融合多尺度深度卷积的轻量级 Transformer 交通场景语义分割算法

谢刚[1,2]，王荃毅[1,2]，谢新林[1,2]，王健安[1,2]

（1. 太原科技大学电子信息工程学院，山西 太原 030024；2. 先进控制与装备智能化山西省重点实验室，山西 太原 030024）

摘 要：针对交通场景语义分割算法中存在的易融入周围背景的纤细条状目标分割不连续、模型参数量大等问题，提出一种融合多尺度深度卷积的轻量级 Transformer 交通场景语义分割算法。首先，基于深度卷积构建多尺度条形特征提取模块，在不同尺度下增强对纤细条状目标特征的表示能力。其次，在浅层网络中利用卷积归纳偏置特性设计空间细节辅助模块，以弥补深层空间细节信息的丢失来优化目标边缘分割。最后，提出基于 Transformer-CNN 框架的非对称编解码网络，编码器结合 Transformer 与 CNN 减少细节信息丢失并降低模型参数量；而解码器采用轻量级的多级特征融合设计来进一步建模全局上下文。所提算法在 Cityscapes 和 CamVid 交通场景公开数据集上分别取得的平均交并比为 78.63%和 81.06%，能够在交通场景语义分割中实现分割精度和模型大小之间的权衡，具备良好的应用前景。

关键词：语义分割；深度学习；注意力机制；轻量级；交通场景

中图分类号：TP391.4

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2023194

Lightweight Transformer traffic scene semantic segmentation algorithm integrating multi-scale depth convolution

XIE Gang[1,2], WANG Quanyi[1,2], XIE Xinlin[1,2], WANG Jian'an[1,2]

1. School of Electronic and Information Engineering, Taiyuan University of Science and Technology, Taiyuan 030024, China
2. Shanxi Key Laboratory of Advanced Control and Equipment Intelligence, Taiyuan 030024，China

Abstract: Aiming at the problems of discontinuous segmentation of thin strip objects that were easy to blend into the surrounding background and a large number of model parameters in the semantic segmentation algorithm of traffic scenes, a lightweight Transformer traffic scene semantic segmentation algorithm integrating multi-scale depth convolution was proposed. First, a multi-scale strip feature extraction module (MSEM) was constructed based on deep convolution to enhance the representation ability of thin strip target features at different scales. Secondly, a spatial detail auxiliary module (SDAM) was designed using the convolutional inductive bias feature in the shallow network to compensate for the loss of deep spatial detail information to optimize object edge segmentation. Finally, an asymmetric encoding-decoding network based on the Transformer-CNN framework (TC-AEDNet) was proposed. The encoder combined Transformer and CNN to alleviate the loss of detail information and reduce the amount of model parameters; while the decoder adopted a lightweight multi-level feature fusion design to further model the global context. The proposed algorithm achieves the mean intersection over union (mIoU) of 78.63% and 81.06% respectively on the Cityscapes and CamVid traffic scene public datasets. It can achieve a trade-off between segmentation accuracy and model size in traffic scene semantic segmentation and has a good application prospect.

Keywords: semantic segmentation, deep learning, attention mechanism, lightweight, traffic scene

收稿日期：2023−07−04；**修回日期：**2023−09−20

通信作者：谢新林，xiexinlin@tyust.edu.cn

基金项目：国家自然科学基金资助项目（No.62006169）；山西省重点研发计划基金资助项目（No.202202010101005）；太原科技大学博士科研启动基金资助项目（No.20192047）

Foundation Items: The National Natural Science Foundation of China (No.62006169), Key Research and Development Plan of Shanxi Province (No. 202202010101005), Taiyuan University of Science and Technology Scientific Research Initial Funding (No.20192047)

0 引言

随着新一代人工智能技术的快速发展、可视化数据的激增以及 GPU（graphics processing unit）硬件设备的不断更新，以深度学习[1]为导向的图像语义分割逐渐成为热点研究内容之一。为了提高驾驶体验，缓解交通压力，自动驾驶技术得到了广泛的研究[2]。相比于单一的视觉任务，图像语义分割能够同时实现对目标的分割和识别，进而为场景理解提供细粒度和高层次的语义信息，在面向城市交通场景理解中起着至关重要的作用，已广泛应用于自动驾驶[3]、3D 点云[4]、智能交通系统[5]等领域。

近年来，卷积神经网络（CNN, convolutional neural network）广泛应用于交通场景语义分割[2-3,5]，利用卷积核的权重和局部感受野能够更好地将局部特征与整个特征图相关联。然而，卷积核的感受野存在局限性，导致难以捕获长距离特征依赖关系。而这些依赖关系通常在处理高分辨率的语义分割任务中至关重要，尤其是在交通场景中提取具有长距离特征的纤细条状目标（如交通标志、路灯、树木、行人等）的特征时，上下文信息的丢失更加明显。为此，Chen 等[6]提出 ASPP（atrous spatial pyramid pooling）模块，通过空洞卷积扩大感受野的方法来解决上下文信息丢失问题，同时进一步证明了丰富的上下文信息可以增强网络的类别区分度，从而更好地提高网络模型的语义分割能力。Dong 等[7]通过减少下采样来获得高分辨率的特征图，并在使用 ASPP 模块前串行加入通道注意力机制（CAM, channel attention mechanism）来更多地提取上下文信息。目前，最先进的 DeepLab-V3+网络模型[8]改进了 ASPP 模块，采用串行多级扩张率的空洞卷积来获取更大的感受野，同时避免池化所引起的图像分辨率低的问题。然而，这些方法未考虑多尺度信息对像素分类的影响以及空洞卷积容易出现梯度消失的问题，没有充分发挥不同阶段网络的特征提取优势，导致具有长距离特征的纤细条状目标的分割结果往往不连续。

针对上述问题，Lv 等[2]提出并行互补模块来增强具有适当补码信息的局部特征；Huynh 等[9]设计了细化模块，使用后一个分割图来细化前一个分割图的目标，以解决纤细条状目标分割不连续的问题；Weng 等[10]提出阶段感知特征对齐模块来对齐和聚合相邻的两层特征映射，以增强空间细节信息。与上述方法相比，所提算法的独特性在于将 Transformer 与 CNN 相融合，在 Transformer 的全局特征表示下通过 CNN 来增强局部特征的提取，并采用多分支不同尺度下的条形卷积（行卷积和列卷积）来聚合多尺度纤细条状特征信息，以增强网络对不同尺度的纤细条状目标的关注。

视觉 Transformer（ViT, vision transformer）首次使用纯 Transformer 结构作为特征提取器来处理图像识别任务[11]，在许多视觉任务上取得了与 CNN 同类模型相当甚至更好的性能，并已成功应用于交通场景语义分割。Transformer 主要对输入特征的上下文信息求加权平均，根据相关像素对之间的相似性函数动态计算自注意力权重。这种灵活的操作可以使注意力模块自适应地关注整个特征图的不同区域。然而，在捕获更多全局特征的同时，其也更容易丢失局部细节特征。为此，Liu 等[12]提出滑动窗口注意力，通过滑动的窗口设计将图像分块再重组，使每个窗口能够关注特征图不同区域的局部信息。为了更加精确地捕获特征图的局部特征，Wang 等[13]将金字塔结构引入 Transformer，对不同尺度的特征图进行特征提取来加强对局部信息的关注。然而，上述方法未考虑高分辨率特征图中存在的低级语义信息。低级语义信息对于局部目标边界细节信息具有较强的提取能力，丢失这些信息容易导致目标边缘分割不平滑。

针对上述问题，Peng 等[14]在 Transformer 主干网络中并行连续耦合 CNN 来解决局部细节信息丢失的问题；Hu 等[15]提出空间细节提取模块，捕获浅层网络多层次局部特征，弥补下采样阶段丢失的几何信息；Xiao 等[16]设计了边界细化模块，利用 Canny 检测器提取的粗略的低级边界特征来细化目标边界。相比上述方法，本文所提算法的独特性在于充分利用高分辨率特征图中的低级语义信息（如纹理等），并利用预训练模型和迁移学习策略来进一步提高模型的语义分割性能。

图像语义分割模型对自动驾驶、智能交通等系统中的有效部署至关重要，而编码器–解码器结构是一种有效的实现方法，该结构由编码器和解码器两部分组成。编码器通常对原图像进行卷积、池化和正则化等操作，构建下采样的主干网络，用于提取图像的高级语义信息；解码器则通过对编码结果进行线性插值、转置卷积或反池化等操作进行上采样，通过解码器能够得到与原图像相同尺寸和语义

信息的输出结果。U-Net[17]和 SegNet[18]采用编码器−解码器结构来实现快速推理。近年来，大部分语义分割方法采用非对称的编码器−解器码结构，以减少模型参数量和计算量。其中，Bai 等[19]采用紧凑的非对称编码器−解码器结构，并加入注意力机制来实现高质量的轻量级语义分割。Dong 等[7]通过采用轻量级主干网络 MobileNet-V2 和减少下采样的方法来降低模型参数量。虽然这些方法能够实时生成分割结果，但其分割精度往往不令人满意。现有的高精度语义分割方法[12-14]都有较高的分割精度，但庞大的模型参数量和计算量不利于在实际交通场景中的部署。因此，如何在分割精度和模型大小之间取得权衡是现有交通场景语义分割任务中重要的考量因素之一[20]。

基于上述讨论和动机，本文提出一种融合多尺度深度卷积的轻量级 Transformer 交通场景语义分割算法，算法框架如图 1 所示，该算法采用编码器−解码器结构，将 Transformer（全局特征处理）和 CNN（通用、易训练、局部特征处理）的各自优势相结合，以实现分割精度与模型大小之间的权衡。1) 编码器采用双分支设计，输入图像先经过语义分支中的 TB（transformer block）获得具有全局特征表示的特征映射，TB 结构如图 1 所示。然后，利用多尺度条形特征提取模块（MSEM, multi-scale strip feature extraction module）以较低的模型参数量来聚合多尺度纤细条状特征信息，以增强网络局部特征表示。而空间分支通过空间细节辅助模块（SDAM, spatial detail auxiliary module）能够保留更多的空间细节信息，以弥补深层空间细节的丢失。2) 解码器分别将语义分支和空间分支生成的高级与低级特征映射相融合，进一步建模全局上下文，并以较低的计算成本进行像素级推理。所提算法能够准确地分割易融入周围背景的纤细条状目标、优化目标边缘，且获得分割精度与模型大小之间的权衡。本文主要贡献如下。

1) 利用深度条形卷积替换标准卷积构建多尺度条形特征提取模块，通过多分支不同尺度下的条形卷积对特征图的提取与融合来增强网络对纤细条状目标的关注。

2) 为了对高分辨率特征图下的空间细节信息加以利用，设计了空间细节辅助模块来保留丰富的空间细节信息，进一步优化目标边缘达到更准确平滑的语义分割结果。

3) 提出基于 Transformer-CNN 的非对称编解码网络（TC-AEDNet, asymmetric encoding-decoding network based on the Transformer-CNN），编码器结合 Transformer 与 CNN 将局部细节信息与上下文信息相关联，同时降低模型参数量。解码器采用轻量级的多

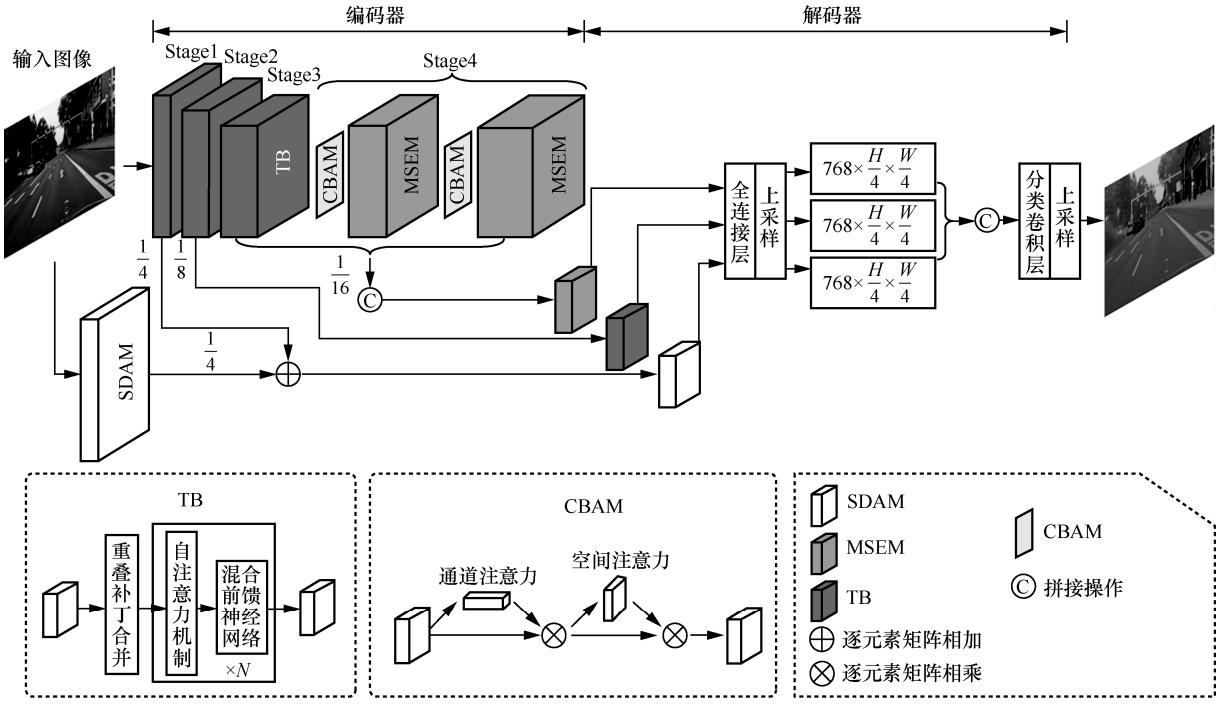

图 1　融合多尺度深度卷积的轻量级 Transformer 交通场景语义分割算法框架

级特征融合设计，分别将多级语义信息和空间细节信息相融合，以提高网络模型分割性能。

1 基于 Transformer-CNN 的轻量级算法框架

针对现有交通场景语义分割算法难以准确平滑地分割易融入周围背景的纤细条状目标、高精度模型参数量大等问题，本文提出融合多尺度深度卷积的轻量级 Transformer 交通场景语义分割算法，所提算法主要由三部分构成： MSEM、SDAM、TC-AEDNet。

1.1 多尺度条形特征提取模块

复杂的交通场景中存在各种各样的目标（如车辆、行人、建筑等），不同类型的目标占据的空间大小、形状和颜色也各不相同。纤细条状目标往往具有长距离和小尺寸的特点，更容易融入周围的背景中难以被准确地识别和分割出来，这是导致交通场景语义分割精度较低的主要原因之一。为了在交通场景中获得高精度分割结果的同时，保证模型轻量化，本文基于深度卷积构建 MSEM，利用多分支下的不同尺度大小的深度条形卷积（MDC, multi-scale depthwise convolution）来提取深层次特征和纤细条状目标特征，以获得更好的语义分割结果。MSEM 的机理在于多尺度卷积能够通过不同卷积核大小的卷积层来处理图像，较小的卷积核可以捕获局部细节信息，而较大的卷积核可以捕获更大范围的上下文信息。此外，多分支设计的 MDC 可以在不引入额外参数的情况下扩大感受野，该方法可以使网络在多尺度下提取更加丰富的语义信息。如表 1 所示，MDC 是由 3 组不同卷积核大小的深度条形卷积（行卷积和列卷积）构成的，卷积核大小分别设置为 7、9 和 11，有助于在多尺度下捕获更多的纤细条状目标特征。其中，深度条形卷积设置了 5 个参数，分别为输入特征图 Input（特征图通道数×特征图高度×特征图宽度）、Scale$_i$（MDC 的 3 个分支）、卷积核的大小 k、步长 s 和零填充的大小 p。

ConvNeXt 与 MSEM 结构如图 2 所示。MSEM 采用与 ConvNeXt[21] 相似的结构，不同的是将 ConvNeXt Block 中的 7×7 大核卷积部分进行改进，如图 2 中虚线框所示。ConvNeXt 在网络设计过程中忽略了多尺度卷积对特征提取的作用，而多尺度特征融合对于语义分割任务至关重要。因此，本文独特性地采用一对深度条形卷积（行卷积和列卷积）代替标准卷积来设计 MSEM。其独特之处如下：1) 使用 5×5 深度卷积细化局部特征；2) 基于多尺度深度卷积构建 MDC，较小的卷积核用于捕捉局部细节特征，而较大的卷积核用于捕捉全局特征，这样可以在不同尺度上提取不同级别的特征映射，并且深度条形卷积是轻量级的；3) 多分支特征融合，将每个分支分别获得的不同特征进行融合，从而得到更丰富的多尺度表示，使模型能够同时关注局部和全局特征。

表 1 MDC 的详细网络配置

MDC	Scale$_i$	Input	s	k	p
MDC$_1$	i=1	320×64×64	1	1×7 7×1	0×3 3×0
	i=2	320×64×64	1	1×9 9×1	0×4 4×0
	i=3	320×64×64	1	1×11 11×1	0×5 5×0
MDC$_2$	i=1	320×64×64	1	1×7 7×1	0×3 3×0
	i=2	320×64×64	1	1×9 9×1	0×4 4×0
	i=3	512×64×64	1	1×11 11×1	0×5 5×0

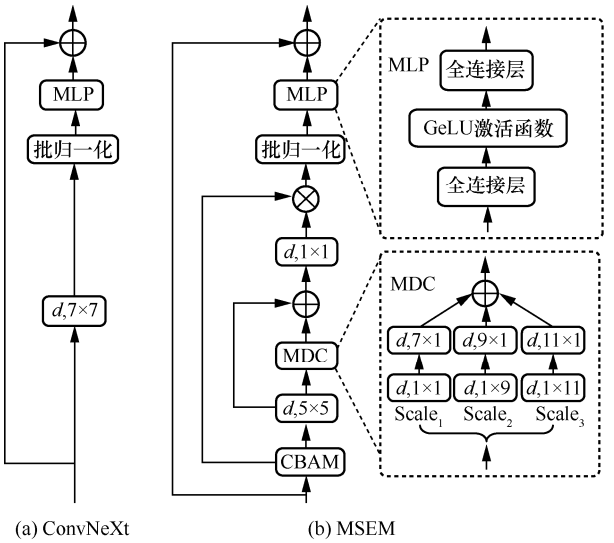

图 2 ConvNeXt 与 MSEM 结构

此外，特征映射的每个通道的重要性往往是不同的，不相关信息会极大地影响重要语义信息的提取，导致语义分割的性能下降。因此，MSEM 在特征图输入前进一步添加通道空间注意力模块（CBAM, convolutional block attention module）[22]，其目的是利用 CBAM 生成的权重来指导网络学习，从而使网络能够选择性地关注重要的目标信息（如信号灯、行人、车辆等）。如图 1 中 CBAM 的网络

结构所示，CBAM 首先采用通道注意力（CAM）获得注意力向量，并根据注意力向量对特征图的不同通道进行加权。然后，将 CAM 的输出作为空间注意力（SAM, spatial attention mechanism）的输入对特征图的空间区域进行加权。最后，通过自适应学习得到的权重来指导网络关注益于语义分割的重要信息。

如图 2(b)所示，MSEM 采用条形卷积来设计，获取更大感受野的同时，能够更多地捕获纤细条状目标的特征。MSEM Block 主要由三部分组成：首先，MSEM 处理的特征图是原图像的$\frac{1}{16}$，故采用 5×5 大核深度卷积用于聚合局部信息，并在使用大核卷积前添加 CBAM 以选择通道维度和空间维度中易于分割的重要信息。其次，纤细条状目标的长度是参差不齐的，因此采用一组深度条形卷积（行卷积和列卷积）来近似表示具有大核的标准卷积（例如，将卷积核大小为 7×7 的标准卷积替换成一对卷积核大小为 7×1 和 1×7 的条形卷积）用于捕获多尺度上下文信息，这样可以有效地提取图像中的纵向和横向目标信息，更有助于提取交通场景中的交通标志、路灯、树木等纤细条状目标特征。最后，使用 1×1 卷积中建模不同通道之间的关系，通过对 1×1 卷积的输出与 CBAM 的输出进行加权来减小语义鸿沟，以进一步改善分割结果。计算式如下

$$x_{\text{Attn}} = \text{Attn}(x_{\text{in}}) \tag{1}$$

$$x_{\text{MDC}} = \sum_{i=0}^{3} \text{Scale}_i\left(\text{DWConv}_{5\times5}(x_{\text{Attn}})\right) \tag{2}$$

$$x_{\text{MSEM}} = x_{\text{Attn}} \otimes \text{DWConv}_{1\times1}(x_{\text{MDC}}) \tag{3}$$

$$x_{\text{out}} = x_{\text{in}} \oplus \text{MLP}\left(\text{Norm}(x_{\text{MSEM}})\right) \tag{4}$$

其中，x_{in} 和 x_{out} 分别为特征图的输入和输出；Attn 表示注意力机制 CBAM；$\oplus$ 和 $\otimes$ 分别表示逐元素矩阵相加和逐元素矩阵相乘；x_{Attn} 和 x_{MDC} 分别表示特征图经过注意力机制和多分支结构设计 MDC 的输出；$\text{DWConv}_{k\times k}$ 表示深度卷积，$k\times k$ 为卷积核大小；$\text{Scale}_i\,(i=0,1,2,3)$ 表示图 2 中 MDC 的第 i 个分支，Scale_0 为 MDC 的残差结构；Norm 表示对输入的特征进行批归一化处理；MLP 表示图 2 中的多层感知机分类器（由 2 个全连接层和 GeLU 激活函数构成）。

1.2　空间细节辅助模块

高性能的语义分割网络模型不仅需要充足的语义信息来支撑，还需要足够精细的空间细节特征来丰富空间信息。对于 CNN 而言，网络不断加深的同时感受野增大，提取的语义信息也增加。现有的大部分语义分割方法通常对特征图多尺度下采样来丰富语义信息，却也丢失了一部分的空间信息，导致分割目标边缘不平滑。然而，空间信息中的低级语义特征在交通场景语义分割任务中对于生成平滑的目标边缘结果至关重要。其中，低级语义特征有助于表征局部结构模式和全局统计特性（如边界、平滑度、规则性），而高级语义特征可能无法很好地解决这些问题。因此，为了防止下采样操作导致的部分低级语义特征丢失，本文设计 SDAM 用于保存丰富的空间信息并生成高分辨率的特征图，通过优化空间路径来辅助网络获得更准确平滑的语义分割结果。为了易于训练和避免过多的计算，本文基于经典的 ResNet-18 网络模型[23]的前两层结构（layer_0 和 layer_1）来搭建 SDAM，并与编码器语义分支中 Stage[1]的输出相融合作为 SDAM 的最终输出，SDAM 结构如图 3 所示。

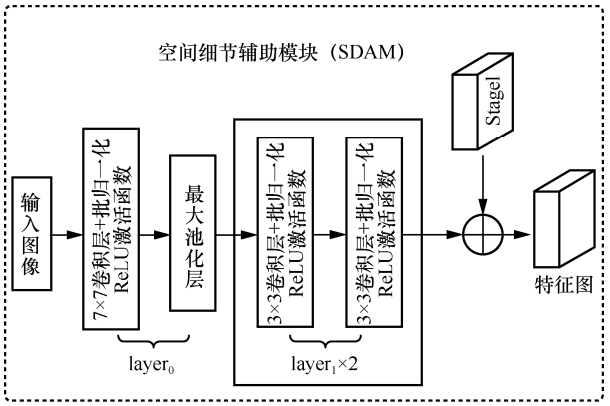

图 3　SDAM 结构

SDAM 的机理在于浅卷积层结构（3×3 卷积层+批归一化+ReLU 激活函数）能够更好地捕获图像中的空间信息，从而在处理高分辨特征图时可以编码比较丰富的空间信息。为了更好地提取空间细节信息，本文循环使用 2 次 layer_1。由于 ResNet 网络中的残差结构会先将分支中的输入结果保存在内存里，直到完成相加操作后才能释放，导致增加内存占用且效率低。因此，SDAM 中去掉残差结构来简化输出。为了方便与 Stage1 的输出相融合，SDAM 的最终输出特征映射为原输入图像大小的$\frac{1}{4}$，包含 64 个通道。虽然 SDAM 获得的特征图分辨率相对较大，但使用浅卷积层会使计算成本较小。此外，

SDAM 在训练过程中采用了迁移学习的方法，将在 ImageNet 数据集上训练过的 ResNet-18 网络所学习到的特征迁移到 SDAM 中，这有助于网络模型的快速收敛和更好地提高网络模型性能。计算式如下

$$\text{layer}_0 = \text{MaxPool}_{3\times3}\left(\text{Conv}_{7\times7}(I)\right) \tag{5}$$

$$\text{layer}_1 = \text{ReLU}\left(\text{Bn}\left(\text{Conv}_{3\times3}(\text{layer}_0)\right)\right) \tag{6}$$

$$x_{\text{SDAM}} = \text{repeat}\left\{\text{layer}_1\right\} \tag{7}$$

其中，$I \in \mathbb{R}^{C\times H\times W}$ 表示初始输入图像（C 为特征图的通道数，H 为特征图的高度，W 为特征图的宽度）；MaxPool 表示最大池化；Bn 表示批归一化；Conv 表示标准卷积；x_{SDAM} 表示特征图经过 SDAM 的最终输出；repeat$\{\ \}$ 表示多次执行相同的计算逻辑，通过将上一次的输出作为下一次的输入来实现重复执行。

1.3 基于 Transformer-CNN 的编解码网络

目前，基于 ViT 的语义分割方法往往具有较高的分割精度，这主要得益于 self-attention 能够更多地捕获上下文信息。然而，该方法也更加容易丢失局部细节信息且计算成本更高，尤其在交通场景中难以准确平滑地分割易融入周围背景的纤细条状目标。因此，针对交通场景中特有的纤细条状目标提取不充分的问题，本文结合 Transformer（全局特征处理）和 CNN（通用、易训练、局部特征处理）的各自优势提出基于 Transformer-CNN 的非对称编解码网络，其结构如图 4 所示。

TC-AEDNet 是基于 SegFormer 网络模型[24]进行改进的。SegFormer 拥有参数量较少、简单高效、鲁棒性强等特点，但在处理图像时会将特征图下采样至原输入图像大小的 $\frac{1}{32}$，导致最后一个阶段（Stage4）的参数量远大于前 3 个阶段（Stage1～Stage3）。这是因为基于纯 Transformer 的网络模型受到 self-attention 和下采样的影响，导致 Stage4 参数量较大且局部细节信息显著丢失，从而难以准确分割交通场景中易融入周围背景的纤细条状目标。因此，为了更好地应用于交通场景语义分割任务，本文对该网络进行了改进，具体如下。

1) TC-AEDNet 的编码器采用双分支结构，语义分支保留 SegFormer 网络编码主干的前 3 个阶段（Stage1～Stage3），以较少的参数量来捕获更多的全局信息依赖关系，并生成较大感受野的特征图。而空间分支通过 SDAM 从原输入图像中提取高分辨率特征以及保留 Transformer 丢失的位置信息，通过丰富空间信息来优化目标边缘分割。

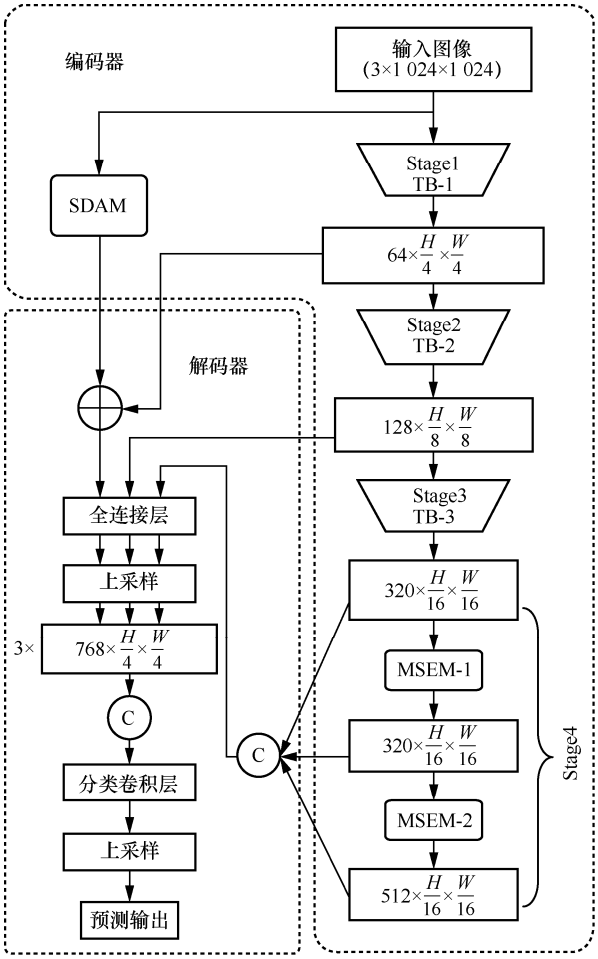

图 4 TC-AEDNet 结构

2) 对于特定的纤细条状目标针对性地构建 MSEM，在编码主干的 Stage4 用于关注不同尺度下的局部细节信息。其中，无论是基于 CNN 还是基于 Transformer 的网络模型，都需要对原图像不断进行下采样操作，以降低显存占用和增加不同尺寸目标特征的可区分性。然而，下采样操作往往会降低特征图像素的数量，从而导致丢失大量的上下文信息。如图 4 所示，为了避免过多的下采样导致纤细条状目标分割不连续问题，TC-AEDNet 进行了 3 次下采样 $\left(\frac{1}{4}、\frac{1}{8}、\frac{1}{16}\right)$，最终输出的特征图仅为输入图像的 $\frac{1}{16}$。而 MSEM 中的 MDC 可以通过不同尺度的深度条形卷积在高分辨率特征图上提取深层次特征和纤细条状目标特征。因此，TC-AEDNet 的编码器结合 Transformer 与 CNN 可以避免全局特

征的丢失，同时能够保持模型的高效性能。

3) TC-AEDNet 的解码器采用轻量级的多尺度特征融合方法分别将语义分支提取的深层特征和空间分支提取的浅层特征进行多尺度融合，进一步提高网络模型的性能。

本文基于 SegFormer-B2 的编码网络主干配置用于语义分割。其中，Stage1～Stage3 的 Transformer 具有相同的网络结构，仅注意力机制的参数不同。本文在 Stage1 ～ Stage3 中采用的 efficient self-attention 参数配置如下。次数 $N=[1,2,5]$，衰减因子 $R=[8,4,2]$，通道数 $C=[64,128,320]$。由于 Stage4 中注意力机制通过 patch embedding 操作将图像投影到一个向量中，更容易破坏图像原有的位置信息且计算量大，对于准确地分割交通场景中的纤细条状目标非常不利。相比于基于深度卷积的方法，卷积核在重叠的特征图上滑动，自身携带的归纳偏置提供了保留位置信息的可能性。此外，深度卷积还具有轻量级的特点。因此，本文在 SegFormer-B2 中的 Stage4 采用 MSEM，以增强对纤细条状目标特征的提取能力，且参数量更小。

TC-AEDNet 的编码器结构如图 4 中的编码器所示，本文先将原图像 I 分别输入编码器的 Transformer Block 和 SDAM 中，经过 Stage1～Stage3 分别生成不同尺寸大小 $\left(\frac{1}{4}、\frac{1}{8}、\frac{1}{16}\right)$ 的特征图，并分别表示为 $F_1 \sim F_3$；然后，将 F_1 和经过 SDAM 生成的特征图 x_{SDAM} 相融合生成 F_{SDAM}；同时，将 Stage3 输出的 F_3 输入 MSEM 中学习纤细条状目标特征，再次生成特征图 F_{MSEM}；最后，将编码器生成的多尺度特征图分别输入解码器进行融合。TC-AEDNet 编码器结构可表示为

$$F_1 = f^{\text{Transformer}}(I) \tag{8}$$

$$x_{\text{encoder}} = \begin{cases} F_{\text{SDAM}} = x_{\text{SDAM}} \oplus F_1 \\ F_{i+1} = f^{\text{Transformer}}(F_i), i = 1,2 \\ F_{\text{MSEM}} = f^{\text{MSEM}}(F_3) \end{cases} \tag{9}$$

TC-AEDNet 的解码器结构如图 4 中的解码器所示，解码器采用轻量级的上采样方法，先将 F_{MSEM} 和 F_3 进行 Concat 特征融合生成特征图 F_4；然后，通过 Linear 来改变通道数，再采用双线性插值分别将 F_4 和 F_2 上采样至与 F_1 尺寸相同，并分别生成特征图 F_5 和 F_6；最后，再次 Concat 特征融合 F_{SDAM}、F_5 和 F_6，并通过全连接层完成分类和输出原尺寸的分割结果 x_{encoder}。TC-AEDNet 解码器结构可表示为

$$F_4 = \text{Concat}(F_{\text{MSEM}}, F_3) \tag{10}$$

$$(F_5, F_6) = \text{Linear}\big(\text{Up}(F_4, F_2)\big) \tag{11}$$

$$x_{\text{decoder}} = \text{Up}\Big(\text{Linear}\big(\text{Concat}(F_6, F_5, F_{\text{SDAM}})\big)\Big) \tag{12}$$

其中，x_{encoder} 和 x_{decoder} 分别表示编码器和解码器的最终输出；$f^{\text{Transformer}}$ 和 f^{MSEM} 分别表示通过 Transformer 和 MSEM 学习得到的目标特征；Concat 表示相同尺寸大小的特征图在通道维度上进行拼接；Up 表示双线性插值方法的上采样；Linear 表示全连接层。

2　实验结果与对比分析

2.1　实验环境及参数配置

本文使用深度学习框架 PyTorch 来实现，实验配置如下：操作系统为 64 bit-Ubuntu20.04；数据处理环境为 Python 3.8；CUDA 版本为 10.1；深度学习框架为 Pytorch 1.7.1。本文仅使用单个 16 GB NVIDIA Tesla T4 的显卡（GPU）来训练和评估所提算法，且所有网络模型都经过 200 epoch 的训练。

网络训练时的超参数如下：网络训练采用的是 AdamW 优化器来调整模型参数和更新梯度，优化器动量 β_1 和 β_2 分别设置为 0.9 和 0.999，权重衰减为 10^{-4}；初始学习率（lr, learning ratio）设置为 0.000 05，学习率调整策略采用默认系数为 1.0 的 poly 方式，最大迭代次数设置为 160 000，随着训练迭代次数的增加，lr 会先下降，然后上升到一个较小的值，以使模型在训练后期能够做出更好的预测。同时，本文使用交叉熵损失函数和 Dice 损失函数，这 2 种损失函数直接被加权求和作为总的损失函数。其中，Dice 损失函数用于促进模型对小物体的检测，而交叉熵损失函数用于加强模型对大物体的检测。在网络训练时，采用了迁移学习的方法，将在 ImageNet 数据集上训练过的 SegFormer-B2 网络学习到的特征迁移到所提网络 TC-AEDNet 中，获得了较快的收敛速度。

2.2　数据集与评价指标

2.2.1　数据集

CamVid 数据集[25]是具有目标类别语义标签的城市道路场景视频集合。该数据集共由 701 幅逐像素标注的图像组成，分别包括训练集、测试集和验

证集，共 32 个语义类别真实标签。本文选取其中常用的 11 个语义类别，并采用分辨率为 960 像素×720 像素的图像进行网络训练。

Cityscapes 数据集[26]主要来自城市街道环境下的驾驶场景，包含 20 000 幅粗略注释图像和 5 000 幅精细标注的像素级图像。本文选取精细标注的图像进行语义分割评估，并采用分辨率为 1 024 像素×1 024 像素的图像来进行网络训练，这是因为 Transformer 处理的图像需要裁剪为正方形图像才能够进行图像切块处理。交通场景数据集对比如表 2 所示。

2.2.2 评价指标

为了更好地评估模型的性能，本文采用像素准确率（PA, pixel accuracy）、平均像素准确率（mPA, mean pixel accuracy）、平均交并比（mIoU, mean intersection over union）对模型分割精度进行评估，以及采用参数量（Params, parameter number of a model）对模型大小进行评估。

1) 像素准确率是正确分类的像素点与像素点总和的比值。计算式如下

$$\mathrm{PA}=\frac{\sum_{i=0}^{k}P_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}P_{ij}} \tag{13}$$

2) 平均交并比是各类别预测分割区域和真实分割区域间交集与并集比值的平均值，是图像语义分割领域的标准性能衡量指标。计算式如下

$$\mathrm{mIoU}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}+\sum_{j=0}^{k}P_{ji}-P_{ii}} \tag{14}$$

其中，$k+1$ 为图像类别总数，P_{ii} 为真实标签为 i 被预测为 i 类的像素数量，P_{ij} 为真实标签为 i 并被错误预测为 j 类的像素数量，P_{ji} 为真实标签为 j 被错误预测为 i 类的像素数量。

2.3 消融实验

为了验证 MSEM、SDAM 和 TC-AEDNet 的有效性，模块之间是否存在相互影响，本文在城市场景数据集 Cityscapes 上进行了消融实验。

2.3.1 MSEM 对网络性能的影响

本文 TC-AEDNet 与原网络 SegFormer-B2 模型进行对比，并验证了 SegFormer 中所提到的单个 Stage4 参数量和分割性能。本文还将 SegFormer 中的 Stage4 分别替换为 ASPP 和 MSEM，并对其分割性能的影响进行了比较，分别表示为 Stage1～Stage3+ASPP 和 Stag1～Stage3+MSEM。不同编码器结构的性能对比如表 3 所示。

如表 3 中 SegFormer-B2 所示，在解码器相同的情况下，Stage4 以远大于 Stage1～Stage3 的参数量来获取丰富的上下文信息。然而，由于 Stage4 被下采样至原图像的 $\frac{1}{32}$，这对于交通场景的纤细条状目标的分割十分不友好，导致分割精度显著下降。因此，本文将 SegFormer 中的 Stage4 分别替换成 ASPP 和 MSEM 并在保持 Stage3 输入的尺寸大小（即原图像的 $\frac{1}{16}$）进行处理，并生成新的网络框架 TC-AEDNet 来解决 Transformer 丢失的局部细节信息。在 TC-AEDNet 中采用 ASPP，受到空洞卷积的影响提取的上下文信息不充分，在分割交通场景中的纤细条状目标时容易出现不连续的情况。与 ASPP 相比较，采用本文的 MSEM，通过多尺度深度条形卷积（行卷积和列卷积）来提取特征，能够

表 2 交通场景数据集对比

数据集	图像大小	类别	训练集/幅	测试集/幅	验证集/幅
CamVid	960 像素×720 像素	11	367	100	233
Cityscapcs	2 048 像素×1 024 像素	19	2 975	500	1 525

表 3 不同编码器结构的性能对比

算法	编码器	参数量	PA	mPA	mIoU
SegFormer-B2	Stage1～Stage4	27.36×10^6	96.06%	86.15%	77.39%
	Stage4	24.56×10^6	95.74%	81.83%	73.46%
TC-AEDNet	Stage1～Stage3+ASPP	15.31×10^6	95.83%	82.70%	74.40%
	Stage1～Stage3+MSEM	18.55×10^6	96.18%	84.53%	77.49%

增强对纤细条状目标的关注，并获得的上下文信息更加丰富，且以较小的参数量超过了 SegFormer-B2 的分割精度。实验验证了 MSEM 能够有效扩大感受野，同时增强对纤细条状目标的特征提取能力。

为了验证不同注意力机制对 MSEM 分割性能的影响，本文设立了 4 组对比实验，MSEM 中不同注意力机制的性能对比如表 4 所示。相比于实验 1，加入注意力机制模块的实验（实验 2～实验 4）各项指标均得到提升，验证了在 MSEM 中加入注意力层能够增强特征提取能力。实验 2 和实验 3 分别验证了 CAM 和 SAM 对语义分割性能的影响。其中，CAM 将一个通道内的信息直接进行全局处理，通常会忽略非常重要的空间位置信息，而 SAM 更容易忽略通道间的信息交互。两者相比较，加入 CAM 的性能比 SAM 更好，主要是因为对 Transformer 处理后的特征图通过通道选择重要的信息能够更好地增强特征表示，从而有效提高语义分割性能。实验 4 采用 CBAM 先在通道维度进行关注，再在空间维度进行关注，从而实现对特征图的自适应调整。实验 4 的各项性能指标达到了最好的结果，验证了 MSEM 融合注意力机制模块更有助于提高语义分割的性能。

表 4　MSEM 中不同注意力机制的性能对比

实验序号	注意力机制	参数量	PA	mPA	mIoU
1	无	18.15×10^6	96.00%	82.58%	75.10%
2	SAM	18.53×10^6	95.17%	84.05%	75.57%
3	CAM	18.15×10^6	96.06%	84.11%	76.37%
4	CBAM	18.55×10^6	96.18%	84.53%	77.49%

2.3.2　SDAM 对网络性能的影响

本文将 TC-AEDNet+SDAM 表示为 SDAM 与 TC-AEDNet 并行输出，并通过常用的逐元素相加策略将其特征图相融合。这种方法实现了在同一特征维度下增加信息量的目的，这使在全局特征表示下的局部空间信息得到了增强。相比于 Concat 策略，逐元素相加不仅显存占用和计算量更低，而且能够提高模型的性能。融合与不融合 SDAM 的性能对比如表 5 所示。其中，↑表示加入 SDAM 后各评价指标绝对提升值。在 SegFormer-B2 和 TC-AEDNet 的基础上加入 SDAM，各项性能指标也均得到提升，且模型参数量仅增加了 0.09×10^6。其中，SegFormer-B2 和 TC-AEDNet 的编码器采用基于 Transformer 的多尺度深度网络，在下采样操作中不可避免地丢失了大量空间细节信息，而且 Transformer 架构中的图像切块处理操作又丢失了部分位置信息。相比之下，SDAM 能够利用卷积在浅层网络中特有的感受野优势和归纳偏置来辅助网络保留大量的空间细节信息和位置信息。实验表明，SDAM 能够提供丰富的空间细节信息，这使交通场景中细微且重要的边界细节信息得到了增强，验证了融合 SDAM 能够有效地提高分割精度。

表 5　融合与不融合 SDAM 的性能对比

算法	参数量	PA	mPA	mIoU
SegFormer-B2	27.36×10^6	96.06%	86.15%	77.39%
SegFormer-B2+SDAM	27.45×10^6	96.14% (↑0.08%)	86.64% (↑0.49%)	78.43% (↑1.04%)
TC-AEDNet	18.56×10^6	96.05%	86.09%	77.91%
TC-AEDNet+SDAM	18.65×10^6	96.16% (↑0.11%)	87.21% (↑1.12%)	78.63% (↑0.72%)

2.3.3　TC-AEDNet 的解码器对网络性能的影响

本文将 Stage1～Stage4 同时上采样至原图像的 $\frac{1}{4}$ 再进行特征融合与先将 Stage3～Stage4 进行特征融合再与 Stage1～Stage2 多尺度特征融合（用 Stage[1, 2, 3-4]表示）对于解码器输出结果的影响进行比较。多级语义信息融合用于解码器的性能对比如表 6 所示，2 种特征融合方式相比较，融合 Stage[1, 2, 3-4]比直接融合 Stage1～Stage4 的分割精度提高了 0.42%。这主要是因为 Stage3～Stage4 中分别包含来自 Transformer 和 CNN 处理得到的特征图，两者相融合能够有效地将 Transformer 捕获长距离依赖的优势以及 CNN 建模局部关系的优点相结合。实验验证了多级特征融合方法能够兼顾 Transformer 和 CNN 模型的优点，并有效地融合了不同层次的特征信息。上述实验表明 TC-AEDNet 对交通场景语义分割的性能提高是有效的。

表 6　多级语义信息融合用于解码器的性能对比

特征融合	参数量	PA	mPA	mIoU
Stage1～Stage4	18.55×10^6	96.18%	84.53%	77.49%
Stage[1, 2, 3-4]	18.56×10^6	96.05%	86.09%	77.91%

2.4　网络性能分析与比较

为了进一步验证所提算法的泛化性和有效性，将所提算法 TC-AEDNet 在 Cityscapes、CamVid 数

据集上与 FCN[27]、SegNet[18]、RtHp[7]、DeepLab-V3+[8]、Swin-B[12]、SegFormer-B2[24]、MagNet[9]、JPANet[15]等现有语义分割网络模型的性能分别进行对比。

2.4.1 Cityscapes 数据集上的实验结果

不同算法在 Cityscapes 数据集上的性能对比如表 7 所示，TC-AEDNet 在精度评价指标 PA、mPA 及 mIoU 下均取得了较好的结果。与经典算法 1、2 对比，FCN 和 SegNet 过度的下采样损失了大量空间信息，导致分割精度较低；与算法 3～算法 5 对比，RtHp、MagNet、JPANet 虽然具有较小的模型参数量，但也牺牲了较大的分割精度；与现有高精度算法 6～算法 8 对比，TC-AEDNet 采用 Transformer 与 CNN 的融合结构在提高精度的同时，降低了参数量，能够充分提取图像的语义信息和空间信息，并取得更高的分割精度。

为了直观地验证所提网络的有效性，对 TC-AEDNet 与高精度网络模型 DeepLab-V3+、SegFormer-B2 在 Cityscapes 数据集上的分割结果进行比较，如图 5 所示，图 5 中方框标记为不同算法针对纤细条状目标的分割结果。如图 5(c)所示，尽管

表 7 **不同算法在 Cityscapes 数据集上的性能对比**

序号	算法	参数量	PA	mPA	mIoU
1	FCN	18.64×10^6	94.15%	63.12%	55.32%
2	SegNet	29.45×10^6	93.53%	64.49%	57.01%
3	MagNet	6.37×10^6	94.23%	69.71%	68.20%
4	RtHp	6.20×10^6	94.95%	78.82%	73.67%
5	JPANet	3.49×10^6	94.53%	78.42%	72.43%
6	DeepLab-V3+	58.75×10^6	95.67%	83.49%	75.04%
7	Swin-B	88.23×10^6	96.30%	85.74%	78.54%
8	SegFormer-B2	27.36×10^6	96.06%	86.15%	77.91%
9	TC-AEDNet+SDAM	18.65×10^6	96.16%	87.21%	78.63%

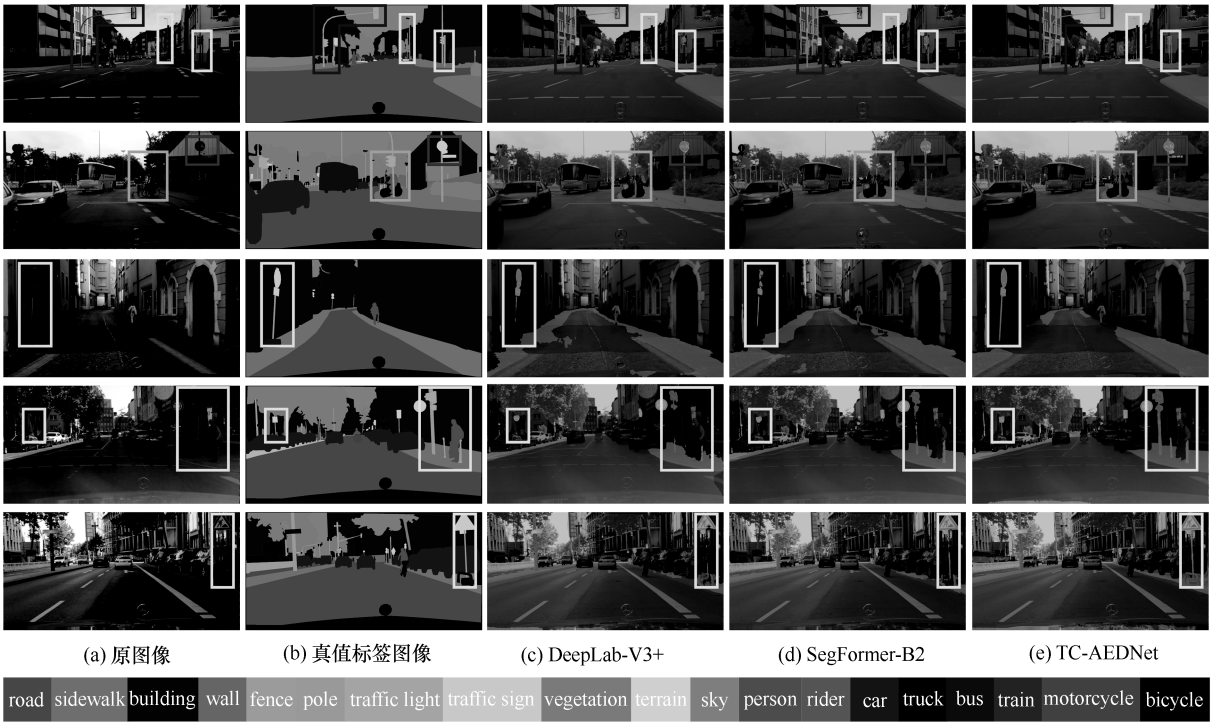

图 5 不同算法在 Cityscapes 数据集上的分割结果

DeepLab-V3+能够识别路灯杆、树干、行人等小目标，但纤细条状目标的分割结果往往是不连续的。如图 5(d)所示，SegFormer-B2 整体的分割结果较准确，能够准确分割显著的目标，但纤细条状目标较多时无法实现精确分割，且目标边缘分割不平滑。如图 5(e)所示，TC-AEDNet 能够连续分割纤细条状目标且边缘分割较平滑。从图 5 中第 2 行信号灯与行人交界处可以看出，DeepLab-V3+、SegFormer-B2 没有将行人与交通灯准确分割出来，TC-AEDNet 能够准确分割易融入周围背景的纤细条状目标。从图 5 中第 2～3 行纤细条状图标可以看出，DeepLab-V3+、SegFormer-B2 无法准确分割横向的条状目标，而 TC-AEDNet 准确地分割了十字路口的信号灯以及带有横向条状特征的交通标志物，且目标边缘分割结果较平滑。根据以上分析，验证了 TC-AEDNet 能够有效分割纤细条状目标、目标边缘以及其他细节信息。

2.4.2　CamVid 数据集上的实验结果

不同算法在 CamVid 数据集上的性能对比如表 8 所示，对比算法 2、算法 5、算法 6 可知，所提算法 TC-AEDNet 具有较少的参数量，且能够实现更好的分割精度。SegNet 算法中采用记录最大池化索引的方法，DeepLab-V3+通过设计 ASPP 模块的方法，SegFormer 通过设计高效的 self-attention 的方法，来增强网络对上下文信息的提取。尽管上述网络模型对精度的提升是有效的，但没有考虑到交通场景分割任务中存在的易融入周围背景的纤细条状目标特征。TC-AEDNet 通过 MSEM 中的多尺度条形卷积丰富了的纤细条状目标特征信息，基于深度卷积方法来减少参数量，设计 SDAM 以解决边缘信息丢失的问题，对比结果进一步证明了 TC-AEDNet 的有效性。

为了直观地验证本文所提算法的有效性，将所提算法 TC-AEDNet 与对比算法 DeepLab-V3+、SegFormer-B2 在 CamVid 数据集上的结果进行定性分析。图 6 中第 1 行所示，DeepLab-V3+未能准确分割行人与路灯，SegFormer-B2 未能连续分割路灯且目标边缘分割不平滑，只有 TC-AEDNet 将行人与路灯连续平滑地分割了出来。如图 6 中第 2～4 行所示，存在易融入周围背景的纤细目标较多时，DeepLab-V3+和 SegFormer-B2 中出现了错误分割的问题，而 TC-AEDNet 能比较准确地分割纤细条状目标。根据以上分析，验证了所提算法 TC-AEDNet 能够有效地优化纤细条状目标的分割结果，更好地提升图像语义分割性能。

3　结束语

本文提出一种融合多尺度深度卷积的轻量级 Transformer 交通场景语义分割算法 TC-AEDNet。所提算法改善了交通场景中易融入环境背景的纤细条状目标分割不连续以及目标边缘分割不平滑的问题，能够权衡交通场景语义分割算法的精度和模型大小。在 Cityscapes 和 CamVid 数据集上的实验结果表明，所提算法的语义分支能够充分发挥对纤细条状目标特征提取优势，结合 Transformer 模型对全局关注和远程依赖建模的优势和 CNN 模型对局部上下文提取和计算效率的优势来共同建模全局和局部关系，并降低了模型参数量，而空间分支能够提高轻量级模型的边缘分割准确率。最后，将语义分支和空间分支的多级特征相融合，进一步提高了模型分割精度。所提算法在分割纤细条状目标、平滑分割目标边缘以及降低模型参数量等方面具有优势。

参考文献：

[1] 周鑫, 何晓新, 郑昌文. 基于图像深度学习的无线电信号识别[J]. 通信学报, 2019, 40(7): 114-125.
ZHOU X, HE X X, ZHENG C W. Radio signal recognition based on image deep learning[J]. Journal on Communications, 2019, 40(7): 114-125.

[2] LV Q X, SUN X, CHEN C R, et al. Parallel complement network for real-time semantic segmentation of road scenes[J]. IEEE Transactions

表 8　**不同算法在 CamVid 数据集上的性能对比**

序号	算法	参数量	PA	mPA	mIoU
1	FCN[13]	18.64×10^6	91.89%	74.05%	64.75%
2	SegNet[15]	29.45×10^6	89.38%	74.25%	65.60%
3	RtHp[7]	6.2×10^6	93.86%	78.87%	68.14%
4	JPANet[25]	3.49×10^6	93.44%	78.27%	67.45%
5	DeepLab-V3+[8]	58.75×10^6	94.22%	85.16%	78.03%
6	SegFormer-B2[18]	27.36×10^6	95.11%	87.34%	80.55%
7	TC-AEDNet	18.65×10^6	96.47%	87.64%	81.06%

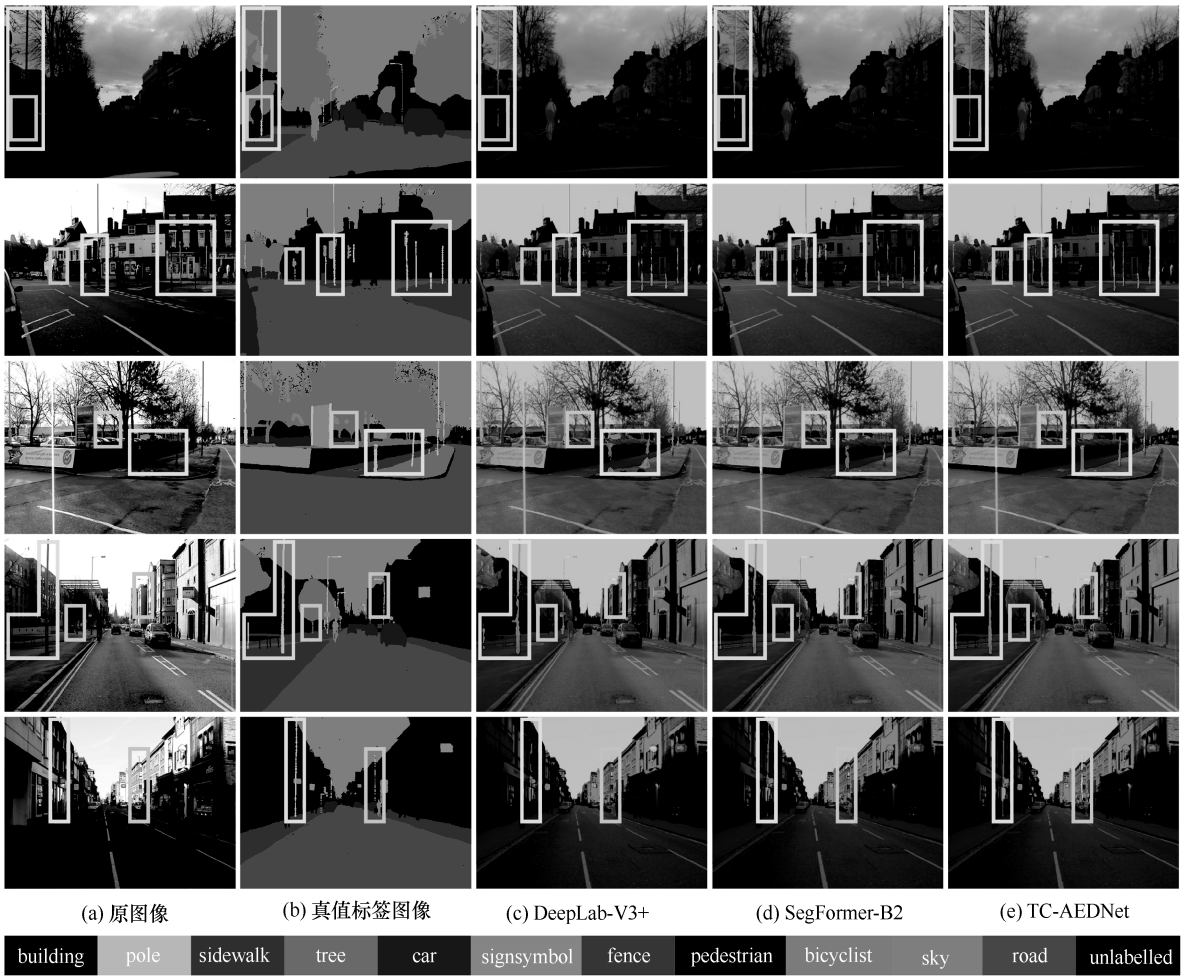

图 6　不同算法在 CamVid 数据集上的分割结果

on Intelligent Transportation Systems, 2022, 23(5): 4432-4444.

[3] 李琳辉, 钱波, 连静, 等. 基于卷积神经网络的交通场景语义分割方法研究[J]. 通信学报, 2018, 39(4): 123-130.
LI L H, QIAN B, LIAN J, et al. Study on traffic scene semantic segmentation method based on convolutional neural network[J]. Journal on Communications, 2018, 39(4): 123-130.

[4] 杨军, 党吉圣. 基于上下文注意力 CNN 的三维点云语义分割[J]. 通信学报, 2020, 41(7): 195-203.
YANG J, DANG J S. Semantic segmentation of 3D point cloud based on contextual attention CNN[J]. Journal on Communications, 2020, 41(7): 195-203.

[5] CHEN B K, GONG C, YANG J. Importance-aware semantic segmentation for autonomous vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(1): 137-148.

[6] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[7] DONG G S, YAN Y, SHEN C H, et al. Real-time high-performance semantic image segmentation of urban street scenes[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(6): 3258-3274.

[8] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 801-818.

[9] HUYNH C, TRAN A T, LUU K, et al. Progressive semantic segmentation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 16750-16759.

[10] WENG X, YAN Y, CHEN S, et al. Stage-aware feature alignment network for real-time semantic segmentation of street scenes[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(7): 4444-4459.

[11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[J]. arXiv Preprint, arXiv: 2010.11929, 2020.

[12] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 9992-10002.

[13] WANG W H, XIE E Z, LI X, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolu-

tions[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 548-558.

[14] PENG Z L, HUANG W, GU S Z, et al. Conformer: local features coupling global representations for visual recognition[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 357-366.

[15] HU X G, JING L Y, SEHAR U. Joint pyramid attention network for real-time semantic segmentation of urban scenes[J]. Applied Intelligence, 2022, 52(1): 580-594.

[16] XIAO X, ZHAO Y, ZHANG F, et al. BASeg: Boundary aware semantic segmentation for autonomous driving[J]. Neural Networks, 2023, 157: 460-470.

[17] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241.

[18] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[19] BAI W. An ENet semantic segmentation method combined with attention mechanism[J]. Computational Intelligence and Neuroscience, 2023, 2023: 1-9.

[20] PAN H H, HONG Y D, SUN W C, et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(3): 3448-3460.

[21] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 11966-11976.

[22] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of Computer Vision - ECCV 2018: 15th European Conference. New York: ACM Press, 2018: 3-19.

[23] WU Z F, SHEN C H, VAN DEN HENGEL A. Wider or deeper: revisiting the ResNet model for visual recognition[J]. Pattern Recognition, 2019, 90(C): 119-133.

[24] XIE E, WANG W, YU Z, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[25] 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.

[26] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 3213-3223.

[27] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 3431-3440.

[作者简介]

谢刚（1972− ），男，山西五台人，博士，太原科技大学教授、博士生导师，主要研究方向为计算机视觉、智能控制等。

王荃毅（1999− ），男，山西长治人，太原科技大学硕士生，主要研究方向为语义分割、深度学习等。

谢新林（1990− ），男，山西运城人，博士，太原科技大学副教授、硕士生导师，主要研究方向为语义分割、深度学习等。

王健安（1984− ），男，江西九江人，博士，太原科技大学教授、硕士生导师，主要研究方向为智能信息系统、复杂网络控制等。