第 61 卷 第 10 期/2024 年 5 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于TransMANet的遥感图像语义分割算法

宋熙睿[1,2]，葛洪伟[1,2*]

[1]江南大学人工智能与计算机学院，江苏 无锡 214122；

[2]江苏省模式识别与计算智能工程实验室(江南大学)，江苏 无锡 214122

摘要 针对multiattention network(MANet)算法与图像语义信息关联不足、全局特征提取不充分和分割精度较低的问题，基于Transformer与注意力机制，提出一种增强浅层网络语义信息，具有融合局部和全局上下文的双分支解码器的网络结构，即Transformer multiattention network(TransMANet)。首先，引入局部注意力嵌入机制，增强上下文信息的嵌入，并将高级特征的语义信息嵌入低级特征；然后，设计基于Transformer与卷积神经网络的双分支解码器，分别提取全局上下文信息和不同尺度的细节信息，对全局与局部信息建模；最后，改进原有的损失函数，缓解遥感数据集类别不平衡的问题，提高分割准确度。实验结果表明，TransMANet在UAVid、LoveDA、Potsdam和Vaihingen数据集上均取得了较MANet及其他有竞争力的先进方法更优的交并比指标，有较好的泛化能力。

关键词 图像处理；语义分割；注意力机制；Transformer；高分辨率遥感影像

中图分类号 TP751.1 **文献标志码** A **DOI**：10.3788/LOP232052

Remote Sensing Image Semantic Segmentation Algorithm Based on TransMANet

Song Xirui[1,2], Ge Hongwei[1,2*]

[1]*School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi* 214122, *Jiangsu, China*;

[2]*Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi* 214122, *Jiangsu, China*

Abstract Herein, we propose a Transformer multiattention network (TransMANet), a network structure based on Transformer and attention mechanisms, to address the issues of low segmentation accuracy, inadequate global feature extraction, and insufficient association between the multiattention network (MANet) algorithm and image semantic information. This network structure features a dual-branch decoder that combines local and global contexts and enhances the semantic information of shallow networks. First, we introduce a local attention embedding mechanism that enhances the embedding of context information and semantic information of high-level features into low-level features. Then, we design a dual-branch decoder that combines Transformer and convolutional neural networks, which extracts global context information and detailed information with different scales, thereby modeling global and local information. Finally, we improve the original loss function and use a joint loss function that combines cross-entropy loss and Dice loss to address the class imbalance problem often encountered in remote sensing datasets and thus improve segmentation accuracy. Our experimental results demonstrate the superiority of TransMANet over MANet and other advanced methods in terms of intersection over union on UAVid, LoveDA, Potsdam, and Vaihingen datasets. This indicates the strong generalization capability of TransMANet and its effectiveness in achieving accurate segmentation results.

Key words image processing; semantic segmentation; attention mechanism; Transformer; high-resolution remote sensing image

1 引 言

遥感图像的语义分割是遥感数据自动分析的关键环节。随着遥感传感器硬件性能不断提升，获得的高分辨率遥感影像具有丰富的几何细节、纹理特征和复杂的上下文特征，并在多个领域得到广泛应用，如国土

收稿日期：2023-09-05；**修回日期**：2023-10-05；**录用日期**：2023-10-13；**网络首发日期**：2023-11-07

基金项目：江苏高校优势学科建设工程资助项目、111引智计划项目(B12018)

通信作者：*ghw8601@163.com

资源调查[1]、城市建设和规划[2]等。遥感影像的语义分割在遥感影像解译中起着重要作用。

针对遥感图像的语义分割问题，现有的方法大体分为两类：基于手工特征描述符的传统方法和深度学习方法。传统方法多采用阈值分割[3]和边缘检测[4]等，但难以提取到深层的语义信息，鲁棒性也较差。深度学习方法可以自动提取特征，克服了手工提取特征的局限，从而在与传统方法的比较中取得了显著的进步。机器学习可以在少量计算资源的情况下，通过学习实验数据，找到最优的激光脉冲参数，从而实现激光脉冲的优化[5]。在激光等离子体物理中，也应用机器学习方法进行预测和分析[6]。

进行遥感图像语义分割的深度学习方法主要分为卷积神经网络（CNN）、注意力机制和Transformer三类方法。随着FCN[7]、U-Net[8]、DeepLab[9]等CNN模型的提出，Zhao等[10]于2017年提出了金字塔池化模块，并将其应用到基于ResNet的架构中，通过多尺度池化层解析全局信息，获取语义类别分布线索。基于金字塔池化模块，Lou等[11]提出了CFPNet，在特征图中引入了通道维度的金字塔池化操作，从而实现了通道级别的特征提取和汇聚。在高功率激光实验中，也利用深度神经网络进行对象检测[12]。注意力机制可以使模型更加关注有价值的特征。Zhao等[13]引入了金字塔注意力池模块，将注意力机制嵌入到多尺度模块，实现特征的自适应细化，提高了高分辨率航空图像的标注准确性。白宗宝等[14]引入一种三叉戟式的注意力机制提取多尺度信息并进行精简，对目标通道的权重进行优化，增强目标区域像素的重要性，实现了空间和通道权重的重新分配。Dosovitskiy等[15]将Transformer引入到图像分类任务，提出了Vision Transformer（ViT）模型。这个模型采用了完全基于自注意力机制的结构，具备全局感受野的优势，能够有效地学习高分辨率遥感图像中远距离空间上下文信息，在大规模数据集上表现出色。Wang等[16]使用CNN作为编码器，在解码器中结合Transformer设计了一种高效的全局-局部注意力机制，对全局和局部信息进行建模。目前使用的遥感图像语义分割方法中，multiattention network（MANet）[17]是研究人员受点积注意力（dot-product attention）[18]的启发设计的深度学习网络模型，在遥感图像语义分割任务中有良好的效果，但是MANet仍然存在着一些问题。首先，物体像素的类别判别是一项综合任务，受上下文的影响，而遥感图像数据集中存在着很大的类内不一致问题，导致难以汇集上下文信息；其次，CNN固定感受野的卷积操作缺乏对全局上下文信息的建模能力，导致每个像素的分类往往是模糊的；最后，遥感图像数据集中不同类别的样本数量差异明显，有些类别的样本数量远远多于其他类别，训练过程中会使模型过度聚焦于数量大的类别，导致小样本分割精度较差。

本文针对遥感图像语义分割任务对MANet进行改进，提出基于MANet的遥感图像语义分割算法，即Transformer multiattention network（TransMANet）。首先，对MANet的跳跃连接进行了优化，引入局部注意力嵌入机制，增强上下文信息的嵌入并将深层网络提取的语义信息嵌入浅层特征中，丰富低级特征的语义信息。其次，对MANet的解码器进行优化，提出了基于Transformer与CNN的双分支解码器，分别提取全局上下文信息和不同尺度的细节信息，对全局与局部信息进行解码，获得更多尺度的信息。最后，使用交叉熵损失函数与Dice损失函数联合的损失函数，缓解遥感数据集中样本类别不平衡的问题，提高分割准确度。

2　基本原理

2.1　MANet基本理论

2022年Li等[17]提出一种计算复杂度低且性能优越的高分辨率遥感图像语义分割的深度学习网络模型。MANet的网络结构如图1（a）所示，是基于编码器-解码器结构而设计的。Tsai等[19]从卷积内核的角度提出了一个新的方法表示注意力机制，将注意力看作是对输入应用内核平滑器。在此研究的基础上，Li等提出了一种线性复杂度的核注意力机制（KAM），这一创新在保持性能的同时减小了计算成本。

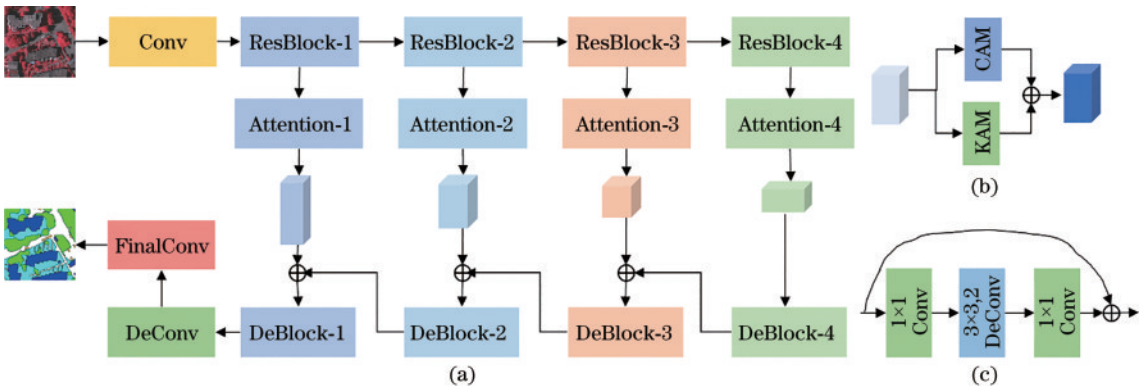

图1　MANet示意图。（a）MANet的结构；（b）注意力模块；（c）解码器模块

Fig. 1　Schematic of MANet. (a) Architecture of MANet; (b) attention block; (c) decoder block

在空间维度上，由于点积注意力的计算复杂度与输入的大小呈二次方关系，因此MANet设计了一种计算复杂度为线性的KAM。对于通道维度，由于输入通道的数量通常远小于特征图中包含的像素数量，函数的复杂度并不高，因此，在通道上使用基于点积注意力的通道注意力机制(CAM)[18]。KAM和CAM分别对位置和通道的长距离依赖进行建模，利用这两个模块设计了一个注意力模块，如图1(b)所示，以增强对各层提取的特征图的判别能力。采用在ImageNet上预训练的ResNet-50模型来提取特征图，最底层的特征通过DeBlock-4直接得到上采样，然后将主干网络输出的不同尺度的特征图输入到相应的注意力模块进行细化，将这些经过细化的特征图与上采样的下层特征图相加。接着，经过融合的特征被相应的DeBlock上采样。最后，通过去卷积操作，将最后一个DeBlock的输出上采样到与输入相同的空间分辨率，并输入到最后的卷积层，得到预测的分割图。

2.2　Swin Transformer基本理论

ViT[15]依靠大规模数据集的预训练在计算机视觉领域展现出了惊人的能力，然而，在执行密集预测任务时仍须付出巨大的训练成本。因此，Liu等[20]提出了基于移位窗口策略的Swin Transformer，首先提出了窗口划分操作，将多头自注意力(MSA)的计算限制在非重叠窗口，进行基于窗口的多头自注意力(W-MSA)；同时为了建立跨越窗口的依赖信息，允许跨窗口信息交互，放弃了传统Transformer中的MSA，而是利用基于移位窗口的多头自注意力(SW-MSA)。Swin Transformer的计算复杂度仅为线性，在图像分类、物体检测和语义分割等各种视觉任务中都能获得先进的性能。Swin Transformer块结构如图2所示，MLP和LN分别代表多层感知器和层归一化。所提方法借鉴了Swin Transformer中的W-MSA。

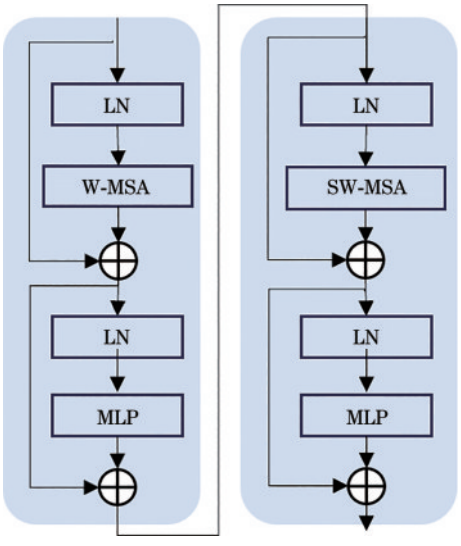

图2　两个连续的Swin Transformer块

Fig. 2　Two consecutive Swin Transformer blocks

3　TransMANet

所提方法是在MANet的基础上提出的：使用ResNet-50作为骨干网络提取特征，使用注意力机制优化跳跃连接，丰富低级特征的语义信息；设计了双分支解码器，在多个尺度上捕捉全局和局部上下文；改进了损失函数，以缓解遥感数据集上常见的类别不平衡问题。

3.1　局部注意力嵌入机制

在编码器-解码器的结构中，编码器常常基于堆叠卷积和池化操作不断缩小特征的空间尺寸，以增强其语义表征，但丢失了详细的空间信息。而在利用解码器的低级特征找回丢失的空间信息时，低级特征和高级特征在语义信息和空间分布上都存在显著差异，例如，低级特征对边缘、纹理更敏感，而高级特征更关注物体的形状、结构，因此将它们简单融合在一起并不能显著提高分割精度。

受Ding等[21]提出的LANet启发，引入了补丁注意力模块(PAM)和注意力嵌入模块(AEM)，以加强上下文信息的聚合，图3展示了PAM的设计。在遥感图像领域中，全局平均池化技术难以有效，因为遥感图像的空间尺寸通常比自然图像大得多，而对象类别的数量却往往较少。例如，Potsdam数据集中的每幅图像都有6000×6000像素，只有6个对象类别。因此，几乎每幅图像都包含所有对象类别，无法在全局层面嵌入明确的全局场景信息。所以将平均池化限制在局部区域上，为每个补丁的通道生成一个描述符，局部池化窗口大小设置为10。

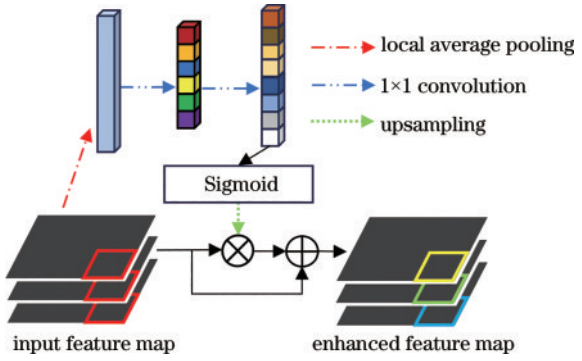

图3　补丁注意力模块

Fig. 3　Architecture of PAM

由于低级特征和高级特征的差异，很难直接有效利用低级特征。最常用的低级特征利用方法是将它们与高级特征串联起来，但效果不佳。为了充分利用低级特征，受LANet启发，使用AEM来丰富它们的语义信息。这一操作在高级和低级特征之间架起了一座桥梁，同时又不会牺牲后者的空间细节信息，图4显示了AEM的设计。高级特征经过PAM强化后，再经过卷积操作和Sigmoid函数，上采样到与低级特征空间分

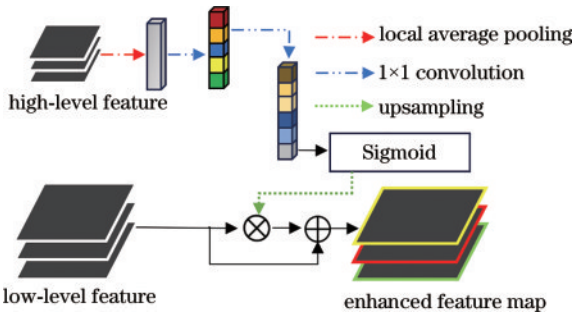

图 4　注意力嵌入模块

Fig. 4　Architecture of AEM

辨率大小相等的尺寸，进行按位相乘操作，最后与原来的特征进行残差连接。这种方法将高层特征的语义信息嵌入到低层特征中。这样，低级特征就被嵌入了超越其感受野限制的上下文信息，其空间细节则得以保留。

3.2　双分支解码器

复杂的人造物体经常出现在高分辨率的城市遥感图像中，如果没有全局语义信息很难实现精确分割。为了捕捉全局上下文，主流解决方案主要是在网络末端附加单个注意力块[22]或引入 Transformer 作为编码器[23]，前者无法捕捉多尺度的全局特征，而后者则大大增加了网络的复杂性并丢失了空间细节信息。相比之下，使用 4 个双分支解码器能在多个尺度上捕捉全局和局部上下文，同时保持高效率。如图 5 所示，受 Swin Transformer 启发设计了双分支解码器，双分支解码器由全局注意力模块、局部注意力模块、多层感知器、两个归一化层和两个跳跃连接组成。虽然全局上下文对复杂场景的语义分割至关重要，但局部信息对保留丰富的空间细节仍然必不可少。为此，设计了全局注意力模块和局部注意力模块，构建了两个平行分支，分别提取全局和局部上下文。

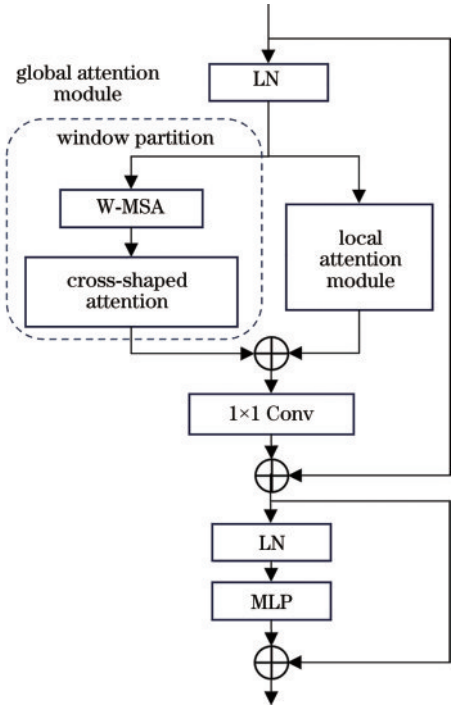

图 5　双分支解码器

Fig. 5　Architecture of the dual-branch decoder

在局部注意力模块中，标准卷积分解成非对称形式以构建特征金字塔（FP）通道，如图 6（a）所示，这样可以使用更少的参数，同时保留从原来相同大小的感受野中学习特征信息的能力，并使用跳跃连接对非对称卷积块提取的特征进行拼接，从而创建多尺度的特征图。遥感图像的细节信息大小不一致，建筑边缘、人类、车辆等细节信息有着不同的尺度，为了局部分支能够学习不同尺度的细节信息，局部分支中设计了多个 FP 通道，并设置了不同的卷积膨胀率。局部注意力分支结构如图 6（b）所示，包含 4 个 FP 通道，每个 FP 通道拥有不同的卷积膨胀率，分别为 1、1、2 和 3。解码器输入的特征维度设为 64，使用 1×1 卷积将高维特征图投影到低维，输出的维度设置为 8。然后，将特征输入到具有不同膨胀率的多 FP 通道中。对于输出的特征采取分层特征融合的方法，因为简单的融合方法会产生棋盘格效应，影响了分割掩码的准确性和质量。从第二个通道开始，采用逐步的求和运算逐渐合并特征图，然后将它们拼接起来，并使用 1×1 卷积来激活输出，最后使用残差连接，建立最终的特征图。

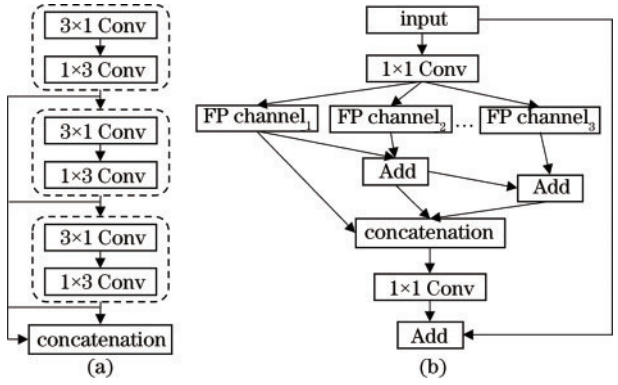

图 6　局部注意力模块示意图。（a）特征金字塔通道；（b）局部注意力模块结构

Fig. 6　Schematic of local attention module. (a) Feature pyramid channel; (b) architecture of local attention module

全局分支采用 W-MSA 来捕捉全局上下文。首先使用标准的 1×1 卷积将输入的特征图的通道维度扩展为原来的 3 倍。然后，使用窗口划分操作将特征分割为查询（$\boldsymbol{Q}$）、键（$\boldsymbol{K}$）和值（$\boldsymbol{V}$）向量进行 W-MSA 计算。解码器通道维度设为 64。窗口大小（window size）和头数量均设为 8。在非重叠的局部窗口中进行自注意力计算虽然效率高，但由于缺乏窗口间交互，会破坏遥感场景的空间一致性。Swin Transformer 引入了一个额外的基于移位窗口的 Transformer 块来挖掘窗口之间的关系，虽然捕捉跨窗口关系的能力增强了，但计算量也相应大幅增加。在本文中，受 UNetFormer[16] 的启

发，使用十字交叉注意力来建立跨越窗口的全局信息。如图 7 所示，在计算 W-MSA 后，使用大小为 $s_{\text{window}}\times 1$ 和 $1\times s_{\text{window}}$ 的水平卷积和垂直卷积对特征图进行水平平均池化和垂直平均池化，再将二者之和按位相加，其中 s_{window} 为窗口尺寸。十字交叉注意力模块融合了水平平均池化层和垂直平均池化层产生的两个特征图，从而捕捉了全局上下文。具体来说，水平平均池化层建立了跨越窗口间的水平关系，对于窗口 1 中的任意点 $P_1^{(m,n)}$，其与窗口 2 中的 $P_2^{(m+w-j,n)}$ 的依赖关系可模拟为

$$P_1^{(m,n)}=\frac{\sum_{i=0}^{w-m-1}P_1^{(m+i,n)}+\sum_{j=0}^{m}P_2^{(m+w-j,n)}}{w},\tag{1}$$

$$P_1^{(m+i,n)}=D_i\left(P_1{}^{(m,n)}\right),\tag{2}$$

$$P_2^{(m+w-j,n)}=D_j\left(P_2{}^{(m+w,n)}\right),\tag{3}$$

$$P_1^{(m,n)}=\frac{\sum_{i=0}^{w-m-1}D_i\left(P_1{}^{(m,n)}\right)+\sum_{j=0}^{m}D_j\left(P_2{}^{(m+w,n)}\right)}{w},\tag{4}$$

式中：m 和 n 是 P_1 的横和纵坐标；w 是窗口大小；D 表示自注意力计算，可以模拟局部窗口中像素对的依赖关系。因此，对于窗口 1 的 a 路径中的任何其他点 $P_1^{(m+i,n)}$，其与 $P_1^{(m,n)}$ 的依赖关系可以用式(2)来模拟。对于窗口 2 的 b 路径上的任何其他点 $P_2^{(m+w-j,n)}$，其与 $P_2{}^{(m+w,n)}$ 的依赖关系可用式(3)来模拟。式(1)可重写为式(4)，即 $P_1^{(m,n)}$ 与 $P_2{}^{(m+w,n)}$ 之间的依赖关系已被模拟。根据这种跨窗口的像素依赖关系，可以确定窗口 1 和窗口 2 之间的水平关系。同样，窗口 1 和窗口 3 之间的垂直关系也可以用同样的方法确定，即 $W_{\text{win1}}=V(W_{\text{win3}})$，而对于窗口 4，$W_{\text{win1}}=V\left[H(W_{\text{win4}})\right]+H\left[V(W_{\text{win4}})\right]$，其中 W_{win} 为十字交叉注意力建立依赖关系。通过连接更多的窗口，可以模拟任意两个窗口之间的长距离依赖关系。因此，十字交叉注意力可以模拟窗口间的长距离依赖关系，从而捕捉全局上下文。

此外，全局分支中的全局上下文与局部分支中的局部上下文进一步聚合，产生全局-局部上下文。

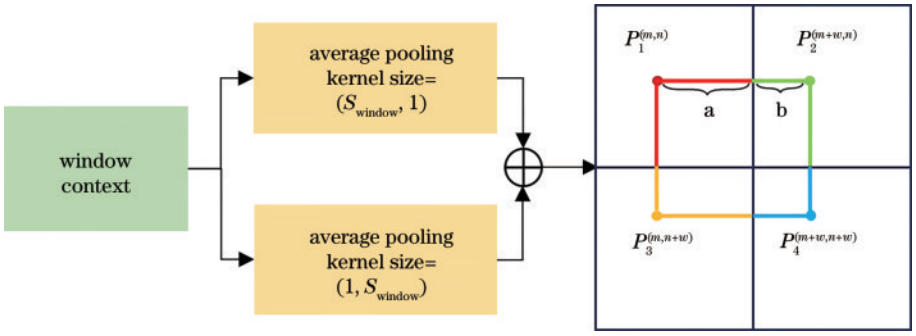

图 7　十字交叉注意力

Fig. 7　Cross-shaped attention

3.3　损失函数

交叉熵损失函数广泛应用于图像分割问题，但是如果数据类别不平衡，交叉熵损失函数可能会使模型过度聚焦于数量多的类别，忽视数量少的类别。对于像素级别的分割任务，交叉熵损失函数无法很好地处理类别不平衡的问题。为了缓解此问题，将 MANet 原来的交叉熵损失函数重新设计为交叉熵损失函数与 Dice 损失函数联合的损失函数。损失函数表示为

$$L_{\text{CE}}=-\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}y_k^{(n)}\log_{\text{e}}\hat{y}_k^{(n)},\tag{5}$$

$$L_{\text{Dice}}=1-\frac{2}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\hat{y}_k^{(n)}y_k^{(n)}}{\hat{y}_k^{(n)}+y_k^{(n)}},\tag{6}$$

$$L=L_{\text{CE}}+L_{\text{Dice}},\tag{7}$$

式中：N 和 K 分别表示样本数和类别数；$y_k^{(n)}$ 和 $\hat{y}_k^{(n)}$ 表示真实标签和网络的输出；L_{CE} 代表交叉熵损失函数；L_{Dice} 代表 Dice 损失函数；L 代表联合损失函数。

3.4　TransMANet 结构

基于 MANet 提出的遥感图像语义分割算法 TransMANet 的结构如图 8 所示。网络结构由编码器、双分支解码器、MANet 中注意力模块跳跃连接、PAM 和 AEM 构成。采用在 ImageNet 上经过预训练的 ResNet-50 作为编码器，并采用 ResBlock-1、ResBlock-2、ResBlock-3、ResBlock-4 输出的 4 个不同尺度的特征图。

ResBlock-4 输出的特征图输入到 PAM 后，直接输入到解码器 4。为了节约计算资源并保持分割精度，解码器 4 首先对特征进行降维，将通道数降为 64 后将特征输入到 Transformer 和 CNN 双分支中，分别提取全局和局部上下文。对处理后的特征以双线性插值的方式进行两倍上采样，并与 ResBlock-3 层输出的特征进行融合，融合采用串联的方式，对堆叠后的特征使用 3×3 的卷积进行降维，降维后的特征再经过 Batch Norm 和 ReLU 层。将融合后的特征输入到解码器 3 中。将 ResBlock-2 层输出的特征输入到注意力模块后，输出再与解码器 3 的输出经过融合模块后，输入到解码器 2 中。

ResBlock-1 层输出的特征首先经过 PAM，增强局部上下文特征的嵌入，但此层输出的特征中语义信息

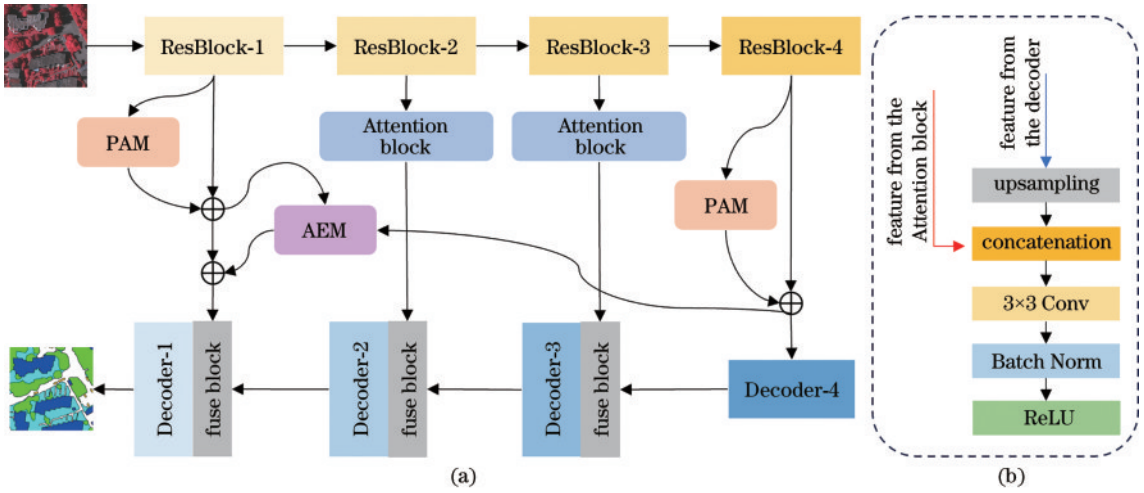

图 8　TransMANet 示意图。(a) TransMANet 的结构；(b) 融合模块

Fig. 8　Schematic of TransMANet. (a) Architecture of TransMANet; (b) fuse block

不充分，将 ResBlock-4 层输出的高级特征输入到 AEM，网络深层的语义信息被嵌入到浅层网络中。为了避免受到高级特征过多的干扰，增加了一个残差设计，以强调低级特征的重要性。对嵌入后的特征与解码器 2 的输出进行融合，将融合后的特征输入到解码器 1 中。最后，将解码器 1 的输出上采样到与输入相同的空间分辨率，经过卷积层后，得到预测的分割图。

4　分析与讨论

4.1　数据集

1) UAVid。如图 9(a)所示，UAVid 数据集侧重于城市街道场景，有 2 种空间分辨率(3840×2160 和 4096×2160)和 8 个类别。由于 UAVid 图像空间分辨率高、空间差异大、类别模糊且场景普遍复杂，因此对其进行分割具有挑战性。数据集中共 420 张图像，按照官方的建议，其中 200 张图像用于训练，70 张图像用于验证，150 张图像用于测试。

2) LoveDA。如图 9(b)所示，LoveDA 数据集包含来自南京、常州、武汉共 5987 幅分辨率为 1024×1024 像素的精细光学遥感图像，空间分辨率为 0.3 m，包括城市和农村两个场景。因此，多尺度物体、复杂的背景和不一致的类别分布带来了相当大的挑战。按照官方的建议，2522 幅图像用于训练，1669 幅图像用于验证，1796 幅图像用于测试。

3) Vaihingen。如图 9(c)所示，Vaihingen 数据集包括 33 张图像，平均尺寸为 2494×2064 像素，空间分辨率为 5 cm。数据集提供了近红外、红色和绿色通道以及数字表面模型(DSM)。使用 ID 为 2、4、6、8、10、12、14、16、20、22、24、27、29、31、33、35 和 38 的图像进行测试，ID 为 30 的图像进行验证，其余 15 幅图像用于训练。在实验中只使用红色、绿色和蓝色通道和有侵蚀边界的真实标签。

4) Potsdam。如图 9(d)所示，Potsdam 数据集包含 38 幅 6000×6000 像素的精细分辨率图像，空间分辨率为 5 cm。数据集提供近红外、红、绿、蓝通道以及 DSM 和归一化 DSM(NDSM)。使用 ID 为 2_13、2_14、3_13、3_14、4_13、4_14、4_15、5_13、5_14、5_15、6_13、6_14、6_15 和 7_13 的图像进行测试，利用 ID 为 2_10 的图像进行验证，并利用除 7_10 外(带有错误标注)的 22 幅图像进行训练。

4.2　实验环境与参数

实验环境为 Python 3.6，PyTorch 1.10.0，单个 GeForce RTX 2080 显卡。使用 AdamW 优化器训练实验中的所有模型。基本学习率设定为 0.0006，并采用余弦策略进行学习率的动态调整。

对于 UAVid 数据集，每张图像都被填充和裁剪成 8 个 1024×1024 像素大小的图像，在训练期间使用随机垂直翻转、随机水平翻转和随机亮度作为数据增强策略；训练回合数(epoch)设置为 40，批大小为 2；在测试过程中，使用了垂直翻转和水平翻转等测试时间增强(TTA)策略。对于 Vaihingen、Potsdam 和 LoveDA 数据集，图像被裁剪成 512×512 像素的图片，在训练过程中，采用了随机缩放、随机垂直翻转、随机水平翻转和随机旋转等数据增强策略；训练回合数设置为 100，批大小为 8；在测试阶段，使用了多尺度和随机翻转数据增强策略。

4.3　评价指标

TransMANet 在 4 个数据集上的性能通过总体准确率(OA)、平均交并比(mIoU)和 F1 分数进行评估，它们的表达式分别为

$$P_{\mathrm{OA}}=\frac{\sum_{k=1}^{K}N_{\mathrm{TP}\,k}+N_{\mathrm{TN}\,k}}{\sum_{k=1}^{K}N_{\mathrm{TP}\,k}+N_{\mathrm{FP}\,k}+N_{\mathrm{TN}\,k}+N_{\mathrm{FN}\,k}},\tag{8}$$

$$R_{\mathrm{mIoU}}=\frac{1}{K}\sum_{k=1}^{K}\frac{N_{\mathrm{TP}\,k}}{N_{\mathrm{TP}\,k}+N_{\mathrm{FP}\,k}+N_{\mathrm{FN}\,k}},\tag{9}$$

$$F_1=2\times\frac{r_{\mathrm{precision}}\times r_{\mathrm{recall}}}{r_{\mathrm{precision}}+r_{\mathrm{recall}}},\tag{10}$$

图 9 数据集展示。(a)UAVid 数据集;(b)LoveDA 数据集;(c)Vaihingen 数据集;(d)Potsdam 数据集

Fig. 9 Dataset presentation. (a) UAVid dataset; (b) LoveDA dataset; (c) Vaihingen dataset; (d) Potsdam dataset

式中:$N_{\mathrm{TP}\,k}$、$N_{\mathrm{FP}\,k}$、$N_{\mathrm{TN}\,k}$和$N_{\mathrm{FN}\,k}$分别表示索引为k类的对象的真阳性、假阳性、真阴性和假阴性;$r_{\mathrm{precision}}$是精确率,r_{recall}是召回率。

4.4 对比实验

为了评估所提模型的分割性能,将 TransMANet 与其他常用遥感图像语义分割模型进行对比,包括 LANet[21]、DeepLabV3+[9]、SwiftNet[24]、ABCNet[25]、UNetFormer[16]、Segmenter[26]等。为了保证公平性,所有方法都使用相同的测试代码,表中黑体字表示结果最优。

UAVid 是一个大规模的城市场景分割数据集,其中的图像是由无人驾驶飞行器在不同城市和不同光照条件下拍摄的,因此在该数据集上获得高分是一项挑战。在官方 UAVid 测试集上对 TransMANet 与其他先进网络进行了详细比较,结果如表 1 所示,TransMANet 获得了最佳的 mIoU(69.9%),同时保持了每个类上 IoU 的优势。具体来说:TransMANet 不仅在 mIoU 方面比基于 CNN 的 ABCNet 高出 6.1 百分点,而且比最近推出的基于 Transformer 和 CNN 的混合网络 UNetFormer 高出 2.1 百分点;尤其在静止车辆和人类分割上,TransMANet 展现出卓越的优势,静止车辆和人类是非常小的对象,并且在很多场景下这两类严重受光照影响,因此非常难处理,尽管如此,TransMANet 仍然在这两个类别的 IoU 比其他方法分别至少高出 3.1 百分点和 2.2 百分点。此外,图 10 所示的 UAVid 测试集的可视化结果也证明了 TransMANet 的有效性。MANet 不能很好地捕捉全局上下文信息,对道路不能进行精确的分割,对极小的物体也不能进行精确分割,而 TransMANet 可以实现精确分割。

在 LoveDA 数据集上进行了对比实验,以进一步评估 TransMANet 的性能,结果如表 2 所示。由于可

表 1　不同方法在 UAVid 数据集上的性能比较

Table 1　Performance comparison of different methods on UAVid dataset

Model	IoU/%								mIoU/%
	Background	Building	Road	Tree	Vegetation	Static car	Moving car	Human	
Segmenter[26]	64.2	84.4	79.8	76.1	57.6	34.5	59.2	14.2	58.7
SwiftNet[24]	64.1	85.3	61.5	78.3	76.4	62.1	51.1	15.7	61.1
ABCNet[25]	67.4	86.4	81.2	79.9	63.1	48.4	69.8	13.9	63.8
SegFormer[27]	66.6	86.3	80.1	79.6	62.3	52.5	72.5	28.5	66.0
MANet[17]	67.4	88.0	79.5	79.4	63.0	64.6	67.4	21.7	66.3
UNetFormer[16]	68.4	87.4	81.5	80.2	63.5	56.4	73.6	31.0	67.8
TransMANet	**69.9**	**88.7**	**81.8**	**80.7**	**63.6**	**67.7**	**73.9**	**33.2**	**69.9**

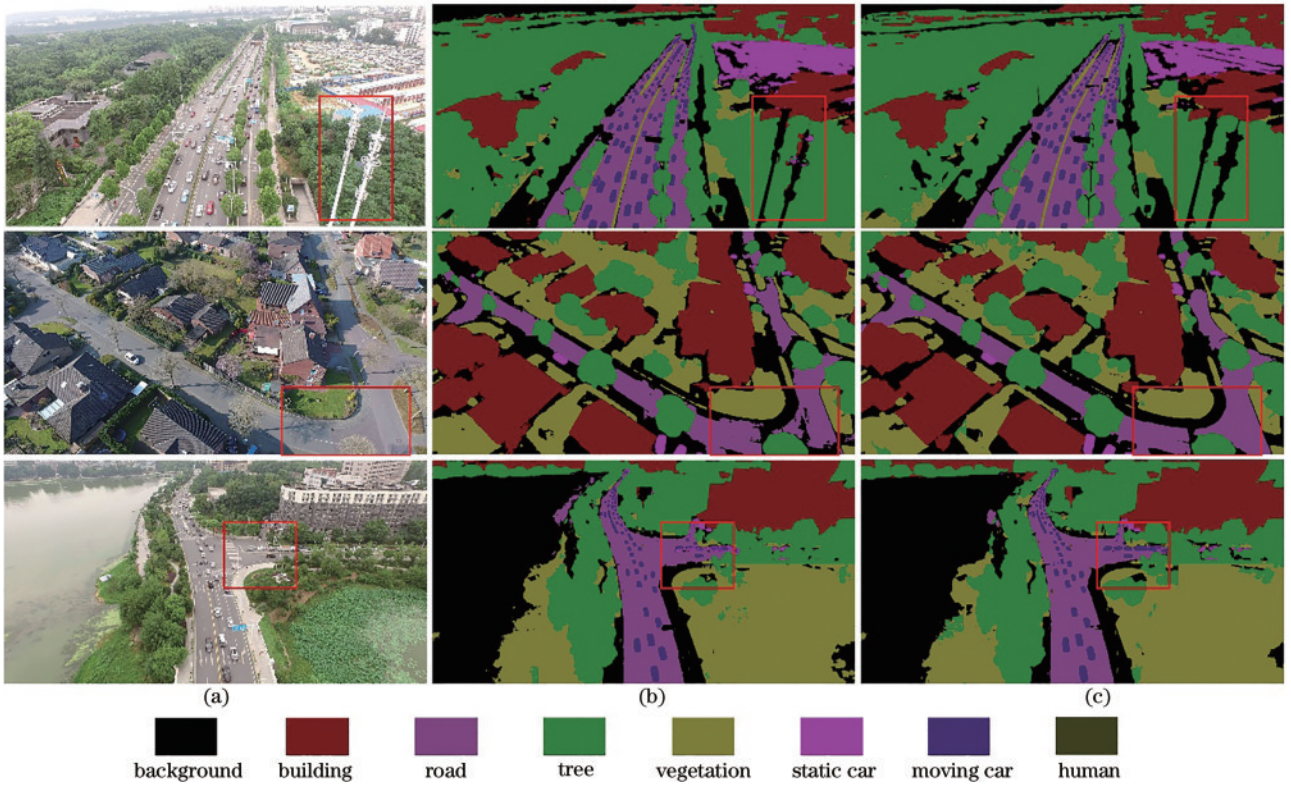

图 10　UAVid 测试集的可视化结果。(a)原始图像;(b)MANet;(c)TransMANet

Fig. 10　Visualization results on the UAVid test set. (a) Original image; (b) MANet; (c) TransMANet

以更好地捕捉全局特征的同时保留不同尺度的细节信息,TransMANet 可以更好地分割 LoveDA 数据集中的城市和乡村场景,处理不同的场景的分割任务。从表 2 可以看出,TransMANet 取得最高的 mIoU(54.8%),尤其在道路类别上的分割,展现出了惊人的优势,遥遥领先其他方法,IoU 至少高出 4.7 百分点。

表 2　不同方法在 LoveDA 数据集上的性能比较

Table 2　Performance comparison of different methods on LoveDA dataset

Model	IoU/%							mIoU/%
	Background	Building	Road	Water	Barren	Forest	Agriculture	
TransUNet[23]	43.0	56.1	53.7	78.0	9.3	44.9	56.9	48.9
Segmenter[26]	38.0	50.7	48.7	77.4	13.3	43.5	58.2	47.1
SwinUperNet[20]	43.3	54.3	54.3	78.7	14.9	45.3	59.6	50.0
DC-Swin[28]	41.3	54.5	56.2	78.1	14.5	47.2	62.4	50.6
MANet[17]	**46.5**	57.5	53.2	79.3	18.3	47.0	**67.2**	52.7
UNetFormer[16]	44.7	58.8	54.9	79.6	**20.1**	46.0	62.5	52.4
TransMANet	46.2	**59.7**	**60.9**	**82.3**	19.7	**49.9**	65.3	**54.8**

LoveDA 测试集的可视化结果如图 11 所示。

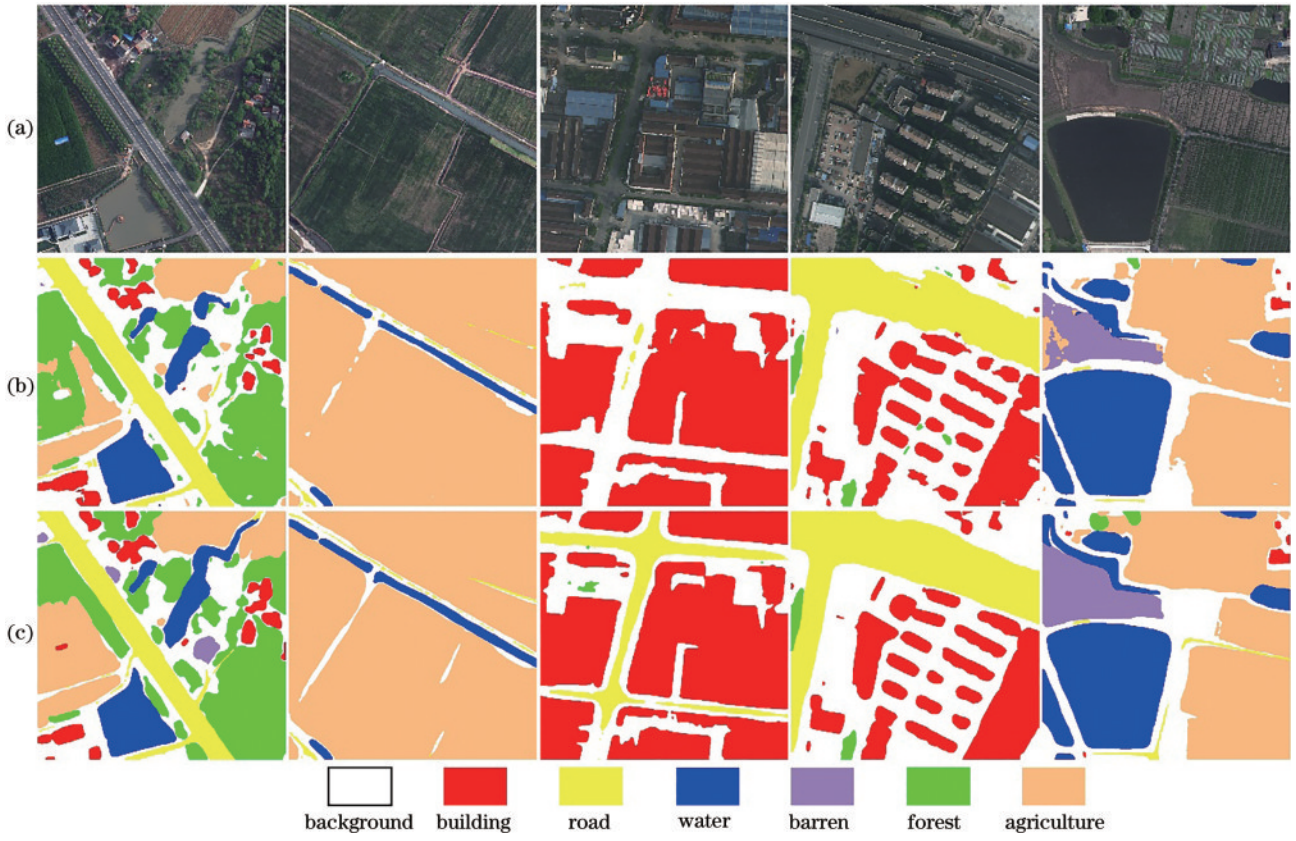

图 11　LoveDA 测试集的可视化结果。(a)原始图像;(b)MANet;(c)TransMANet

Fig. 11　Visualization results on the LoveDA test set. (a) Original image; (b) MANet; (c) TransMANet

为了进一步证实所提 TransMANet 的有效性,将所提方法与其他先进方法在 Vaihingen 和 Potsdam 数据集上进行了比较。在 Vaihingen 数据集上的实验结果如表 3 所示。所提方法在 Vaihingen 测试集获得了最佳的平均 F1 分数(91.3%)、OA(91.5%)和 mIoU(84.9%),超过了其他基于 CNN 和 Transformer 的网络,mIoU 比其他网络至少高出 2.2 百分点。并且值得注意的是,所提方法在汽车类别上获得了 91.0% 的 F1 分数,比其他网络高出 2.1 百分点以上。在 Potsdam 数据集上的实验结果如表 4 所示。TransMANet 超过了最近基于 Transformer 的网络,如 UNetFormer,也超过了基于 CNN 的 ABCNet。TransMANet 在各个类别上的分割结果均表现良好,取得了 93.1% 的平均 F1 分数、91.7% 的 OA 和 87.3% 的 mIoU。

表 3　不同方法在 Vaihingen 数据集上的性能比较

Table 3　Performance comparison of different methods on Vaihingen dataset

Model	F1/%					Mean F1/%	OA/%	mIoU/%
	Impervious surface	Building	Low vegetation	Tree	Car			
LANet[21]	92.4	94.9	82.9	88.9	81.3	88.1	89.8	
DeepLabV3+[9]	91.6	94.1	82.5	88.0	77.7	86.7	89.1	77.1
SwiftNet[24]	92.2	94.8	84.1	89.3	81.2	88.3	90.2	79.6
ABCNet[25]	92.7	95.2	84.5	89.7	85.3	89.5	90.7	81.3
MANet[17]	93.0	95.4	84.6	90.0	88.9	90.4	91.0	82.7
UNetFormer[16]	92.7	95.3	84.9	90.6	88.5	90.4	91.0	82.7
TransMANet	**93.4**	**95.9**	**85.6**	**90.7**	**91.0**	**91.3**	**91.5**	**84.9**

对所提方法与其他方法进行关于参数量(Parameters)和浮点运算数(FLOPs)的对比,如表 5 所示。相比轻量级网络,所提方法在增加极少的参数量和浮点运算数的前提下,大幅度提升了分割精度。相比 MANet,所提方法不仅参数量和浮点运算数都有明显减少,而且分割精度得到大幅度提高。

表 4 不同方法在 Potsdam 数据集上的性能比较

Table 4 Performance comparison of different methods on Potsdam dataset

Model	F1 /%					Mean F1 /%	OA /%	mIoU /%
	Impervious surface	Building	Low vegetation	Tree	Car			
LANet[21]	93.1	**97.2**	87.3	88.0	94.2	92.0	90.8	
DeepLabV3+[9]	92.1	95.3	85.6	86.5	94.8	90.9	89.2	84.2
SwiftNet[24]	91.8	95.9	85.7	86.8	94.5	91.0	89.3	83.8
ABCNet[25]	93.5	96.9	87.9	89.1	95.8	92.7	91.3	86.5
MANet[17]	93.4	97.0	**88.3**	89.4	96.5	92.9	91.3	87.0
UNetFormer[16]	93.6	**97.2**	87.7	88.9	96.5	92.8	91.3	86.8
TransMANet	**93.7**	96.9	**88.3**	**89.6**	**96.9**	**93.1**	**91.7**	**87.3**

表 5 参数量和计算量对比

Table 5 Comparison of parameter number and calculation amount

Model	Parameters /10^6	FLOPs /10^9
Segmenter[26]	6.7	26.8
SwiftNet[24]	11.8	51.6
ABCNet[25]	14.0	62.9
SegFormer[27]	13.7	63.3
MANet[17]	35.9	311.6
UNetFormer[16]	11.7	46.9
TransUNet[23]	90.7	803.4
SwinUperNet[20]	60.0	349.1
DC-Swin[28]	45.6	183.8
LANat[21]	23.8	22.0
DeepLabV3+[9]	39.7	30.7
TransMANet	27.2	112.5

4.5 消融实验

为了验证注意力机制、双分支解码器和联合损失函数的有效性，在 LoveDA、UAVid、Vaihingen 和 Potsdam 数据集上进行了消融实验。联合损失函数记为①，局部注意力嵌入机制记为②，双分支解码器记为③。

表 6 展示了在 LoveDA 数据集上进行消融实验的结果。结果表明：与 MANet 相比，使用局部注意力嵌入机制可以显著提高 mIoU，尤其对建筑、道路和森林，提升更明显，使用局部注意力嵌入机制可以更好地聚合上下文信息，mIoU 提高 0.8 百分点；双分支解码器可以显著提高分割精度，mIoU 提高 0.9 百分点，尤其是对建筑、道路、水面和森林，其中对建筑和道路的提升尤为突出，IoU 提升了 1.3 百分点和 4.2 百分点，这与解码器可以更好地建模长距离依赖有关；当使用了所有模块后，mIoU 显著提高了 2.1 百分点，在道路类别上更是具有极大的突破，IoU 提升了 7.7 百分点。结果表明，各个模块从不同角度利用了全局和局部上下文信息，为语义分割带来了重大突破。

表 7 展示了在 UAVid 数据集上进行消融实验的结果。与 MANet 相比，使用联合损失函数可以显著提升 mIoU，尤其是对小物体，如移动车辆和人类，IoU 提升 2.3 百分点和 11.1 百分点，验证了 Dice 损失函数可以极大地缓解遥感数据集的类别不平衡问题，解决了分割对象中的小物体像素量占比较少这一问题；使用了局部注意力嵌入机制后，增强了语义信息，显著提升了对背景、建筑和道路的分割精度，对小目标物体，如人类和移动车辆，也有一定提升，mIoU 提升了 1.3 百分点，OA 提升了 1.2 百分点；单独使用双分支解码器对大物体，如建筑和道路，指标有一定提升，OA 提升了 0.7 百分点；当使用所有模块后，网络可以更好地捕捉长距离依赖信息，对大物体指标均有不同程度的提升，对小物体指标提升极为显著，对静止车辆、移动车辆和人类，IoU 分别提升了 3.1 百分点、6.5 百分点和 11.5 百分点，mIoU 提升了 3.6 百分点，OA 提升了 1.3 百分点。

表 6 LoveDA 数据集上的消融实验

Table 6 Ablation study on LoveDA dataset

Model	IoU /%							mIoU /%
	Background	Building	Road	Water	Barren	Forest	Agriculture	
Base	46.5	57.6	53.2	79.3	18.3	47.0	67.2	52.7
Base+①	45.7	57.6	54.8	80.0	16.5	47.4	65.8	52.5
Base+②	46.4	60.2	56.4	80.7	18.4	48.9	63.5	53.5
Base+③	45.5	58.9	57.4	80.2	18.7	48.1	66.5	53.6
Base+①②③	46.2	59.7	60.9	82.3	19.7	49.9	65.3	54.8

表 7 UAVid 数据集上的消融实验

Table 7 Ablation study on UAVid dataset

Model	IoU /%								mIoU /%	OA /%
	Background	Building	Road	Tree	Vegetation	Static car	Moving car	Human		
Base	67.4	87.7	79.5	79.4	63.0	64.6	67.4	21.7	66.3	86.1
Base+①	68.8	88.2	81.5	80.2	64.3	64.1	69.7	32.8	68.7	86.9
Base+②	69.6	89.1	81.3	80.7	63.8	63.4	68.7	24.7	67.6	87.3
Base+③	68.6	89.0	80.8	80.2	63.4	59.1	63.0	23.9	66.0	86.8
Base+①②③	69.9	88.7	81.7	80.7	63.6	67.7	73.9	33.2	69.9	87.4

表 8 和表 9 分别展示了在 Vaihingen 和 Potsdam 数据集上的消融实验结果。基于联合损失函数的网络在 Vaihingen 数据集上的指标有一定程度的提升。相比 MANet，基于局部注意力嵌入机制的网络的各个评价指标显著，在 Vaihingen 和 Potsdam 数据集上的 mIoU 分别提升了 0.9 百分点和 0.1 百分点。相比 MANet，双分支解码器在 Vaihingen 和 Potsdam 数据集上的 OA 分别提升了 0.5 百分点和 0.3 百分点。

表 8 Vaihingen 数据集上的消融实验

Table 8 Ablation study on Vaihingen dataset

Model	F1 /%					Mean F1 /%	OA /%	mIoU /%
	Impervious surface	Building	Low vegetation	Tree	Car			
Base	93.0	95.4	84.6	90.0	88.9	90.4	91.0	82.7
Base+①	93.0	95.8	85.8	90.7	86.9	90.5	91.4	82.8
Base+②	93.4	96.0	85.7	90.7	89.1	91.0	91.6	83.6
Base+③	93.2	96.0	85.5	90.7	88.7	90.8	91.5	83.4
Base+①②③	93.4	95.9	85.6	90.7	91.0	91.3	91.5	84.9

表 9 Potsdam 数据集上的消融实验

Table 9 Ablation study on Potsdam dataset

Model	F1 /%					Mean F1 /%	OA /%	mIoU /%
	Impervious surface	Building	Low vegetation	Tree	Car			
Base	93.4	97.0	88.3	89.4	96.5	92.9	91.3	87.0
Base+①	93.1	96.7	88.3	89.6	96.5	92.9	91.3	86.9
Base+②	93.5	97.2	88.0	89.7	96.4	93.0	91.6	87.1
Base+③	93.6	97.0	88.2	89.5	96.3	92.9	91.6	87.0
Base+①②③	93.7	96.9	88.3	89.6	96.9	93.1	91.7	87.3

5 结 论

提出一种基于 MANet 的遥感图像语义分割算法 TransMANet。为了解决卷积神经网络的低级特征语义信息不足、高级特征空间细节模糊的问题，提出了局部注意力嵌入机制，可以增强上下文信息的嵌入，丰富了低级特征的语义信息。为了解决 CNN 提取全局上下文信息不充分的问题，设计了基于 Transformer 与 CNN 的双分支解码器，可以更好地提取全局上下文和局部上下文，对长距离依赖进行建模。最后，使用交叉熵损失函数与 Dice 损失函数联合的损失函数，使网络更加关注数量少的类别，从而缓解遥感数据集的类别不平衡问题，提高分割准确度。大量的实验结果表明：所提算法与 MANet 算法在多个数据集上相比，具有更优的分割效果和上下文信息建模能力；与其他算法相比，所提算法的各项指标均有显著提升，且应对复杂场景、小目标等问题的效果较好。然而，所提算法的网络结构复杂，网络深度较大，计算量相比轻量级网络还是较大，后续将改进网络结构，进一步优化编码器和解码器，以提升算法性能。

参 考 文 献

[1] 陈玲，贾佳，王海庆. 高分遥感在自然资源调查中的应用综述[J]. 国土资源遥感，2019，31(1)：1-7.
Chen L，Jia J，Wang H Q. An overview of applying high resolution remote sensing to natural resources survey[J]. Remote Sensing for Land & Resources，2019，31(1)：1-7.

[2] 潘银，邵振峰，程涛，等. 利用深度学习模型进行城市内涝影响分析[J]. 武汉大学学报(信息科学版)，2019，44(1)：132-138.
Pan Y，Shao Z F，Cheng T，et al. Analysis of urban waterlogging influence based on deep learning model[J]. Geomatics and Information Science of Wuhan University，2019，44(1)：132-138.

[3] Huang D Y, Wang C H. Optimal multi-level thresholding using a two-stage Otsu optimization approach[J]. Pattern Recognition Letters, 2009, 30(3): 275-284.

[4] Al-Amri M S S, Kalyankar D N, Dr Khamitkar S D. Image segmentation by using edge detection[J]. International Journal on Computer Science and Engineering, 2010, 2(3): 804-807.

[5] Wu F Y, Yang X H, Ma Y Y, et al. Machine-learning guided optimization of laser pulses for direct-drive implosions [J]. High Power Laser Science and Engineering, 2022, 10: e12.

[6] Döpp A, Eberle C, Howard S, et al. Data-driven science and machine learning methods in laser-plasma physics[J]. High Power Laser Science and Engineering, 2023, 11: e55.

[7] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(4): 640-651.

[8] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W M, et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.

[9] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-06-17)[2023-06-05]. https://arxiv.org/abs/1706.05587.

[10] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[11] Lou A G, Loew M. CFPNET: channel-wise feature pyramid for real-time semantic segmentation[C]//2021 IEEE International Conference on Image Processing (ICIP), September 19-22, 2021, Anchorage, AK, USA. New York: IEEE Press, 2021: 1894-1898.

[12] Lin J P, Haberstroh F, Karsch S, et al. Applications of object detection networks in high-power laser systems and experiments[J]. High Power Laser Science and Engineering, 2023, 11: e7.

[13] Zhao Q, Liu J H, Li Y W, et al. Semantic segmentation with attention mechanism for remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 5403913.

[14] 白宗宝, 张俊举, 高原, 等. 基于注意力机制的航拍图像目标检测算法[J]. 激光与光电子学进展, 2023, 60(12): 1215003.
Bai Z B, Zhang J J, Gao Y, et al. Attention mechanism-based object detection algorithm in aerial images[J]. Laser & Optoelectronics Progress, 2023, 60(12): 1215003.

[15] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22)[2023-06-05]. https://arxiv.org/abs/2010.11929.

[16] Wang L B, Li R, Zhang C, et al. UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 190: 196-214.

[17] Li R, Zheng S Y, Zhang C, et al. Multiattention network for semantic segmentation of fine-resolution remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5607713.

[18] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 3141-3149.

[19] Tsai Y H H, Bai S J, Yamada M, et al. Transformer dissection: an unified understanding for transformer's attention via the lens of kernel[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China. Stroudsburg: Association for Computational Linguistics, 2019: 4344-4353.

[20] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 9992-10002.

[21] Ding L, Tang H, Bruzzone L. LANet: local attention embedding to improve the semantic segmentation of remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(1): 426-435.

[22] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7794-7803.

[23] Chen J N, Lu Y Y, Yu Q H, et al. TransUNet: transformers make strong encoders for medical image segmentation[EB/OL]. (2021-02-08) [2023-06-05]. https://arxiv.org/abs/2102.04306.

[24] Oršić M, Šegvić S. Efficient semantic segmentation with pyramidal fusion[J]. Pattern Recognition, 2021, 110: 107611.

[25] Li R, Zheng S Y, Zhang C, et al. ABCNet: attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 181: 84-98.

[26] Strudel R, Garcia R, Laptev I, et al. Segmenter: transformer for semantic segmentation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 7242-7252.

[27] Xie E Z, Wang W H, Yu Z D, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[EB/OL]. (2021-05-31) [2023-06-05]. https://arxiv.org/abs/2105.15203.

[28] Wang L B, Li R, Duan C X, et al. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 6506105.