DOI: 10.13382/j.jemi.B2407866

基于双分支多尺度特征融合的跨模态语义分割算法*

陈广秋 任天蓉 段 锦 黄丹丹
(长春理工大学电子信息工程学院 长春 130022)

摘 要:针对单模态可见光 RGB 图像语义分割在夜晚或光线变化环境下存在分割效果差、目标边缘分割不清晰等问题,以及现有的跨模态语义分割在获取全局上下文信息和融合跨模态特征时还存在大量不足。为此提出了一种基于双分支多尺度特征融合的跨模态语义分割算法。采用 Segformer 作为主干网络提取特征,捕获长距离依赖关系,采用特征增强模块提升浅层特征图的对比度和边缘信息的判别性,利用有效注意力增强模块和跨模态特征融合模块,对不同模态特征图像素点间的关系进行建模,聚合互补信息,发挥跨模态特征优势。最后,采用轻量级的 All-MLP 解码器重建图像,预测分割结果。相比较于已有主流算法,该算法在 MFNet 城市街景数据集上的各项评估指标均为最优,平均准确率(mAcc)和平均交并比(mIoU)分别达到了 76.9% 和 59.8%。实验结果表明,该算法在处理复杂场景时,能够有效改善目标边缘轮廓分割不清晰的问题,提高图像的分割精度。

关键词: 多模态深度学习;语义分割;特征融合;跨模态;Segformer
中图分类号: TP391.41; TN215 **文献标识码**: A **国家标准学科分类代码**: 510.40

Cross-modal semantic segmentation algorithm based on dual-branch multi-scale feature fusion

Chen Guangqiu Ren Tianrong Duan Jin Huang Dandan
(School of Electronic Information Engineering, Changchun University of Science and Technology, Changchun 130022, China)

Abstract: To solve the problems of poor segmentation effect and unclear target edge segmentation of single-modal visible RGB image semantic segmentation at night or in the environment of light change, and there are still many shortcomings in the existing cross-modal semantic segmentation networks when obtaining global context information and fusing cross-modal features. This paper proposed a cross-modal semantic segmentation algorithm based on dual-branch multi-scale feature fusion. The Segformer is used as the backbone network to extract features and capture long-distance dependencies. The feature enhancement module is used to improve the contrast of shallow feature maps and the discrimination of edge information. The effective attention enhancement module and cross-modal feature fusion module are used to model the relationship between pixels of different modal feature maps, aggregate complementary information, and give full play to the advantages of cross-modal features. Finally, the lightweight All-MLP decoder was used to reconstruct the image and predict the segmentation result. Compared with the mainstream algorithms in the existing literature, the proposed algorithm has the best evaluation indicators on the MFNet urban street view dataset, and the mAcc and mIoU reach 76.9% and 59.8% respectively. Experimental results show that the proposed algorithm can effectively improve the problem of unclear target edge contour segmentation and improve the accuracy of image segmentation when dealing with complex scenes.

Keywords: multimodal deep learning; semantic segmentation; feature fusion; cross-modal; Segformer

0 引 言

图像语义分割是指对像素点按其属性进行分类,获得目标的大小、形状、位置等信息,从而将视觉场景分解为不同语义类别实体,实现对图像的细粒度理解和分析[1]。目前,语义分割在多个领域都有广泛应用,如自动驾驶、地质监测、道路安全检测、军事侦察等[2]。

收稿日期: 2024-09-29 Received Date: 2024-09-29
*基金项目:国家自然科学基金重大仪器专项(62127813)、吉林省科技发展计划项目(20210203181SF)资助

传统的图像语义分割方法是根据图像本身的特征，通过人工预先设定的规则来划分区域，将图像中具有相似特性的像素划分为一个类别，主要包括基于阈值、基于边缘检测以及基于区域的分割方法[3-5]。这类算法参数少，复杂度低，但只用到了图像表层信息，无法实现端到端的图像语义分割，面对较为复杂的图像分割任务时，在一些细节处，人工设定的分割规则不能根据图像特征自动进行调整，导致分割效果并不理想，鲁棒性差。

随着深度学习技术的发展，其强大的自动特征学习和数据拟合能力在图像语义分割领域表现出了更好的性能和泛化能力。基于深度学习的图像语义分割主要包括特征编码与增强、像素或区域分类及边界优化与精细处理 3 个步骤。主要包括基于区域的图像语义分割方法、全监督学习图像语义分割方法和弱监督学习图像语义分割方法[6]。基于深度学的图像语义分割方法能够深度挖掘图像的像素特征，自动学习图像的高层语义信息，根据图像自身的场景推理出图像所表达的信息，随着网络模型的不断优化，在分割准确度和效率方面要优于传统方法。

上述图像语义分割方法都是基于单模态可见光 RGB 图像实现的，受限于其成像机理，可见光相机难以在不良光照条件以及恶劣天气情况下捕获足够有效的场景信息，导致在分割时难以区分有着相似外观或形状的不同类别物体，对于尺度过小的物体，难以识别出其具体轮廓[7]。为了克服上述缺陷，近年来学者们逐渐关注利用红外与可见光 RGB 图像进行跨模态语义分割。红外成像传感器对热辐射敏感，能够在低照度和恶劣天气环境中获取目标信息，而可见光成像传感器能够捕捉到丰富的颜色和细节信息[8]，为了充分利用两者的互补信息，弥补各自在特定场景下的不足，2017 年，Ha 等[9]提出了基于卷积神经网络的红外与可见光 RGB 跨模态语义分割算法 MFNet，并公开了 RGB-Thermal 数据集；2019 年 Sun 等[10]针对城市场景提出了跨模态语义分割算法 RTFNet；2020 年，Shivakumar 等[11]设计了一种双路卷积神经网络(convolutional neural networks, CNN)结构来融合红外图像与可见光 RGB 图像，提出了语义分割算法 PSTNet；2021 年 Deng 等[12]提出了在融合红外与可见光 RGB 图像前采用空间注意力和通道注意力提取不同模态图像互补信息的语义分割算法 FEANet；2022 年，Dong 等[13]提出了通过图像的边缘信息指导特征融合的语义分割算法 EGFNet。2023 年，Zhang 等[14]提出一种通过计算不同通道的信息权重得到红外与可见光 RGB 图像间互补特征的语义分割方法 MS-IRTNet；2023 年 Yi 等[15]提出了一种使用轻量级模型 MobileNetv2 作为主干网络，采用自适应特征融合策略的语义分割方法 HAFFseg，轻量级模型显著提升了推理速度，但在处理复杂场景时，算法捕捉上下文信息的能力不足，牺牲了分割的精确度。

上述基于跨模态的语义分割算法都是采用卷积神经网络作为网络模型的主体框架，取得了很好的分割性能，但由于 CNN 感受野受限，只能提取局部特征信息，影响后续特征提取的可区分性，从而导致分割性能受到限制。为解决 CNN 对全局信息的忽视，Dosovitskiy 等[16]提出具有多层 Transformer 结构的 ViT(visual transformer)，利用自注意力机制对序列间的关系进行建模，增强获取全局上下文信息的能力。Xie 等[17]提出基于 Transformer 的 Segformer 算法，通过改进自注意力和使用多层感知机(multilayer perceptron, MLP)解码器，在显著减少参数量的同时保证了语义分割精度。但现有的基于 Transformer 的语义分割网络大多针对单模态可见光图像，没有考虑到跨模态特征的融合，在夜间或光照变化环境下分割精度受限。

针对上述问题，本文提出了一种基于双分支多尺度特征融合的跨模态语义分割算法，命名为 DMSFNet，采用可提取多尺度特征的 Segformer[17]作为主干网络，采用特征增强模块和有效注意力增强模块来加强网络对跨模态之间互补特征信息的提取能力，利用跨模态特征融合模块，对两种模态图像进行多尺度融合，生成包含跨模态信息的特征图。最后利用 MLP 解码器聚合多级特征并预测分割掩码，得到语义分割图像。

1　DMSFNet 网络模型

1.1　网络架构

本文提出的 DMSFNet 网络采用编-解码架构，如图 1 所示。编码器包括两个分支的 Segformer 主干网络、特征增强模块、有效注意力增强模块和跨模态特征融合模块，解码器采用轻量级的 All-MLP 解码器，降低计算的复杂度。

在编码器部分，主干网络对输入图像由浅至深的提取不同尺度的特征图，第 1 个尺度的浅层特征图通过特征增强模块(FEM)，增强图像中细节特征的可判别性，提升浅层特征的空间信息表达能力，增强后的特征图和其他尺度的特征图统一表示成 $F_i^j, j \in \{1,2,3,4\}$ 为尺度索引，$i \in \{rgb, thermal\}$ 为图像模态索引；有效注意力增强模块(EAEM)和跨模态特征融合模块(CFFM)对各尺度特征图进行特征增强和融合，合并不同模态特征图间的互补信息，融合后各尺度特征图表示为 F_{Merged}^j。

在解码器部分，利用轻量化的 MLP 解码器，将融合后的各尺度特征图分别进行上采样、级联、融合和预测，得到语义分割结果。

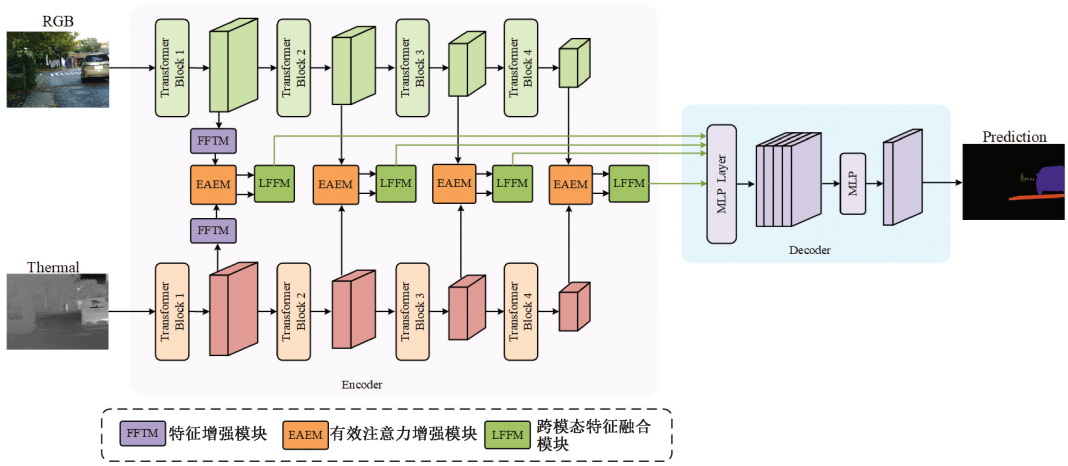

图 1　DMSFNet 网络架构

Fig. 1　Network architecture of DMSFNet

1.2　Segformer

Segformer 是一种分层的 Transformer[18] 结构，通过多层 Transformer 块生成高分辨率的细节特征和低分辨率的语义特征，捕捉输入图像中的长距离依赖关系，提高网络的感知能力，结构如图 2 所示。

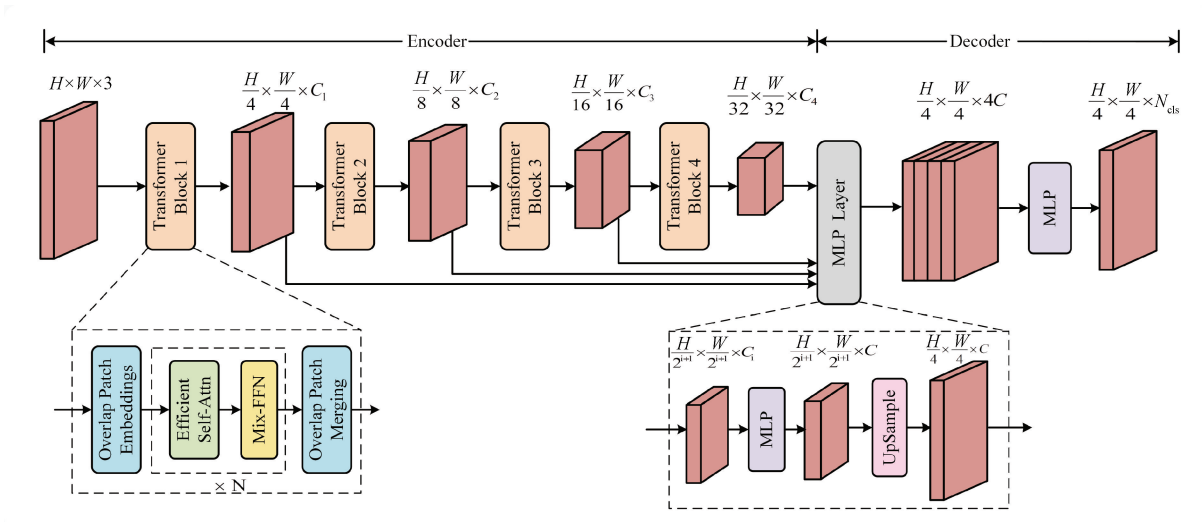

图 2　Segformer 结构

Fig. 2　Structure of Segformer

编码器中的 Transformer 模块首先采用重叠块嵌入(overlap patch embeddings, OPE)模块将输入图像分解成重叠的图像块；然后通过高效自注意力层和混合前馈神经网络层捕捉全局上下文信息和局部细节信息。高效自注意力层可以捕捉序列中任意两个元素之间的关系，有利于建立全局信息的依赖关系，同时添加了一个缩放因子 R，降低了计算复杂度；混合前馈神经网络层通过在前馈神经网络层的两层线性变换层中间加入 3×3 二维卷积层来获取图像空间上的位置信息，有利于捕捉到更长范围的依赖关系和语义信息，更好地处理局部细节和细微变化；最后通过重叠块合并层(overlap patch merging, OPM)合并成无重叠的完整特征图。主干网络中 4 层 Transformer 模块中的重叠块合并层是通过设定两组不同的超参数来获取不同尺度特征图。超参数分别为 $K=7$, $S=4$, $P=3$ 和 $K=3$, $S=2$, $P=1$，其中 K 为重叠图像块大小，S 为相邻块之间的步长，P 为填充大小。特征图的尺寸为 $\frac{H}{2^{j+1}}\times\frac{W}{2^{j+1}}\times C_j$，其中 $j\in\{1,2,3,4\}$ 为尺度索引；H 为输入图像的高度；W 为输入图像的宽度；C_j 为第 j 尺度上的通道数。

解码器中采用轻量级的 All-MLP 解码器，先通过 MLP 层统一通道数，然后将各级特征分别进行上采样到相同尺寸并进行级联，最后通过一层 MLP 进行预测，得到语义分割结果。

1.3　特征增强模块

为了提高浅层特征图的对比度和边缘信息的判别性，本文在第 1 层 Transformer 模块输出端嵌入 FEM，对浅层特征图进行增强，得到第 1 个尺度的特征图，结构如图 3 所示。

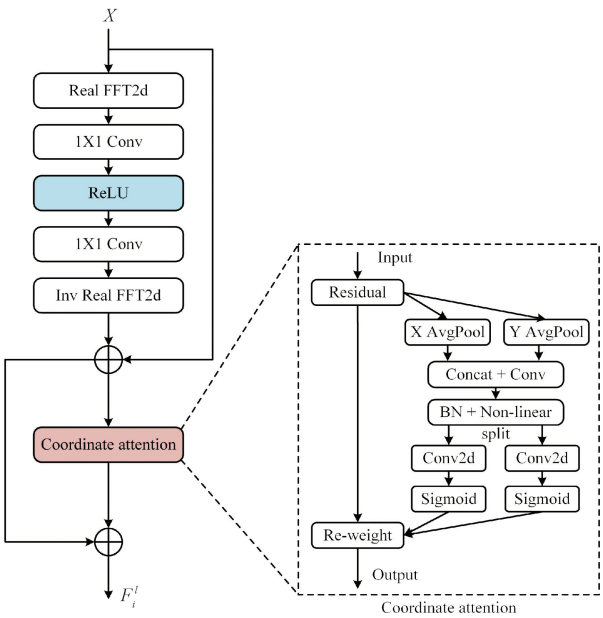

图 3　特征增强模块

Fig. 3　Feature enhancement module

特征增强模块由具有残差结构的快速傅里叶变换模块和坐标注意力机制模块构成，其中快速傅里叶变换操作将特征从空间域转换到频域，由于边缘信息和纹理信息对应着频域中的高频信息，通过在频域中进行卷积操作增强特征中细节信息的表示，同时利用残差连接保留了原始浅层特征，防止信息丢失。最后通过坐标注意力机制聚焦特征中的关键信息，减少之前操作带来的噪声，细化提取的浅层特征。

浅层特征图经过 1 层二维快速傅里叶变换(FFT)、1 个 1×1 卷积层、ReLu 激活函数、1 个 1×1 卷积层、快速傅里叶逆变换(IFFT)层和坐标注意力机制，得到增强后的第 1 个尺度特征图，如式(1)~(5)所示。

$$FFT_{conv} = Conv_{1\times1}(RELU(Conv_{1\times1}(BN(FFT2D(X))))) \tag{1}$$

$$IFFT = BN(InvFFT2D(FFT_{conv})) \tag{2}$$

$$FFT_{res} = RES(IFFT, X) \tag{3}$$

$$FFT_{ca} = CA(FFT_{res}) \tag{4}$$

$$F_i^1 = RES(FFT_{ca}, FFT_{res}) \tag{5}$$

式中：X 为第 1 层 Transformer 模块输出的浅层特征图；$FFT2D(\cdot)$ 为二维快速傅里叶变换；$Conv_{1\times1}(\cdot)$ 为 1×1 卷积层；$RELU(\cdot)$ 为 ReLu 激活函数层；FFT_{conv} 为经过快速傅里叶变换和卷积操作的输出；$InvFFT2D(\cdot)$ 为二维逆快速傅里叶变换；$BN(\cdot)$ 为标准归一化层；$IFFT$ 为快速傅里叶逆变换的输出；$RES(\cdot)$ 为残差连接操作；FFT_{res} 为快速傅里叶变换的结果与输入特征进行残差连接得到的输出；$CA(\cdot)$ 为坐标注意力机制；FFT_{ca} 为经过坐标注意力操作的输出；F_i^1 为特征增强模块输出的第 1 个尺度特征图。

坐标注意力机制是在通道注意力机制中嵌入位置信息，捕获跨通道信息、方向感知和位置敏感信息，定位和识别感兴趣的目标[19]。首先将特征图 FFT_{res} 在宽和高两个方向上分别进行全局平均池化，获得两个方向上的特征图，特征图尺寸从 $\frac{H}{4}\times\frac{W}{4}\times C_1$ 变为 $\frac{H}{4}\times\frac{1}{4}\times C_1$ 和 $\frac{1}{4}\times\frac{W}{4}\times C_1$，然后将宽度方向的特征图进行转置与高度方向特征图进行级联，通过标准归一化(batch normalization，BN)与非线性处理(non-linear，NL)对空间信息进行编码，利用 split 操作分别得到宽、高维度特征，通过 1×1 卷积调整通道数；最后使用 Sigmoid 函数进行归一化加权。特征增强模块利用快速傅里叶变换操作与坐标注意力机制最大限度的保留了浅层特征中的细节信息，有助于优化分割目标的边缘轮廓。

1.4　有效注意力增强模块

为了增强可见光 RGB 图像与红外图像中的关键特征，本文设计了一种有效注意力增强模块，利用通道注意力机制跨模态捕捉特征关系，生成通道权重向量，增强不同模态显著信息的表示。该模块主要由 3 支路组成，分别为注意力关系支路、共有特征信息支路、互补特征信息支路，如图 4 所示。

注意力关系支路旨在利用通道注意力机制得到两种模态特征通道权重。首先，利用全局池化层(global average pooling，GAP)分别将红外特征图和可见光特征图聚合到特定类别的通道向量中，然后使用由全连接层和 sigmoid 激活函数组成的多层感知器得到两种模态的通道权重向量，最后通过逐元素相乘得到跨模态注意力关系权重向量，计算过程如式(6)~(8)所示。

$$\boldsymbol{R}^j = (MLP(Pooling(F_{rgb}^j))) \tag{6}$$

$$\boldsymbol{T}^j = (MLP(Pooling(F_{thermal}^j))) \tag{7}$$

$$\boldsymbol{S}^j = \boldsymbol{R}^j \otimes \boldsymbol{T}^j \tag{8}$$

式中：$\boldsymbol{R}^j$ 为可见光 RGB 图像的通道权重向量；$\boldsymbol{T}^j$ 为红外特征的通道权重向量；$\boldsymbol{S}^j$ 为跨模态注意力关系向量；$j \in \{1,2,3,4\}$ 为尺度索引；$MLP(\cdot)$ 为全连接层操作；

$Pooling(\cdot)$ 为全局池化操作;⊗为逐元素相乘操作。

共有特征信息支路旨在捕获两种模态之间共同的特征信息。首先按通道将输入的红外特征图、可见光特征图分别与注意力关系权重向量相乘,得到特征图 G_i^j ,然后将两种模态特征图级联,经过一个 7×7 深度可分离卷积层、通道注意力机制(全局平均池化层、全连接层、Sigmoid 层),再利用 split 操作得到两种模态的通道权重向量,与对应的 G_i^j 相乘得到跨模态共有特征信息增强的特征图 H_i^j ,过程如式(9)~(11)所示。

$$G_i^j = \sigma(a \times S^j) \otimes F_i^j \tag{9}$$

$$G_{rgb,t}^j = Cat(G_{rgb}^j, G_{thermal}^j) \tag{10}$$

$$H_i^j = G_i^j \otimes split(CAM(DM_{7\times7}(G_{rgb,thermal}^j))) \tag{11}$$

式中:a 为常数,是一个调节跨模态注意力关系权重值分配缩放因子;$\sigma(\cdot)$ 为 Sigmoid 激活函数层;$Cat(\cdot)$ 为级联操作;$DM_{7\times7}(\cdot)$ 为 7×7 深度可分离卷积层;$CAM(\cdot)$ 为通道注意力机制操作;$split(\cdot)$ 为通道分离操作;⊗为逐元素相乘操作。

互补信息支路旨在获取两种模态之间的互补特征信息。首先反转注意力关系权重向量值,得到反映跨模态互补信息的通道权重向量,然后按通道分别与输入的红外特征图、可见光特征图相乘,得到包含跨模态互补特征信息的特征图 E_i^j, 如式(12)所示。

$$E_i^j = (1 - \sigma(\boldsymbol{a} \times (\boldsymbol{R}^j \otimes T^j))) \otimes F_i^j \tag{12}$$

最后将包含跨模态互补特征信息的特征图 E_i^j 与包含共有特征信息的特征图 H_i^j 逐像素相加,得到最终的特征图 $\overline{F}_i^j$,如式(13)所示。

$$\overline{F}_i^j = H_i^j + E_i^j \tag{13}$$

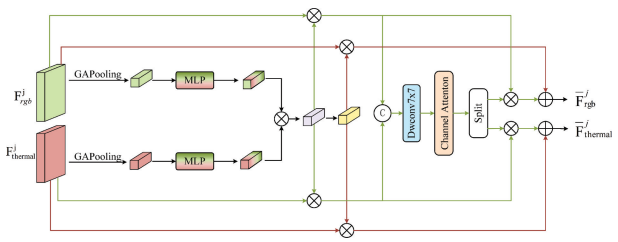

图 4　有效注意力增强模块

Fig. 4　Effective attention enhancement module

1.5　跨模态特征融合模块

为了充分聚合红外与可见光 RGB 图像的特征信息,本文从行和列两个维度上度量像素的活性测度,设计了一种基于 L_1 范数的特征图加权融合策略。首先,对两种模态的特征图 $\overline{F}_{rgb}^j(x,y)$ 和 $\overline{F}_{thermal}^j(x,y)$,计算行(列)向量的 L_1 范数,通过 softmax 获得行(列)方向上的活性测度 $\varphi_{rgb}^{j,k}$ 和 $\varphi_{thermal}^{j,k}$,$k \in \{row, col\}$ 表示方向索引,row 表示行方向,col 表示列方向;然后以活性测度作为权重值,对特征图进行加权融合,计算过程如式(14)和(15)所示。

$$\varphi_{rgb}^{j,k} = \frac{\exp(\|\overline{F}_{rgb}^{j,k}\|_1)}{\exp(\|\overline{F}_{rgb}^{j,k}\|_1) + \exp(\|\overline{F}_{thermal}^{j,k}\|_1)} \tag{14}$$

$$\varphi_{thermal}^{j,k} = \frac{\exp(\|\overline{F}_{thermal}^{j,k}\|_1)}{\exp(\|\overline{F}_{rgb}^{j,k}\|_1) + \exp(\|\overline{F}_{thermal}^{j,k}\|_1)} \tag{15}$$

式中:$\|\cdot\|_1$ 为 L_1 范数计算。

然后,将活性测度与对应的输入特征相乘,从行(列)向量方向得到融合后的特征 $\overline{F}_f^{j,k}(x,y)$,如式(16)所示。

$$\overline{F}_f^{j,k} = \overline{F}_{rgb}^{j,k} \times \varphi_{rgb}^{j,k} + \overline{F}_{thermal}^{j,k} \times \varphi_{thermal}^{j,k} \tag{16}$$

将 $\overline{F}_f^{j,row}$ 和 $\overline{F}_f^{j,col}$ 进行级联得到多维特征图 $\overline{F}_f^j$,利用 1 个 1×1 卷积层调整通道数,再通过一个由步长为 1 的 3×3 深度可分离卷积层、1 个 ReLU 激活函数层、1 个步长为 1 的 1×1 卷积层和 1 个跳跃连接的 1×1 卷积层组成的残差块,得到不同尺度上的融合特征图 F_{Merged}^j ,过程如式(17)和(18)所示。

$$\overline{F}_f^j = Cat(\overline{F}_f^{j,row}, \overline{F}_f^{j,col}) \tag{17}$$

$$F_{Merged}^j = Conv_{1\times1}(RELU(DW_{3\times3}(Conv_{1\times1}(\overline{F}_f^j)))) + Conv_{1\times1}(\overline{F}_f^j) \tag{18}$$

式中:$Cat(\cdot)$ 为级联操作;$Conv_{1\times1}(\cdot)$ 为 1×1 卷积层;$DW_{3\times3}(\cdot)$ 为 3×3 深度可分离卷积层;$RELU(\cdot)$ 为 ReLU 激活函数层。

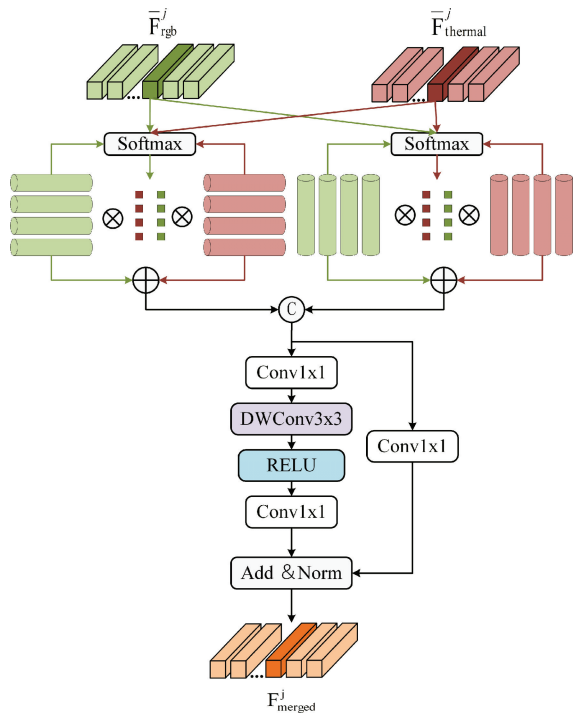

图 5　跨模态特征融合模块

Fig. 5　Cross-modal feature fusion moulde

1.6　解码器

解码器采用轻量级的 All-MLP 解码器,避免了冗余的计算,结构如图 6 所示。首先将各尺度的融合特征图 F^j_{Merged} 通过一个 MLP 层调整为统一的通道数,然后将各尺度的融合特征图进行上采样,特征图尺寸统一为 $\frac{H}{4}\times\frac{W}{4}\times C$,再进行级联,利用另一个 MLP 层得到预测分割结果 F_{out},其过程如式(19)所示。

$$F_{out}=MLP(Cat(Upsample(MLP(F^j_{Merged}))))\quad(19)$$

式中:$MLP(\cdot)$ 为全连接层;$Upsample(\cdot)$ 为上采样操作;$Cat(\cdot)$ 为级联操作。

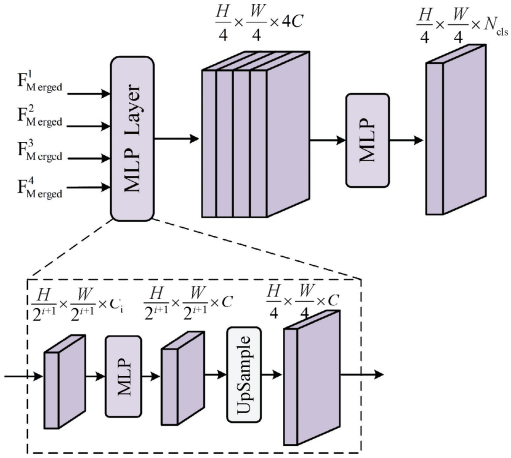

图 6　解码器

Fig. 6　Decoder

2　实验结果与分析

2.1　数据集和实验环境

本文实验使用公开数据集 MFNet[9]。该数据集是利用 InfReCR500 摄像机针对城市街景采集的可见光 RGB 图像和与红外图像对。数据集共包含 1 569 对图像,其中日间图像有 820 对,夜间图像有 749 对。除背景类别外,目标类别有 8 种,分别为色锥(color cone)、车挡(car stop)、行人(person)、曲线(curve)、自行车(bike)、护栏(guardrail)、车辆(car)和凸起(bump),图像的分辨率均为 640×320。

数据集的划分为训练集为 784 对图像,其中包含 410 对日间图像和 374 对夜间图像;验证集为 394 对图像,其中包含 205 对日间图像和 189 对夜间图像;测试集为 391 对图像,其中包含 205 对日间图像和 186 对夜间图像。本文实验环境及参数设置如表 1 所示。

2.2　评价指标

在本文实验中,采用平均交并比(mean intersection over union, mIoU)和平均准确率(mean accuracy, mAcc)作为测试结果的客观评价指标。平均交并比是语义分割的标准度量,作为分割网络的总体精度;交并比(IoU)是真实值和预测值两个集合的交集与并集之比,其值越高,表明两者重合部分越大,网络预测性能越好;平均交并比为所有类别交并比的均值,其表达式如式(20)所示。

$$mIoU=\frac{1}{N}\sum_{m=1}^{N}\frac{TP_m}{TP_m+FN_m+FP_m}\quad(20)$$

表 1　实验环境及参数设置

Table 1　Experimental environment and parameter setting

实验环境	操作系统	处理器	内存	显卡	CUDA 版本	Pytorch 版本
	Win10	Intel(R) Core(TM) i5-8250U	24 GB	NVIDIA RTX3090	11.3	1.10.0
实验参数	输入图像分辨率	迭代次数	Batch size	初始学习率	学习衰减权重	优化器选择
	640×320	300	4	6×10^{-5}	0.01	AdamW

式中:N 为识别的类别数,实验中 N 值为 8;TP_m 为被正确识别为第 m 类的像素点数量;FN_m 为未被正确识别出的第 m 类像素点数量;FP_m 为不是该类别但被识别为第 m 类像素点的数量。

平均准确率衡量的是网络分类正确的像素点数量占总像素点的数量,其表达式如式(21)所示。

$$mAcc=\frac{1}{N}\sum_{m=1}^{N}\frac{TP_m}{TP_m+FN_m}\quad(21)$$

2.3　实验结果与分析

为了验证本文 DMSFNet 网络算法的有效性,将本文算法与其他 9 种主流算法在 MFNet 数据集上进行实验对比,测试集包含 205 对日间图像和 186 对夜间图像。对比算法包括基于 Transformer 的高效语义分割网络算法 Segformer(B2)[16]、面向多光谱自动驾驶场景的实时语义分割网络算法 MFNet[9]、面向城市场景语义分割的 RGB-T 融合网络算法 RTFNet[10]、面向自动驾驶场景的 RGB-T 融合网络算法 FuseSeg[20]、基于 RGB-T 语义分割的模态自适应的空间特征融合网络算法 SFAF-MA[21]、针对多光谱场景的边缘感知引导的语义分割网络算法 EGFNet[13]、基于 RGB-T 语义分割的通道和空间关系传播网络算法 CSRPNet[22]、基于 Transformer 的 RGB-X 语

义分割跨模态融合网络算法 CMX(B2)[23]和基于 RGB-T 语义分割的上下文感知交互网络算法 CAINet[24],其中 Segformer(B2)为典型单模态语义分割算法,MFNet、RTFNet 和 FuseSeg 为典型跨模态语义分割算法,SFAF-MA、EGFNet、CSRPNet、CMX(B2)和 CAINet 为最新跨模态语义分割算法。实验结果如表 2 所示,其中圆圈内双实线加粗字体表示该网络在 IoU 指标和准确率(Acc)指标上取得了最好的结果。

表 2 对比实验结果

Table 2 Quantitative comparison results (%)

模型	Color Cone	Person	Car Stop	Curve	Bike	Guardrail	Car	Bump	
	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	mAcc
	IoU	IoU	IoU	IoU	IoU	IoU	IoU	IoU	mIoU
Segformer(B2)	–	–	–	–	–	–	–	–	–
	50.9	62.8	25.6	31.7	63.2	9.8	87.4	49.6	53.2
MFNet	30.3	67.0	12.5	36.2	53.9	0.1	77.2	30.0	45.1
	25.2	58.9	9.9	29.9	42.9	0.0	65.9	27.7	39.7
RTFNet	45.5	79.3	38.5	60.7	76.8	0.0	93.0	74.7	63.1
	29.1	70.3	29.8	45.3	62.7	0.0	87.4	55.7	53.2
FuseSeg	55.8	81.4	29.1	68.4	78.5	63.7	93.1	66.4	70.6
	46.9	71.7	22.7	44.8	64.6	6.4	87.9	47.9	54.5
SFAF-MA	57.9	82.5	37.5	63.6	73.9	42.2	94.0	74.4	69.6
	45.7	73.0	29.5	45.6	61.3	5.5	88.1	53.8	55.5
EGFNet	65.3	89.0	48.7	71.5	80.6	33.6	95.8	71.1	72.7
	48.3	69.8	33.8	42.8	58.8	7.0	87.6	47.1	54.8
CSRPNet	58.5	86.5	40.2	69.8	78.1	36.1	93.7	71.5	70.4
	46.3	72.3	29.8	45.0	60.7	6.4	87.6	53.1	55.5
CMX(B2)	–	–	–	–	–	–	–	–	–
	52.4	74.8	30.1	47.3	64.7	8.1	89.4	59.4	58.2
CAINet	55.6	74.6	34.7	65.9	85.2	**65.6**	93.0	85.0	73.2
	48.9	66.3	31.5	55.4	68.7	9.0	88.5	**60.7**	58.6
DMSFNet(本文)	**65.4**	**89.2**	48.4	**71.5**	**85.2**	64.7	**96.3**	**86.1**	**76.9**
	54.8	**76.5**	**36.8**	50.3	**68.7**	**10.1**	**91.1**	59.9	**59.8**

由表 2 可知,DMSFNet 在几乎所有类别上,都取得了最好的结果。相较于其他算法,DMSFNet 在 Color Cone 这一类别上的分割结果最好,是由于 DMSFNet 通过 Transformer 块对上下文信息进行建模,并利用特征增强模块对浅层特征进行增强,很好的保留了细节特征,提高了算法对小目标在复杂场景下小目标的分割能力。CMX 也采用了 Transformer 结构,但对跨模态之间的互补信息挖掘不够,导致对小目标的检测能力低于本文算法。EGFNet 使用先验边缘图辅助算法进行语义分割,因此对 Car stop 的分割准确度最好,而本文算法在这方面有进一步提升的空间。CAINet 采用多标签监督增强算法对物体边缘的分割,因此在 Bump 上的分割结果略高于 DMSFNet。所有算法在 Guardrail 这一类别上的结果都非常低,这是因为数据集中此类别的数据非常少,训练不够充分,无法取得较好的识别效果。

为了更好的对实验进行分析,本文将所有数据绘制成散状图,如图 7 所示。可以看出,本文算法相对于其他算法 mIoU 值有显著提高,总体分割效果优于其他算法。

为进一步验证 DMSFNet 算法的有效性,本文分别对数据集中的日间图像与夜间图像进行独立测试,日间图像测试集包含 205 对图像,夜间图像测试集包含 186 对图像。实验结果如表 3 所示。可以看出本文算法在两种场景中分割性能都是最好的,在夜晚场景尤其明显。Segformer 由于是单模态语义分割算法,夜间场景的分割结果明显低于日间分割结果。其他 9 种跨模态语义分割算法利用红外图像特征辅助语义分割,夜间场景下的分割精度更高,再次证明了跨模态语义分割算法针对复杂

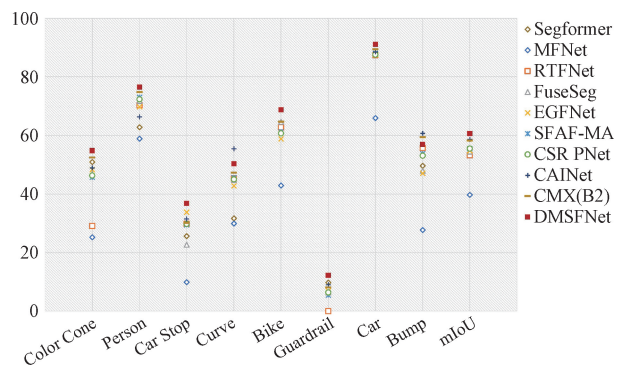

图 7 对比实验结果

Fig. 7 Quantitative comparison results

光照场景的分割具有一定的优越性。

表 3　日间数据与夜间数据实验结果

Table 3　Results from daytimes and nighttime　(%)

模型	日间		夜间	
	mAcc	mIoU	mAcc	mIoU
Segformer	–	50.6	–	43.2
MFNet	42.6	36.1	41.4	36.8
RTFNet	60.0	45.8	60.7	54.8
FuseSeg	62.1	47.8	67.3	54.6
SFAF-MA	71.1	47.0	68.7	54.9
EGFNet	74.4	47.3	68.0	55.0
CSRPNet	72.9	49.1	68.2	55.6
CMX	70.2	51.3	67.4	57.8
CAINet	74.5	56.3	73.2	58.6
DMSFNet(本文)	**77.0**	**56.9**	**74.9**	**59.7**

2.4　可视化分析

为了更直观说明本文算法的有效性和优越性,对本文算法的分割结果进行可视化,并选取 5 种主流跨模态语义分割算法进行可视化结果对比,分别为 Segformer[16]、MFNet[9]、CAINet[24]、SFAF-MA[21] 和 EGFNet[13]。每个算法分别可视化了 8 组图像,其中包含 4 组日间图像和 4 组夜间图像,如图 8 所示。

为了便于观察,在图 8 中矩形标注了值得关注的部分。在所有的可视化结果中,DMSFNet 的结果都与标签值最为接近。在日间场景的第 1 组图像中,最右侧的车辆由于距离太远且受到行人的遮挡,分割难度较大,只有本文算法和 Segformer 算法分割成功,这得益于 Segformer 中的 transformer 块强大的上下文信息提取能力,但由于 Segformer 是单模态语义分割算法,没有借助红外图像提供特征信息,因此没有分割出远处的行人,而本文算法利用 transformer 块和跨模态互补信息成功分割出了最右侧车辆与行人。

由图 8(i)可见,本文算法相较其他网络对行人的边缘轮廓识别的更加清晰,如第 4 组和第 5 组图像中明显分割出了行人腿部的轮廓,这是因为 DMSFNet 对浅层特征进行了增强,最大限度地保留细节特征,提升了算法对目标边缘轮廓的分割能力;在第 3 组图像中,只有本文算法正确分割出了被遮挡的车辆,图 8(h)中,EGFNet 利用边缘标签监督网络也分割出了车辆,但由于主干结构采用 ResNet-152,缺乏提取上下文信息的能力,车辆的分割结果不完整、存在明显缺陷;在第 8 组图中,EGFNet 和 SFAF-MA 由于提取跨模态互补信息的能力较弱,导致分割行人与自行车的边缘不清晰。由图 8(f)可见,CAINet 也存在相似的问题,导致行人与车辆的分割结果出现边缘重叠的情况。可以看到 DMSFNet 对相邻的不同类别边缘的分割能力最强,分割结果也最准确,类别之间边缘最清晰,再次证明了本文算法的优越性。

如图 8(d)所示,本文选取了单模态语义分割算法 Segformer 进行了对比,可以看到单模态算法在日间场景分割能力与早期的跨模态算法 MFNet(图 8(e))接近,第 1 组图像的分割结果甚至优于 MFNet,而在面对夜间场景时分割能力极差,几乎难以分割出任何目标。证明了引入红外图像可以大大提高复杂夜间环境下语义分割的准确性,证明了跨模态语义分割算法研究的有效性。

2.5　消融实验

为了进一步的验证所提出的各个模块的有效性,本文在实验参数设置相同的情况下对 DMSFNet 进行消融实验,如表 4 所示。实验分为 4 组:第 1 组为不添加任何模块,仅采用元素相加的方式融合特征信息;第 2 组为只添加特征增强模块但未添加有效注意力增强模块和跨模态特征融合模块;第 3 组为加入特征增强模块和有效注意力增强模块未添加跨模态特征融合模块;第 4 组为同时加入特征增强模块、有效注意力增强模块以及跨模态特征融合模块。实验结果如表 2 所示,第 1 组与第 2 组比较可知,算法在加入特征增强模块后 mIoU 提高了 1.2%。第 2 组与第 3 组可知,继续引入 EAEM 模块后,算法对特征信息的提取能力显著提升,算法的 mIoU 提高 1.9%。由第 2 组、第 3 组与第 4 组比较可知,同时加入所有模块时算法的性能达到最佳,mIoU 可以达到 59.8%。以上实验证明了算法加入各个模块的合理性。

表 4　消融实验分析

Table 4　Analysis of ablation experiments　(%)

组别	Baseline	FEM	EAEM	CFFM	mAcc	mIoU
第 1 组	√	—	—	—	68.3	56.2
第 2 组	√	√	—	—	72.0	57.4
第 3 组	√	√	√	—	74.9	59.3
第 4 组	√	√	√	√	76.9	59.8

为了验证有效注意力增强模块结构的合理性和有效性,本文对有效注意力增强模块的内部结构进行了消融对照实验,如表 5 所示。第 1 组为去掉了模块中共有信息支路的通道注意力机制(CAM)和互补信息支路(CIB),第 2 组和第 3 组为分别去掉模块中的通道注意力机制和互补信息支路,第 4 组使用完整的有效注意力增强模块。实验表明,有效注意力增强模块在同时使用通道注意力机制和互补增强模块时效果最好,mIoU 可以达到 59.8%。有效注意力增强模块旨在通过利用通道注意力机制获取跨模态之间的特征关系,增强不同模态特征中显著性信息的表达,利用互补信息支路捕捉不同模态之间的互补信息,将跨模态互补信息融合到单模态特征中,增强了跨模态特征,提升了 EAEM 模块的效果,有效提高了整体网络的分割精度。

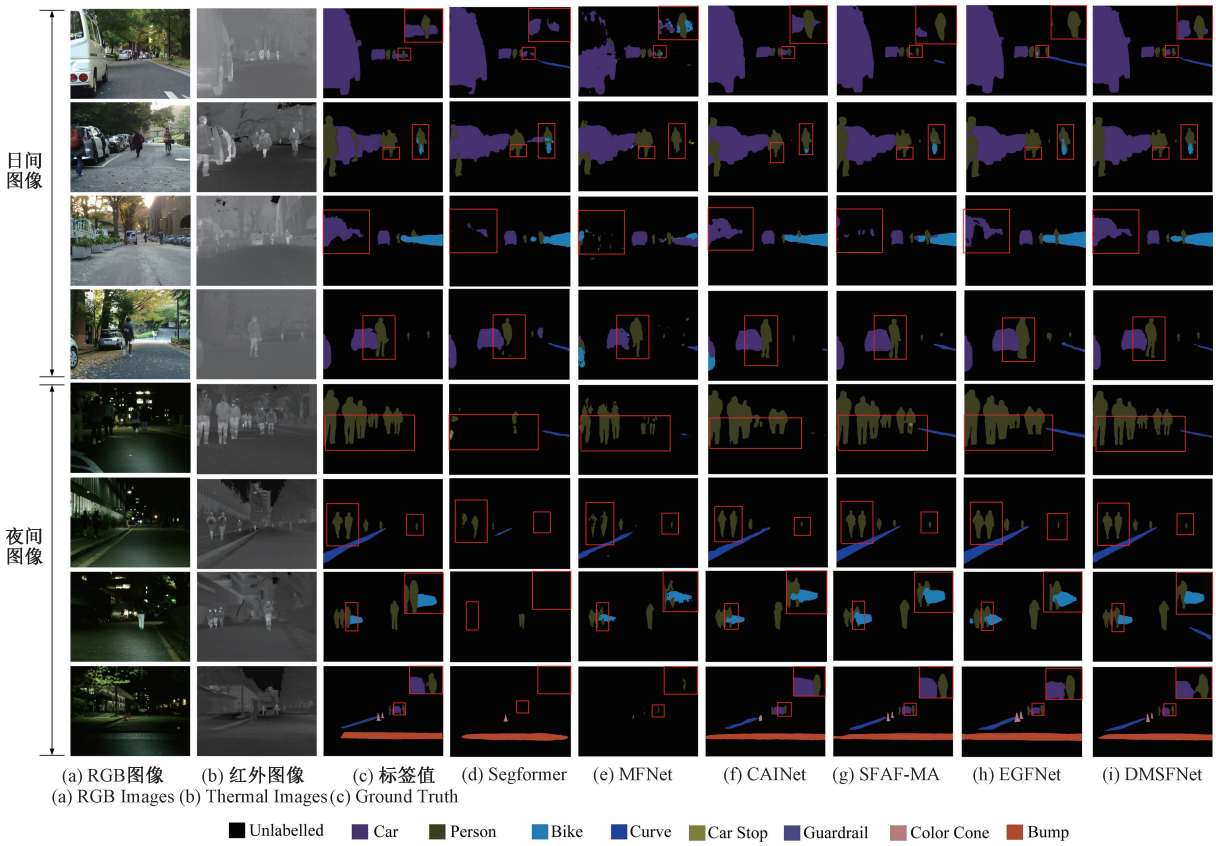

图 8　可视化对比结果

Fig. 8　Visualization comparison results

表 5　有效注意力增强模块有效性验证

Table 5　Validation of effective attention enhancement module　(%)

组别	EAEM		mAcc	mIoU
	CAM	CIB		
第 1 组	—	—	72.2	57.6
第 2 组	√	—	73.0	57.9
第 3 组	—	√	73.2	59.4
第 4 组	√	√	74.9	59.8

3　结　论

本文提出了一种基于双分支多尺度特征融合的跨模态语义分割算法，采用 Segformer 作为主干网络，分别对可见光 RGB 图像和红外图像进行特征提取，利用本文设计的特征增强模块、有效注意力增强模块以及跨模态特征融合模块对特征进行增强和融合。首先通过特征增强模块对浅层特征进行增强，并引入坐标注意力机制加强细节特征的表达。然后利用有效注意力增强模块和跨模态特征融合模块，在行和列向量方面测量权重，对两种模态的互补特征信息进行增强，融合跨模态特征。随后利用轻量级解码器重构图像，得到预测分割掩码。本文采用 MFNet 数据集进行实验，在测试集上 mAcc 和 mIoU 分别达到了 76.9%和 59.8%。实验结果表明，本文算法较已有主流算法，语义分割性能得到显著提高，能够适应夜间或光照变化环境，有效改善目标边缘分割模糊的问题。

在未来的工作中，将加强算法对跨模态特征的提取，探究算法如何在提高分割精度的同时减少计算成本，进一步提升分割的准确度和实时性，推动跨模态语义分割在智能交通、自动驾驶及城市基础设施监测领域实现更精准的故障诊断和环境监测。

参考文献

[1] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition, Long Beach: IEEE, 2019: 3146-3154.

[2] ROMERA E, ALVAREZ J M, BERGASA L M, et al. ERFNet: Efficient residual factorized convnet for real-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2017(1):1-10.

[3] 李长春，李元金. 基于多阈值分割和 B 样条插值的 CT 图像金属伪影去除方法研究[J]. 电子测量与仪器学报，2020，34(7)：128-132.

LI CH CH, LI Y J. Research on metal artifact removal in CT images based on multi-threshold segmentation and B-spline interpolation [J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(7): 128-132.

[4] MOHD S, ZURINA M, NORAIDAH A, et al. Image segmentation for lung region in chest X-Ray images using edge detection and morphology[C]. IEEE International Conference on Control System, Computing and Engineering, Penang, 2014:46-51.

[5] 严南, 姚捃, 黄宇. 基于改进区域分割遥感图像的航天器目标自动识别方法[J]. 计算机测量与控制, 2020, 28(10): 151-164.
YAN N, YAO J, HUANG Y. Automatic spacecraft target recognition method based on improved region segmentation remote sensing image[J]. Computerized Measurement and Control, 2020, 28(10): 151-164.

[6] 李美丽,杨传颖,石宝. 基于语义分割的图像风格迁移技术研究[J]. 计算机工程与应用, 2020, 56(24): 207-213.
LI M L, YANG CH Y, SHI B. Research on image style transfer based on Semantic SegmentSegFormer: Simple and efficient design for semantic segmentation with transformersation [J]. Computer Engineering and Applications, 2019,56(24):207-213.

[7] 王嫣然, 陈清亮, 吴俊君. 面向复杂环境的图像语义分割方法综述[J]. 计算机科学, 2019, 46(9):11.
WANG Y Y, CHEN Q L, WU J J. Survey of image semantic segmentation methods for complex environments [J]. Computer Science, 2019, 46(9):11.

[8] XU J, SHI X, QIN S, et al. LBP-BEGAN: A generative adversarial network architecture for infrared and visible image fusion[J]. Infrared Physics & Technology, 2019, 104: 103144.

[9] HA Q, WATANABE K, KARASAWA T, et al. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes[C]. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) IEEE, 2017.

[10] SUN Y X, ZUO W X, AND LIU M. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes [J]. IEEE Robotics and Automation Letters, 2019:2576-2583.

[11] SHIVAKUMAR S S, RODRIGUES N, ZHOU A, et al. Pst900: Rgb-thermal calibration, dataset and segmentation network [J]. 2020, DOI: 10. 1109/ICRA40945. 2020. 9196831.

[12] DENG F, FENG H, LIANG M, et al. FEANet: Feature-enhanced attention network for RGB-thermal real-time semantic segmentation [J]. ArXiv preprint arXiv. 2110. 08988,2021.

[13] DONG S H, ZHOU W J, XU C, et al. EGFNet: Edge-aware guidance fusion network for rgb-thermal urban scene parsing [J]. IEEE Transactions on Intelligent Transportation Systems, 2024.

[14] ZHANG Z, LIU Y, XUE W. MS-IRTNet: Multistage information interaction network for RGB-T semantic segmentation[J]. Information Sciences: An International Journal, 2023:647,DOI:10. 1016/j. ins. 2023. 119442.

[15] YI S, CHEN M, LIU X, et al. HAFFseg: RGB-thermal semantic segmentation network with hybrid adaptive feature fusion strategy [J]. Signal Processing. Image Communication: A Publication of the European Association for Signal Processing, 2023: 117, DOI: 10. 1016/j. image. 2023. 117027.

[16] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale [C]. International Conference on Learning Representations, 2021.

[17] XIE E, WANG W, YU Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers [J]. ArXiv preprint arXiv. 2105. 15203,2021.

[18] VASWANI A, SHAZEER N, PARMAR K, et al. Attention is all you need [J]. Neural Information Processing Systems, 2017, DOI: 10. 48550/arXiv. 1706. 03762.

[19] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design [J]. ArXiv preprint arXiv. 2103. 02907,2021.

[20] SUN Y X, ZUO W X, YUN P, et al. Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion [J]. IEEE Transactions on Automation Science and Engineering, 2020, DOI: 10. 1109/TASE. 2020. 2993143.

[21] HE X, WANG M, LIU T, et al. SFAF-MA: Spatial feature aggregation and fusion with modality adaptation for rgb-thermal semantic segmentation [J]. IEEE Transactions on Instrumentation and Measurement, 2023, DOI:10. 1109/tim. 2023. 3267529.

[22] ZHOU ZH, WU S K, ZHU G Q, et al. Channel and spatial relation-propagation network for rgb-thermal semantic segmentation[J]. ArXiv preprint arXiv:2308. 12534,2023.

[23] LIU H, ZHANG J, YANG K, et al. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers[J]. 2022,

DOI:10.1109/TITS.2023.3300537.

[24] LYU Y, LIU Z, LI G. Context-aware interaction network for rgb-t semantic segmentation[J]. IEEE Transactions on Multimedia, 2023, DOI:10.1109/TMM.2023.3349072.

作者简介

陈广秋,1999年于吉林大学获得学士学位,2006年于吉林大学获得硕士学位,2015年于吉林大学获得博士学位,现为长春理工大学副教授,主要研究方向为图像处理与机器视觉。

E-mail:gaungqiu_chen@126.com

Chen Guangqiu received his B. Sc. degree from Jilin University in 1999, M. Sc. degree from Jilin University in 2006 and Ph. D. degree from Jilin University in 2015, respectively. He is now an associate professor in Changchun University of Science and Technology. His main research interests include image processing and machine vision.

任天蓉,2020年于哈尔滨理工大学获得学士学位,现为长春理工大学硕士研究生,主要研究方向为图像处理与机器视觉。

E-mail:shanghuo1997@163.com

Ren Tianrong received her B. Sc. degree from Harbin University of Science and Technology in 2020. She is now a M. Sc. candidate at Changchun University of Science and Technology. Her main research interests include image processing and machine vision.

段锦(通信作者),1993年于北京理工大学获得学士学位,1998年于沈阳工业学院获得硕士学位,2004年于吉林大学获得博士学位,现为长春理工大学教授,主要研究方向为偏振成像探测、图像处理与模式识别、数字光学环境仿真。

E-mail:duanjin@vip.sina.com

Duan Jin (Corresponding author) received his B. Sc. degree from Beijing Institute of Technology in 1993, M. Sc. degree from Shenyang Institute of Technology in 1998 and Ph. D. degree from Jilin University in 2004, respectively. He is now a professor in Changchun University of Science and Technology. His main research interests include polarization imaging detection, image processing and pattern recognition, digital optical environment simulation.

黄丹丹,2007年于长春理工大学大学获得学士学位,2009年于东北大学获得硕士学位,2014年于大连理工大学获得博士学位,现为长春理工大学讲师,主要研究方向为计算机视觉和机器学习。

E-mail:hdd@cust.edu.cn

Huang Dandan received her B. Sc. degree from Changchun University of Science and Technology in 2007, M. Sc. degree from Northeastern University in 2009 and Ph. D. degree from Dalian University of Technology in 2014, respectively. Now she is a lecturer in Changchun University of Science and Technology. Her main research interests include computer vision and machine learning.