智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于多模态互补特征学习的遥感影像语义分割

王兴武, 雷涛, 王营博, 耿新哲, 张月

引用本文:
王兴武,雷涛,王营博,耿新哲,张月. 基于多模态互补特征学习的遥感影像语义分割[J]. 智能系统学报, 2022, 17(6): 1123–1133.
WANG Xingwu,LEI Tao,WANG Yingbo,GENG Xinzhe,ZHANG Yue. Semantic segmentation of remote sensing image based on multimodal complementary feature learning[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(6): 1123–1133.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202201025

DOI: 10.11992/tis.202201025
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20221009.0956.006.html

基于多模态互补特征学习的遥感影像语义分割

王兴武[1,2], 雷涛[1,2], 王营博[1,2], 耿新哲[1,2], 张月[1,2]

（1. 陕西科技大学 陕西省人工智能联合实验室, 陕西 西安 710021; 2. 陕西科技大学 电子信息与人工智能学院, 陕西 西安 710021）

摘　要: 在遥感影像语义分割任务中, 数字表面模型可以为光谱数据生成对应的几何表示, 能够有效提升语义分割的精度。然而, 大部分现有工作仅简单地将光谱特征和高程特征在不同的阶段相加或合并, 忽略了多模态数据之间的相关性与互补性, 导致网络对某些复杂地物无法准确分割。本文基于互补特征学习的多模态数据语义分割网络进行研究。该网络采用多核最大均值距离作为互补约束, 提取两种模态特征之间的相似特征与互补特征。在解码之前互相借用互补特征, 增强网络共享特征的能力。在国际摄影测量及遥感探测学会 (international society for photogrammetry and remote sensing, ISPRS) 的 Potsdam 与 Vaihingen 公开数据集上验证所提出的网络, 证明了该网络可以实现更高的分割精度。

关键词: 计算机视觉; 遥感影像; 图像分割; 卷积神经网络; 语义分割; 多模态特征融合; 深度学习; 互补特征学习
中图分类号: TP183　　**文献标志码:** A　　**文章编号:** 1673−4785(2022)06−1123−11

中文引用格式: 王兴武, 雷涛, 王营博, 等. 基于多模态互补特征学习的遥感影像语义分割 [J]. 智能系统学报, 2022, 17(6): 1123–1133.
英文引用格式: WANG Xingwu, LEI Tao, WANG Yingbo, et al. Semantic segmentation of remote sensing image based on multimodal complementary feature learning[J]. CAAI transactions on intelligent systems, 2022, 17(6): 1123–1133.

Semantic segmentation of remote sensing image based on multimodal complementary feature learning

WANG Xingwu[1,2], LEI Tao[1,2], WANG Yingbo[1,2], GENG Xinzhe[1,2], ZHANG Yue[1,2]

(1. Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi ’an 710021, China; 2. School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi’an 710021, China)

Abstract: In the semantic segmentation of remote sensing images, the digital surface model can provide a corresponding geometric representation of the spectral data, which can effectively increase segmentation accuracy. However, most literature studies simply add or merge spectral and elevation features at different stages, ignoring the correlation and complementarity between multimodal data. This makes the network unable to accurately segment some complex features. This paper studies a multimodal data semantic segmentation network based on complementary feature learning. The network uses the multicore maximum mean distance as a complementary constraint to extract similar and complementary features between two modal features. The complementary features are borrowed from each other before decoding to enhance the feature sharing capability of the network. The proposed network is verified on the Potsdam and Vaihingen datasets of ISPRS and achieves higher segmentation accuracy.

Keywords: computer vision; remote sensing image; image segmentation; convolutional neural network; semantic segmentation; multimodal feature fusion; deep learning; complementary feature learning

遥感图像语义分割是将对象类标签分配给遥感图像中每个像素的一项技术, 被广泛应用于地质的量化分析、城市规划、环境监测和保护等领域。因此, 语义分割一直是遥感领域的研究热点。然而, 建筑物、道路和树木等地物具有较高的类内方差和相似的类间外观, 因此遥感图像语义分割具有一定的挑战性。

遥感图像的光谱信息, 例如红、绿、蓝三通道图像（red, green, blue, RGB）或近红外、红、绿（infrared, red, green, IRRG）图像, 通常是语义分割任

收稿日期: 2022−01−16.　**网络出版日期:** 2022−10−09.
基金项目: 国家自然科学基金项目 (61871259; 61861024; 62201334); 陕西省重点研发计划项目 (2021ZDLGY08-07); 陕西省人工智能联合实验室项目（2020SS-03）.
通信作者: 雷涛.E-mail: leitao@sust.edu.cn.

务的第一个数据源。近年来，深度学习在遥感图像的光谱图像语义分割方面取得了重大进展[1-2]。基于全卷积网络的方法和编码架构已被广泛应用并取得了较好效果。Fu 等[3] 设计了基于全卷积神经网络的遥感图像语义分割算法，实现了遥感图像端到端的语义分割。在此基础上，研究者们还提出了空间关系模块[4] 和空间信息推理模块[5]。通过对特征图通道及空间上下文关系进行有效建模，使网络聚焦于目标区域，抑制其他类别的影响。Zhang 等[6] 在编解码结构中采用金字塔池化模块以聚合不同区域的上下文信息，从而提高网络获取全局信息的能力。此外，在损失函数的设计上，为了解决遥感图像中严重的类不平衡问题，Dong 等[7] 提出了加权损失和像素级交叉熵损失相结合的多类损失，提升了小样本量地物分割精度。Liu 等[8] 将边界损失引入网络中得到了更精细的地物边界分割结果。这些方法在遥感图像语义分割中均取得了显著效果。但在某些特定场景下，如物体的外观因阴影和天气条件而改变，某些地物在光谱信息上高度相似的类间外观（高低植被）等，仅仅利用光谱信息作为单一信息源会导致分割性能低，在某些地物上错分严重。

随着航空拍摄和卫星成像技术的不断发展，获取到的遥感图像也愈加多样，如数字地形模型（DTM）、数字表面模型（DSM），数字高程模型（DEM）等。这些高程数据通过向二维光谱数据提供三维几何信息，自动补充光谱信号，这对光照变化、阴影等具有鲁棒性，合理地使用高程数据可以大幅提高遥感图像的语义分割性能[9]。目前，简单地将近红外、红色、绿色（IRRG) 光谱和 DSM 合并为四通道作为网络输入的像素级融合方法，但这种方法已被证明不能充分融合异构信息之间的关系，需要额外的网络结构单独对高程信息进行特征提取[10]。

基于多模态信息融合的遥感图像语义分割算法可以总结为三类结构，即早期融合、中期融合和后期融合。

Cao 等[11] 提出的中期融合策略用两个独立编码器分别提取光谱和高程分支特征，在上采样之前或期间进行融合。融合特征仅存在于解码器端，这种中期融合的优势在于：在网络中间阶段进行融合，多模态特征语义信息丰富，可以避免某单一源的低级特征将噪声信息带入融合特征。Hazirbas 等[12] 提出的早期融合策略，将高程分支特征作为光谱分支特征的补充，并在不同的下采样阶段将它们融合在一起。这种方法存在高程分支只处理高程信息的问题，而光谱分支实际上处理的是融合数据，高程分支带来有效信息的同时也引入了噪声信息。因此，这种方法会破坏原始光谱分支的特征流，使网络在对某些地物分类时陷入对某单一模态数据的过分依赖，造成误判。Audebert 等[13] 提出了后期融合策略。他们为光谱数据和 DSM 数据训练了两个独立的深度网络，然后将两个深度网络的最终特征映射概率图输入到残差校正网络中进行训练。然而，存在的不足是：一方面，仅在语义层面的融合忽略了语义分割任务对低级融合特征的需求；另一方面，大量参数和复杂训练过程限制了该方法的实际应用。

虽然上述方法在多模态数据遥感图像语义分割任务中均取得了不错的效果，但这些网络在进行特征融合时均未考虑多模态数据之间的关系，容易导致复杂地物的误分。在本文中，假设从光谱和高程信息计算的中层特征（网络中特征提取阶段结束时提取的特征）有相同特征和互补特征，那么从光谱和高程计算的信息将相互作用，从而为分类提供更独特和互补的特征。基于这一假设，本文设计了用于多模态遥感图像语义分割的双分支互补特征学习 (complementary features learning) 网络。在该网络中，通过添加额外的损失重建共有和互补特征，可以产生更鲁棒的融合特征以提高分割精度。

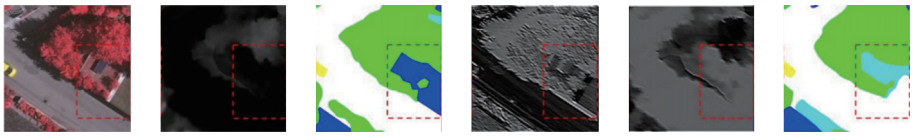
1　相关知识

为了充分利用光谱数据与高程数据之间的异构信息，大多数现有算法均采用两个深度网络分别提取多模态信息特征，在网络不同阶段将提取到的多模态特征融合，融合特征比原始单一模态特征更具鲁棒性，从而产生更精细的分割结果。然而，现有方法存在以下问题：一方面，以直接相加或级联的方式通过建模实现多模态信息融合，或简单地以多模态特征自身所含信息量决定该模态特征在融合特征中所占比例[14]，这两种方法均忽略了多源信息间的关系，在某些复杂地物上造成严重错分；另一方面，融合策略的选取也会严重影响分割结果，几种融合策略各有优劣[15]。基于此，我们在网络训练过程中引入互补特征学习约束，并根据该约束的特点选取中期融合策略。提出的网络根据多模态数据之间的关系，选择性地提取对分割任务更有效的部分，可以有效减少误分，提升分割精度。

1.1　互补特征的提取

现有算法在构建多模态特征融合时，仅仅将多源特征在网络的不同阶段简单合并或相加，而

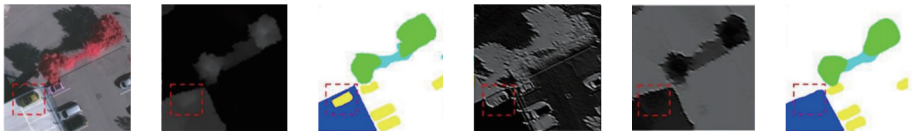
忽略了多源信息间的互补性, 对于一些光谱上表现相似的地物, 如高低植被、建筑物阴影遮挡后的路面与正常路面等, 网络偏向于高程信息时, 对提升分割精度是有利的; 但是对于车辆、路面而言, 高程信息难以体现情况, 分割精度较低。为了更清晰地展示过分依赖单一信息源造成的误分现象, 利用网络提取到的特征图解释这一现象, 图 1 和图 2 给出了裁剪出更小区域时的分割结果, 并用红色框标记重点误分区域。图 1 和图 2 中的(a)、(b)、(c)分别为光谱图、归一化 DSM 图以及手工标注标签, (d)、(e)分别为 DSMFNet 编码器中光谱分支与高程分支第一次下采样后得到的特征图, (f)为 DSMFNet 分割结果。图 1 的地物背景是两块在光谱上表现完全不同的建筑物, 其中一块光谱上表现为低植被类, 另一块在光谱上表现为建筑物类别, 但是在高程数据中, 两块建筑物区域数值并不高, 与道路特征类似, 由于大量数据样本下的建筑物高程上数值都很高, 在这种情况下网络偏向于信任高程信息, 将该区域错分为低植被类。在图 2 中, 位于建筑物顶部的汽车, 在光谱图像中特征明显, 但是在高程图像中几乎表现不出来, 高程分支提取到的特征图在该区域也十分模糊, 导致网络仍偏向于高程特征, 最终分割结果并没有将汽车类别准确分割。



(a) 正射影像　(b) DSM　(c) 手工标注　(d) 光谱采样　(e) 高程采样　(f) DSMFNet

图 1　建筑区域的可视化特征图

Fig. 1　Visual feature map of the buliding region



(a) 正射影像　(b) DSM　(c) 手工标注　(d) 光谱采样　(e) 高程采样　(f) DSMFNet

图 2　不透水和汽车的可视化特征图

Fig. 2　Visual feature map of water impervidle and automobile region

由上述分析可得, 光谱和高程数据在相同地物上会表现出不同的特征。如何有效地识别它们的差异并将两种类型信息统一表示为语义特征是十分重要的。我们认为光谱和高程信息计算的中层特征存在相同信息和互补信息, 相同信息包括重合的建筑物、植被的边缘信息; 光谱数据中的互补信息有车辆颜色、纹理信息, 高程数据中的互补信息有可区分的高、低植被高度以及轮廓信息。光谱和高程计算的信息需要相互作用, 这为分类提供更独特和互补的信息。基于这一结论, 我们把多模态数据提取到的中层特征分别划分为相同的和互补的特征, 在多模态特征融合时将提取到的互补特征与另一模态所有特征合并, 补充后的互补特征可以有效避免上述误分。

1.2　中期融合策略

目前基于多模态信息融合的语义分割网络中, 应用最为广泛的是早期融合策略。该策略采用两个深度网络分别对多模态数据进行特征提取, 将高程分支下采样过程中产生的不同尺寸的特征图作为辅助与光谱分支中的特征图进行合并, 再由共用的解码器完成特征重建[16]。该方法的理论依据是: 网络的浅层主要提取边缘特征, 而光谱图像和高程图像所表现的边缘特征有很大不同。主要表现为: 无结构的边缘主要依靠颜色区分, 无纹理的边缘主要依靠结构高度区分[17]。因此, 早期融合策略是从浅层开始对特征进行融合。这种方法虽然应用广泛且取得了不错的分割结果, 但仍存在一些问题: 一方面, 在网络浅层提取到的低级特征在提供融合特征细节信息的同时也会带来大量冗余信息, 例如, 光谱数据中地物阴影的边界, 高程数据中建筑物锯齿状的边缘等噪声, 均会干扰网络模型对该区域的识别; 另一方面, 由于需要从多模态特征中提取出对应的互补特征, 该互补特征是依靠最大化多模态特征之间的距离得到的, 如果在网络浅层就开始采用互补特征约束最大化多模态特征之间的距离, 会导致网络过于注重提取到的多模态特征间的差异性, 而忽略了对分割任务语义信息的提取。

基于上述分析, 本文先采用中期融合策略, 利

用光谱和高程数据分别训练两个结构相同的深度网络，再对编码器的最后一层输出特征图进行融合。这种方法虽然舍弃了一定的细节信息，但网络中添加的互补特征约束的语义级特征融合会对融合特征中的错误部分进行修正。此外，在下采样过程中丢失的细节信息会通过跳跃连接补充到解码器中。

2　主要研究内容和关键技术

多模态高分辨率遥感语义分割任务中，输入为 IRRG 光谱图像和相应的 DSM 高程图像，输出是每个像素的语义类别标签。本文利用一种模态数据改进另一种模态的特征提取过程：一方面，由于一些语义特征在这两种模态中都是可见的，可以从光谱图像和相应的高程信息中提取一组相似特征；另一方面，由于光谱图主要捕获外观信息，高程图主要捕获形状信息，可以分别从它们中提取一些特定的特征。

在本文中，我们通过最大化共享信息之间的相似性和特定模态信息之间的互补性，将每个模态的特征分别分解为共同特征和互补特征。得到这些特征后，一种模态通过借用另一种模态中的互补特征，以增强它们共享信息的能力。这种共享机制在单一模态特征没有被很好提取时是十分有效的，最后通过融合这两种模态输出的概率图得到分割结果。

2.1　网络结构

如图 3 所示，互补特征学习网络由两个相同的并行基础分割网络和互补特征提取部分组成，并行基础分割网络包括光谱信息分支和高程信息分支。互补特征提取部分位于编码器 (encoder) 与解码器 (decoder) 之间，由 4 个并行的补充性特征融合 (complementary feature fusion) 卷积层组成。

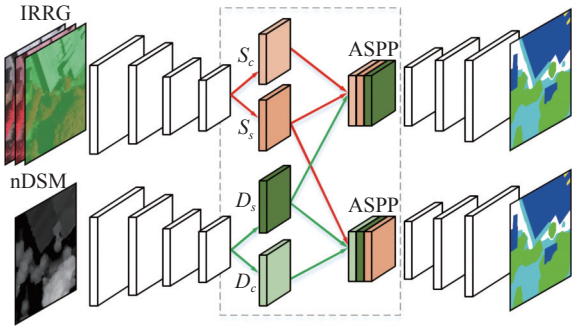


图 3　CFL-Net 网络结构

Fig. 3　CFL-Net network structure

两个基础分割网络均以 ResNet-50[18] 作为 backbone，并有完成特征重建的解码器与之对应。为了防止网络过深造成的梯度消失问题，在解码器中添加残差结构。为了将低阶特征利用起来，将每一个 res-block 输出特征图以跳跃连接的方式补充到解码器部分。为了避免信息冗余，我们在每个跳跃连接上添加了额外的卷积块，以自适应地过滤由低级特征带来的冗余信息。

此外，为适应遥感图像特殊的纹理特征，对骨干网络 ResNet-50 做出以下改进。原始 ResNet-50 网络用于自然图像分类，由一个步长为 2 的 7×7卷积层、3×3池化层以及 4 个残差块组成。此处保持 4 个残差块不变，将最初步长为 2 的7×7卷积层替换为3×3卷积层，步长保持不变，后续的3×3池化层被删除。通过在第一块中使用较小的卷积核和丢弃最大池化，可以避免特征图过于平滑，从而保留更清晰的边界信息，以区分通常具有尖锐边界信息的对象，如建筑物、车辆等。

该网络主要创新在于互补特征提取部分，光谱数据与高程数据经过相同的编码器处理得到中间的语义特征，为了不破坏空间信息，本文分别用两个卷积对它们进一步处理，将单个特征图划分为共同特征与互补特征，以此达到不同模式的共同特征是相似的，而互补特征是不同的。此外，采用额外的互补特征学习损失约束共同特征和相似特征之间的距离。在融合策略上，以光谱分支为例，将光谱分支获取的共同特征与互补特征、高程分支获取的互补特征合并在一起，构成新的融合特征。得到的光谱分支的融合特征U_s及高程分支的融合特征U_d可以分别表示为

$$U_s = S_s + S_c + D_s \tag{1}$$

$$U_d = D_s + D_c + S_s \tag{2}$$

式中：S_s和S_c分别表示光谱、高程分支的共同特征；D_s和D_c分别表示光谱、高程分支的互补特征。在特征融合之后，利用 ASPP 模块[19] 扩大感受野，获取多尺度融合特征。最后，解码器使用具有残差结构的反卷积模块重建语义标记结果。

该结构中，这两种模态可以通过学习到的互补特征相互促进。若其中一种模态的数据提取到的特征较差，丢失了部分关键信息时，另外一种模态的互补特征将是十分有利的。由于不同模态的数据特征通过不同网络获取，且通过额外的损失约束两组特征之间的距离，在保证互补特征的差异性基础上又可以保证对分割任务是有效的。

2.2　多尺度特征融合模块

为了减少池化操作引起的细节信息丢失问题，研究人员提出了使用空洞卷积代替池化操作[20]。

这种空洞卷积与普通卷积不同的地方在于引进了空洞率的思想, 在卷积取点时将空洞率作为取点步长。空洞卷积在不增加计算量、不添加网络层数的基础上有效扩大感受野。另一种在池化操作上的改进是 Zhao 等[21] 提出的空间金字塔池化模块, 该模块对不同核大小卷积得到的多尺度特征进行融合, 大幅提升了对过大或者过小目标的分割性能。

结合空洞卷积和多尺度金字塔池化块的优点, Chen 等[22] 提出了一种可以大幅提升分割精度的空洞空间金字塔池化模块。然而, ASPP 模块在图像分割中仍然存在两个问题: 1)在原始空洞率组合下, 任意一个空洞卷积对某些点均复发采样, 造成局部信息严重丢失, 从分割结果上表现为网格效应严重, 分割结果不完整[23]; 2)原始 ASPP 模块中高达 18 的空洞率在遥感图像提取到的特征中是不适用的, 这种空洞率下的卷积提取到的信息在很大的距离上是不相关的。原始输入图片的尺寸为256×256, 经过 ResNet-50 中的 5 次下采样之后, 特征图尺寸变为8×8, 只需要较小的空洞率即可获得全局的特征。因此对 ASPP 模块中的空洞率部分的超参做出调整, 具体结构如图 4 所示。

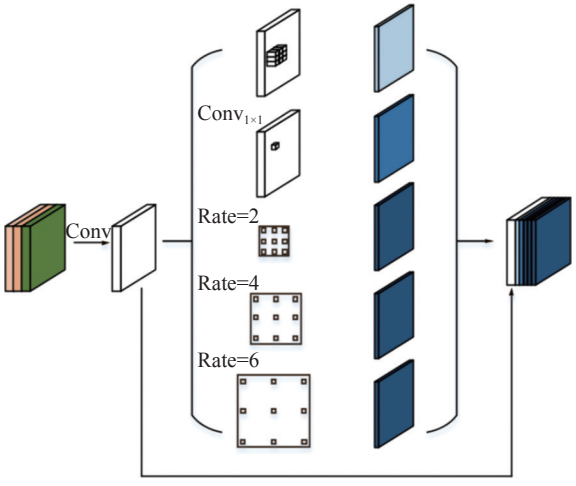


图 4　多尺度特征融合模块

Fig. 4　Multi-scale feature fusion module

由于显存限制, 首先将融合特征用1×1卷积进行降维, 再经过优化后的 ASPP 模块处理, 该特征分别由最大池化 (max-pooling)、1×1卷积、空洞率为 2、4、6 的空洞卷积进行处理, 得到 5 个多尺度特征图, 并在通道上均有一定的维度减少。最终将多尺度特征统一尺寸之后与原始特征图进行合并, 通过这种融合, 可以保留对象的空间上下文和边界细节, 以极小的计算增加获取更大范围的感受野, 产生更精细的分割结果。

2.3　互补特征学习损失

为了获得相似的光谱和高程的共同特征, 可以简单地最小化它们的欧氏距离。然而, 欧氏距离对不相似的共同特征的异常值很敏感。将两种模态的共同(互补)特征看作两个分布的样本, 将问题建模为计算分布之间的距离。为了获得两个相似分布的公共特征和不同分布的互补特征: 一方面需要衡量两个分布之间的距离; 另一方面, 需要将这一距离作为损失函数并约束训练阶段, 提取共同特征时要最小化这个损失, 相反, 提取互补特征时需要最大化该损失。

当前有许多计算分布之间相似性的技术, 如熵、互信息或 JS、KL 散度等[24]。然而, 这些信息论方法依赖于密度估计, 或复杂的空间划分、偏差校正策略, 这些策略对于高维数据通常是不可行的。还有一些利用生成对抗的思想, 间接最小化分布之间的 JS 散度[25]。这种方法需要额外的网络且难以收敛。

在迁移学习领域中, 最大均值差异(MMD)损失[26] 常用来衡量多模态特征之间的距离, 表现出优异的性能, 通过最小化子空间特征的分布差异, 从而使源域与目标域的特征分布尽可能相似。我们将 MMD 应用在多模态特征融合任务中, 一方面要求从两种模态数据中提取的共同特征分布相同时需要最小化该差异, 另一方面提取互补特征时需要最大化该差异。

假设分别存在一个满足P分布的光谱特征$X^s=[x_1^s\ x_2^s\ \cdots\ x_n^s]$和一个满足$Q$分布的 DSM 特征$X^d=[x_1^d\ x_2^d\ \cdots\ x_n^d]$, 令$H$表示再生希尔伯特空间, $\varphi(\cdot): X\to H$表示原始特征映射到希尔伯特空间的映射函数, 当$n_s\to\infty, n_t\to\infty$时, X^s和X^d在希尔伯特空间中的最大均值差异可表示为

$$\mathrm{MMD}\left(X^s,X^d\right)=\left\|\frac{1}{n_s}\sum_{i=1}^{n_s}\phi(x_i^s)-\frac{1}{n_d}\sum_{i=1}^{n_d}\phi\left(x_i^d\right)\right\| \tag{3}$$

MMD 的作用效果很大程度上取决于核函数k的选取, 不同的核函数会得到不同的结果。Gretton 等[27] 在两个测试样本中提出了基于多核的 MMD 方法, 通过生成基于核族的核函数, 多核 MMD 可以提高测试能力, 并成功地应用于域适应中。目前, 以多个不同高斯核的线性组合为核函数的多核 MMD 损失已经被集成在 pytorch 工具包中。

训练阶段, 在 batchsize 维度上分组计算特征之间 MK-MMD 的无偏估计, 以共同特征之间的损失函数$d(S_c,D_c)$为例:

$$d(S_c,D_c)=\frac{2}{n}\sum_{i=1}^{n/2}\varphi(u_i) \tag{4}$$

式中：n为 batchsize；S_c和D_c分别为解码器得到的光谱和高程的共同特征，可以计算出它们互补特征之间的相似性。

在本文的网络中，期望共同特征S_c和D_c尽可能地相似，而互补特征S_s和D_s期望分布不同。因此，应最小化$d(S_c,D_c)$，同时最大化$d(S_s,D_s)$。网络的损失函数为

$$L=\mathrm{CE}_s+\mathrm{CE}_d+\lambda(d(S_c,D_c)-d(S_s,D_s)) \tag{5}$$

$$\varphi(u_i)=k\left(S_c^{2i-1},S_c^{2i}\right)-k\left(S_c^{2i-1},D_c^{2i}\right)+k\left(D_c^{2i-1},D_c^{2i}\right)-k\left(D_c^{2i-1},S_c^{2i}\right) \tag{6}$$

式中：CE_s和CE_d是标签与网络输出之间的像素级交叉熵损失。使用参数λ平衡像素级损失与互补学习的损失，λ参数由交叉验证得到。在反向传播中，从反卷积特征和 MK-MMD 距离两个不同的来源计算了公共特征和互补特征的梯度。

3　实验分析及结果

3.1　数据集及预处理

为了验证提出方法的有效性，使用公开的 Potsdam 和 Vaihingen 数据集进行了实验测试。这 2 个数据集由 ISPRS 语义分割委员会提供[28]。这两个数据集主要覆盖城市及其周边地区，均标注了 6 个常见类别，包括不透水表面（白色）、建筑物 (蓝色)、低植被 (青色)、树木 (绿色)、汽车 (黄色) 和背景 (红色)，其中背景类较特殊，包括集装箱、网球场和水池等地物，根据文献 [29] 本文将背景忽略。此外，为了防止不精确的边缘标注对评估模型精度造成的影响，这两个数据集还提供了用半径为 3 像素的圆盘腐蚀类别边界的标签。

Potsdam 是德国的一座历史名城，有着巨大的建筑和密集的街道。Potsdam 数据集包括地面采样距离 (GSD) 为 5 cm 的 24 幅图像，包括近红外、红色、绿色、蓝色和归一化数字表面模型 (nDSM) 的 5 个通道，分辨率均为 6 000×6 000。在实验中，本文使用了 17 幅图像进行训练，7 幅图像进行测试。另外，Vaihingen 是一个小而分散的村庄，Vaihingen 数据集包含 3 波段 IRRG（红外、红色和绿色）光谱图像和相应的数字表面模型（DSM）。其中，GSD 为 9 cm，平均分辨率为 2 500×2 500。根据之前的工作[30]，本文选择 11 幅图像进行训练，其余 5 幅图像进行测试。

在数据预处理中，考虑到 GPU 内存有限，使用步长为 64 像素的滑动窗口将图像分割成大小为 256×256 的较小块。为了减少可能出现的过拟合现象，采用 4 种形式的数据增强：噪声干扰（高斯噪声）、0.8~1.2 倍的随机非比例缩放、0°~360°的随机旋转和 90°、180°或 270°的随机翻转。

3.2　实验设置

所提出方法使用 pytorch 工具库实现，互补特征学习网络在两块显存为 32 GB 的 GTX 3 090 显卡上训练，batchsize 设置为 16。初始学习率设置为 0.001，每 10 个 epoch 学习率衰减 10%。在优化器设置上，采用动量为 0.9 的随机梯度下降进行优化。此外，在训练阶段本文使用没有腐蚀边界的标签，而在计算指标时使用腐蚀边缘的标签，以避免不确定的边缘标注对模型评估的影响。其他实验细节说明如下。

1）多核 MMD 损失设置：使用 pytorch 中的多核 MMD 损失，其核函数为多个高斯核的线性组合，将核函数个数设置为 11，以保证可以区分互补特征与共同特征。

2）Fine-tuning：根据之前的工作[31]，网络模型在 ImageNet 及 PASCAL VOC 2012 语义分割数据集上的预训练模型基础上进行 fine-tuning，可以收敛得更快且获得更精确的结果。因此，本文用 backbone 为 Resnet-50 的 Deeplab V3+模型在 PASCAL VOC 2012 数据集上训练之后，将该 Resnet-50 及 ASPP 部分的 checkpoint 作为预训练模型。

3）训练设置：训练阶段，为了避免一开始网络就将重点放在最小化互补学习损失中而忽略对语义信息的学习，本文先将两个网络分开训练 20 个 epoch，当两个网络都产生较稳定的语义标记输出后，再将两个网络进行联合训练。

4）测试设置：测试阶段，采用重叠滑窗采样方法截取预测的小块图，获取预测结果后对于重叠部分取平均值，可以有效纠正拼接边界的小错误，进一步减少拼接带来的边界效应。

3.3　实验结果

为了更好地评估本文提出的互补特征学习网络的性能，本文与 5 种 SOTA 方法进行了比较，将这 5 种方法分为两类，即光谱数据网络和多模态数据融合网络，前者包括 FCN8s[32]、Deeplab v3+[22]，后者包括 VFuse-Net[33]、DP-DCN[34]、DSMFNet[11]。根据数据集 guideline。在本实验中使用 3 个指标全面评估每个网络的分割质量，分别是总体准确率 (OA)、平均 F_1 分数和每个类别的 F_1 分数。在 Potsdam 数据集上的实验结果如表 1 所示。

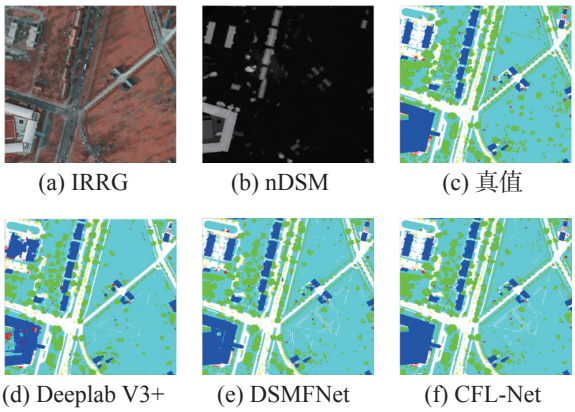
从结果可以看出，所提出的 CFL-Net 在 OA 中达到 91.21%，在平均 F_1 分数中达到了 92.47%。与 DSMFNet 相比，平均 OA 与 F_1 分别提高了 0.48% 和 0.87%，相较于 SOTA 方法，本文的模型在更加细化的 5 个类别上的 F_1 分数都有所提升。

表 1　Potsdam 数据集分割精度对比

Table 1　Comparison of segmentation accuracies for the Potsdam dataset　　%

方法	不透水表面	建筑	低植被	树木	车	平均F_1	总体
FCN8s	88.61	93.12	83.29	79.83	93.02	87.85	85.59
Deeplabv3+	90.81	94.87	85.57	83.84	94.50	90.16	89.72
V-fuseNet	92.42	95.20	86.15	86.27	94.45	90.90	90.15
DP-DCN	92.53	95.36	87.21	86.32	95.42	91.37	90.45
DSMFNet	93.10	95.87	86.60	87.10	95.35	91.60	90.73
CFL-Net	93.35	96.51	88.01	88.62	95.88	92.47	91.21

为了更直观、更清晰地比较本文提出的 CFL-Net 分割性能, 分别选取两大类分割网络中性能较好的方法 DeeplabV3+、DSMFNet 与本文的结果进行可视化对比。图 5 展示了在 Potsdam 数据测试集上 3 种网络在整张图上分割的结果, 从整张图的分割结果来看, 由于在测试时均采用了小步长滑动裁剪的方法, 这几张图像的分割结果都没有明显的边界效应, 相邻拼接块之间物体边界都比较平滑。



(a) IRRG　(b) nDSM　(c) 真值

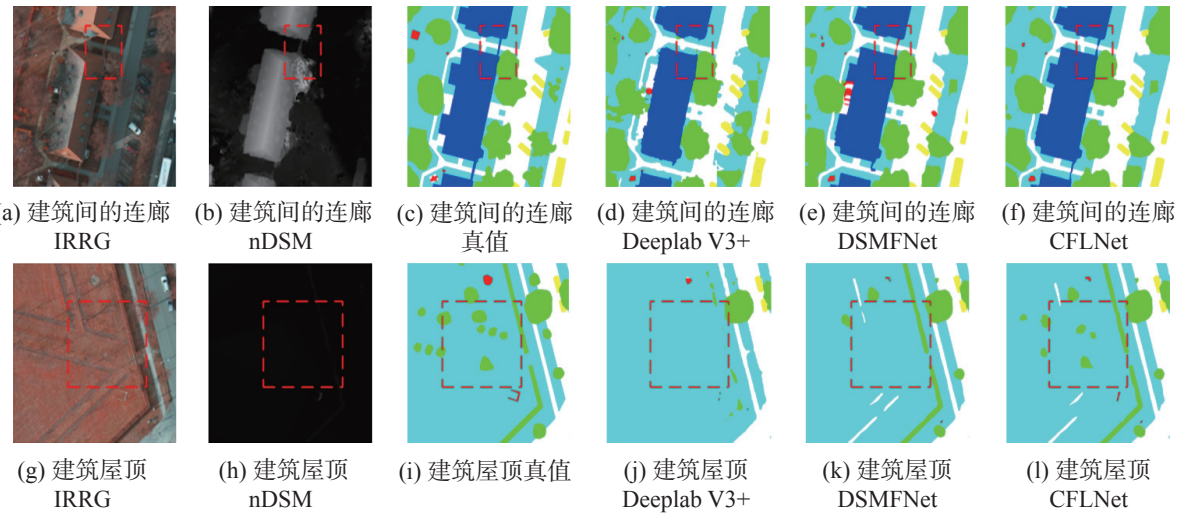
(d) Deeplab V3+　(e) DSMFNet　(f) CFL-Net

图 5　Potsdam 数据集整图预测结果对比

Fig. 5　Comparison of the whole graph prediction results for the Potsdam dataset

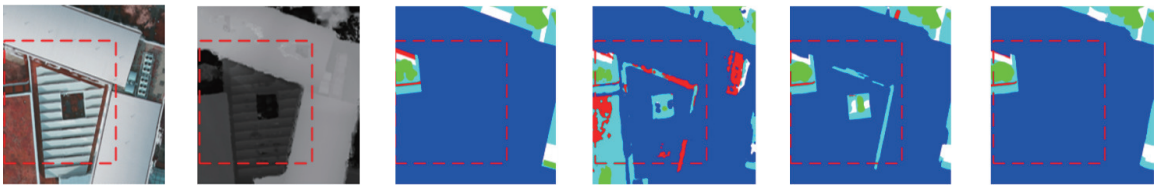
此外, 上述方法在防水表面类别的分割中都表现良好, 均有较明显的道路网, 这是由于在 Potsdam 数据集中大多防水表面都有高对比度的边界和可区分的光谱信息。可以看到, CFL-Net 的分割结果边界往往更加精确且光滑, 尤其是建筑和草地的边缘及拐角。此外, 对于一些零星的小目标 CFL-Net 也能够实现较精准的分割。

在图 6 中给出了 Postdam 数据测试集上更精细的矫正错误分类区域的例子, 并截取了较小的尺寸以便对比观察。由图 6 可见, 两栋楼房之间的连廊区域较小, 样本量也十分少, 由于阴影、遮挡等因素在光谱上难以区分, 因此仅光谱数据输入的 Deeplab V3+网络的分割结果丢失了连廊区域, DSMFNet 虽然考虑到了高程信息, 但简单地合并两者信息并不能够充分利用到多模态信息之间的相似性与互补性, 造成对连廊的错分, 将其分割为背景类。本文的 CFL-Net 在该区域的分割上表现出良好的性能。图 6 中, nDSM 信息不能够完全提供满足对高低植被分类的高程信息时, CFL-Net 也能够充分利用光谱信息, 并得到了比较好的分割结果。



(a) 建筑间的连廊 IRRG　(b) 建筑间的连廊 nDSM　(c) 建筑间的连廊 真值　(d) 建筑间的连廊 Deeplab V3+　(e) 建筑间的连廊 DSMFNet　(f) 建筑间的连廊 CFLNet

(g) 建筑屋顶 IRRG　(h) 建筑屋顶 nDSM　(i) 建筑屋顶真值　(j) 建筑屋顶 Deeplab V3+　(k) 建筑屋顶 DSMFNet　(l) 建筑屋顶 CFLNet



(m) 建筑楼梯 IRRG　(n) 建筑楼梯 nDSM　(o) 建筑楼梯真值　(p) 建筑楼梯 Deeplab V3+　(q) 建筑楼梯 DSMFNet　(r) 建筑楼梯 CFLNet

图 6　Potsdam 数据集裁剪的小区域分割结果对比

Fig. 6　Comparison of small region segmentation results for the Potsdam dataset cropping

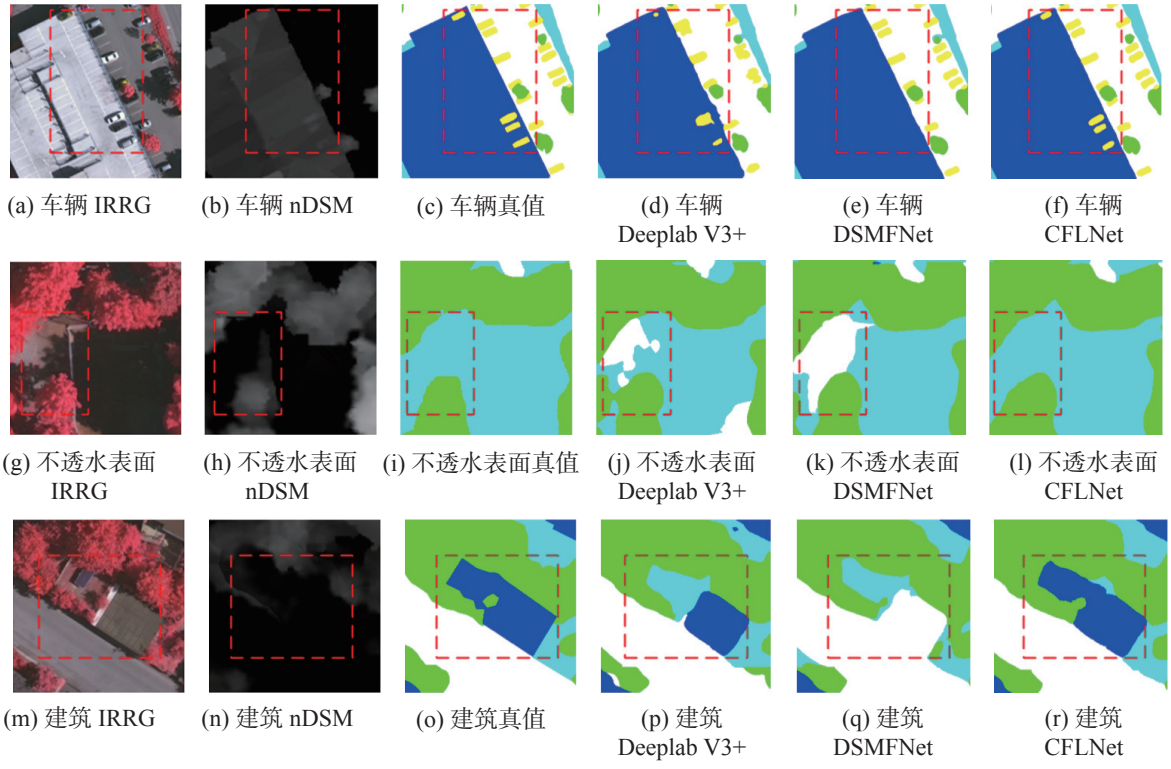
本文在 Vaihingen 数据集上进行了同样的测试，实验结果如表 2 所示。

表 2　Vaihingen 数据集分割精度对比

Table 2　Comparison of segmentation accuracies for the Vaihingen dataset　%

方法	不透水表面	建筑	低植被	树木	车	平均F_1	总体
FCN8s	88.76	92.38	76.54	85.85	74.77	83.67	86.57
Deeplabv3+	90.32	92.89	77.57	87.85	79.02	85.53	88.41
V-fuseNet	90.52	93.21	79.26	88.12	78.79	85.98	89.08
DP-DCN	91.47	94.55	80.13	88.02	80.25	86.83	89.32
DSMFNet	91.78	95.83	82.03	89.52	81.21	88.02	90.05
CFL-Net	92.23	95.80	83.71	90.25	81.45	88.69	90.83

从图 7 中裁切的可视化结果可以看出，本文模型针对于之前提到的问题有了很好的解决，在提取融合互补特征之后，多模态输入网络模型不再偏向于某单一模态数据提供的信息，而是针对语义分割任务以及多模态特征之间的关系选择性地从多模态数据中提取特征。



(a) 车辆 IRRG　(b) 车辆 nDSM　(c) 车辆真值　(d) 车辆 Deeplab V3+　(e) 车辆 DSMFNet　(f) 车辆 CFLNet

(g) 不透水表面 IRRG　(h) 不透水表面 nDSM　(i) 不透水表面真值　(j) 不透水表面 Deeplab V3+　(k) 不透水表面 DSMFNet　(l) 不透水表面 CFLNet

(m) 建筑 IRRG　(n) 建筑 nDSM　(o) 建筑真值　(p) 建筑 Deeplab V3+　(q) 建筑 DSMFNet　(r) 建筑 CFLNet

图 7　Vaihingen 数据集裁剪的小区域分割结果对比

Fig. 7　Comparison of small region segmentation results for Vaihingen dataset cropping

通过图 7 可以发现, 在对一些低矮的建筑物, 以及光谱上表现为植被的建筑物进行分类时, 本文模型表现出很好的性能, 证明该模型具有较好鲁棒性。整体来说, 本文提出的网络具有更好的能力区分具有相似光谱性能的物体, 如不透水表面和建筑物、低植被和树木。同时又可以避免陷入对某一种模态信息的过分信任, 减少可能的误判。

从得到的结果来看, Vaihingen 数据集上 CFL-Net 的 OA 为 90.83%, 平均 F_1 为 88.69%, 分别比最近的竞争方法 DSMFNet 高 0.78% 和 0.67%。尽管 Vaihingen 数据集上的精度集小于 Potsdam 数据集, 但本文提出的方法仍然能够获得更好的性能。Vaihingen 数据集中整张图像分割的可视化结果如图 4 所示。从各个类别的分割结果来看, 本文提出的方法相较于其他方法可以获得更加清晰的地物轮廓, 对于形状多变、纹理特征复杂、结构复杂的建筑物类别分割得也更加完整。另外, 对于一些面积较小的目标, 如车辆、单体植被等, 本文的分割方法在相邻的小目标之间没有粘连现象。

3.4　消融实验

本文对所提出的网络进行了分解和组合, 利用 OA 和 F_1-score 指标验证了每个模块的有效性。消融实验在 Vaihengen 数据集上完成。首先, 为了验证互补特征学习的有效性, 分别单独训练了完整的光谱分支与高程分支, 对得到的模型进行验证。其次, 为了验证多核 MMD 的有效性, 本文在原有网络框架基础上, 使用欧氏距离以及余弦相似度衡量特征之间的距离, 并设置对应的损失函数参与训练。采用 IRRG-branch 表示光谱分支, DSM-branch 表示高程分支, dual path (DP)表示将两个分支联合训练, DP+ED 与 DP+consine 分别表示将欧氏距离与余弦相似度作为互补特征约束的双分支网络, 最终实验结果如表 3 所示。

表 3　消融实验结果

Table 3　Results of the ablation experiments　%

方法	总体精度	F_1
DSM-branch	65.21	64.86
IRRG-branch	88.41	85.53
DP+ED	89.25	87.02
DP+consine	89.39	87.21
DP+MK-MMD(本文)	90.83	88.69

由表 3 可得出结论: 单一 DSM 分支由于缺少车辆、路面等地物的细节信息, 精度最低。同样, 单一光谱分支虽然细节信息较多, 但对于光谱上表现相似的地物识别能力有限, 仍然无法得到较高的精度。联合训练后, 欧氏距离和余弦相似度的表示能力有限, 无法完全衡量分布之间的距离。本文提出的多核 MMD 方法取得了最好的结果, 有效性最好。

4　结束语

本文对目前主流多模态数据融合遥感图像语义分割存在的问题进行了分析, 针对该问题提出了基于互补特征学习的多模态数据语义分割网络。一方面, 本文从互补特征约束中设计损失函数参与网络训练, 使多模态数据在特征提取过程中可以相互学习, 并将它们建模为共同特征与互补特征。另一方面, 本文将一种模态数据提取到的互补特征补充给另外一种模态, 在单一模态特征不足以完成语义重建时, 另一模态中提取到的互补特征可以对其进行补充, 产生更鲁棒的融合特征。

参考文献:

[1] YUAN X, SHI J, GU L. A review of deep learning methods for semantic segmentation of remote sensing imagery[J]. Expert systems with applications, 2021, 169: 114417–114430.

[2] DING Lei, TANG Hao, 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.

[3] FU Gang, LIU Changjun, ZHOU Rong, et al. Classification for high resolution remote sensing imagery using a fully convolutional network[J]. Remote sensing, 2017, 9(5): 498–518.

[4] LI Jinglun, XIU Jiapeng, YANG Zhengqiu, et al. Dual path attention net for remote sensing semantic image segmentation[J]. ISPRS international journal of geo-information, 2020, 9(10): 571–591.

[5] LI Haifeng, QIU Kaijian, CHEN Li, et al. SCAttNet: semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images[J]. IEEE geoscience and remote sensing letters, 2021, 18(5): 905–909.

[6] ZHANG Jing, LIN Shaofu, DING Lei, et al. Multi-scale

context aggregation for semantic segmentation of remote sensing images[J]. Remote sensing, 2020, 12(4): 701–716.

[7] DONG Rongsheng, PAN Xiaoquan, LI Fengying. DenseU-net-based semantic segmentation of small objects in urban remote sensing images[J]. IEEE access, 2019, 7: 65347–65356.

[8] LIU Shuo, DING Wenrui, LIU Chunhui, et al. ERN: edge loss reinforced semantic segmentation network for remote sensing images[J]. Remote sensing, 2018, 10(9): 1339.

[9] CHEN Kaiqiang, FU Kun, GAO Xin, et al. Effective fusion of multi-modal data with group convolutions for semantic segmentation of aerial imagery[C]//2019 IEEE International Geoscience and Remote Sensing Symposium. Yokohama: IEEE, 2019: 3911−3914.

[10] SUN Weiwei, WANG Ruisheng. Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM[J]. IEEE geoscience and remote sensing letters, 2018, 15(3): 474–478.

[11] CAO Zhiying, FU Kun, LU Xiaode, et al. End-to-end DSM fusion networks for semantic segmentation in high-resolution aerial images[J]. IEEE geoscience and remote sensing letters, 2019, 16(11): 1766–1770.

[12] HAZIRBAS C, MA Lingni, DOMOKOS C, et al. FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture[M]//Computer Vision-ACCV 2016. Cham: Springer International Publishing, 2017: 213−228.

[13] AUDEBERT N, LE SAUX B, LEFÈVRE S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks[M]//Computer Vision-ACCV 2016. Cham: Springer International Publishing, 2017: 180−196.

[14] QIN Rongjun, FANG Wei. A hierarchical building detection method for very high resolution remotely sensed images combined with DSM using graph cut optimization [J]. Photogrammetric engineering & remote sensing, 2014, 80(9): 873–883.

[15] CAI Ziyun, HAN Jungong, LIU Li, et al. RGB-D datasets using microsoft kinect or similar sensors: a survey[J]. Multimedia tools and applications, 2017, 76(3): 4313–4355.

[16] ZHANG Wenkai, HUANG Hai, SCHMITZ M, et al. Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling[J]. Remote sensing, 2017, 10(2): 52–65.

[17] WEINMANN M, WEINMANN M. Geospatial computer vision based on multi-modal data—how valuable is shape information for the extraction of semantic information? [J]. Remote sensing, 2017, 10(2): 2–21.

[18] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016 : 770−778.

[19] WANG Yuhao, LIANG Binxiu, DING Meng, et al. Dense semantic labeling with atrous spatial pyramid pooling and decoder for high-resolution remote sensing imagery[J]. Remote sensing, 2018, 11(1): 20–37.

[20] 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.

[21] ZHAO Hengshuang, SHI Jianping, QI Xiaojuan, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6230−6239.

[22] CHEN L C, ZHU Yukun, PAPANDREOU G, et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation[C]//European Conference on Computer Vision. Cham: Springer, 2018: 833−851.

[23] YANG Maoke, YU Kun, ZHANG Chi, et al. DenseASPP for semantic segmentation in street scenes[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018 : 3684−3692.

[24] SHI Lukui, WANG Ziyuan, PAN Bin, et al. An end-to-end network for remote sensing imagery semantic segmentation *via* joint pixel- and representation-level domain adaptation[J]. IEEE geoscience and remote sensing letters, 2021, 18(11): 1896–1900.

[25] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139–144.

[26] YAN Hongliang, DING Yukang, LI Peihua, et al. Mind the class weight bias: weighted maximum mean discrepancy for unsupervised domain adaptation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017 : 945−954.

[27] GRETTON A, SEJDINOVIC D, STRATHMANN H, et al. Optimal kernel choice for large-scale two-sample

tests[C]//Annual Conference on Neural Information Processing Systems. Lake Tahoe: NIPS, 2012: 1205−1213.

[28] ROTTENSTEINE F, SOHN G, GEREK M, et al. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction[J]. ISPRS journal of photogrammetry and remote sensing, 2014, 93: 256–271.

[29] LIU Yifan, ZHU Qigang, CAO Feng, et al. High-resolution remote sensing image segmentation framework based on attention mechanism and adaptive weighting[J]. International journal of geo-information, 2021, 10(4): 241–258.

[30] CAO Zhiying, DIAO Wenhui, SUN Xian, et al. C3Net: cross-modal feature recalibrated, cross-scale semantic aggregated and compact network for semantic segmentation of multi-modal high-resolution aerial images[J]. Remote sensing, 2021, 13(3): 528–545.

[31] LIU Siyu, HE Changtao, BAI Haiwei, et al. Light-weight attention semantic segmentation network for high-resolution remote sensing images[C]//IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium. Waikoloa: IEEE, 2020: 2595−2598. .

[32] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3431−3440.

[33] AUDEBERT N, LE SAUX B, LEFÈVRE S. Beyond RGB: very high resolution urban remote sensing with multimodal deep networks[J]. ISPRS journal of photogrammetry and remote sensing, 2018, 140: 20–32.

[34] PENG Cheng, LI Yangyang, JIAO Licheng, et al. Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation[J]. IEEE journal of selected topics in applied earth observations and remote sensing, 2019, 12(8): 2612–2626.

作者简介:

王兴武, 硕士研究生, 主要研究方向为人工智能、深度学习。

雷涛, 教授, 博士生导师, 陕西科技大学电子信息与人工智能学院副院长, IEEE 高级会员, 主要研究方向为计算机视觉、机器学习。发表学术论文 90 余篇。

王营博, 讲师, 博士, 主要研究方向为散射环境下图像复原与场景感知。