_{1/4} \in \mathbf{R}^{C_3\times H_2\times W_2}$和深层特征$\boldsymbol{M}_{1/8} \in \mathbf{R}^{C_1\times H_1\times W_1}$共同输入到轻量级Transformer中，将两个特征图下采样到输入图像的1/16大小，获取丰富的语义信息。

在标准的多头注意力中，输入序列会被多个不同的线性变换映射到不同的$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$空间，然后分别计算注意力分数，并通过一个线性层投影到输出空间。如图5所示，LDMA共有两个输入，分别为$\boldsymbol{D} \in \mathbf{R}^{C\times H\times W}$和$\boldsymbol{S} \in \mathbf{R}^{C\times H\times W}$。首先，将不同层级特征的通道数降为输入的一半，利用深层特征获取$\boldsymbol{Q} \in \mathbf{R}^{HW\times C}$，利用浅层特征获取$\boldsymbol{K} \in \mathbf{R}^{C\times HW}$和$\boldsymbol{V} \in \mathbf{R}^{HW\times C}$。将$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分为$x$组，分别利用每一组的$\boldsymbol{Q}_i$、$\boldsymbol{K}_i$、$\boldsymbol{V}_i$进行自注意力计算，计算过程为

$$\boldsymbol{S}_i = softmax\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i}{\sqrt{d}}\right)\boldsymbol{V}_i \tag{9}$$

上述过程会得到x组结果，然后将得到的结果拼接，最后进行通道数恢复，在保持全局信息建模能力的同时，减少了模型的内存占用和计算成本。

通过多头注意力机制，模型能够更好地理解车辆、行人和道路标志等在不同视角下的相对位置。此外，实验结果表明，通过结合CNN和轻量级Transformer，模型能够在提取局部和全局特征之间实现更好的平衡，从而提升语义分割的精度。

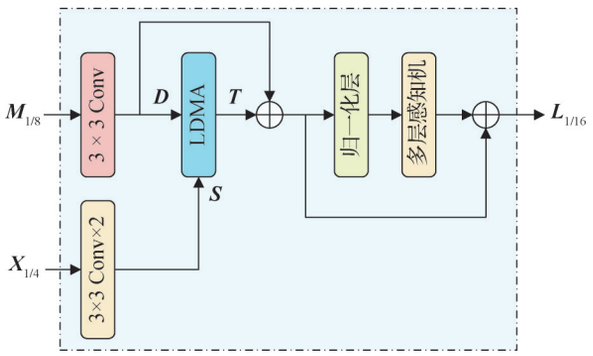


图4 轻量级Transformer

Fig. 4 Lightweight Transformer (LWT)

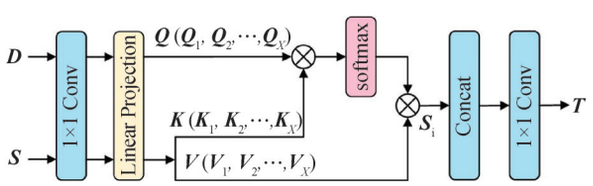


图5 轻量级双输入多头注意力

Fig. 5 Lightweight dual-input multi-head attention (LDMA)

2 实 验

2.1 数据集

为了证明所提算法的有效性，本节展示了在Cityscapes（Cordts 等，2016）和CamVid（Brostow 等，2009）数据集上的实验。

1）Cityscapes是语义分割领域得到广泛应用的数据集，包含30个类别的精细标注，其中19个类别专用于该领域的研究。数据集由两部分构成：5 000幅精细标注图像与20 000幅粗略标注图像，全面覆盖了城市道路、建筑、行人及车辆等多元语义实体。本文仅采用分辨率为1 024×2 048像素的精细标注图像集，其中2 975幅用于模型训练，500幅用于验证模型性能，剩余的1 525幅作为测试集，以评估模型的泛化能力。

2）CamVid数据集专门设计用于场景理解和语义分割任务，标签包含32个语义类别，其中11个类用于语义分割任务。整个数据集包括701幅图像，其中367幅用于训练模型，101幅用于验证模型的性

能，233幅则作为测试集。每个图像的分辨率为720×960像素。

2.2 实验环境和参数设置

本文实验在Ubuntu 16.04系统上使用一块NVIDIA GeForce RTX 3090进行模型训练，并且基于Python 3.7、PyTorch 1.8.1和CUDA 11.1等实验环境。

本文算法采用Baseline的解码器结构，并且LCTNet不进行任何预训练。对于Cityscapes数据集，采用批量随机梯度下降算法（stochastic gradient descent，SGD）训练网络。初始学习率设置为4.5E-2，权重衰减为1E-4，采用动量为0.9的学习策略，以降低学习率和权重衰减和权重衰减。为了提高模型的鲁棒性，采用随机裁剪、随机水平翻转和随机缩放来增强训练图像。图像被随机裁剪成512×1 024像素，网络经过1 000次迭代训练。对于CamVid数据集，采用适应性动量估计算法（Adam）训练网络。图像被随机裁剪成360×480像素，经过1 000次迭代训练。在推理阶段，Cityscapes和CamVid数据集输入尺寸分别为512×1 024像素和720×960像素。将batch size大小设置为1，在单张RTX 3090 GPU上测量推理速度。

在移动终端实验设备的算力方面，本文算法设计充分考虑了性能与效率的平衡。多尺度融合模块（MFM）采用轻量的深度可分离卷积，以更高效地提取多尺度信息。同时，引入了轻量级Transformer（LWT）获取全局语义信息。最终，所提算法的参数量为1.06 M，计算复杂度为11.9 GFLOPS。相比之下，目标检测领域的YOLO11n-seg模型参数量为2.9 M，计算量为10.4 GFLOPS，其在CPU上的推理速度为15.2帧/s。这意味着本文算法在参数量上仅为YOLO11n-seg模型的2.7倍，计算量差距较小，表明其在硬件平台上的部署潜力。

2.3 消融实验

为了验证所提方法的有效性，在Cityscapes数据集上对3个模块进行了消融实验，包括小波增强卷积块（WECB）、多尺度信息融合模块（MFM）和轻量级的Transformer（LWT）模块。采用平均交并比（mean intersection over union，mIoU）、每秒帧数（frames per second，FPS）、浮点运算次数（floating-point operations per second，FLOP）和模型参数量（parameter）指标评估相应的模型，分析、验证各模块对提升算法整体的性能表现。

2.3.1 模块性能分析

为了验证小波增强卷积块（WECB）、多尺度信息融合模块（MFM）和多轻量级的Transformer（LWT）模块的有效性，进行了消融实验，结果如表1所示。

第1组实验为复现Baseline的结果；第2组实验在此基础上加入WECB模块，mIoU值达到74.9%；第3组实验加入MFM，相较于Baseline，mIoU值提高0.7%；第4组实验加入LWT模块，mIoU值为75.0%。前4组实验结果表明，WECB模块、MFM模块和LWT模块对于分割性能提升的有效性。第5组实验为WECB模块和MFM模块的结合，mIoU值达到75.3%；第6组实验结合了WECB和LWT，mIoU值为75.5%，与Baseline相比，mIoU值提高1.1%；

表1 LCTNet各个模块在Cityscapes数据集上的消融实验研究

Table 1 Ablation experiment study of each module of this algorithm on Cityscapes dataset

实验组号	Baseline	WECB	MFM	LWT	参数量/M	FLOPs/G	mIoU/%		速度/(帧/s)	
							Cityscapes	CamVid	Cityscapes	CamVid
1	√	–	–	–	0.7	9.8	74.4	67.9	83	120
2	√	√	–	–	0.8	10.7	74.9	68.3	72	100
3	√	–	√	–	0.8	10.6	75.1	68.2	78	105
4	√	–	–	√	0.9	10.3	75.0	68.5	74	103
5	√	√	√	–	0.9	11.5	75.3	69.1	66	87
6	√	√	–	√	1.0	11.1	75.5	69.5	68	88
7	√	–	√	√	1.0	11.1	75.4	69.3	70	93
8	√	√	√	√	1.1	11.9	75.9	69.7	63	84

注：“√”表示采用，“–”表示未采用。

第7组实验结合MFM和LWT，mIoU值为75.4%；第8组实验是3种方法的结合，相较于Baseline，mIoU值提高1.5%，进一步证实了LCTNet的有效性。

2.3.2 WECB性能分析

小波变换有不同的类型，为了使下采样块提取到更加精细的边缘和纹理特征，设计了如表2所示消融实验。

第1组实验是Baseline的实验结果；第2组实验在Baseline的基础上，采用Daubechies小波变换与卷积结合的下采样块，mIoU为74.1%；第3组为Coiflets小波变换与卷积相结合，mIoU达到74.6%；最后一组为Haar小波与卷积结合的方法，mIoU为74.9%。在参数量相同的情况下，前两种小波变换的计算量比Haar小波方法稍高一些，FPS与Haar小波持平或偏低，因此，本文选取Haar小波与卷积结合的方法作为下采样块，从而提升编码器对边缘纹理特征的捕捉能力。

表2 小波卷积增强块在Cityscapes上的消融实验研究

Table 2 Ablation experiment study of wavelet-enhanced convolution block on Cityscapes

方法	mIoU/%	参数量/M	FLOPs/G	速度/(帧/s)
Baseline	74.4	0.67	9.88	83
Daubechies	74.1	0.75	10.78	72
Coiflets	74.6	0.75	10.80	70
Haar(本文)	74.9	0.75	10.76	72

2.3.3 MFM性能分析

为了验证多尺度信息融合模块(MFM)的有效性，针对MFM进行了4组实验，如表3所示。

表3 多尺度融合模块在Cityscapes数据集上的消融实验研究

Table 3 Ablation experiment study of multi-scale fusion module on Cityscapes dataset

实验组号	Baseline	DAPPM	Stride = 1	Fusion	mIoU/%	参数量/M	FLOPs/G	速度/(帧/s)
1	√	–	–	–	74.4	0.67	9.8	83
2	√	√	–	–	74.6	0.92	11.6	72
3	√	√	√	–	74.9	0.78	10.6	79
4	√	√	√	√	75.1	0.78	10.6	78

注："√"表示采用，"–"表示未采用。

第1组实验为本文算法的Baseline，mIoU仅有74.4%；第2组实验加入了DAPPM模块用于提取多尺度信息，mIoU达到74.6%；为了减少信息损失同时避免双线性插值上采样引入冗余信息，第3组实验将DAPPM模块中的池化操作步长全部设置为1，减少了上采样步骤，参数量降低为0.78 M，mIoU值达到74.9%；第4组实验在第2组实验的基础上，将提取到的多尺度信息进行融合，采用了密集连接的融合方式，相较于第2组实验，参数量和计算量都没有改变，但是mIoU值提升至75.1%。以上实验结果可以看出，MFM可以更加高效地提取并融合多尺度信息，从而提升分割精度。

2.4 定量分析

2.4.1 Cityscapes

表4提供了LCTNet在Cityscapes数据集上与近年的轻量型分割模型的比较。本文在单个RTX3090GPU上测试，速度测试没有任何加速策略。LCTNet的参数量为1.06 M，mIoU为75.9%，推理速度为63帧/s。如表4所示，比较了经典的轻量型分割算法，LCTNet的精度均高于其他算法。虽然一些算法的推理速度比LCTNet高，但LCTNet在分割精度上更具优势，并且在速度和精度上取得了较好的平衡。

为了直观展示分析结果，图6提供了本文算法在Cityscapes测试集上准确性与速度比较的散点图。总体而言，LCTNet在综合性能上也表现出显著优势。

2.4.2 CamVid

为进一步验证本文算法的有效性，图7展示了在CamVid测试集上准确性与速度比较的散点图，本文算法在分割方面表现出更高的精确度，并且在综合性能上也表现出显著优势。表5提供了本文算法在CamVid数据集上与一些实时分割模型的比较。本文算法的mIoU为69.7%，推理速度为84帧/s，在

表4　Cityscapes数据集上的准确性和速度比较

Table 4　Accuracy and speed comparison on Cityscapes dataset

方法	文献来源	分辨率/像素	mIoU/%	参数量/M	速度/(帧/s)
DABNet (Li等, 2019)	arXiv2019	512 × 1 024	70.1	0.76	104
LEDNet (Wang等, 2019)	ICIP2019	1 024 × 512	70.6	0.94	71
FDDWNet (Liu等, 2020)	ICASSP2020	1 024 × 512	71.5	0.8	60
LRNNet (Jiang等, 2020)	ICMEW2020	512 × 1 024	72.2	0.68	71
LRDNet (Zhuang等, 2021)	IJON2021	1 024 × 512	70.1	0.66	77
Lite-HRNet (Yu等, 2021)	CVPR2021	512 × 1 024	72.8	1.1	–
EACNet (Li等, 2021)	SPL2021	512 × 1 024	74.2	1.1	113
BiAttnNet (Li等, 2022)	SPL2022	512 × 1 024	74.7	2.2	89.2
LAANet (Zhang等, 2022b)	NCA2022	512 × 1 024	73.6	0.67	95.8
FBSNet (Gao等, 2023)	TMM2023	512 × 1 024	70.9	0.62	90
SFRSeg (Singha等, 2023)	PR2023	512 × 1 024	70.6	1.6	194
LBARNet (Hu和Zhou, 2023)	CG2023	512 × 1 024	70.8	0.6	72
ELANet (Yi等, 2023)	NPL2023	512 × 1 024	74.4	0.67	83
LETNet (Xu等, 2023)	T-ITS2023	512 × 1 024	72.8	0.95	150
LCFNet_slim(Yang等, 2024)	IEEE TIV2024	1 024 × 512	76.5	7.28	171
FCFNet(Ye等, 2025)	J FIELD ROBOT2025	1 024 × 512	70.7	3.29	110
LCTNet (本文)	JIG2025	512 × 1 024	75.9	1.06	63

注:“–”表示原文没有提供该数据。

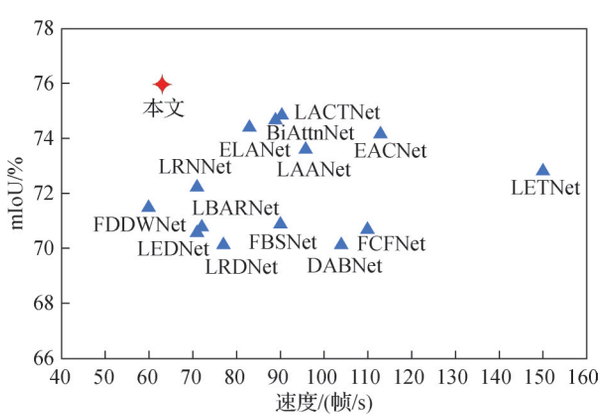


图6　在Cityscapes数据集上的准确性—速度散点图

Fig. 6　Accuracy-speed scatter plot on Cityscapes dataset

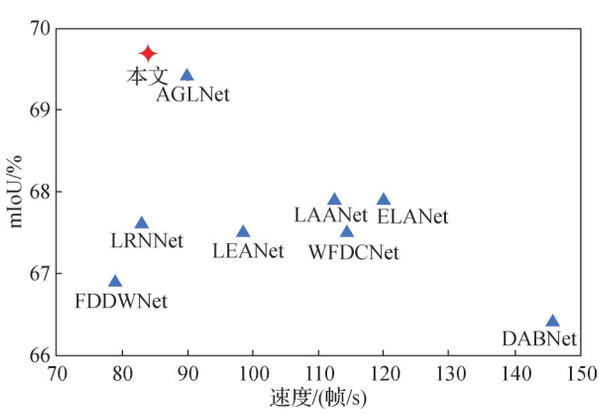


图7　在CamVid数据集上的准确性—速度散点图

Fig. 7　Accuracy-speed scatter plot on CamVid dataset

准确度方面占据一定优势,同时,在综合性能上有一定的优势。

2.5　定性分析

在Cityscapes和CamVid数据集上,图8和图9展示了LCTNet与Baseline的定性分析结果对比,使用白色方块标记有挑战性的区域。

2.5.1　Cityscapes

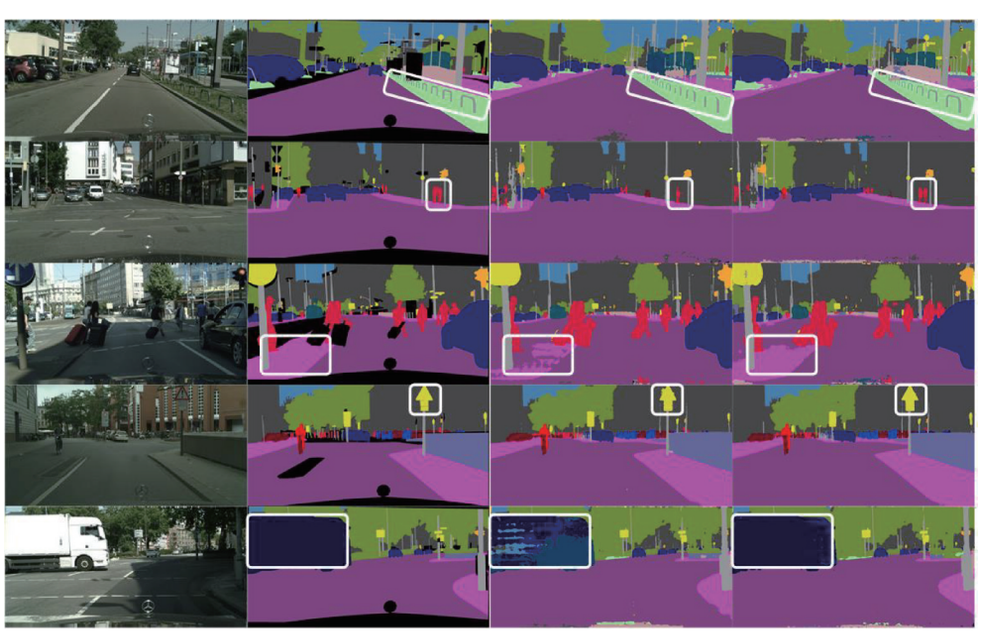
图8展示了LCTNet与Baseline之间的定性比较,并用白色方框标记出了具有挑战性的区域。Baseline对于一些区域的分割不够准确,而LCTNet能够更加精确地分割,验证了本文算法的有效性。例如第1行图像中,LCTNet能完整地分割出每一个对象,并且对于边缘的分割更加精细;第2行中,GroundTruth中有两个行人,而Baseline只分割出一个,LCTNet准确地分割出了两个行人;在第3行图像中,LCTNet对于地面的分割更加完整;在第4行图像

表 5　CamVid 测试集的准确性和速度比较

Table 5　Comparison of accuracy and speed on CamVid testset

方法	来源	分辨率/像素	mIoU/%	参数量/M	速度/(帧/s)
FPENet（Liu 等，2020）	arXiv2019	480 × 360	65.4	0.4	–
DABNet（Li 等，2019）	arXiv2019	360 × 480	66.4	0.76	146
CGNet（Wu 等，2021）	TIP2020	480 × 360	65.6	0.5	–
FDDWNet（Liu 等，2020）	ICASSP2020	960 × 720	66.9	0.8	79
LRNNet（Jiang 等，2020）	ICMEW2020	360 × 480	67.6	0.67	83
AGLNet（Zhou 等，2020）	ASC2020	480 × 360	69.4	1.12	90.1
WFDCNet（Hao 等，2021）	IVC2021	360 × 480	67.5	0.5	114.4
LRDNet（Zhuang 等，2021）	IJON2021	960 × 720	69.7	0.66	–
LEANet（Zhang 等，2022a）	APIN2022	360 × 480	67.5	0.74	98.6
LAANet（Zhang 等，2022b）	NCA2022	360 × 480	67.9	0.67	112.5
FBSNet（Gao 等，2023）	TMM2022	360 × 480	68.9	0.62	–
LBARNet（Hu 和 Zhou，2023）	CG2023	720 × 960	67.2	0.6	–
FCFNet（Ye 等，2025）	J FIELD ROBOT2025	720 × 960	68.1	3.2	–
LCTNet（本文）	JIG2025	360 × 480	68.5	1.06	106
		720 × 960	69.7	1.06	84

注：“–”表示原文没有提供该数据。



输入图像　真实值　基线可视化　LCTNet 可视化

图 8　Cityscapes 数据集可视化分割结果

Fig. 8　Visual segmentation results on Cityscapes dataset

的标志牌上，可以看出本文算法对于物体边缘的分割更加平滑；最后一行图像中，Baseline 将卡车内部错分为其他类别，而 LCTNet 的分割更加准确。LCTNet 结合了 CNN 和 Transformer 的特性，能更好地捕捉全局语义信息，对复杂背景下的对象分割更具优势。

2.5.2　CamVid

如图 9 所示，在 CamVid 数据集上 LCTNet 能更

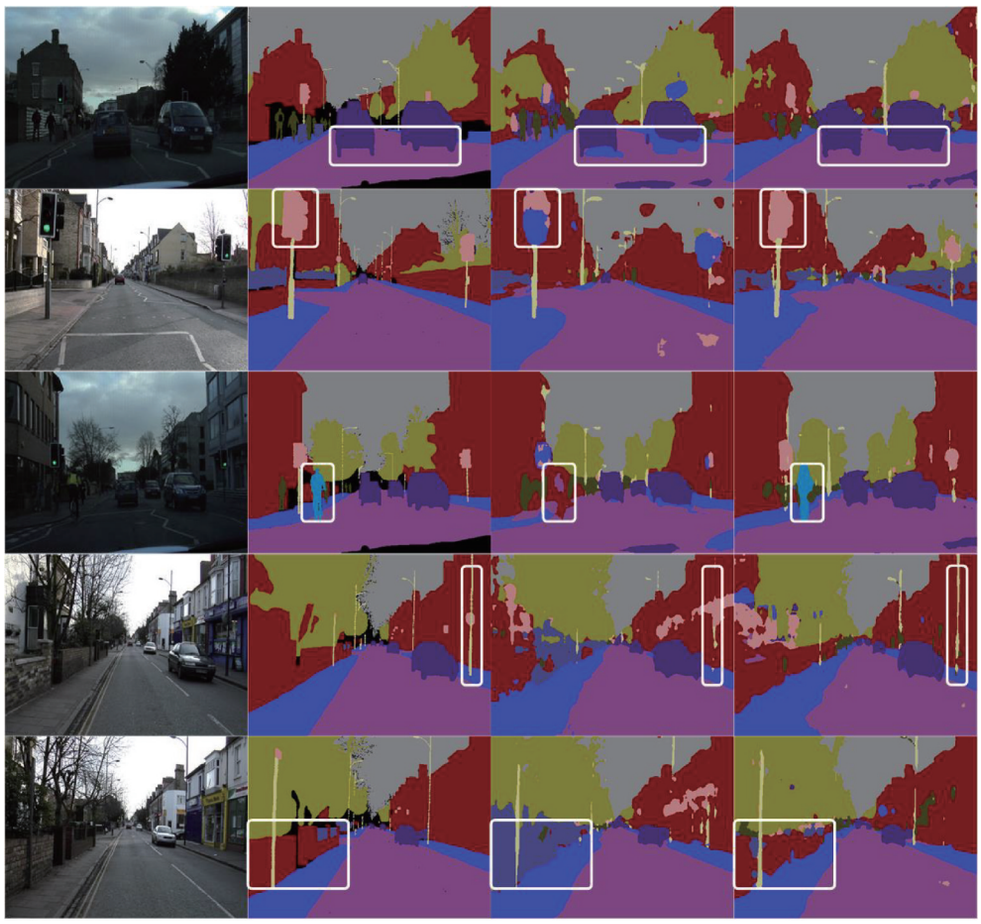


图9　CamVid数据集可视化分割结果

Fig. 9　Visual segmentation results on CamVid dataset

准确地判别物体所属类别并分割。在第1行图像中，对于汽车和地面这两类对象，LCTNet能更准确且完整地区分它们并进行准确的分割；第2行图像中，本文算法对于交通标志牌的分割更加准确；在第3行图像中，LCTNet准确地分割出了行人，而Baseline将行人和房屋错分为一类对象；第4行图像中对于路灯杆的分割，LCTNet更加精细；最后一行图像中，LCTNet对于墙面的分割更加准确。图9的对比结果进一步验证了本文算法在全局理解方面的优势。

3　结　论

本文提出一种轻量级CNN-Transformer相结合的实时语义分割网络（LCTNet），结合了CNN和Transformer的优点，旨在解决当前语义分割任务中面临的精度、计算资源和速度之间的平衡问题。通过引入小波增强卷积块、多尺度融合模块以及轻量级Transformer结构，LCTNet在准确提取局部细节的同时，有效捕获了全局语义信息，实现了对复杂城市街景的精准分割。实验结果表明，LCTNet在Cityscapes和CamVid数据集上表现出了卓越的性能，达到了较高的分割精度，同时保持了较低的参数量和快速的推理速度。尽管LCTNet在实验中表现出色，但仍存在一些局限性。例如，模型在处理极端尺度变化或复杂背景时表现有待进一步提升。此外，如何进一步降低模型的计算开销和内存占用，使其更适用于嵌入式系统或移动设备也是未来研究的一个方向。未来的研究工作将着重于以下几个方面：1）优化LCTNet的网络结构，进一步提升模型的分割精度，同时降低计算复杂度；2）探索更多高效的特征提取和融合方法，以提高模型在不同场景下的适应性；3）改进模型的解码器部分，更有效地融合网络不同层次的信息。

参考文献(References)

Brostow G J, Fauqueur J and Cipolla R. 2009. Semantic object classes in video: a high-definition ground truth database. Pattern Recognition Letters, 30(2): 88-97 [DOI: 10.1016/j.patrec.2008.04.005]

Cao W J, Duan X H, Xu Z W and Sheng S. 2024. Application of U-Net channel transformation network in gland image segmentation. Journal of Image and Graphics, 29(3): 713-724 (曹伟杰, 段先华, 许振伟, 盛帅. 2024. U-Net通道变换网络在腺体图像分割中的应用. 中国图象图形学报, 29(3): 713-724) [DOI: 10.11834/jig.230233]

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848 [DOI: 10.1109/TPAMI.2017.2699184]

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B. 2016. The cityscapes dataset for semantic urban scene understanding//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3213-3223 [DOI: 10.1109/CVPR.2016.350]

Fan S L, Bai Z Y, Lu Q J and Zhou X. 2023. A transformer network based CT image segmentation for COVID-19-derived lung disease. Journal of Image and Graphics, 28(10): 3203-3213 (樊圣澜, 柏正尧, 陆倩杰, 周雪. 2023. 基于Transformer网络的COVID-19肺部CT图像分割. 中国图象图形学报, 28(10): 3203-3213) [DOI: 10.11834/jig.220865]

Gao G W, Xu G A, Li J C, Yu Y, Lu H M and Yang J. 2023. FBSNet: a fast bilateral symmetrical network for real-time semantic segmentation. IEEE Transactions on Multimedia, 25: 3273-3283 [DOI: 10.1109/TMM.2022.3157995]

Hao X C, Hao X J, Zhang Y R, Li Y Y and Wu C. 2021. Real-time semantic segmentation with weighted factorized-depthwise convolution. Image and Vision Computing, 114: #104269 [DOI: 10.1016/j.imavis.2021.104269]

Hofmarcher M, Unterthiner T, Arjona-Medina J, Klambauer G, Hochreiter S and Nessler B. 2019. Visual scene understanding for autonomous driving using semantic segmentation//Samek W, Montavon G, Vedaldi A, Hansen L K and Müller K R, eds. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham, Schweiz: Springer: 285-296 [DOI: 10.1007/978-3-030-28954-6_15]

Hong Y D, Pan H H, Sun W C and Jia Y S. 2021. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes [EB/OL]. [2024-08-24]. https://arxiv.org/pdf/2101.06085.pdf

Hu X G and Zhou B M. 2023. LBARNet: lightweight bilateral asymmetric residual network for real-time semantic segmentation. Computers and Graphics, 116: 1-12 [DOI: 10.1016/j.cag.2023.07.039]

Jiang W H, Xie Z Z, Li Y Y, Liu C and Lu H T. 2020. LRNNET: a light-weighted network with efficient reduced non-local operation for real-time semantic segmentation//Proceedings of 2020 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). London, UK: IEEE: 1-6 [DOI: 10.1109/ICMEW46912.2020.9106038]

Li G, Yun I and Kim J. 2019. DABNet: depth-wise asymmetric bottleneck for real-time semantic segmentation [EB/OL]. [2024-08-24]. https://arxiv.org/pdf/1907.11357.pdf

Li G L, Li L and Zhang J W. 2022. BiAttnNet: bilateral attention for improving real-time semantic segmentation. IEEE Signal Processing Letters, 29: 46-50 [DOI: 10.1109/LSP.2021.3124186]

Li Y Q, Li X K, Xiao C J, Li H B and Zhang W M. 2021. EACNet: enhanced asymmetric convolution for real-time semantic segmentation. IEEE Signal Processing Letters, 28: 234-238 [DOI: 10.1109/LSP.2021.3051845]

Liu J, Zhou Q, Qiang Y, Kang B, Wu X F and Zheng B Y. 2020. FDDWNet: a lightweight convolutional neural network for real-time semantic segmentation//Proceedings of 2020 ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 2373-2377 [DOI: 10.1109/ICASSP40776.2020.9053838]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Mehta S, Rastegari M, Caspi A, Shapiro L and Hajishirzi H. 2018. ESPNet: efficient spatial pyramid of dilated convolutions for semantic segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 561-580 [DOI: 10.1007/978-3-030-01249-6_34]

Paszke A, Chaurasia A, Kim S and Culurciello E. 2016. ENet: a deep neural network architecture for real-time semantic segmentation// Proceedings of the 5th International Conference on Learning Representations. Toulon, France: ICLR: 1-10

Shi M, Lin S W, Yi Q M, Weng J, Luo A W and Zhou Y C. 2024. Lightweight context-aware network using partial-channel transformation for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems, 25(7): 7401-7416 [DOI: 10.1109/TITS.2023.3348631]

Shi W T, Xu J and Gao P. 2022. SSformer: a lightweight transformer for semantic segmentation//2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP). Shanghai, China: IEEE: 1-5 [DOI: 10.1109/MMSP55362.2022.9949177]

Siam M, Elkerdawy S, Jagersand M and Yogamani S. 2017. Deep semantic segmentation for automated driving: taxonomy, roadmap and challenges//Proceedings of the 20th IEEE International Confer-

ence on Intelligent Transportation Systems (ITSC). Yokohama, Japan: IEEE: 1-8 [DOI: 10.1109/ITSC.2017.8317714]

Singha T, Pham D S and Krishna A. 2023. A real-time semantic segmentation model using iteratively shared features in multiple sub-encoders. Pattern Recognition, 140: #109557 [DOI: 10.1016/j.patcog.2023.109557]

Strudel R, Garcia R, Laptev I and Schmid C. 2021. Segmenter: transformer for semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7242-7252 [DOI: 10.1109/ICCV48922.2021.00717]

Wang Y, Zhou Q, Liu J, Xiong J, Gao G W, Wu X F and Latecki L J. 2019. Lednet: a lightweight encoder-decoder network for real-time semantic segmentation//Proceedings of 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China: IEEE: 1860-1864 [DOI: 10.1109/ICIP.2019.8803154]

Wu T Y, Tang S, Zhang R, Cao J and Zhang Y D. 2021. CGNet: a light-weight context guided network for semantic segmentation. IEEE Transactions on Image Processing, 30: 1169-1179 [DOI: 10.1109/TIP.2020.3042065]

Xie E Z, Wang W H, Yu Z D, Anandkumar A, Alvarez J M and Luo P. 2021. SegFormer: simple and efficient design for semantic segmentation with transformers//Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021). [s.l.]: Curran Associates Inc: 1-14

Xu G A, Li J C, Gao G W, Lu H M, Yang J and Yue D. 2023. Lightweight real-time semantic segmentation network with efficient transformer and CNN. IEEE Transactions on Intelligent Transportation Systems, 24 (12): 15897-15906 [DOI: 10.1109/TITS. 2023.3248089]

Yang L, Bai Y W, Ren F L, Bi C K and Zhang R H. 2024. LCFNets: compensation strategy for real-time semantic segmentation of autonomous driving. IEEE Transactions on Intelligent Vehicles, 9(4): 4715-4729 [DOI: 10.1109/TIV.2024.3363830]

Ye X, Pan J C, Chen J C and Zhang J B. 2025. A multiscale feature fusion-guided lightweight semantic segmentation network. Journal of Field Robotics, 42(1): 272-286 [DOI: 10.1002/rob.22406]

Yi Q M, Dai G S, Shi M, Huang Z K and Luo A W. 2023. ELANet: effective lightweight attention-guided network for real-time semantic segmentation. Neural Processing Letters, 55(5): 6425-6442 [DOI: 10.1007/s11063-023-11145-z]

Yu C Q, Xiao B, Gao C X, Yuan L, Zhang L, Sang N and Wang J D. 2021. Lite-HRNet: a lightweight high-resolution network//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10435-10445 [DOI: 10.1109/CVPR46437.2021.01030]

Zhang X L, Du B C, Luo Z C and Ma K. 2022a. Lightweight and efficient asymmetric network design for real-time semantic segmentation. Applied Intelligence, 52 (1): 564-579 [DOI: 10.1007/s10489-021-02437-9]

Zhang X L, Du B C, Wu Z Y and Wan T B. 2022b. LAANet: lightweight attention-guided asymmetric network for real-time semantic segmentation. Neural Computing and Applications, 34(5): 3573-3587 [DOI: 10.1007/s00521-022-06932-z]

Zhao H S, Qi X J, Shen X Y, Shi J P and Jia J Y. 2018. ICNet for real-time semantic segmentation on high-resolution images//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 418-434 [DOI: 10.1007/978-3-030-01219-9_25]

Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 6230-6239 [DOI: 10.1109/CVPR.2017.660]

Zhou Q, Wang L J, Gao G W, Kang B, Ou W H and Lu H M. 2024. Boundary-guided lightweight semantic segmentation with multi-scale semantic context. IEEE Transactions on Multimedia, 26: 7887-7900 [DOI: 10.1109/TMM.2024.3372835]

Zhou Q, Wang Y, Fan Y W, Wu X F, Zhang S F, Kang B and Latecki L J. 2020. AGLNet: towards real-time semantic segmentation of self-driving images via attention-guided lightweight network. Applied Soft Computing, 96: #106682 [DOI: 10.1016/j.asoc.2020.106682]

Zhuang M X, Zhong X Y, Gu D B, Feng L Y, Zhong X G and Hu H S. 2021. LRDNet: a lightweight and efficient network with refined dual attention decorder for real-time semantic segmentation. Neurocomputing, 459: 349-360 [DOI: 10.1016/j.neucom.2021.07.019]

作者简介

侯志强,男,教授,博士生导师,主要研究方向为图像处理、计算机视觉、无人机应用和信息融合。
E-mail: hou-zhq@sohu.com
屈敏杰,通信作者,女,硕士研究生,主要研究方向为计算机视觉与图像语义分割。E-mail: qmj_mu@163.com
李俊歌,女,硕士研究生,主要研究方向为计算机视觉与图像语义分割。E-mail: ljg163yx@163.com
马素刚,男,硕士生导师,主要研究方向为计算机视觉和机器学习。E-mail: msg@xupt.edu.cn
王昀琛,女,讲师,主要研究方向为遥感图像处理方法和应用。E-mail: wangyunchen@lzb.ac.cn
杨小宝,男,副教授,主要研究方向为计算机视觉、图像处理和深度学习。E-mail: y78h11b09@xupt.edu.cn