# 支持模态扩展的高效遥感图像分割网络

张众维[1]，朱国庆[1]，秦海洁[2]，刘树东[1]
（1. 天津城建大学计算机与信息工程学院，天津 300384；2. 深圳宏芯宇电子股份有限公司，广东 深圳 518000）

**摘　要：**语义分割是遥感领域中的一项基础且关键的技术。传统的单模态方法不能全面捕捉地物的多样化特征，而多模态研究多采用双模态输入，难以全面挖掘更多模态数据之间的潜在互补性。为此，提出了一种支持模态扩展的高效遥感图像分割网络。首先，在设计前期提取了注意头模块，以区分不同模态在融合阶段的重要程度；其次，为了校正偏移并有效捕获各模态异质性特征，设计了跨模态对齐多头注意力模块；再次，为了提升模型对全局信息的建模能力，设计了维度协同线性注意力模块；最后，引入了辅助损失函数监督解码端的中间特征学习。该网络利用早期融合策略，实现了超过两种模态的输入，且在扩展至多模态时，计算成本和参数开销仅小幅增长。在 Potsdam 和 Vaihingen 数据集上进行的若干实验表明，所提网络优于几种现有主流单模态和多模态遥感分割网络。

**关键词：**语义分割；多模态融合；遥感图像；注意力机制；模态扩展
**中图分类号：**TP751
**文献标志码：**A
**doi：**10.11959/j.issn.2096−3750.WLW25146

# Efficient remote sensing image segmentation network with modality expansion capability

Zhang Zhongwei[1], Zhu Guoqing[1], Qin Haijie[2], Liu Shudong[1]
1. School of Computer and Information Engineering, Tianjin Chengjian University, Tianjin 300384, China
2. Shenzhen HosingGlobal Electronics Co., Ltd., Shenzhen 518000, China

**Abstract:** Semantic segmentation is key in remote sensing, but traditional unimodal methods fail to capture diverse object features fully. Most multimodal studies were confined to bimodal inputs, limiting the exploration of potential complementarity between multimodal data. To address this, an efficient remote sensing image segmentation network with modality expansion capability was proposed. First, a pre-extraction attention head module was designed to distinguish the importance of different modalities in the fusion stage. Second, a cross-modal alignment multi-head attention module was designed to correct offsets and effectively capture heterogeneous features of each modality. Third, a dimensional collaborative linear attention module was designed to enhance the model's global information modeling capability. Finally, an auxiliary loss function was introduced to supervise intermediate feature learning in the decoder stage. The design employs an early fusion strategy, supporting inputs from more than two modalities, and when extended to multiple modalities, the computational cost and parameter overhead increase only slightly. Experiments on the Potsdam and Vaihingen datasets demonstrate that the proposed method outperforms several compared mainstream unimodal and multimodal remote sensing segmentation networks.

**Key words:** semantic segmentation, multimodal fusion, remote sensing image, attention mechanism, modal expansion

**收稿日期：**2025−09−12；**修回日期：**2025−12−08
**通信作者：**刘树东，liushudongtcu@163.com
**基金项目：**天津市哲学社会科学规划项目（No. TJGL19XSX-045）
**Foundation Item:** Tianjin Philosophy and Social Science Planning Project (No. TJGL19XSX-045)

## 0 引言

基于遥感图像的土地资源分割作为遥感和计算机领域的关键研究方向，具有极其重要的现实意义。它借助先进的遥感技术获取丰富的地表信息，再运用计算机领域的智能算法，精准识别出不同的土地覆盖类型，如耕地、林地、水域、建设用地等[1-2]。这种精准分割能够为环境监测提供基础数据，及时发现环境变化趋势。在灾害评估中，该技术可快速掌握受灾土地情况，还能为土地管理部门的科学规划与合理决策提供有力支持，促进土地资源的可持续利用。

传统的分割方法包括以下几种：（1）基于边缘的分割方法，通过检测图像中的边缘信息来识别物体的轮廓[3]；（2）基于域的方法，通过定义特定区域或结构来提取图像的语义信息；（3）基于图的方法，利用图结构表示像素之间的关系并进行分割[4]；（4）基于机器学习的方法，借助分类器对像素进行分类并赋予相应的语义标签[5]。然而，遥感图像通常面临尺度变化大、类内差异大和类间差异小等挑战，传统方法难以有效应对。相比之下，基于深度学习的方法通过端到端的数据驱动特征学习，在处理复杂遥感图像时展现出显著优势。

## 1 相关工作

在常见的深度学习方法中，全卷积网络（fully convolutional network，FCN）[6]是使用卷积神经网络（convolutional neural network，CNN）进行单模态语义分割的开创性尝试。U 型网络（U-shaped network，UNet）[7]通过跳跃连接将编码器中的高级特征与解码器的相应层相连接，能够精确识别遥感场景中的小目标和复杂地形。金字塔场景解析网络（pyramid scene parsing network，PSPNet）[8]通过在特征图上进行多尺度池化，并将池化后的特征连接起来，成功整合了全局与局部信息。DeepLab 系列[9]模型利用膨胀卷积，并行捕获不同尺度的上下文信息，能够在复杂场景中高效处理多样化地物，提升分割的精度。高分辨率网络（high-resolution network，HRNet）[10]通过并行保留低分辨率和高分辨率分支，实现对不同层次遥感地物特征的感知和学习。视觉 Transformer（vision transformer，ViT）[11]首次将 Transformer[12]模型引入计算机视觉领域，将图像转化为补丁序列并引入多头自注意力机制。Swin Transformer[13]引入了滑动窗口自注意力机制，极大地降低了自注意力机制的计算复杂度。上述研究使用单一图像模态作为模型输入，但在一些复杂场景中，如具有相似纹理和颜色的对象，由于感知信息的局限，单模态模型通常难以准确区分它们。

随着数据采样技术的不断发展，为了实现高精度的分割，越来越多的应用场景能够提供更加多样化的数据信息，尤其是在遥感领域，通过结合光学图像三原色（red green blue，RGB）、数字表面模型（digital surface model，DSM）、激光雷达（light detection and ranging，LiDAR）等多模态数据，可以更全面地捕捉图像中的语义信息，显著提升模型的分割性能。为此，模态信息交换与融合网络（modal information exchange and fusion network，MIEFNet）[14]将 RGB 与 DSM 数据结合，有助于增强类间差异；空间注意力融合网络（spatial attention fusion network，SAFN）[15]则将 RGB 与 LiDAR 数据结合，增强多模态数据中判别性信息的获取能力。高效深度引导泛化网络（efficient depth-guided generalization network，EDGNet）[16]利用 DSM 信息和边界信息来调整各模态的特征权重，缓解多模态特征之间的差异。多模态全局上下文网络（multi-modal global context network，MGCNet）[17]在编码器中设计了差分融合模块，提取 RGB 和 DSM 两模态的差分权重，实现了两模态的特征融合。上述多模态网络多采用中期融合策略，为每个模态分配独立的主干网络，导致了模型参数量显著增加，计算成本随之大幅增长，这对硬件性能提出了巨大的挑战，难以满足轻量化场景的需求。

早期融合策略通过在特征提取前统一处理多模态信息，能够在较低计算成本下充分利用不同模态的互补特征，并避免了为每个模态独立设计主干网络所带来的结构冗余与参数开销。但早期融合往往仅采用特征级的简单拼接、通道叠加等方式，缺乏对模态间相关性与互补性的深入建模。因此，如何设计一个既能充分利用多模态信息、结构灵活可调，又能在计算成本上保持可控的网络，成为亟待解决的问题。

受早期融合策略的启发，本文提出了一种支持模态扩展的高效遥感图像分割网络——高效多模态

分割网络（efficient multimodal segmentation network，EMSegNet）。该网络采用早期融合策略，在较低计算成本下将模态数量扩展，即外推到$N$个（$N$≥1），充分利用了不同模态的信息。本文主要贡献如下。

（1）设计了多模态遥感图像分割网络EMSegNet，在无须为每个模态单独分配主干网络的前提下，实现了模态数量的外推，有效减少了模型的参数量和计算复杂度。

（2）设计前期提取注意头（early extraction attention head, E2HA）模块，通过模态身份令牌和通道注意力机制，更好地区分不同模态在融合过程中的重要性。

（3）设计跨模态对齐多头注意力（cross-modal multi-head attention，CMHA）模块，校正不同模态传感器成像时产生的空间偏移和几何失真，同时搜索多模态特征库中的互补信息。

（4）设计维度协同线性注意力（dimensional collaborative linear-attention block，DCA）模块，实现跨维度的特征协调，使模型能够更好地理解空间维度和通道维度的全局依赖关系。

## 2　方法

在网络模型设计过程中，主要考虑了以下几个关键问题：第一，如何有针对性地表达模态的重要性权重，平衡模态间的信息利用；第二，如何在确保信息完整性的同时筛选和提取互补特征，避免冗余信息干扰；第三，如何增强模型的全局与局部上下文建模能力，最终提出了EMSegNet。

### 2.1　总体框架

EMSegNet整体架构如图1所示，该架构采用了改进的早期融合策略[18-19]，支持$N$种模态的数据作为输入（$N$≥1，此处$N$=4）。网络整体分为4个部分：早期融合与跨模态对齐部分（E2HA和CMHA）、编码部分、解码部分和辅助损失部分。

在早期融合与跨模态对齐部分中，每一个输入模态被分配了一个轻量级的E2HA模块，用以提取该模态内的特征分布。随后，对所有E2HA的输出进行拼接，使CMHA模块的输入实现特征对齐。根据需要调整输入模态数量，当每引入一种模态时，仅需要额外分配一个E2HA模块，而无须大幅度增加网络的复杂度。在编码部分，DCA与在图像网络（image network，ImageNet）上预训练的ResNet-50各阶段形成4个骨干块，分别提取并输出特征，随后通过跳跃连接到对应的解码端，以获得全局和局部特征信息。解码部分通过转置卷积分阶段进行上采样来逐步恢复原图大小。在辅助损失部分，设计

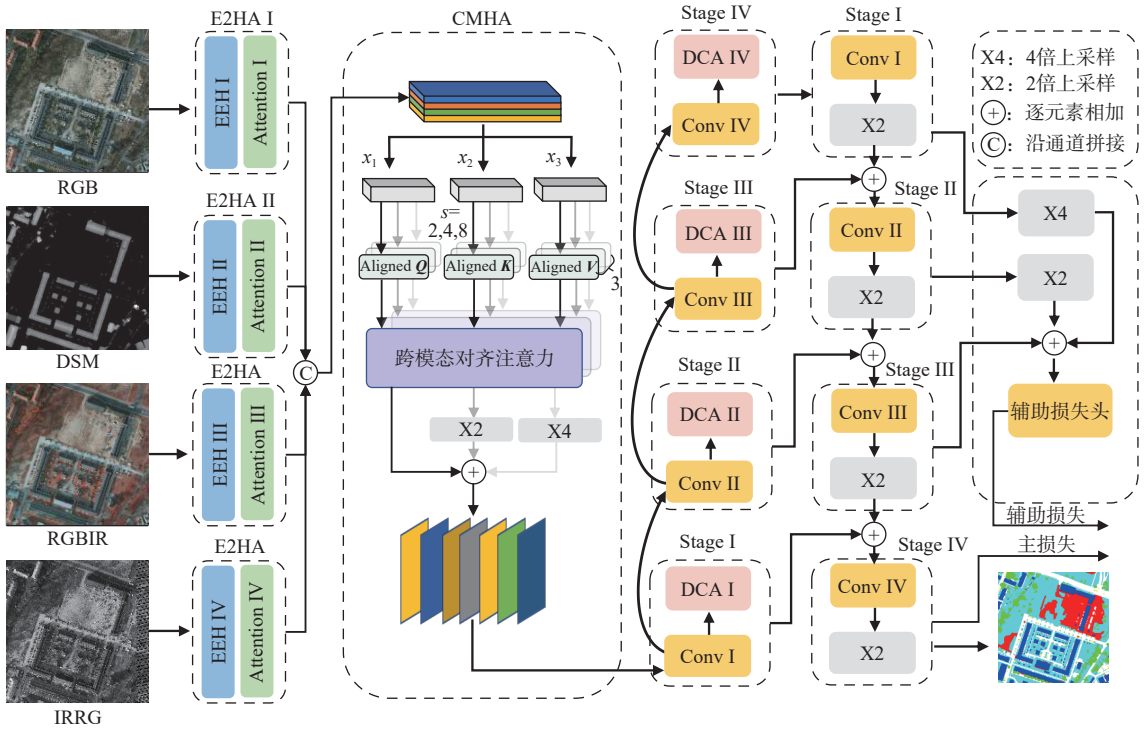



图1　EMSegNet整体架构

辅助头来缩小深层特征和浅层特征之间的语义差距。最终通过主损失和辅助损失共同监督网络的训练，从而提高整体分割性能和网络的泛化能力。

### 2.2 E2HA模块

该模块利用多模态学习中的模态对齐和自适应融合思想，能够在解决传统融合模态冲突问题的同时，通过对各个模态自适应地调整权重，有效区分不同模态在特征融合过程中的重要性。E2HA模块结构如图2所示。此外，E2HA的轻量级架构设计显著降低了额外的计算量和参数开销，使模型能够在低成本的前提下外推至多模态场景（$N \geqslant 1$）。

对于某一模态的E2HA，具体细节如下。

首先，对输入原始模态使用卷积操作进行升维：

$$C(\boldsymbol{x}) = \mathrm{Conv}_{3*3}(\boldsymbol{x}) \tag{1}$$

其中，$\boldsymbol{x}$为输入特征图。

其次，以归纳偏差的形式为该模态设计一个唯一的身份令牌，用以区分不同模态。这些标识符通过嵌入层进行学习，为模型引入关于每个模态的先验知识，使得模型能够识别并区分不同模态。假设模态索引集合为$M=\{m_0, m_1, \cdots, m_{M-1}\}$，模态索引$m_i \in \{0, 1, \cdots, M-1\}$，定义嵌入矩阵$\boldsymbol{W}_{\mathrm{embed}}$：

$$\boldsymbol{W}_{\mathrm{embed}} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{bmatrix}, w_m \in \mathbf{R}^d \tag{2}$$

其中，$d$为特征图通道数。生成对应的嵌入向量集合如下：

$$E[i,:] = \boldsymbol{W}_{\mathrm{embed}}[m_i] \tag{3}$$

将所有模态的嵌入向量批处理后得到嵌入向量：

$$\boldsymbol{E}(m) = \mathrm{Embedding}(M) \tag{4}$$

最后，通过广播操作将模态的嵌入向量扩展到与特征图的空间维度一致：

$$B(m) = \mathrm{Broadcast}\big(E(m)\big) \tag{5}$$

其中，Broadcast( )为广播操作。

模型采用组卷积操作，将特征图沿着通道维度分为$N$组，并且各组之间不进行跨组通信。通过残差连接，将升维后的特征信息与组卷积结果相加，以此作为后续通道注意力模块的输入。组卷积如下：

$$G(\boldsymbol{x}) = \mathrm{GroupConv}_{n=4}(\boldsymbol{x}) \in \mathbb{R}^{h*w*d} \tag{6}$$

其中，$h$、$w$、$d$分别表示特征图的高、宽和通道数，$n$表示组卷积分组数。

E2HA的输出如下：

$$F(\boldsymbol{x}) = G\big(\mathrm{Conv}_{1*1}\big(C(\boldsymbol{x})\big) + B(m)\big) + C(\boldsymbol{x}) \tag{7}$$

由于特征图中不可避免地包含了冗余或无关的信息（通常表现为负值较多的低频通道），在融合过程中可能会引入噪声。为了抑制这些信息，利用通道注意力机制，重点关注高频信息，选择高质量的模态内特征。通道注意力如下：

$$A(\boldsymbol{x}) = \sigma\big(\mathrm{MLP}\big(\mathrm{AvgPool}(\boldsymbol{x})\big) + \mathrm{MLP}\big(\mathrm{MaxPool}(\boldsymbol{x})\big)\big) \tag{8}$$

其中，$\sigma$( )表示ReLU6激活函数；AvgPool( )、MaxPool( )分别表示全局平均池化和全局最大池化；MLP( )为线性变换。

最终，E2HA模块的输出如下：

$$Y(\boldsymbol{x}) = A\big(F(\boldsymbol{x})\big) \tag{9}$$

### 2.3 CMHA模块

与单模态情况不同，由于传感器的视场差异和多模态图像的特性，在处理双模态或多模态输入时，需要重点关注模态间像素对齐与互补特征的提取问题。为此，本文参照Transformer中的多头自注意力机制，设计了CMHA模块，以并行地捕捉输入数据的特征。如图1所示，CMHA可以细分为两个子过程，即多头跨模态对齐和多头自注意力。

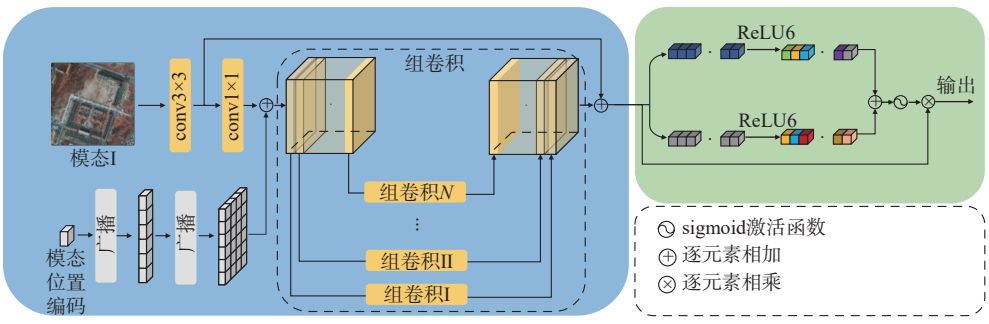



图2　E2HA模块结构

### 2.3.1　多头跨模态对齐

多头跨模态对齐的核心思想是将拼接后的多模态特征图划分为多个大小为$M\times M$的窗口，并计算得出每个窗口内的代表值作为该窗口的描述符。这样即使图像存在整体的偏移，只要各模态的偏移量不超过划定的窗口范围$M$，多模态关键信息就可以大致保持对齐。为确保描述符能够充分代表每个窗口内的重要信息，将窗口的代表值定义为窗口内像素的最大值，因此采用最大池化操作，这一过程可表示为：

$$K(X)=\mathrm{MaxPool}(\boldsymbol{X},N)=\left\{\max_{(i,j)\epsilon K}X(i,j)\,|K\epsilon\mathrm{Windows}(N)\right\} \tag{10}$$

其中，MaxPool( ,$N$)表示步长和窗口大小均为$N$的最大池化操作，Windows($N$)表示将特征图划分为$N\times N$的窗口集合。

为了有效校正不同模态传感器成像时产生的空间偏移，跨模态对齐模块设计为多头机制，采用可变窗口大小的对齐方式。较小的窗口适合于对齐高分辨率模态的细粒度信息，而较大的窗口可以用于对齐低分辨率或广域模态的信息，从而为不同分辨率和尺度的模态数据提供更合适的对齐方式。

CMHA模块的输入为每个模态的E2HA输出沿通道维度进行的堆叠，CMHA模块输入记为$\boldsymbol{X}=[x_1,x_2,\cdots,x_n]$，$n$表示输入模态数。首先将输入特征映射到适合跨模态对齐的子空间中，使用3个卷积核大小为1×1的卷积操作将$\boldsymbol{X}$映射为3个不同的副本，分别表示为$\boldsymbol{X}_1$、$\boldsymbol{X}_2$和$\boldsymbol{X}_3$：

$$\begin{cases}\boldsymbol{X}_1=\boldsymbol{X}\boldsymbol{W}_1\\ \boldsymbol{X}_2=\boldsymbol{X}\boldsymbol{W}_2\\ \boldsymbol{X}_3=\boldsymbol{X}\boldsymbol{W}_3\end{cases} \tag{11}$$

其中，$\boldsymbol{W}_1$、$\boldsymbol{W}_2$、$\boldsymbol{W}_3\in\mathbf{R}^{D\times D_k}$，$D$是输入特征的维度或通道数，$D_K$是映射到新特征空间的维度。

接着，为每个副本划分$n$ ($n$=3)个头，每个头分别使用不同窗口大小的池化操作提取特征图的代表值，对特征图进行对齐。具体而言，这3个头的池化窗口分别为2×2、4×4和8×8，对应的最大池化步长也为2、4和8。然后，对池化后特征图的高和宽进行重塑，将三维张量展平为二维向量，分别得到对齐后的多头Aligned_$Q_i$( )、Aligned_$K_i$( )和Aligned_$V_i$( )：

$$\mathrm{Align_Q}_i(\boldsymbol{X}_1)=\mathrm{Flatten}(\mathrm{MaxPool}(\boldsymbol{X}_1,N)_{N=i})\in\mathbf{R}^{D_k,M_i} \tag{12}$$

$$\mathrm{Align_K}_i(\boldsymbol{X}_2)=\mathrm{Flatten}(\mathrm{MaxPool}(\boldsymbol{X}_2,N)_{N=i})\in\mathbf{R}^{D_k,M_i} \tag{13}$$

$$\mathrm{Align_V}_i(\boldsymbol{X}_3)=\mathrm{Flatten}(\mathrm{MaxPool}(\boldsymbol{X}_3,N)_{N=i})\in\mathbf{R}^{D_k,M_i} \tag{14}$$

其中，Flatten( )为展平操作；$M_i=H_i\times W_i$，$H_i$和$W_i$是第$i$个头特征图的高度和宽度；$N$为池化窗口大小，$N=i,\ i\in[2,4,8]$。

### 2.3.2　多头自注意力

DCA模块结构如图3所示。为了避免相似特征的冗余堆叠，捕获各模态之间的异质性，本文引入了空间注意力机制，为每个头对齐后的结果自动调整注意力权重，使模型能够在不同模态之间自适应地选择最具信息性的特征区域，动态聚焦在对任务

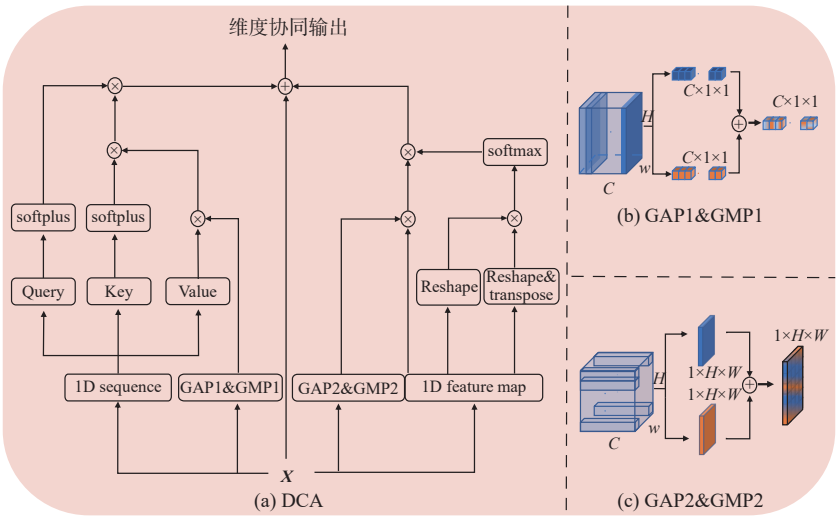



图3　DCA模块结构

最有贡献的模态特征上。同时，为了平衡计算成本和模型复杂性，本文选用了DCA模块中的空间分支（图3(a)左半分支）作为捕获各模态间异质性的核心组件，多头自注意力表示如下：

$$A(\boldsymbol{x})=\frac{\mathrm{Softplus}\left(\mathrm{Aligned_Q}\right)*(\mathrm{Softplus}\left(\mathrm{Aligned_K}\right)^{\mathrm{T}})*(\mathrm{Aligned_V}*D_{\mathrm{c}}(\boldsymbol{x}))}{\sqrt{D}} \tag{15}$$

其中，Softplus()为激活函数。

每个头使用的池化核大小不同，导致相应的注意力结果必然在维度上存在差异，为了使这些结果在融合时能够匹配，CMHA模块对第二个头的输出进行2倍上采样，对第3个头的输出进行4倍上采样，使其空间分辨率与第一个头保持一致。最终，将3个头的结果进行逐元素相加，得到CMHA模块的整体输出。

### 2.4　DCA模块

DCA模块对交叉注意力进行扩展，引入空间维度和通道维度的双向注意力计算，通过信息的相互融合来增强全局的特征表示。DCA采用两种不同的协同增强策略。首先，对于输入的特征映射$\boldsymbol{X}$，通过3个权重矩阵生成向量$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$，以便进行后续的注意力计算，如下：

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{X}\\ \boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}_k\\ \boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_v\end{cases} \tag{16}$$

其中，$\boldsymbol{X}\epsilon\mathbf{R}^{B\times N\times D}$表示输入的特征向量；$N=W\times H$，$W$和$H$分别表示输入图像的高度和宽度；$D$表示输入特征向量的维度或通道数；$\boldsymbol{W}_{\mathrm{q}}\epsilon\mathbf{R}^{D_x\times D_k}$、$\boldsymbol{W}_{\mathrm{k}}\epsilon\mathbf{R}^{D_x\times D_k}$和$W_{\mathrm{v}}\epsilon\,\mathbf{R}^{D_x\times D_k}$为权重矩阵；$\boldsymbol{Q}\in\mathbf{R}^{B\times N\times D_k}$、$\boldsymbol{K}\in\mathbf{R}^{B\times N\times D_k}$、$\boldsymbol{V}\in\mathbf{R}^{B\times N\times D_k}$。

在空间注意部分（图3(a)左半分支），通过引入核注意力机制[18]，将空间注意部分的计算复杂度降到$O$（$N$），并对其进行优化。具体而言，先对特征向量进行归一化操作，然后进行点积运算：

$$A_s(x)=\frac{\mathrm{Softplus}\left(\boldsymbol{Q}\right)*\left(\mathrm{Softplus}\left(\boldsymbol{K}\right)^{\mathrm{T}}\boldsymbol{V}\right)}{\sqrt{D}} \tag{17}$$

为了增强空间注意部分对通道维度的特征表达能力，本文设计了GAP1&GMP1模块（如图3(b)所示），负责提取并整合通道维度的信息。结构上，GAP1&GMP1分别沿通道维度对特征映射$\boldsymbol{X}$执行全局平均池化和全局最大池化操作，并将二者的输出融合在一起：

$$D_{\mathrm{c}}(\boldsymbol{x})=\mathrm{GAP}_{\mathrm{channel}}(\boldsymbol{x})+\mathrm{GMP}_{\mathrm{channel}}(\boldsymbol{x}) \tag{18}$$

功能上，这一融合结果被作为通道权重系数，显式地添加到值向量$\boldsymbol{V}$中，从而提升通道信息在空间注意力中的影响力。协同通道维度的线性空间核注意力机制表示如下：

$$A_{\mathrm{p}}(\boldsymbol{x})=\frac{\mathrm{Softplus}\left(\boldsymbol{Q}\right)*(\mathrm{Softplus}\left(\boldsymbol{K}\right)^{\mathrm{T}}(\boldsymbol{V}D_{\mathrm{c}}(\boldsymbol{x}))}{\sqrt{D}} \tag{19}$$

在通道注意力部分（图3(a)右半分支），输入通道数$C$远小于特征图中像素数($C<N$)，因此通道注意力的复杂度为$O(C_2)$远小于$O(N_2)$。通过基于点积的通道注意力机制捕获特征图通道维度信息：

$$A_{\mathrm{k}}(\boldsymbol{x})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{D}}\right)\boldsymbol{V} \tag{20}$$

其中，Softmax()为激活函数。

同时，为了增强通道注意力部分对空间维度的特征表达能力，设计GAP2&GMP2模块（如图3(c)所示）对空间维度信息进行整合。与空间注意力部分类似，GAP2&GMP2分别沿空间维度对特征映射$\boldsymbol{X}$执行全局平均池化和全局最大池化操作，并将二者的输出融合在一起：

$$D_{\mathrm{p}}(\boldsymbol{x})=\mathrm{GAP}_{\mathrm{position}}(\boldsymbol{x})+\mathrm{GMP}_{\mathrm{position}}(\boldsymbol{x}) \tag{21}$$

然后，显式地添加到值向量$\boldsymbol{V}$中。协同空间维度的通道注意力机制如下：

$$A_{\mathrm{c}}(\boldsymbol{x})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{D}}\right)\left(\boldsymbol{V}D_{\mathrm{p}}(\boldsymbol{x})\right) \tag{22}$$

最后，将两个分支的输出相加，实现全局信息感知的维度协同DCA模块输出：

$$A(\boldsymbol{x})=A_{\mathrm{p}}(\boldsymbol{x})+A_{\mathrm{c}}(\boldsymbol{x})+\boldsymbol{x} \tag{23}$$

在处理遥感图像时，模型需要具备同时捕捉局部细节和感知全局上下文信息的能力。将所提出的DCA模块与在ImageNet数据集上预训练的ResNet-50各阶段进行组合，形成4个骨干块。ResNet-50的骨干网络通过卷积操作提取多层次的局部特征，DCA模块通过空间和通道维度的协同注意力机制，增强了每个阶段特征的全局感知，二者结合使特征提取更全面。

### 2.5　辅助损失头

在解码过程中，由于要从多个模态中提取和融合特征，信息传递和上采样的复杂性可能使一些模态重要特征丢失或被弱化。尤其是在深层网络中，

部分模态的信息可能无法在最终的输出中得到完整保留。如果仅依靠主损失来优化模型，模型可能难以有效协调不同模态特征的融合。因此，引入辅助损失头，通过对解码过程的中间阶段引入监督信号，可确保每个解码阶段的输出不仅依赖最终的主损失，还能够得到更多的中间反馈。这有助于保证每个模态的特征在解码过程中得到充分利用和表达。

如图1所示，将解码端的Stage I输出和Stage II输出分别上采样4倍和2倍，以匹配Stage III尺寸，随后将它们相加融合。在辅助头中，融合后的特征经过卷积和上采样操作，进一步进行特征学习，最终将通道维度映射为数据集类别维度，作为辅助损失函数的输入。

## 3　实验

### 3.1　数据集

本文在两个多模态遥感数据集ISPRS Potsdam和ISPRS Vaihingen上进行了对比与消融实验。这两个数据集均由6个类别的数据组成：建筑物、树木、汽车、不透水表面、低矮植被和杂项。

（1）Potsdam数据集：包含4种不同模态的遥感数据，有RGB、DSM、红绿蓝+近红外（RGBIR）和近红外红绿（IRRG）。每种模态包含38张高分辨率影像，图像的空间分辨率为5 cm，图像尺寸为6 000×6 000像素。数据集中包含小目标物体、狭窄的道路和密集的交通场景，对分割网络提出了较高的要求和挑战。

（2）Vaihingen数据集：包含两种不同模态的遥感数据，即IRRG和DSM。每种模态包含33张高分辨率影像，图像的空间分辨率为9 cm。与Potsdam相比，Vaihingen的空间分辨率较低，这意味着每个像素代表更大的地面面积，在处理过程中更容易丢失细节和上下文信息。

### 3.2　损失函数

在训练过程中，本文采用交叉熵损失与骰子损失的联合损失作为主损失函数($\mathcal{L}_{\text{main}}$)，以同时优化分类准确性和分割精度。其中，交叉熵损失($\mathcal{L}_{\text{ce}}$)用于优化模型的逐像素分类能力，能够有效提升模型的类别判别能力，特别适用于类别边界模糊或存在噪声标注的遥感分割任务：

$$\mathcal{L}_{\text{ce}} = -\sum_{c=1}^{C} \tilde{y}_c \log p_c \tag{24}$$

其中，$C$表示类别数，$\tilde{y}_c$为真实标签的一位有效编码（one-hot）表示，$p_c$为预测概率。

骰子损失($\mathcal{L}_{\text{dice}}$)可用于优化模型的区域一致性与小目标识别能力，在类别不平衡（如树木、汽车类较少）时可显著提升模型性能，如下：

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum Y\check{Y}_{\text{main}}}{\sum Y + \sum \check{Y}_{\text{main}}} \tag{25}$$

其中，$Y$、$\check{Y}_{\text{main}}$分别为真实标签掩码和主分支预测掩码。

主损失函数通过加权方式组合上述两种损失：

$$\mathcal{L}_{\text{main}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} \tag{26}$$

其中，$\lambda_{\text{ce}}$与$\lambda_{\text{dice}}$分别表示两部分损失的权重，在本文实验中设为$\lambda_{\text{ce}}$=1.0，$\lambda_{\text{dice}}$ = 1.0。

此外，单独引入软交叉熵损失作为辅助损失函数($\mathcal{L}_{\text{aux}}$)加速模型的收敛并监督中间特征，如下：

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{ce}} \tag{27}$$

最终，用于训练的总损失$\mathcal{L}_{\text{total}}$为：

$$\mathcal{L}_{\text{total}} = \begin{cases} \mathcal{L}_{\text{main}} + 0.4\mathcal{L}_{\text{aux}}, \text{处于训练阶段且Logit为2} \\ \mathcal{L}_{\text{main}}, \text{其他} \end{cases} \tag{28}$$

### 3.3　评价指标

实验选取了语义分割任务中常用的3项指标进行评估，分别为平均交并比（mean intersection over union，mIoU）、F1分数（F1-score）和总体精度（overall accuracy，OA）。

mIoU是图像分割中常用的评价指标，用来衡量模型的预测与实际标签之间的交集与并集的平均比率。该指标通过计算每个类别的交集与并集的比值，并取其平均值，来评估模型在多个类别上的分割精度：

$$\text{mIoU} = \frac{1}{N}\sum_{k=1}^{N}\frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k + \text{FN}_k} \tag{29}$$

其中，$\text{TP}_k$、$\text{FN}_k$和$\text{FP}_k$分别表示类别$k$的真阳性、假阴性和假阳性像素的个数。

精确率（precision）是评估模型将正类样本预测为正类的准确性，即在模型预测为正的样本中，有多少比例是真正的正类样本：

$$\text{precision}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k} \tag{30}$$

召回率（recall）表示模型能够从所有真实的正类样本中识别出多少，即模型能找到所有实际为正类的样本的比例：

$$\text{recall}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k} \tag{31}$$

F1 指标是由精确率和召回率综合得到的调和平均数，用于衡量模型在某一类别上的整体分割性能：

$$\text{F1}_k = 2 \times \frac{\text{precision}_k \times \text{recall}_k}{\text{precision}_k + \text{recall}_k} \tag{32}$$

OA 是语义分割任务中最基本、最直观的评价指标，表示模型在所有类别的所有像素上总体分类正确的比例。OA 衡量的是模型整体性能，而非逐类表现，因此在类别分布不平衡的数据集中，OA 可能会受主类别像素占比较大的影响：

$$\text{OA} = \frac{\sum_{k=1}^{N} \text{TP}_k}{\sum_{k=1}^{K} \text{TP}_k + \text{FP}_k + \text{TN}_k + \text{FN}_k} \tag{33}$$

其中，$\text{TN}_k$表示类别$k$的真阴性像素的个数。

### 3.4　实验配置

实验采用 AdamW 优化器，初始学习率设置为 $8\times10^{-5}$，并通过余弦退火策略动态调整学习率，以加速模型的收敛。所有模型均训练 100 个轮次（epoch），批次大小（batch size）设置为 4。每个数据集的多模态图像被切分为若干张 512×512 子图，步长为 400。由于数据集提供模态的限制，对于 Potsdam 数据集，选择 RGB、DSM、RGBIR 和 IRRG 这 4 种模态作为输入，数据集划分比例为 9∶1；对于 Vaihingen 数据集，使用 IRRG 和 DSM 两种模态作为输入，数据集划分比例为 8∶2。为增强模型的泛化能力，对数据进行了随机数据增强操作，包括随机缩放、翻转和高斯模糊，以增加数据的多样性。实验中所有模型均使用 PyTorch-Lightning 框架实现，并在单个 NVIDIA RTX 3080Ti GPU 上进行训练和验证，操作系统为 Ubuntu 18.04，CUDA 版本为 12.1，Pytorch 和 Pytorch-Lightning 版本均为 2.3.1。

### 3.5　实验结果

为验证所提模型的有效性，本文在 Potsdam 数据集和 Vaihingen 数据集上与多种单模态、双模态网络进行了定性与定量的比较。单模态网络包括 DeepLabV3+[20]、多注意力网络（multi-attention network，MANet）[21]、密集空洞卷积合并网络（dense dilated convolutions merging，DDCM）[22]、注意力聚合特征金字塔网络（attention aggregation feature pyramid network，A2FPN）[23]、注意力双边上下文网络（attentive bilateral contextual network，ABCNet）[24] 和 U 型 Transformer 网络（unet-like transformer，UNetFormer）[25]；双模态网络包括残差编码器解码器网络（residual encoder-decoder network，RedNet）[26]、跨模态融合网络（cross-modal fusion network，CMX）[27]、MGCNet[17]、密集残差网络（dense residual network，DRNet）[28] 和渐进相邻层协调对称级联网络（progressive adjacent-layer coordination symmetric cascade network，PACSCNet）[29]。随后，通过消融实验分析各模块及辅助损失函数的贡献，以验证各模块对模型性能的提升效果。

#### 3.5.1　Potsdam 数据集对比实验

Potsdam 数据集上各模型对比实验结果见表 1。从表 1 可知，EMSegNet 在 mIoU（88.29%）、mF1 分数（94.19%）和 OA（94.54%）方面均优于其他单模态和双模态网络，相较于第二位的 PACSCNet，在 mIoU、mF1 和 OA 分别提升了 0.35、0.68 和 0.78 个百分点。在具体类别上，建筑和不透水表面类别表现最佳，分别达到 96.95% 和 94.40%。低矮植被和

**表 1　Potsdam 数据集上各模型对比实验结果**

| 网络 | 模态 | 交并比 | | | | | mIoU | mF1 | OA |
|---|---|---|---|---|---|---|---|---|---|
| | | 不透水表面 | 建筑 | 低矮植被 | 树木 | 车辆 | | | |
| DeepLabV3+ | RGB | 88.23% | 93.27% | 79.36% | 76.65% | 83.66% | 85.20% | 91.91% | 91.93% |
| DDCM | RGB | 91.57% | 95.17% | 84.83% | 81.95% | 85.84% | 86.91% | 92.93% | 93.15% |
| A2FPN | RGB | 92.34% | 95.15% | 87.30% | 84.15% | 86.29% | 87.26% | 93.11% | 93.35% |
| ABCNet | RGB | 91.49% | 95.34% | 84.74% | 82.16% | 85.70% | 84.76% | 91.32% | 89.82% |
| UnetFormer | RGB | 92.54% | 95.70% | 86.64% | 83.96% | 86.86% | 87.74% | 93.30% | 93.51% |
| MANet | RGB | 91.53% | 95.26% | 84.80% | 81.12% | 85.99% | 86.73% | 92.82% | 93.43% |
| RedNet | RGB+DSM | 92.26% | 96.07% | 85.79% | 83.27% | 85.49% | 87.88% | 93.50% | 93.77% |
| CMX | RGB+DSM | 92.25% | 95.94% | 84.80% | 81.12% | 85.99% | 87.75% | 93.42% | 93.73% |
| MGCNet | RGB+DSM | 90.85% | 95.23% | 84.13% | 81.30% | 85.28% | 87.71% | 93.45% | 93.63% |
| DRNet | RGB+DSM | 90.89% | 94.11% | 82.19% | 85.42% | 85.19% | 87.28% | 93.08% | 93.12% |
| PACSCNet | RGB+DSM | 93.01% | 96.37% | 86.91% | 85.36% | **87.37%** | 87.94% | 93.51% | 93.76% |
| 本文所提 | 多模态 | **94.40%** | **96.95%** | **88.82%** | **85.56%** | 87.21% | **88.29%** | **94.19%** | **94.54%** |

树木类别的IoU较其他算法得到了显著提升。在车辆类别上，分割结果与PACSCNet模型差距不大，但相较于其他单模态和多模态模型，仍取得了显著提升。这些性能的提升不仅归因于多模态数据的引入为模型提供了额外光谱信息与高度信息，更得益于所设计模块的相互作用，包括E2HA模块的自适应模态权重调整、CMHA的多头对齐与互补特征搜索和DCA模块的全局上下文感知，上述机制提升了模型在各个类别的分割性能。

为了更直观地展示所提方法的优越性，本文从数据集中选取了两个容易产生错误分割的样例进行可视化比较。Potsdam数据集可视化结果如图4所示。从图4可以看出，EMSegNet在多个类别上都表现出了更精确的分割结果。在细节区域，如建筑物附近的细小区域，其他模型均出现错误分类情况，而EMSegNet则能正确分割。在低矮植被和树木这两种易混淆类别上，本文算法也能够更准确地区分，分割结果更加细致。相比之下，其他模型在这些区域中存在明显的混淆，导致边界模糊且分割不准确。

### 3.5.2　Vaihingen数据集对比实验

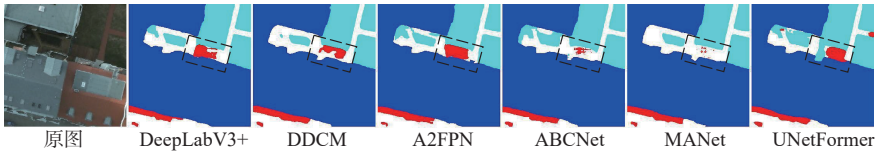
Vaihingen数据集上各模型对比实验结果见表2。从表2可知，从整体上看，EMSegNet在mIoU上达到了86.73%，mF1为92.12%，OA为94.74%；在具体类别上，EMSegNet在低矮植被和树木类别中表现突出，分别达到了84.69%和86.89%的交并比，明显领先于其他对比模型。这两个类别在语义特征上极为相似，但在高度上有显著差异，单靠IRRG模态难以准确区分。模型通过引入DSM模态，能够利用高度信息弥补IRRG模态的不足，同时得益于E2HA和CMHA模块的设计，能够更好地利用模态间的互补特征，显著提升了这两个类别的分割精度。同样地，与Potsdam数据集相同，Vaihingen数据集可视化结果如图5所示。可以看出，EMSegNet在复杂边界和细节区域的分割上具有显著的优势。反观其他模型，在建筑物周围的细小区域，容易出现不连续的分割，导致边界模糊不清。同时，EMSegNet能够更准确地捕捉低矮植被与树木之间的细微差异，而其他模型在这些区域则容易出现混淆、错分和漏分等问题。

### 3.5.3　消融实验

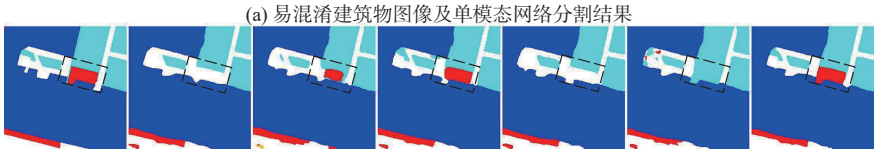
为验证EMSegNet中各模块的有效性与兼容性，本文在Potsdam数据集和Vaihingen数据集上进行了3组消融实验。

（1）实验1：模态外推成本分析

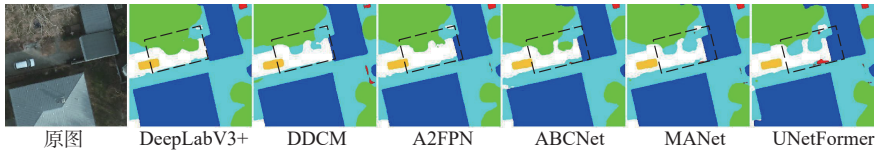
Potsdam数据集采用四模态输入，而Vaihingen



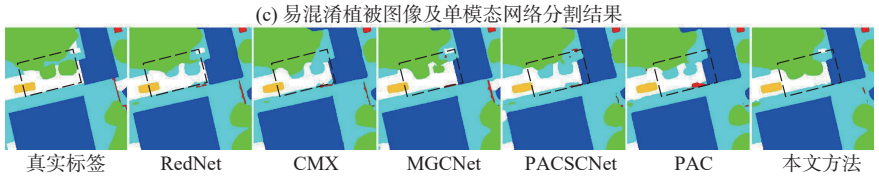
(a) 易混淆建筑物图像及单模态网络分割结果



(b) 易混淆建筑物图像真实标签及多模态网络分割结果



(c) 易混淆植被图像及单模态网络分割结果



(d) 易混淆植被图像真实标签及多模态网络分割结果

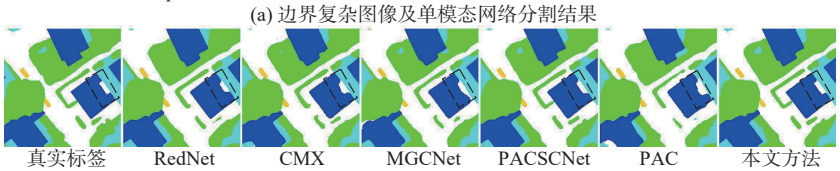
图4　Potsdam数据集可视化结果

**表 2　Vaihingen 数据集上各模型对比实验结果**

| 网络 | 模态 | 交并比 | | | | | mIoU | mF1 | OA |
|---|---|---|---|---|---|---|---|---|---|
| | | 不透水表面 | 建筑 | 低矮植被 | 树木 | 车辆 | | | |
| DeepLabV3+ | IRRG | 86.66% | 89.62% | 80.09% | 82.20% | 76.97% | 83.23% | 90.72% | 92.14% |
| DDCM | IRRG | 85.84% | 88.87% | 78.71% | 81.34% | 74.20% | 82.85% | 90.47% | 92.11% |
| A2FPN | IRRG | 86.40% | 89.44% | 78.86% | 81.79% | 73.34% | 84.29% | 91.35% | 92.83% |
| ABCNet | IRRG | 86.79% | 89.95% | 78.81% | 81.34% | 76.28% | 83.17% | 90.66% | 92.15% |
| UnetFormer | IRRG | 88.78% | 92.56% | 78.04% | 82.40% | 76.97% | 84.90% | 91.21% | 92.51% |
| MANet | IRRG | 87.45% | 89.82% | 82.16% | 83.93% | 76.56% | 84.77% | 91.63% | 93.24% |
| RedNet | IRRG+DSM | 88.89% | 92.96% | 77.48% | 82.34% | 76.60% | 85.32% | 92.21% | 93.28% |
| CMX | IRRG+DSM | 87.91% | 90.67% | 82.09% | 83.70% | 75.33% | 84.26% | 91.33% | 93.12% |
| MGCNet | IRRG+DSM | 88.16% | 90.02% | 85.10% | 85.63% | 79.20% | 86.37% | **92.59%** | 93.79% |
| DRNet | IRRG+DSM | 87.55% | 89.69% | 83.54% | 84.69% | 78.65% | 85.34% | 91.98% | 93.34% |
| PACSCNet | IRRG+DSM | 90.09% | 93.39% | 83.60% | 86.09% | 70.02% | 85.42% | 91.91% | 94.33% |
| 本文所提 | IRRG+DSM | **91.34%** | **94.08%** | **84.69%** | **86.89%** | **79.63%** | **86.73%** | 92.12% | **94.74%** |



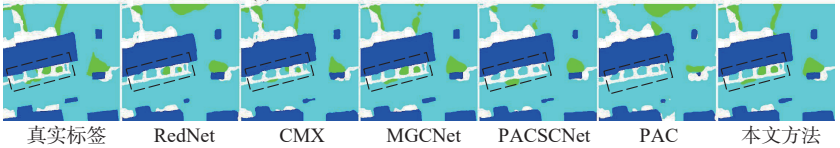
(a) 边界复杂图像及单模态网络分割结果



(b) 边界复杂图像真实标签及多模态网络分割结果



(c) 细节对比图像及单模态网络分割结果



(d) 细节对比图像真实标签及多模态网络分割结果

图 5　Vaihingen 数据集可视化结果

数据集则为双模态输入，除模态数量外，其他变量在两组实验中保持一致。因此，通过这两组实验对比，能够准确评估模态外推所带来的成本增量。模态外推成本分析结果见表 3。由表 3 可以看出，Potsdam 数据集相比 Vaihingen 数据集，虽然增加了两个模态，但参数量仅增加约 $1.2\times10^5$，计算量仅增加 17.67 GFLOPS。这表明外推一个模态时，仅需额外增加参数量约 $6\times10^4$，计算量约增加 8.8 GFLOPS。由此可见，EMSegNet 外推模态的成本低，在硬件传感器资源允许的情况下，可轻松外推至更多模态。

（2）实验 2：多头机制对 CMHA 性能的影响

在该实验中，主要通过调整 CMHA 模块中对齐头的数量来探索多模态对齐在不同窗口尺度下的效果。CMHA 头数量实验结果见表 4。随着多头数

**表 3　模态外推成本分析结果**

| 数据集 | 模态 | 参数量 | 计算量/GFLOPS |
|---|---|---|---|
| Potsdam | RGB+DSM+IRRG+RGBIR | 3.628 526 $5\times10^8$ | 101.25 |
| Vaihingen | IRRG+DSM | 3.616 332 $9\times10^8$ | 83.58 |

**表 4　CMHA 头数量实验结果**

| 数据集 | Head | mIoU | mF1 | OA |
|---|---|---|---|---|
| Potsdam | 1($s$=2) | 87.20% | 92.91% | 93.12% |
| Potsdam | 2($s$=2,4) | 87.91% | 93.93% | 94.15% |
| Potsdam | 3($s$=2,4,8) | 88.29% | 94.19% | 94.54% |

量的增加，模型的性能（包括mIoU、mF1和OA）均有显著提升，这表明多头机制能够更好地适应多模态数据的异质性，不同窗口大小的配合提高了跨模态对齐的效果，特别是三头机制（$s$=2, 4, 8）通过在不同尺度上对齐信息，达到了最佳的融合效果和分割精度。

（3）实验3：模块累加消融分析

实验在Potsdam数据集和Vaihingen数据集上分别逐步累加各模块及辅助损失函数，来评估它们对模型性能的贡献，各模块累加实验结果见表5。由表5可以看出，通过逐步加入各模块（DCA、E2HA、CMHA）及辅助损失函数，模型性能（mIoU、mF1、OA）得到了显著提升；仅加入DCA模块后，Potsdam和Vaihingen数据集上的mIoU分别提升了1.99和1.16个百分点，这表明DCA模块通过提供全局感受野在提升模型表现方面发挥了至关重要的作用；进一步加入E2HA模块、CMHA模块和辅助损失函数后，模型性能达到了最优，表明E2HA、CMHA模块在多模态特征提取、多模态特征对齐和捕捉各模态异质性过程中起到了关键作用，而辅助损失函数则通过对中间层特征的监督学习，确保各模态特征在解码过程中的充分利用。

**表5　各模块累加实验结果**

| 数据集 | DCA | E2HA | CMHA | 辅助损失 | mIoU | mF1 | OA |
|---|---|---|---|---|---|---|---|
| Potsdam | | | | | 84.92% | 91.91% | 91.93% |
| Potsdam | √ | | | | 86.91% | 92.93% | 92.15% |
| Potsdam | √ | √ | | | 87.86% | 93.71% | 93.35% |
| Potsdam | √ | √ | √ | | 88.09% | 93.92% | 94.32% |
| Potsdam | √ | √ | √ | √ | 88.29% | 94.19% | 94.54% |
| Vaihingen | | | | | 82.81% | 89.68% | 91.89% |
| Vaihingen | √ | | | | 83.97% | 91.04% | 92.28% |
| Vaihingen | √ | √ | | | 85.26% | 91.67% | 93.48% |
| Vaihingen | √ | √ | √ | | 86.17% | 91.82% | 94.17% |
| Vaihingen | √ | √ | √ | √ | 86.73% | 92.12% | 94.74% |

## 4　结束语

本文提出了一种支持模态扩展的高效遥感图像分割EMSegNet，旨在充分利用不同模态数据的互补特征。具体而言，本文针对前期融合策略对各模态权重一致的问题，设计了E2HA模块，以更好地区分不同模态在融合过程中的重要性；针对多模态数据中语义像素错位和相似特征堆叠的问题，设计了CMHA模块，以校正偏移并有效捕捉异质性特征；为了提升模型对全局上下文信息的感知能力，设计了维度协同线性注意力模块，并引入辅助损失函数以监督解码端的中间特征学习。EMSegNet支持超过两种模态的输入，且在扩展至更多模态时，计算成本和参数开销相对较小。在两个多模态遥感数据集上进行的一系列对比实验表明，EMSegNet优于目前主流的单模态、多模态分割网络。此外，本文通过一系列消融实验，验证了所提各模块的有效性。未来的研究可在其他领域、其他多模态数据集上验证方法的有效性，以进一步提升模型的适应性和泛化能力。

## 参考文献：


[1] Du S H, Du S H, Liu B, et al. Mapping large-scale and fine-grained urban functional zones from VHR images using a multi-scale semantic segmentation network and object based approach[J]. Remote Sensing of Environment, 2021, 261: 112480.

[2] Zhang C, Sargent I, Pan X, et al. Joint Deep Learning for land cover and land use classification[J]. Remote Sensing of Environment, 2019, 221: 173-187.

[3] Canny J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, PAMI-8(6): 679-698.

[4] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation[J]. International Journal of Computer Vision, 2004, 59(2): 167-181.

[5] Noble W S. What is a support vector machine?[J]. Nature Biotechnology, 2006, 24(12): 1565-1567.

[6] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[7] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[M]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Cham-Springer International Publishing, 2015: 234-241.

[8] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 6230-6239.

[9] 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.

[10] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 5686-5696.

[11] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16× 16 words: transformers for image recognition at scale[C]//International Conference on Learning Representations. Virtual: ICLR, 2021.

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//

Advances in Neural Information Processing Systems 30 (NeurIPS 2017). 2017: 5998-6008.

[13] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 9992-10002.

[14] Fan J Y, Li J J, Hua Z, et al. Elevation information-guided multimodal fusion robust framework for remote sensing image segmentation[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 2501005.

[15] 刘晓敏, 余梦君, 乔振壮, 等. 面向多源遥感数据分类的尺度自适应融合网络[J]. 电子与信息学报, 2024, 46(9): 3693-3702.
Liu X M, Yu M J, Qiao Z Z, et al. Scale adaptive fusion network for multimodal remote sensing data classification[J]. Journal of Electronics & Information Technology, 2024, 46(9): 3693-3702.

[16] Jin J H, Zhou W J, Yang R W, et al. Edge detection guide network for semantic segmentation of remote-sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 1-5.

[17] Yang E Q, Zhou W J, Qian X H, et al. MGCNet: multilevel gated collaborative network for RGB-D semantic segmentation of indoor scene[J]. IEEE Signal Processing Letters, 2022, 29: 2567-2571.

[18] Gadzicki K, Khamsehashari R, Zetzsche C. Early vs late fusion in multimodal convolutional neural networks[C]//Proceedings of the 2020 IEEE 23rd International Conference on Information Fusion (FUSION). Piscataway: IEEE Press, 2020: 1-6.

[19] Pereira L M, Salazar A, Vergara L. A comparative analysis of early and late fusion for the multimodal two-class problem[J]. IEEE Access, 2023, 11: 84283-84300.

[20] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]// Computer Vision-ECCV 2018. ChamSpringer International Publishing, 2018: 833-851.

[21] 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.

[22] Liu Q H, Kampffmeyer M, Jenssen R, et al. Dense dilated convolutions' merging network for land cover classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(9): 6309-6320.

[23] Hu M, Li Y L, Fang L, et al. A2-FPN: attention aggregation based feature pyramid network for instance segmentation[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 15338-15347.

[24] 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.

[25] 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.

[26] Jiang J D, Zheng L N, Luo F, et al. RedNet: residual encoder-decoder network for indoor RGB-D semantic segmentation[PP]. V2. arXiv (2018-08-06)[2026-05-06] arXiv: 1806.01054.

[27] Zhang J M, Liu H Y, Yang K L, et al. CMX: cross-modal fusion for RGB-X semantic segmentation with transformers[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(12): 14679-14694.

[28] Yang E Q, Zhou W J, Qian X H, et al. DRNet: dual-stage refinement network with boundary inference for RGB-D semantic segmentation of indoor scenes[J]. Engineering Applications of Artificial Intelligence, 2023, 125: 106729.

[29] Fan X M, Zhou W J, Qian X H, et al. Progressive Adjacent-layer coordination symmetric cascade network for semantic segmentation of multimodal remote sensing images[J]. Expert Systems with Applications, 2024, 238: 121999.

[作者简介]



**张众维**（1986– ），女，博士，天津城建大学计算机与信息工程学院讲师，主要研究方向为深度学习、图像处理、模式识别。



**朱国庆**（2003– ），男，天津城建大学计算机与信息工程学院硕士生，主要研究方向为深度学习、目标分割。



**秦海洁**（1997– ），男，硕士，深圳宏芯宇电子股份有限公司工程师，主要研究方向为计算机视觉、嵌入式系统开发。



**刘树东**（1965– ），男，博士，天津城建大学计算机与信息工程学院教授、院长，主要研究方向为图像处理、无线传感网络。