基于多通道自注意力网络的遥感图像场景分类

岳泓光[1]，韩龙玫[2]，王正勇[1]，卿粼波[1]
(1. 四川大学电子信息学院，成都 610065；
2. 成都市规划设计研究院，成都 610041)

摘　要：高分辨率遥感图像场景分类广泛应用于土地监测、环境保护及城市规划等诸多领域. 现有场景分类方法不能很好地结合局部纹理信息和全局语义信息，同时各通道特征之间的关系没有得到有效挖掘. 因此，本文提出了一种基于多通道自注意力网络的遥感图像场景分类模型. 通过卷积网络提取遥感图像的多尺度特征；随后采用特征融合单元建立多尺度特征间的局部-全局关系，基于多头自注意力机制的 Inter-Channel Transformer 在通道维度对融合后的特征建模，并推导特征在通道间的关系，进一步扩大全局感受野，以捕捉其语义结构信息，有效提高了网络的分类精度. 在数据集 AISC 和 SIRI-WHU 上，本文所提算法的整体分类准确率(OA)分别为 95.70%和 94.00%，超过了当前最新的研究算法，证明了所提模型在高分辨率遥感图像场景分类任务中的有效性.

关键词：高分辨率遥感图像场景分类；卷积神经网络；自注意力机制；多通道特征
中图分类号：TP391.4　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2023.023002

Remote sensing image scene classification based on multi-channel self-attention network

YUE Hong-Guang[1], *HAN Long-Mei*[2], *WANG Zheng-Yong*[1], *QING Lin-Bo*[1]
(1. College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China;
2. Chengdu Institute of Planning & Design, Chengdu 610041, China)

Abstract: High resolution remote sensing image (HRRSI) scene classification is widely used in many fields such as land monitoring, environment protection, urban planning and so on. The existing scene classification methods cannot fuse the local-texture and global-semantic information well, and the relationship between the features of each channel has not been effectively explored. Therefore, this paper proposed a new method based on multi-channel self-attention network for HRRSI scene classification. Firstly, the multi-resolution features are extracted by Convolutional Neural Network(CNN); then, a feature fusion unit is used to establish the local-global relationship between multi-scale features. In addition, Inter-Channel Transformer, which is based on multi-head self-attention mechanism, models the merged representations in the channel dimension, and reasons the relationship between the features of each channel, further expands the global receptive field to capture its semantic structure information. Finally, the proposed method improves the classification accuracy. This paper also designs series of experiments on AISC and SIRI-WHU datasets to demonstrate the validity of the proposed algorithm for

收稿日期：2022-05-24
基金项目：国家自然科学基金(61871278)
作者简介：岳泓光(1996—)，男，四川巴中人，硕士研究生，研究方向为计算机视觉. E-mail：610381446@qq.com
通讯作者：王正勇. E-mail：690728634@sina.com

HRRSI scene classification task. The OA(Overall Accuracy) performance are 95.70% and 94.00% on AISC and SIRI-WHU respectively. It has surpassed the state-of-the-art algorithms.

Keywords: HRRSI scene classification; Convolutional neural network; Self-attention mechanism; Multi-channel feature

1 引　言

遥感技术主要是通过卫星等现代化设施捕捉地表物体的电磁辐射信息,进而产生包含大量纹理特征、光谱信息及颜色信息等多种复杂特征的遥感图像[1,2]. 得益于遥感技术的高速发展,高分辨率遥感图像被广泛应用于土地监测[3-6]、环境保护[7]、自然灾害检测[8]和城市规划[9]等众多领域.

研究表明,相对于传统自然图像,遥感图像包含了大量的浅层纹理特征和颜色信息[10]、中层语义信息及高层的视觉信息等[11]. 同时图像中地物目标尺度不一,前/背景特征丰富,如何表达并在特征空间中有效地传递特征信息,成为遥感图像实现精准场景分类的关键. 基于 CNN 的深度学习方法如 VGG[12]、ResNet[13]等由于其强大的特征提取能力而不断兴起,并广泛应用于图像分类[13]、目标检测[14]等机器视觉领域,如张意等[15]学者通过注意力机制及感知损失,构建了基于残差自编码器的遥感图像去噪网络. 考虑到光谱域的非线性特征的提取对于高光谱图像分类至关重要,因此池涛等[16]学者基于多层感知器卷积层和批归一化层构建了七层网络结构,在不同类别物体的光谱特征提取上取得了较好的结果. 然而 CNN 不能很好地对全局信息进行建模,从而对图像语义信息的挖掘不够深入. 得益于 Visual Transformer(ViT)[17]捕捉长距离特征依赖关系的能力,使其经过海量数据训练后的性能超越了 CNN,但 ViT 在全局建模时会损失局部信息,对图像精准分类有一定的损失.

2 相关工作

大量学者通过不断挖掘卷积神经网络的潜力,一定程度上提升了模型对于遥感图像场景分类任务的性能[18-30]. 如李彦甫等[18]在残差网络中融合了自注意力机制,对图像的全局信息建模,实现了较好的分类性能. Li[19]根据图卷积神经网络和神经结构搜索的思想,设计了随机拓扑器和随机映射单元这一轻量级模型用于提高分类精度. 遥感图像中场景的几何形变较大,对深度网络的学习能力有一定的影响,因此,施慧慧等[20]通过引入可形变卷积,构建了深度迁移可变形卷积神经网络(Deep Transfer Deformable Convolutional Neural Networks, DTDCNN),一定程度上增强了对空间位置的采样能力. 王李祺等[21]在遥感图像空间特性如纹理特征的提取上引入了灰度共生矩阵(Gray-Level Cooccurrence Matrix, GLCM)和局部二值模式(Local Binary Pattern,LBP),随后采用 AlexNet 对浅层特征进一步抽象,提取更具表达能力的深层特征,最后融合浅层特征与深层特征用于场景分类. 为了解决传统机器学习在提取图像特征效率较低以及对图像中的空间信息和通道信息无法精准把握等问题,乔星星等[22]应用 ResNet101[13]为基础网络,并在其中融入了空间注意力和通道注意力,进而高效地完成了特征的提取与重建. 然而遥感图像中地物目标的尺度变化较大,上述方法并未考虑在图像特征的学习中引入多尺度学习. 王协等[25]为实现输入图像和特征图像的多尺度学习,引入了多尺度神经网络和空洞卷积;刘美等[26]提出了基于卷积神经网络的多分辨率特征融合网络(Multi-Resolution Feature Network,MRFN),并在 AISC[26]和 SIRI-WHU[27]数据集上取得了较好的性能. 上述学者虽然在模型的设计之初考虑了多尺度特征对模型性能的影响,但是对于多尺度特征的融合依然有所欠缺,未充分利用大尺度特征中所蕴含的纹理信息以及小尺度特征中的语义信息. 因此不能很好地对多尺度特征进一步挖掘其中的深层次属性,同时也不能在各层通道中对特征建模,以此长距离捕捉特征依赖关系.

为解决上述问题,本文充分融合基于多尺度特征的卷积神经网络和基于自注意力机制的通道间 Transformer(Inter-Channel Transformer, ICT)模型,提出了多通道自注意力网络. 采用 MRFN[26]作为基础网络提取初始多分辨率特征图(Feature Map, FM);其次,为更好地捕捉图像中多尺度地物目标的特征信息,本文设计了多尺度特征生成器(Multi-Scale Generator, MSG);最后利用 ICT 模块对多尺度特征进行融合,并推理融合后的特征图在不同通道间的关系,以此增强模型的特征表达能力和长距离捕捉能力. 为验证模型在遥感图像场景

分类任务中的有效性，本文分别在数据集 AISC[26] 和 SIRI-WHU[27] 上开展大量实验，证明了所提模型的良好性能.

3　算法设计

3.1　整体结构

针对遥感图像空间信息多样化、地物目标较小以及类间多样性等特点，本文提出了多通道自注意力网络，其总体结构如图 1 所示. 首先通过卷积和瓶颈残差块（Bottleneck）提取高分辨率遥感图像的初始特征，为了防止梯度消失/爆炸，对残差块重复堆叠四次，并采用串行连接的方式，对特征进行降维减少计算量；然后利用两个并行连接的支路以不同的卷积步长提取不同分辨率的特征，抽象出图像的纹理信息、颜色信息及轮廓信息等，这类特征能很好地描述图像的表层属性. 随后在两个分支的基本残差块（BasicBlock）中分别嵌入通道注意力和空间注意力模块（Convolutional Block Attention Module，CBAM）[28]，为更好地构建深度神经网络，以便学习多分辨率特征中的空间信息和通道信息，对 CBAM BasicBlock 重复堆叠 4 次；利用多尺度特征生成器（Multi-Scale Feature Generator，MSFG）进一步提取多分辨率特征中的语义信息，并产生三组多尺度特征图 F_1、F_2、F_3；最后通过基于自注意力机制的 ICT 对多尺度特征进行融合并推理各通道特征间的关系，以此长距离捕捉特征依赖关系，最终通过全连接网络（Fully Connection，FC）实现遥感图像的场景分类.

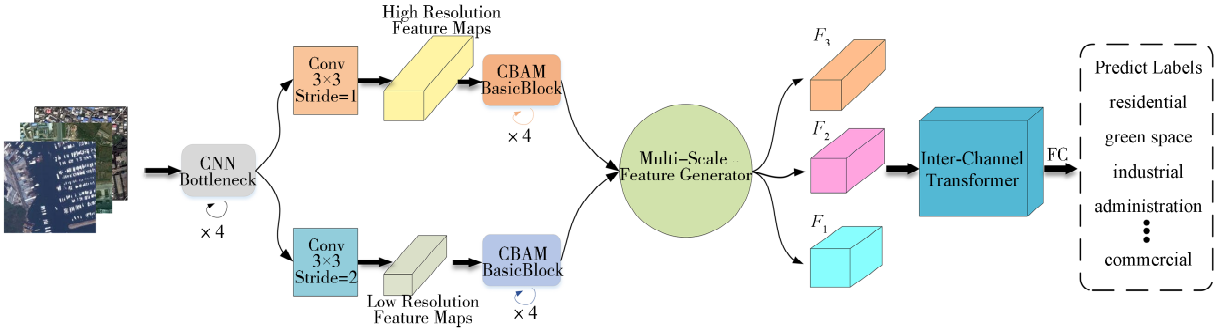

图 1　模型结构

Fig. 1　The architecture ofthe proposed model

3.2　多尺度特征挖掘

为了更好地感知遥感图像中不同尺度地物目标的语义信息，本文设计多尺度特征生成器（Multi-Scale Feature Generator，MSFG），如图 2 所示. 对前文提取的高/低分辨率特征图 F_H 和 F_L 采用卷积和转置卷积（TransConv）以不同的平滑步长 stride 下采样，扩展特征图的感受野. 感受野表征的是特征图的一个图点对应初始图像上某个区域的映射，在深度神经网络中，较大的感受野能提炼出图像中抽象的语义信息，从全局视角下描述图像的本质属性. 本文针对 F_H 和 F_L 分别采用不同的降采样方式产生三组多尺度特征图，为完整保留从 F_H 和 F_L 中抽样出的特征信息，采用张量拼接 Concatenate 方式将 6 个特征图两两融合构成新的特征矩阵，最终产生三个多尺度特征图 F_1、F_2 及 F_3.

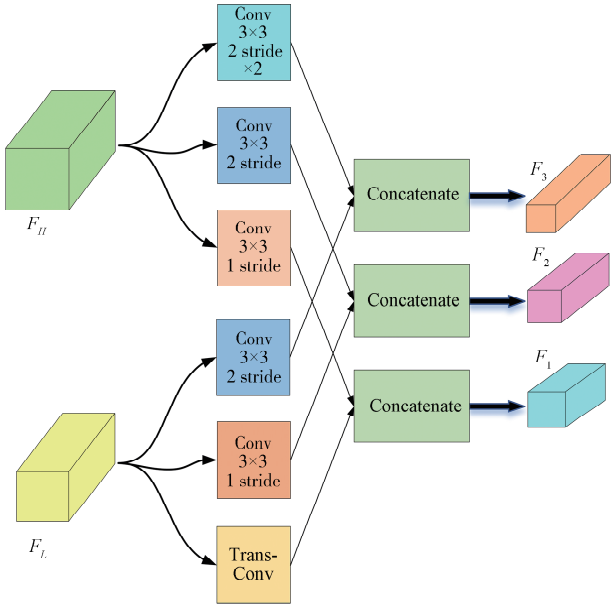

图 2　MSFG 结构图

Fig. 2　The construction of MSFG

3.3　多尺度特征融合及推理

对前文提出的多尺度特征以加权求和的方式进行融合，充分利用各个尺度特征中所聚合的信息，如图 3 中特征融合（Feature Fusion Unit）部分所示. 首先对多尺度特征 F_1、F_2 及 F_3 采用自适应

最大池化将空间维度降为一致,最大池化可以很好地保留图像的纹理信息,因此下采样后的特征图不仅聚合了大量的空间信息,同时有效降低了模型的计算量;其次,相同尺度的特征在各个通道的表达能力各不相同,因此对池化后的特征矩阵采用 Sigmoid 激活函数计算各通道的权重 W_1、W_2 及 W_3. 将权重矩阵与下采样后的特征图相乘,即可得到经通道重校准后新的特征表示 F_1'、F_2' 和 F_3',增强了特征图在通道中的表达能力;最后对加权后的特征图求和得到 F_f,有效融合了低层几何信息和高层语义信息. 相较于本文所提模型,刘美等[26]提出的多分辨率特征融合网络则采用了较为常见的卷积方式,将多分辨率的特征图下采样至尺度一致,然后相互叠加,以实现特征的融合,如此一来便损失了高分辨特征图中所蕴涵的空间几何信息,卷积的引入也带来了额外的参数量;此外未经通道校准的特征也会带来大量的信息冗余,不能突出通道表达能力更强的特征,同时多尺度特征中的语义信息及几何信息未能得到充分的利用,降低了特征的表征能力. 本文所提出的特征融合阶段的计算如式(1)所示.

$$F_f=(\sigma(\mathrm{MaxPool}([F_1,F_2,F_3])))\times(\mathrm{MaxPool}([F_1,F_2,F_3]))^T \tag{1}$$

其中,$[F_1,F_2,F_3]$ 表示多尺度特征矩阵,$\mathrm{MaxPool}([F_1,F_2,F_3])$ 表示计算通道权重后的矩阵 $[W_1,W_2,W_3]$,F_f 为加权后的特征. MaxPool 表示最大池化,σ 表示激活函数 Sigmoid.

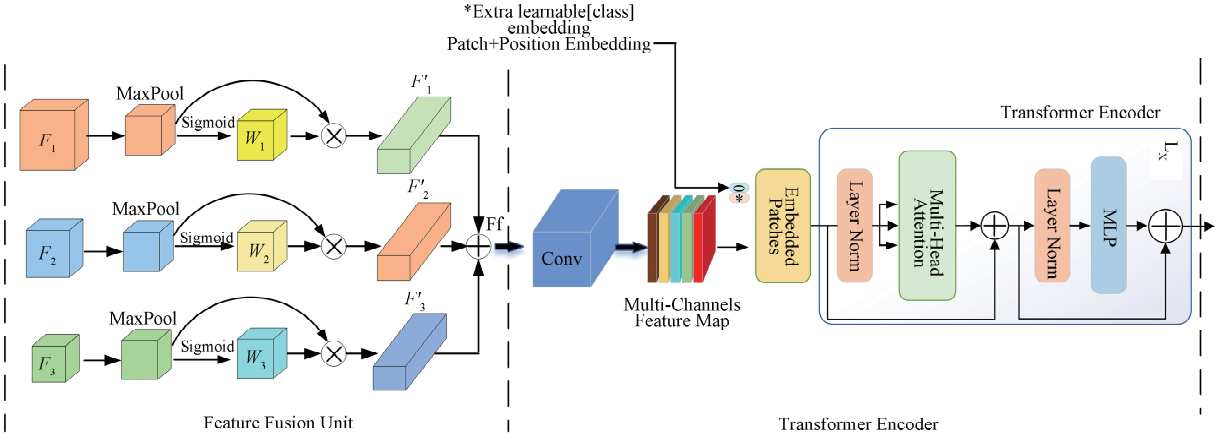

图 3　通道间 Transformer 结构图

Fig. 3　The proposed Inter-Channel Transformer

融合多尺度特征后,对 F_f 作点卷积变换映射到高维空间,充分融合通道信息,产生多通道特征图(Multi-Channel Feature Maps, MCFM). 类似于 Transformer 在自然语言处理领域中将输入切分为多个词向量,在本文中将 MCFM 按通道切分为多个 patches,同时增加一个用于分类输出的可学习向量 class token(X_{class}),为降低模型的学习成本并保留各个 patch 在通道中的位置关系,加入了可学习的位置编码 Position Embedding,将 X_{class} 与 patches 拼接之后加上位置编码构成新的输入特征矩阵 F_t. 在 Transformer Encoder 中,对输入特征进行层归一化(Layer Normalization, LN)处理,利用多头自注意力模块(Multi-Head Self-Attention)在高维空间中对特征建模,产生全局感受野,有效捕捉图像中的长距离依赖关系,多头自注意力机制的计算过程如式(2)所示. 多层感知机(Multi-Layer Perception, MLP)将经注意力加权后的特征图通过线性映射提取其中的高层视觉信息,增强模型的表达能力. 编码器中的残差连接保证了信息流的无损传输,最终实现从低层到高层的跨越. 最后通过聚合了全局语义特征和高层视觉信息的 X_{class} 作最后的分类器.

$$Atten(Q,K,V)=\mathrm{Softmax}\left(\frac{Q\cdot K^T}{\sqrt{d_k}}\right)V \tag{2}$$

$$Q=F_t\times W_Q^T+b_0 \tag{3}$$

$$K=F_t\times W_K^T+b_1 \tag{4}$$

$$V=F_t\times W_V^T+b_2 \tag{5}$$

式(2)中,Q、K、V 分别表示由特征矩阵 F_t 通过不同的权重线性映射得到的 $Query$(查询矩阵)、Key(键矩阵)和 $Value$(值矩阵),$Attn$ 为注意力机制的输出矩阵;式(3)~式(5)中的 W_0、W_1、W_2 为不同的权重矩阵,采用随机初始化并可学习,d_k 代表 K 的维度,b_0、b_1、b_2 表示偏置,$Q\cdot K^T$ 即为矩阵 Q 在矩阵 K^T 上的投影,Softmax 为权值生成函数.

4　实验结果与分析

4.1　实验数据集

本文采用的实验数据集来源于数据集 AISC[26] 和 SIRI-WHU[27]. 其中 AISC 由刘美等[26]提出，数据集覆盖中国二十多个城市，共计 6 种类别，分别为：residential、administration、industrial、road、commercial 以及 green space，图 4 展示了部分场景图. 数据集共有 17 831 张高分辨率遥感图像，每个类别分别有大约 2900 张图片，其空间分辨率为 0.6 m，每张图片为 200×200 像素. SIRI-WHU[27]数据集由武汉大学的 RE-IDEA 团队设计并于 2016 年发布，数据集包括 agriculture、commercial、industrial 等 12 个场景类别，共计 2400 张图片，每类 200 张图片，每张图片的像素尺寸为 200×200，空间分辨率 2 m，部分场景图如图 5 所示.

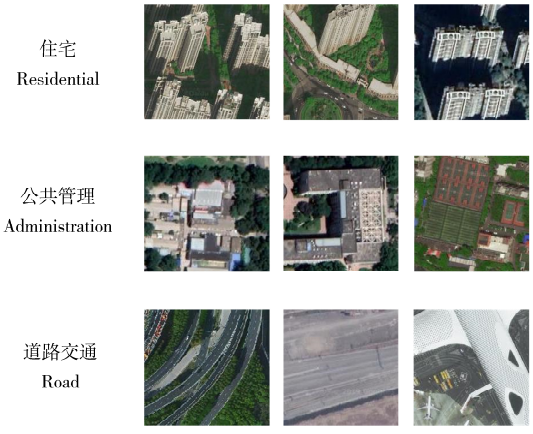

图 4　AISC 数据集部分图片
Fig. 4　Images of AISC dataset

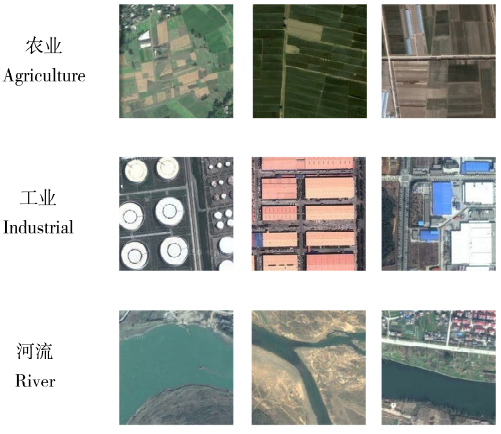

图 5　SIRI-WHU 数据集部分图片
Fig. 5　Images of SIRI-WHU dataset

为避免由实验数据的差异对实验带来的影响，本文在训练集、验证集、测试集的分配与[26]保持相同. 在 AISC[26] 数据集中，随机采用每个类别中 60%的图片作为训练集用于拟合模型，20%的图片作为验证集以验证模型的性能，剩下的 20%作为最终的测试集，测试模型的泛化能力. 在 SIRI-WHU[27]数据集中随机选取每个场景 80%的图片作为训练集，20%的图片作为测试集. 并对两个数据集采用数据预处理，包含随机旋转、随机裁剪及归一化等.

4.2　实验环境及参数配置

为验证本文所提出的模型在高分辨率遥感图像场景分类任务上的有效性，本文在 AISC[26] 数据集和 SIRI-WHU[27] 数据集上开展大量实验. 实验采用 64 位 Ubuntu、Intel(R) Core(TM) i7-9700@3.00 GHz CPU、64 GB 内存以及 11 GB 显存的 NVIDIA RTX2080Ti 显卡，深度学习框架采用的是 PyTorch 1.7.1. 本文采用交叉熵损失函数计算预测值与真实标签间的损失值，并通过 Adam 优化器对模型进行优化，每批次样本的输入量(batch size)设置为 32，总共训练次数(epoch)为 300 轮，初始学习率为 0.0001，每训练 150 轮，学习率乘以 0.5.

4.3　实验结果及分析

4.3.1　所提模型在 AISC[26] 和 SIRI-WHU[27] 数据集上的结果及分析　为验证本文所提出的模型在高分辨率遥感图像场景分类任务中的性能，分别在 AISC[26] 和 SIRI-WHU[27] 数据集上开展与现有经典算法的对比实验. 表 1 和表 2 分别展示了在 AISC[26] 数据集和 SIRI-WHU[27] 数据集上与现有方法在总体分类准确率上的对比. 如表 1 和表 2 所示，本文所提算法在两个数据集上的整体分类准确率达到最高，相较于 baseline 网络[26]在 AISC[26] 和 SIRI-WHU[27] 上分别提高了 7.53%和 9.83%. 而相对于 ResNet-50[13] 和 ViT[17] 算法，也有显著的提高. 由此可见，本文所提算法中的 CBAM 模块能捕捉多尺度特征中的位置信息和空间信息从而将特征精细化，而通过特征融合模块能够有效聚合多尺度特征中的局部纹理信息和全局语义信息，从而产生多通道特征，最后通道间 Transformer 对不同通道的特征信息进行有效的位置推理，在高维空间中捕捉其局部-全局特征间的依赖关系，实现了较好的分类精度. 在参数量、计算量以及推理速度方面，本文所提出的多通道自注意力网络由于引入了 Transformer Encoder，而 Transformer Encoder 主要由多头自注意力机制以及多层感知机(Multi-Layer Perception, MLP)构成，在多头自注意力机

制中模型的复杂度与 tokens 的 HW 成二次关系，其中 HW 表示特征图的高和宽，同时 MLP 主要由全连接层构成，所以带来了一定的计算量及参数量，但是实验结果表明，多头自注意力机制更有利于在高层视觉任务中对全局语义信息的挖掘，从而有效提升了模型性能. 文献[19]引入了图神经网络，由于图神经网络独特的结构，以节点和边界传递特征信息，所以只涉及基本的线性计算，因此不会带来大量的参数. 在推理速度方面，本文计算出了在测试过程中推理一张图片所需要的时间，结果如表 1[26] 及表 2[27] 所示. RTRMM[19] 由于作者未公开源码所以推理时间无法计算.

表 1　本文方法与其他方法在 AISC 上的比较

Tab. 1　Comparison between theproposed method and other methods on AISC

分类算法	整体准确率(OA)/%	Params /M	FLOPs /G	推理时间 /s
AlexNet[28]	71.32	57.03	0.53	1.12
VGG16[3]	74.53	134.31	12.29	3.37
ResNet-50[13]	83.47	23.53	3.52	1.12
Inception-v3[29]	86.26	27.16	2.08	1.68
MRFN[26]	88.17	5.7	1.81	0.56
本文模型	95.70	67.71	14.61	4.49

表 2　本文方法与其他方法在 SIRI-WHU 上的比较

Tab. 2　Comparison between the proposed method and other methods on SIRI-WHU

分类算法	整体准确率(OA)/%	Params /M	FLOPs /G	推理时间 /ms
BoVW[30]	73.93	/	/	/
SPM-SIFT[3]	80.26	/	/	/
MRFN[26]	84.17	5.23	1.08	0.83
RTRMM[19]	92.00	0.82	0.24	/
本文模型	94.00	67.71	14.61	4.16

图 6 和图 7 分别展示了本文所提算法在 AISC[26] 和 SIRI-WHU[27] 数据集上的混淆矩阵. 主对角线表示每类场景的分类准确率，即召回率(Recall). Kappa 系数可用于验证模型的一致性，由混淆矩阵计算得出. 由图 6 可以计算得出，本模型在 AISC[26] 上的 $Kappa=0.949$，表明本模型的多分类精度较高，一致性较强. 同时，在 AISC[26] 数据集上每个场景正确分类的准确率都达到了 93% 以上，但是 residential 和 industrial 分别被错分为 commercial 和 road 的比例均为 3%，对数据样本仔细分析后发现，residential 和 commercial 具有相似的纹理和颜色，同时其地物目标的排列方式也有相同的规律，这就导致了极大的类间相似性，使得模型对这两类场景的识别造成了一定的困难；其次，对比 industrial 样本和 road 样本时，发现两者的颜色特征近乎一致，同时其 industrial 的图像内也包含了众多 road 的特征，这也就造成了少许比例的误判.

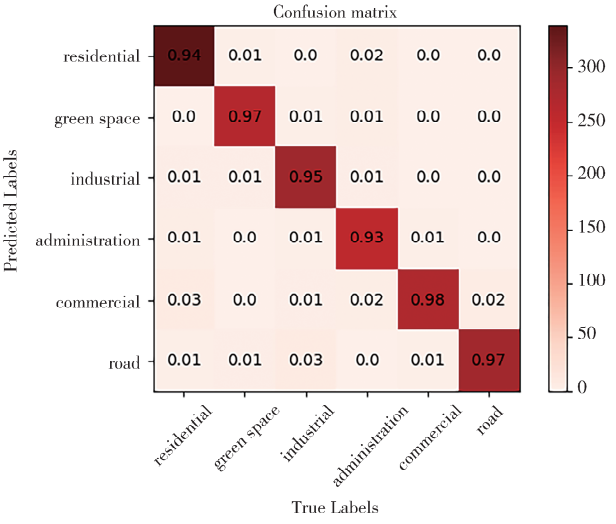

图 6　AISC[26] 上的混淆矩阵

Fig. 6　The confusion matrix on AISC[26]

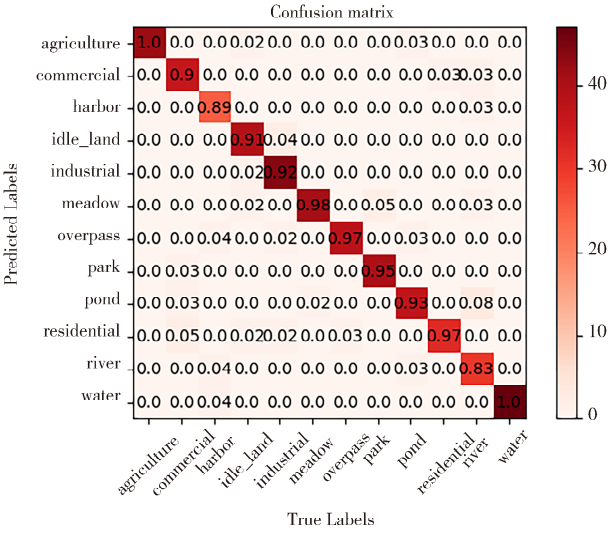

图 7　SIRI-WHU[27] 上的混淆矩阵

Fig. 7　The confusion matrix on SIRI-WHU[27]

在图 7 中，$Kappa=0.934$，同样表明了本文所提算法在公开数据集 SIRI-WHU[27] 上的多分类性能较好，一致性较强，所有场景的分类准确率都在 83%以上，其中 agriculture 和 water 的准确率达到了 100%，表明这两种场景与其他场景表现出

了不同的纹理布局和特殊的颜色及结构特征，然而对于 harbor 场景，与 water 和 river 的颜色信息极为相似，分别导致了 4%的错分率，而 overpass 在结构上与 harbor 相似，同样有 4%的错分率. 由于 river 与 pond 有较大的类间相似性，具体表现在整体结构、场景目标以及颜色分布，导致了分类准确率仅有 83%，被错分为 harbor 的比例达到 8%. 除此之外，其余十类场景都表现出了较高的分类准确率，均达到 90%以上.

4.3.2 通道推理对比实验　为验证基于多头自注意力的通道间 Transformer(Inter-Channel Transformer, ICT)和通道注意力(Convolutional Block Attention Module, CBAM)[28]在推理性能及推理速度方面的差异，本文设计了相关的对比试验，实验结果如表 3 所示. 在验证通道注意力时，本文采用 CBAM 中的通道注意力部分替换 ICT 中的 Transformer Encoder，通道注意力模块首先对融合多尺度特征后的特征表示做全局平均池化和最大池化处理，随后通过多层感知机(Multi-Layer Perception, MLP)以及激活函数求得通道注意力权重，将融合后的特征与权重矩阵相乘得到经通道注意力加权后的特征表示. 由于在通道注意力中为了降低计算量采用了池化操作，因此会损失较多的信息，同时针对场景分类这一高层视觉任务，语义信息对于精准区分类间差异性有着较大的作用，而语义信息的获取与特征图中的感受野有直接关系. 较大的感受野能有效捕捉场景中的语义类，然而通道注意力不能显著增加特征图的感受野，因此在全局语义特征的获取上较 ICT 有一定的差距. 实验结果证明，ICT 得益于 Transformer Encoder 中的多头自注意力机制，在建立长距离特征依赖关系时有着明显的优势，同时通过恒等映射及矩阵缩放点乘产生注意力权重时不会带来信息的大量损失. 在 AISC[26]中，由于数据集较为简单，语义类别仅有 6 类，类间差异性较大. 因此 ICT 和通道注意力在推理性能方面相差不大. 在总体分类准确率(OA)上采用 ICT 的模型仅比采用通道注意力的高 0.4 个百分点. 但是在数据集 SIRI-WHU[27]中，由于场景类别较多，达到 12 类，同时类间的差异性相对较小，对于精准分类带来了较大的困难，因此对模型的推理能力也带来了更大的考验. 此时采用通道注意力相比于采用 ICT，OA 低了 5.42 个百分点. 由此可见，在相对复杂的数据集中，通道间 Transformer(ICT)较于通道注意力在通道推理上有着更强的性能.

表 3　通道推理对比实验

Tab. 3　Contrastive experiments of channel inference

数据集	ICT	通道注意力	OA/%
AISC[26]	✓	×	95.70
	×	✓	95.30
SIRI-WHU[27]	✓	×	94.00
	×	✓	88.58

4.3.3　消融实验　本文设计了系列消融实验，以此验证模型中 CBAM 和 ICT 模块的有效性. 如表 4 和表 5 所示.

表 4　AISC[26]的消融实验

Tab. 4　The ablation study on AISC[26]

Method	OA/%
MRFN[26]	88.17
MRFN[26]+CBAM[28]	92.71
MRFN[26]+ICT	91.70
MRFN[26]+CBAM[28]+ICT	95.70

表 5　SIRI-WHU[27]的消融实验

Tab. 5　The ablation study on SIRI-WHU[27]

Method	OA/%
MRFN[26]	84.17
MRFN[26]+CBAM[28]	90.62
MRFN[26]+ICT	91.04
MRFN[26]+CBAM[28]+ICT	94.00

(1) 即仅使用多分辨率特征融合网络(Multi-Resolution Feature Network, MRFN)，整体分类准确率分别仅为 88.17%和 84.17%.

(2) 在 MRFN 中的 BasicBlock 模块中添加 CBAM 后，可以看出其 OA 分别增加了 4.54 个百分点和 6.45 个百分点，由此可见 CBAM 有效聚合了特征空间中的通道信息和空间信息，增强了模型的特征表达能力.

(3) 当仅用 ICT 对多通道特征在高维空间中建模时，OA 分别从 88.17%和 84.17%提升到了 91.70%和 91.04%，由此可见，ICT 模块有效实现了从局部特征到全局特征的跨越，其捕捉的高维语义结构信息表达了图像最本质的属性，因此提高了整体分类精度. 注意到，在 CBAM 和 ICT 的双重加持下，使得本文模型的性能在 baseline 的基础上

提升较大,整体分类准确率超过了一些经典算法.

5 结　语

遥感图像具有分辨率高、场景信息复杂且地物目标分辨率差异较大等特点.考虑到现有方法不能很好地聚合遥感图像中复杂的空间信息和通道信息,也不能捕捉长距离依赖关系,本文充分融合基于多尺度特征的卷积神经网络和基于自注意力机制的通道间 Transformer(Inter-Channel Transformer, ICT),提出了多通道自注意力网络.利用卷积神经网络提取精细化的多尺度特征,对多尺度特征加权融合之后产生多通道特征,通过 ICT 建立全局感受野捕捉高维结构信息,有效提升了模型的性能.同时在 AISC 和 SIRI-WHU 数据集上分别达到了 95.70%和 94.00%的优异性能,证明了本文所提模型在高分辨率遥感图像场景分类上的可行性.但是对于类间相似性较大的场景,例如结构、颜色信息接近的场景,本文算法会造成一定的误判几率.如何针对相似的场景精确捕捉其中细微的差别,以及融合其他多源大数据设计新的分类模型,从而进一步提高分类精度是下一步研究方向.

参考文献:

[1] 杨帆.航测遥感技术探析[J].科技创新与应用,2017,27:43.

[2] 张裕,杨海涛,袁春慧.遥感图像分类方法综述[J].兵器装备工程学报,2018,39:108.

[3] Fauvel M, Tarabalka Y, Benediktsson J A, *et al*. Advances in Spectral-Spatial classification of hyperspectral images [J]. P IEEE, 2013, 101: 652.

[4] Martha T R, Kerle N, van Westen C J, *et al*. Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis [J]. IEEE T Geosci Remote, 2011, 49: 4928.

[5] Cheng G, Guo L, Zhao T, *et al*. Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA [J]. Int J Remote Sens, 2013, 34: 45.

[6] Lv Z Y, Shi W, Zhang X, *et al*. Landslide inventory mapping from bitemporal high-resolution remote sensing images using change detection and multiscale segmentation [J]. IEEE J-Stars, 2018, 11: 1520.

[7] Zhang Y S, Wu L, Ren H Z, *et al*. Mapping water quality parameters in urban rivers from hyperspectral images using a new self-adapting selection of multiple artificial neural networks [J]. Remote Sens-Basel, 2020, 12: 336.

[8] Veraverbeke S, Dennison P, Gitas I, *et al*. Hyperspectral remote sensing of fire: state-of-the-art and future perspectives [J]. Remote Sens Environ, 2018, 216: 105.

[9] Cheng G, Han J W, Lu X Q. Remote sensing image scene classification: benchmark and state of the art [J]. Proc IEEE, 2017, 105: 1865.

[10] Li W M, Liu H Y, Wang Y, *et al*. Deep learning-based classification methods for remote sensing images in urban built-up areas [J]. IEEE Access, 2019, 7: 36274.

[11] 赵理君,唐娉,霍连志,等.图像场景分类中视觉词包模型方法综述[J].中国图象图形学报,2014,19:333.

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2014-09-04]. https://arxiv.org/abs/1409.1556.

[13] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, USA: IEEE, 2016.

[14] Han J, Zhang D, Cheng G, *et al*. Advanced deep-learning techniques for salient and category-specific object detection: a survey[J]. IEEE Signal Proc Mag, 2018, 35: 84.

[15] 张意,阚子文,邵志敏,等.基于注意力机制和感知损失的遥感图像去噪[J].四川大学学报:自然科学版,2021,58:042001.

[16] 池涛,王洋,陈明.多层局部感知卷积神经网络的高光谱图像分类[J].四川大学学报:自然科学版,2020,57:103.

[17] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale [EB/OL]. [2020-10-22]. https://arxiv.org/abs/2010.11929.

[18] 李彦甫,范习健,杨绪兵,等.基于自注意力卷积网络的遥感图像分类[J].北京林业大学学报,2021,43:81.

[19] Li J, Weinmann M, Sun X, *et al*. Random topology and random multiscale mapping: an automated design of multiscale and lightweight neural network for remote sensing image recognition[J]. IEEE T Geosci Remote, 2021, 60: 1.

[20] 施慧慧,徐雁南,滕文秀,等.高分辨率遥感影像

深度迁移可变形卷积的场景分类法[J]. 测绘学报，2021，50：652.

[21] 王李祺，张成，侯宇超，等. 基于深度学习特征融合的遥感图像场景分类应用[J]. 南京信息工程大学学报：自然科学版，2022，15：1.

[22] 乔星星，施文灶，刘芫汐，等. 基于 ResNet 双注意力机制的遥感图像场景分类[J]. 计算机系统应用，2021，30：243.

[23] 常洪彬，李文举，李文辉. 基于注意力机制的航空图像旋转框目标检测[J]. 吉林大学学报：理学版，2022，60：1363.

[24] 郁强，王宽，王海. 一种多尺度 YOLOv3 的道路场景目标检测算法[J]. 江苏大学学报：自然科学版，2021，42：628.

[25] 王协，章孝灿，苏程. 基于多尺度学习与深度卷积神经网络的遥感图像土地利用分类[J]. 浙江大学学报：理学版，2020，47：715.

[26] 刘美，卿粼波，韩龙玫，等. 基于遥感影像和神经网络的城市用地功能分类[J]. 太赫兹科学与电子信息学报，2021，19：132.

[27] Zhu Q，Zhong Y，Zhao B，*et al*. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery [J]. IEEE Geosci Remote S，2016，13：747.

[28] Woo S，Park J，Lee J Y，*et al*. Cbam：convolutional block attention module[C]//Proceedings of the 2018 European conference on computer vision. Cham：Springer，2018.

[29] Szegedy C，Vanhoucke V，Ioffe S，*et al*. Rethinking the inception architecture for computer vision [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2016.

[30] Liu Y，Zhong Y，Fei F，*et al*. Scene classification based on a deep random-scale stretched convolutional neural network[J]. Remote Sens-Basel，2018，10：444.

引用本文格式：

中　文：岳泓光，韩龙玫，王正勇，等. 基于多通道自注意力网络的遥感图像场景分类[J]. 四川大学学报：自然科学版，2023，60：023002.

英　文：Yue H G，Han L M，Wang Z Y，*et al*. Remote sensing image scene classification based on multi-channel self-attention network [J]. J Sichuan Univ：Nat Sci Ed，2023，60：023002.