计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(5):15−27 [doi: 10.15888/j.cnki.csa.009512]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

双分支注意力与 FasterNet 相融合的航拍场景分类①

杨本臣, 曲业田, 金海波

(辽宁工程技术大学 软件学院, 葫芦岛 125105)
通信作者: 曲业田, E-mail: QuYeTian123@163.com

摘　要: 航拍高分辨率图像的场景类别多且类间相似度高, 经典的基于深度学习的分类方法, 由于在提取特征过程中会产生冗余浮点运算, 运行效率较低, FasterNet 通过部分卷积提高了运行效率但会降低模型的特征提取能力, 从而降低模型的分类精度. 针对上述问题, 提出了一种融合 FasterNet 和注意力机制的混合结构分类方法. 首先采用“十字型卷积模块”对场景特征进行部分提取, 以提高模型运行效率. 然后采用坐标注意力与通道注意力相融合的双分支注意力机制, 以增强模型对于特征的提取能力. 最后将“十字型卷积模块”与双分支注意力模块之间进行残差连接, 使网络能训练到更多与任务相关的特征, 从而在提高分类精度的同时, 减小运行代价, 提高运行效率. 实验结果表明, 与现有基于深度学习的分类模型相比, 所提出的方法, 推理时间短而且准确率高, 参数量为 19M, 平均一张图像的推理时间为 7.1 ms, 在公开的数据集 NWPU-RESISC45、EuroSAT、VArcGIS (10%) 和 VArcGIS (20%) 的分类精度分别为 96.12%、98.64%、95.42% 和 97.87%, 与 FasterNet 相比分别提升了 2.06%、0.77%、1.34% 和 0.65%.

关键词: 遥感场景; 图像分类; 注意力机制; 残差连接; FasterNet

引用格式: 杨本臣,曲业田,金海波.双分支注意力与 FasterNet 相融合的航拍场景分类.计算机系统应用,2024,33(5):15–27. http://www.c-s-a.org.cn/1003-3254/9512.html

Aerial Scene Classification by Fusion of Dual-branch Attention and FasterNet

YANG Ben-Chen, QU Ye-Tian, JIN Hai-Bo

(Software College, Liaoning University of Engineering and Technology, Huludao 125105, China)

Abstract: The scenes in high-resolution aerial images are of many highly similar categories. The classic classification method based on deep learning offers low operational efficiency because of the redundant floating-point operations generated in the feature extraction process. FasterNet improves the operational efficiency through partial convolution but reduces the feature extraction ability and hence the classification accuracy of the model. To address the above problems, this study proposes a hybrid structure classification method integrating FasterNet and the attention mechanism. Specifically, the “cross-shaped convolution module” is used to partially extract scene features and thereby improve the operational efficiency of the model. Then, a dual-branch attention mechanism that integrates coordinate attention and channel attention is used to enable the model to better extract features. Finally, a residual connection is made between the “cross-shaped convolution module” and the dual-branch attention module so that more task-related features can be obtained from network training, thereby reducing operational costs and improving operational efficiency in addition to improving classification accuracy. The experimental results show that compared with the existing classification models based on deep learning, the proposed method has a short inference time and high accuracy. Its number of parameters is

① 基金项目: 国家自然科学基金 (62173171); 国家自然科学基金青年基金 (41801368)
收稿时间: 2023-11-30; 修改时间: 2023-12-29; 采用时间: 2024-01-18; csa 在线出版时间: 2024-04-07
CNKI 网络首发时间: 2024-04-10

19M, and its average inference time for one image is 7.1 ms. The classification accuracy of the proposed method on the public datasets NWPU-RESISC45, EuroSAT, VArcGIS (10%), and VArcGIS (20%) is 96.12%, 98.64%, 95.42%, and 97.87%, respectively, which is 2.06%, 0.77%, 1.34%, and 0.65% higher than that of the FasterNet model, respectively.

Key words: remote sensing scene; image classification; attention mechanism; residual connection; FasterNet

遥感图像是指通过遥感技术获取的地球表面或大气层的图像数据, 获取方式包括卫星遥感、航空摄影、雷达遥感等[1]. 近些年来, 随着遥感技术的发展, 遥感图像的质量在不断提升, 并且图像获取变得更加容易, 这使得遥感图像在自然灾害预测[2]、城乡规划[3]、植被制图[4]和土地覆盖分类[5]等方面得到了日趋广泛的应用.

场景分类是遥感影像领域中一个重要的研究方向. 遥感场景是指通过遥感技术获取到的多种地物目标以特定的空间布局形成的场景区域. 同类场景的不同方位可能分布着多种类型的地物目标, 不同类的场景中可能只包含存在空间布局差异的相同类型地物目标, 这给建立底层地物目标与高层语义之间的联系带来了巨大挑战. 类别多并且不同类之间相似度高, 导致现有模型在分类精度和分类时间上均有进一步提升的空间.

早期研究通过手动设计特征的方法来计算类间相似度, 从而对不同的场景类进行区分. 例如最大似然法[6]、最小距离法[7]和逻辑回归等分类算法[8], 但是这类方法出现了“维度灾难”[9]和分类精度较低的问题. 基于支持向量机 (SVM)、随机森林算法 (RF) 和主成分分析 (PCA) 等机器学习的方法很好地解决了上述问题. 其中基于支持向量机的方法在高维数据和小样本数据上的表现非常良好, 在土地类别的判别[10]中有广泛应用. 但是这些方法非常依赖于手动设计特征的质量, 只能学习到遥感场景图像的浅层特征, 容易忽略其丰富的深层特征, 导致分类精度不高.

随着深度学习的兴起, 遥感技术与深度学习分类器结合的方法[11]能够学习到图像蕴含的更深层次特征, 从而解决了传统的遥感图像分类方法只能学习到浅层特征的问题, 使得分类精度得到一定的提升. 自 AlexNet[12]被提出以来, 卷积神经网络凭借着高效精准的捕获语义信息的能力, 备受计算机视觉领域的关注. Hu 等人[13]采用一维卷积神经网络对光谱信息进行深层次特征提取, 但是由于忽略了空间信息, 导致分类精度不高. Duan 等人[14]将卷积神经网络和超像素算法相结合, 以增强局部区域的平滑特性, 从而获取更丰富的空间信息; Wang 等人[15]将卷积提取的特征通过主成分分析降维, 并获取到具有层次化特点的全局特征, 同时融合底层和中层特征, 从而形成具有判别性特征的图像数据, 以捕捉重要的空间信息.

以上方法在训练微调过程中要求输入图像具有固定尺寸, 这种过程会丢失重要的空间信息, 导致提升分类精度的效果不明显. 为解决此类问题, Xie 等人[16]提出了一种无尺度的 CNN, 以达到保留高分辨率遥感图像中关键空间信息的目的. Liu 等人[17]设计了一种多尺度的 CNN, 构建了一种包含固定尺度网络和可变尺度网络的双分支网络结构来融合不同空间的特征信息. 为了使模型有更好的空间信息捕获能力, 许多研究者将 Transformer 用到图像识别任务中. Roy 等人[18]提出了一种基于 LiDAR 数据的多模态融合变化器网络, 该方法在高光谱遥感图像分类任务中具有良好的表现. 金传等人[19]提出了将 CNN 与 Transformer 相融合的网络, 该方法利用坐标注意力定位到最感兴趣的区域, 在遥感场景分类中取得了良好的效果; Touvron 等人[20]构建了一种具有残差结构的多层感知机 (MLP), 在图像的输入和末尾分别添加两个残差连接, 使每个模块相对独立, 提高了模型对特征的表达能力. 以上基于 Transformer 的分类网络能够取得良好的分类效果得益于 Transformer 拥有强大的上下文语义信息捕捉能力. 但是, Zhu 等人[21]验证了视觉转换器 (Vision Transformer)[22]在小规模数据集上的性能不佳, 除此之外, 基于 Transformer 的模型复杂度较大, 处理视觉任务的时间较长, 容易丢失局部语义信息.

在遥感场景分类任务中, 为了能更快速并且精准地定位到重要的特征, 受到以上文献的启发, 本文构建 FasterNet[23]与坐标注意力[24]相结合的混合网络. 本文的贡献如下.

(1) 对 FasterNet 中的部分卷积进行改进, 融合点卷积, 构建改进的部分卷积模块, 对特征采取部分提取的策略, 减少内存访问次数, 提高了分类速度.

(2) 建立双分支的混合注意力模块, 一个分支对卷积模块提取到的特征进行水平和垂直方向的聚集, 另一分支建立通道注意力聚集特征, 通过两个分支特征的融合, 增加了与任务相关的空间特征信息, 从而提高分类精度.

(3) 将部分卷积和注意力机制进行残差融合, 以解决嵌入注意力机制增加模型复杂度所导致的梯度消失问题.

1 本文模型

1.1 本文模型结构

本文研究的基础模型是 FasterNet, 其结构如图 1 所示, 采用部分卷积的策略, 虽然避免了深度可分离卷积在提取特征过程中由于过多的内存访问而带来的冗余计算, 但是忽略了另一部分特征图中的重要信息, 从而影响整体模型的分类精度.

FasterNet 模型缺少位置编码、窗口移动和相对编码等操作, 处理视觉任务的计算量较小并且时间短. 整个 FasterNet 的结构和 Swin Transformer[25]的结构类似, 堆叠基础模块的次数为 4 次.

在 FasterNet 模型基础之上, 考虑到遥感场景图像具有复杂的纹理特征, 对 FasterNet block 进行拆分并建立更深层次特征提取模块, 分别是部分卷积模块和双分支注意力模块, 两个模块之间通过残差模块连接, 以增强模型对特征的表达能力, 如图 2 所示.

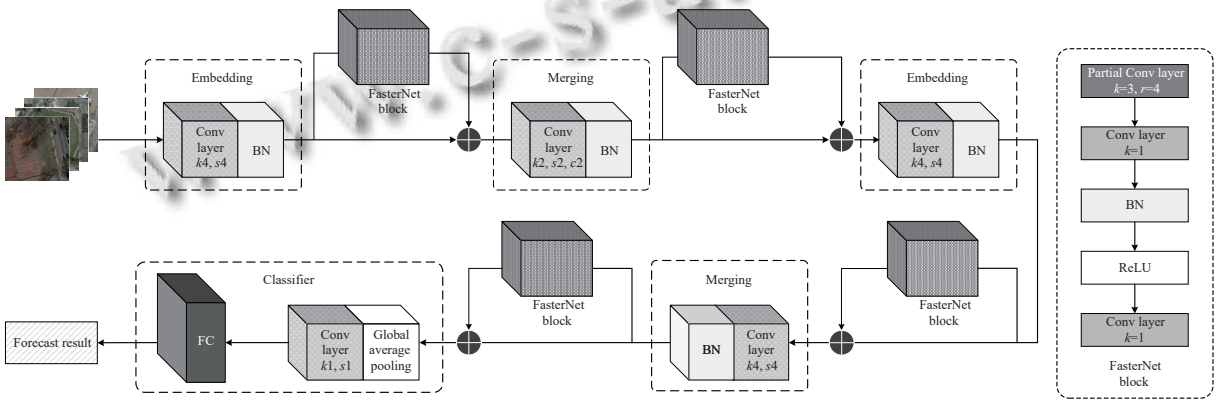


图 1 FasterNet 模型图

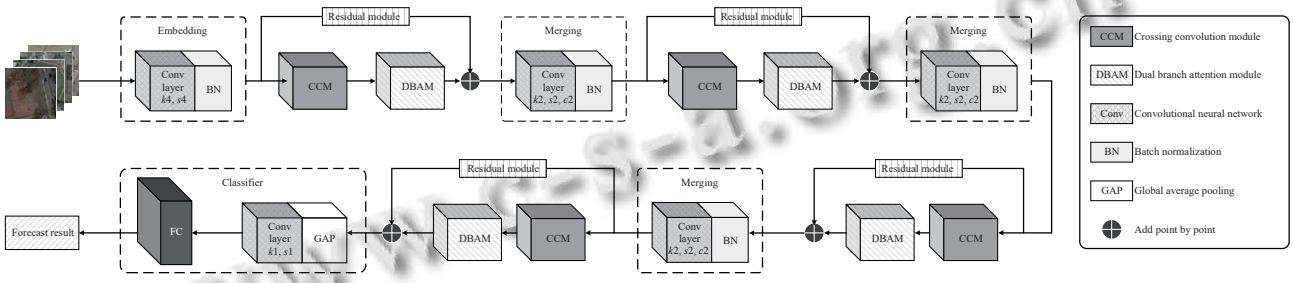


图 2 本文模型图

模型主要由 Embedding 层, 十字型卷积模块 (crossing convolution module)、双分支注意力模块 (dual branch attention module)、Merging 层和 Classifier 层组成. Embedding 层将图像切分成均匀的块, 高效地提取图像信息, 并经过归一化层 (batch normalization, BN) 处理, 将离散的图像数据转为连续的向量序列, 最终输出对应的特征图, 从而提高模型的稳定性; 十字型卷积层和双分支注意力层对得到的特征图进行快速并精准地提取, 得到目标图像的重要特征; Merging 层将输入的原始图像数据增加通道数以降低分辨率, 同时划分若干个均匀的块, 对每个块进行归一化操作, 保留了原始图像的特征, 减小了重要特征信息的损失; Classifier 层接收到最终的特征图后, 得到与类别相关的特征向量, 将得到的特征向量输入到全连接层, 得到类别的数值向量.

1.2 十字型卷积模块

深度可分离卷积是卷积神经网络的一个流行变体, 在神经网络基础模块的构建中得到了应用. 比如 Mobile-

Net[26,27]、ConvNeXt[28]、ShuffleNet[29,30]等. 除了上述基于纯卷积神经网络的架构之外, 还有新型的架构是基于视觉转换器, 例如 MobileVit[31], 结合深度可分离卷积和改进的注意力机制降低纯 Transformer 的计算复杂度. 然而, 这些模型存在深度可分离卷积增加内存访问和重复的冗余计算的问题.

具体来说, 如图 3 所示, 深度可分离卷积对于输入 $I \in \mathbb{R}^{c\times h\times w}$, 需应用 c 个滤波器 $W \in \mathbb{R}^{k\times k}$, 每个滤波器在一个通道上进行空间滑动, 最后输出 $O \in \mathbb{R}^{c\times h\times w}$, 得到的 FLOPs 为 $h\times w\times k^2\times c$. 与图 4 普通卷积使用相同参数的滤波器后的 FLOPs 为 $h\times w\times k^2\times c$ 相比, 减少了 FLOPs. 但用深度可分离卷积来代替普通卷积, 会导致精度的下降. 因此在实际应用中, 通过将深度可分离卷积的通道数增加到 $c'(c'>c)$ 以补偿损失的准确率. 例如, ConvNeXt 将深度可分离卷积的通道数扩大到原来的 4 倍, 即 $c't=4c$. 此时的 FLOPs 增加到了 $h\times w\times k^2\times c'$. 对于 FasterNet 所提出的部分卷积, 如图 5 所示, 它只对输入通道的一部分采用普通卷积来进行特征提取, 其余部分采用恒等映射. 对于连续的 c_p 通道, 其 FLOPs 为:

$$h\times w\times k^2\times c_p^2 \tag{1}$$

其中, $c'=4c$, r $(r>1)$ 为倍率因子. 如本文中 $r=4$, 其 FLOPs 仅为普通卷积的 1/16. 余下的 $c-c_p$ 个通道的 FLOPs 计算量如式 (2) 所示:

$$h\times w\times c\times\left(c-c_p\right) \tag{2}$$

总的 FLOPs 如式 (3) 所示:

$$h\times w\times c\times\left(k^2\times c_p^2+\left(c-c_p\right)\right) \tag{3}$$

为了更有效地提取到遥感场景的纹理特征, 将剩余的通道采用点卷积, 在部分卷积的基础之上, 扩大感受野, 图 6(a) 为本文所采用的十字型卷积模块流程图. 由于文中采用的卷积提取模块由一种点卷积加部分卷积再加点卷积构成, 整个形式呈现出“十字型”.

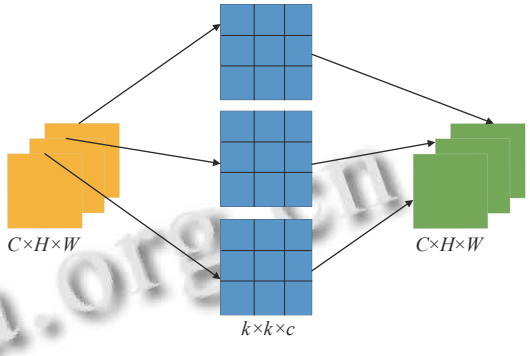


图 3 深度可分离卷积

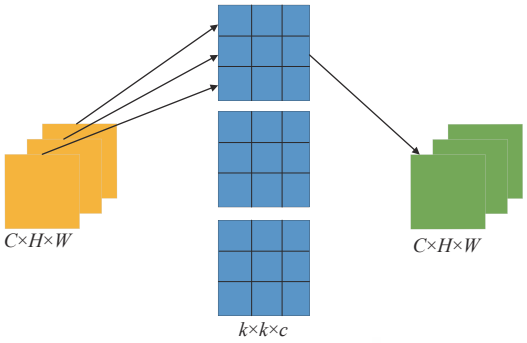


图 4 普通卷积

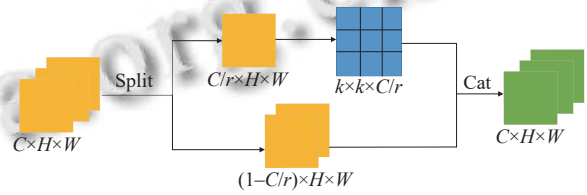
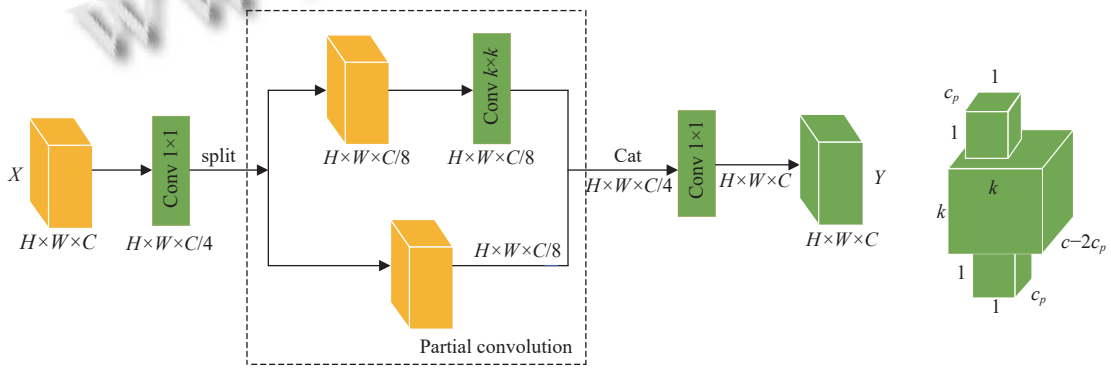


图 5 部分卷积



(a) 流程图 (b) 结构图

图 6 本文十字型卷积模块流程图

为确定本文十字型卷积模块中部分卷积中参数 k 的大小, 选择常用的 1×1、3×3、5×5、7×7 和 9×9 共 5 种不同大小的卷积核大小进行实验, 随机选取 10% 的 NWPU-RESISC45 为训练样本, 实验结果见表 1.

表 1 十字形卷积模块中参数 k 在数据集上的分类结果 (%)

Kernel size	OA
1×1	73.6±0.19
3×3	75.8±0.20
5×5	77.3±0.28
7×7	75.1±0.13
9×9	74.9±0.16

由表 1 所知, 十字形卷积模块在参数 k 选择 5×5 时取得了最好的效果, 达到了 77.3%, 因此本文十字型卷积中的卷积核 k 大小为 5×5.

如图 7 所示残差模块, 受残差模块[26]的启发, 本文模型先用 1×1 卷积进行降维, 然后进入部分卷积模块, 最后采用 1×1 卷积进行升维, 3 个过程以及总共的 FLOPs 如式 (4)–式 (7) 所示:

$$h\times w\times c_p^2 \tag{4}$$

$$h\times w\times c_p\times\left(k^2\times\frac{c_p^2}{4}+\frac{c_p}{2}\right) \tag{5}$$

$$h\times w\times c^2 \tag{6}$$

$$h\times w\times c^2\times\left(\frac{k^2c}{4r^3}+\frac{1}{2r^2}\right) \tag{7}$$

根据式 (3) 得式 (8):

$$h\times w\times c^2\times\left(\frac{k^2c}{r^2}+r-1\right) \tag{8}$$

显然, 对比式 (7) 和式 (8), 本文的十字型卷积模块运行后的 FLOPs 比部分卷积更少.

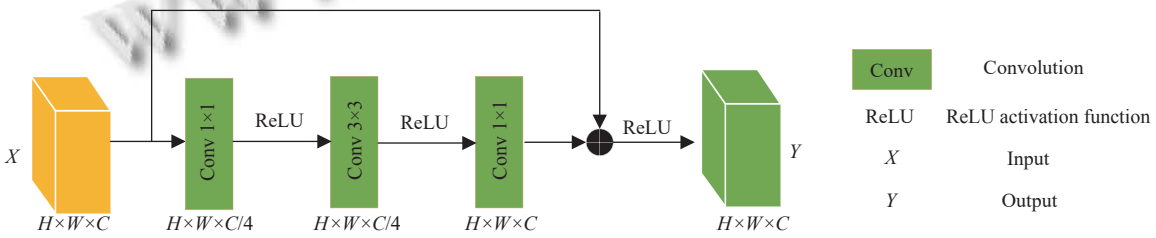


图 7 残差模块

1.3 双分支注意力模块

图 8 为坐标注意力模块的结构示意图. 坐标注意力模块不仅考虑了每个通道的重要性, 而且考虑了空间位置信息, 通过嵌入坐标信息和生成坐标注意力两个步骤, 编码感兴趣区域的精确位置信息通道关系和依赖关系.

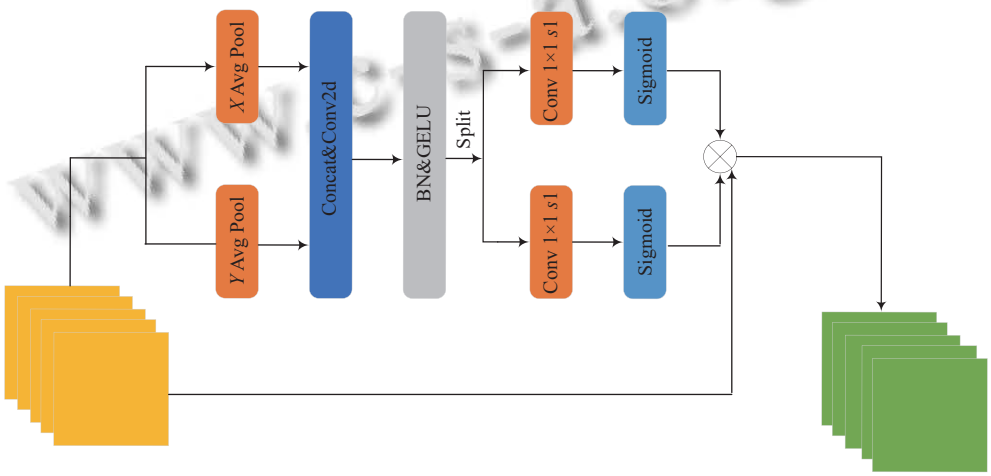


图 8 CA 结构

本文着重考虑通道关系, 在坐标注意力的基础上进一步考虑通道特征与感兴趣区域的依赖关系, 增加另一条嵌入通道处的分支, 最终形成了具有通道与 CA 的双分支混合注意力模型, 其结构如图 9 所示.

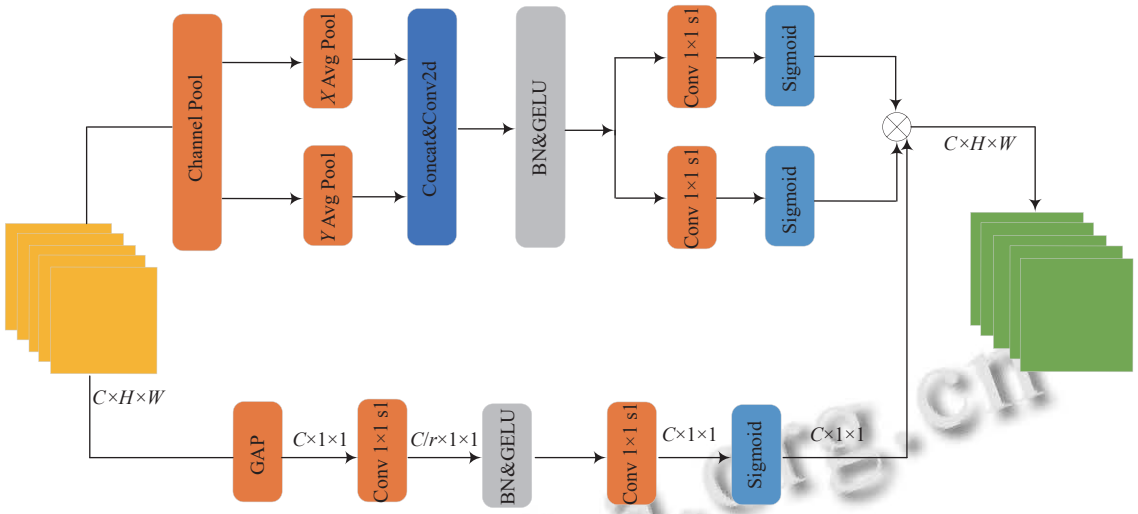


图 9 本文嵌入通道与 CA 的双分支注意力结构图

全局平均池化常用于通道注意力对空间信息的编码, 其流程是将全局空间信息进行压缩, 保留其重要的位置信息. 为了定位到注意力模块在空间上捕获的位置信息, 将全局池化分解成如式 (9) 所示的一对一维度的特征编码操作.

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \tag{9}$$

具体来说, 给定输入X, 使用了两个维度的池化核 $(H, 1)$ 和 $(1, W)$, 分别沿着水平坐标和垂直坐标对每个通道进行编码. 在高度 h 处的第 c 通道的输出可以表示为式 (10). 类似地, 宽度为 w 处的第 c 个通道的输出可以表示为式 (11).

$$z_c^h(h) = \frac{1}{W} \sum_{0 \leqslant i < \mathrm{W}} x_c(h, i) \tag{10}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \leqslant j < H} x_c(j, w) \tag{11}$$

通过这两种变换方式使注意力模块能够在捕获到物体之间在一个空间方向上的依赖关系的同时, 保留另一空间方向上的位置信息, 有助于网络能更准确地定位到图像上感兴趣的区域.

给定由上述输入 X, 通过式 (10) 和式 (11) 产生的特征映射, 将它们进行拼接, 形成聚合特征图, 然后输入到 1×1 的卷积函数 F_1, 如式 (12) 所示. 在另一条分支上的通道处的特征输入到 1×1 的卷积函数 F_1 进行压缩, 如式 (13) 所示.

$$f = \delta(F_1([z^h, z^w])) \tag{12}$$

$$v = \delta(F_1^{c/r}(z_c)) \tag{13}$$

其中, $[\cdot, \cdot]$表示表示沿空间方向的拼接操作, δ是非线性激活函数, $f \in \mathbb{R}^{C/r \times W}$ 表示在水平和垂直方向的空间信息编码后的中间特征图, r 是控制块缩放比. 然后将 f 沿着水平和垂直两个空间维度分裂成 $f^h \in \mathbb{R}^{C/r \times W}$ 和 $f^h \in \mathbb{R}^{C/r \times W}$. 利用 1×1 的卷积 F_h、F_w 和 F_c, 分别将 f^h、f^w 和 v 变换为与输入 X 相同通道数的张量, 如式 (14)–式 (16) 所示.

$$g^h = \sigma(F_h(f^h)) \tag{14}$$

$$g^w = \sigma(F_w(f^w)) \tag{15}$$

$$g^c = \sigma(F_c(v)) \tag{16}$$

记σ为激活函数, 然后将 g^h、g^w 和 g^c 分别作为水平、垂直方向以及通道处的的注意力权重.

混合注意力块 Y 的输出如式 (17) 所示:

$$y_c(i, j) = g^c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{17}$$

不同于 CA 模块仅考虑了物体的空间信息和依赖关系, 本文的双分支注意力加强了物体的空间信息与物体之间的依赖关系. 综上所述, 这种编码过程使得双分支注意力模块能够更精准地定位到物体的位置以及表达出物体之间的依赖关系, 从而能帮助更好地完成计算机视觉任务.

2 模型训练与结果分析

2.1 实验环境与配置

进行本文实验的训练与测试环境均为 Windows 10 操作系统, 使用 PyTorch 深度学习框架完成整个模型

的训练与测试过程, 调用 timm 库完成对比实验, 实验环境的具体参数如表 2 所示.

表 2 实验环境配置表

实验环境	配置
CPU	Intel(R) Core(TM) i5-12400F
GPU	NVIDIA GeForce RTX 3060
Memory	12 GB
Python version	Python 3.9.13
Deep learning framework	PyTorch 1.14.0

2.2 实验数据集

为检验本文方法的有效性和泛化能力, 在公开的高分辨率遥感数据集 EuroSAT[32]、NWPU-RESISC45[33]以及 VArcGIS[34]上开展对比与消融实验; 其中, EuroSAT 是遥感场景领域常用的数据集; VArcGIS 与 EuroSAT 相比, 场景类别和地物目标更加丰富, 具有更大的挑战性; NWPU-RESISC45 遥感图像数据集拥有 45 个类别, 场景类别非常丰富, 类间的相似度较高, 是目前遥感场景领域最具挑战性的数据集; 3 个数据集的特征与部分场景实例如表 3 和图 10 所示.

表 3 实验数据集特征

Dataset	Classes	Number	Total	Resolution (m)	Image size	Year
EuroSAT	10	2 700	2 700	10	64×64	2019
NWPU-RESISC45	45	700	31 500	0.2–30	256×256	2016
VArcGIS	38	1 504–1 904	59 071	0.07–19.11	256×256	2021

图 10 3 种数据集部分场景实例

2.3 训练参数与评价指标

本文选择带有权重的 Adam 算法, 与传统的 Adam 相比, 加有惩罚项的 Adam 在更准确地控制权重衰减强度的同时避免了梯度计算的影响, 可以减少过拟合的风险, 避免陷入局部最优, 从而使得模型更容易收敛. 学习率衰减方式为余弦退火, 输入网络的批量大小为 16, 模型的迭代次数为 300, 考虑到模型训练刚开始时, 受到学习率的影响导致模型的不稳定, 因此选择预热学习率的方式, 模型的训练参数如表 4 所示.

表 4 模型训练参数

参数名称	参数值
Epochs	300
warmup_epochs	30
batch_size	16
learning_rate	0.000 1
Optimizer	AdamW
weight_decay	0.000 5

为了评价模型的有效, 实验采用总体精度 (overall accuracy, *OA*) 和标准差作为评价指标. 总体精度定义为分类正确的样本数占总体样本的比率. 计算公式如下:

$$OA = \frac{T}{T+F} \tag{18}$$

其中, T 为正确样本的个数, F 为错误样本的个数.

标准差的定义是 N 个数方差的算数平方根, 用于描述数据分布离散程度的一种统计量, 计算公式如下:

$$\bar{B} = \frac{1}{N}\sum_{i=1}^{N} B_i \tag{19}$$

$$\tau = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(B_i - \bar{B}\right)^2} \tag{20}$$

其中, B_i 表示第 i 次实验的精度, $\bar{B}$表示 N 次实验的平

均精度. 标准差越小表示越稳定.

2.4 对比实验

2.4.1 模型性能对比实验

为验证模型的有效性, 实验的模型参数均保持一致. 其次, 为了避免实验数据的差异带来的影响. NWPU-RESISC45 数据集与 EuroSAT 数据集的训练集、测试集和验证集分别为 50%、30% 和 20%; VArcGIS 数据集与文献[35]保持一致, 随机选取 10% 和 20% 的场景为训练集, 其余为测试集. 所对比的模型包括现有的遥感场景分类模型 GoogLeNet、ConvNeXt、ResNet50、SDCASA[36]、SCRSISC[37]、Vision Transformer[35]、CNN+Transformer+CA[19]. 其中, SDCASA 是一种基于自蒸馏级联注意力机制的遥感场景分类方法; SCRSISC 是一种基于 ResNet 监督对比学习的遥感场景分类方法; Vision Transformer 是视觉转换器在遥感场景分类领域的应用; CNN+Transformer+CA 是一种基于 CNN 与 Transformer 结构的分类网络. 实验结果如表 5 所示.

由表 5 可知, 本文模型在 NWPU-RESISC45 数据集上的平均精度为 96.12%, 在 EuroSAT 数据集上的平均精度为 98.64%, 在训练占比为 10% 和 20% 的 VArcGIS 数据集上平均精度分别为 95.42% 和 97.84%, 均超过了现有的主流模型.

表 5 不同模型在 3 种数据集上的实验结果对比

Model	Parameters (M)	NWPU-RESISC45 (%)	EuroSAT (%)	VArcGIS (%)	
				10%	20%
GoogLeNet	54.4	78.48±0.26	—	77.33±0.57	86.79±0.47
ConvNeXt	28.579	84.79±0.41	96.62±0.21	85.28±0.36	91.64±0.22
ResNet50	25.557	94.89±0.15	98.28±0.06	92.90±0.15	96.57±0.12
SDCASA	—	—	91.17±0.04	—	—
SCRSISC	—	94.73±0.19	—	—	—
Vision Transformer	59.738	91.57±0.29	97.45±0.15	89.68±0.30	94.23±0.26
CNN+Transformer+CA	20.04	96.00±0.07	—	91.27±0.02	95.01±0.14
Ours	19.609	**96.12±0.09**	**98.64±0.12**	**95.40±0.15**	**97.86±0.08**

2.4.2 热力图对比实验

为验证模型定位重要特征的能力, 对生成模型的特征图上设计热力图实验, 如图 11 所示, 其中图 11(a)–图 11(c) 分别为河流、海滩、人行横道, 本模型能够精准定位到图像中最感兴趣的区域.

2.4.3 收敛性对比实验

为验证本文方法的收敛性, 使用本文方法与 FasterNet 在分别在 VArcGIS 数据集上进行实验, 并分别绘制训练收敛曲线与测试收敛曲线对比, 收敛曲线图如图 12 和图 13 所示.

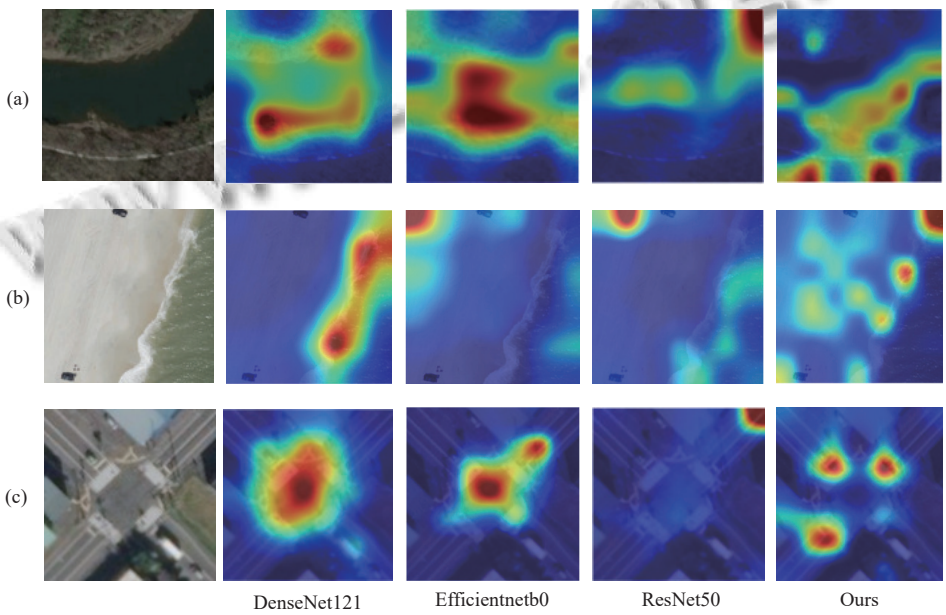
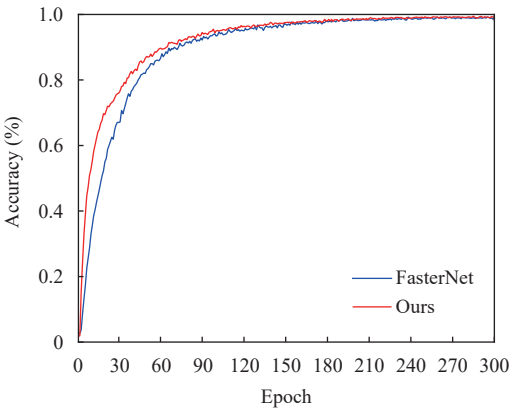
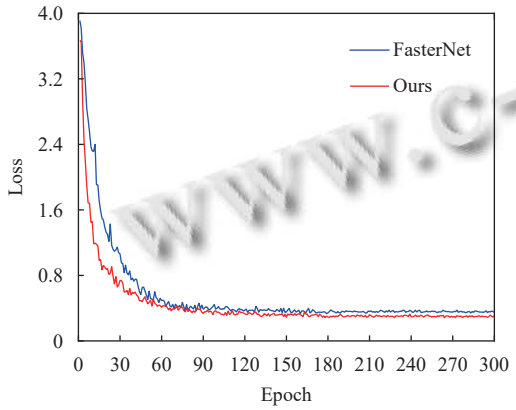


图 11 本文模型与其他模型的热力图对比



(a) 准确率曲线



(b) 损失收敛曲线

图 12 VArcGIS 数据集上收敛性对比图

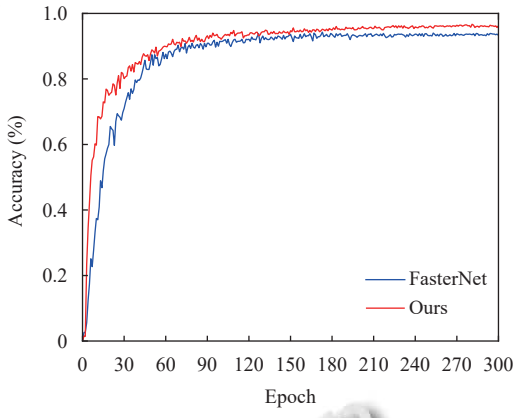
由图 12 可知, 在训练实验过程中, 本文模型的准确率均约为 99%, 损失约为 3.7. 从训练实验过程可以看出本文模型能更快地趋于收敛.

由图 13 所知, 在测试实验过程中, 本文模型的收敛时的准确率约为 95%, 损失值约为 3, 模型表现略优于 FasterNet 模型.

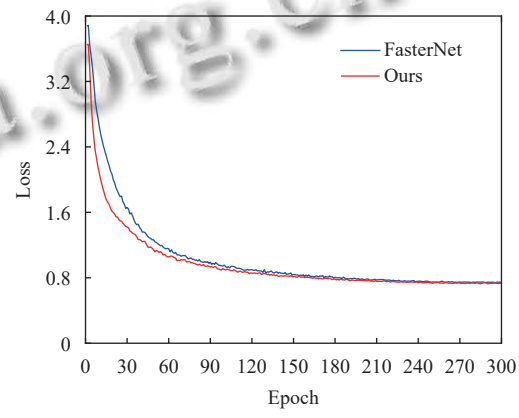
2.5 消融实验

为了验证十字型卷积模块捕捉底层特征的能力和双分支注意力定位物体位置信息的效果, 本节实验从 50% 占比的 NWPU-RESISC45 和 EuroSAT 数据集以及 10% 以及 20% 占比的 VArcGIS 数据集作为训练集, 其余为测试集, 进行 300 次迭代. 其中①为去掉 FasterNet block 的 FasterNet 模型. 在实验环境、图像预处理方式和网络超参数等条件相同的情况下进行实验, 结果如表 6 所示.

由表 6 可知, 单独加入十字型卷积模块和双分支注意力的方法在精度上均得到了提升. 十字型卷积模块主要用于降低内存访问次数, 提高运行速度, 与 FasterNet 模型相比, 平均一张图像的推理时间少了约 0.4 ms. 双分支注意力加强了特征的表达能力, 实验结果表明, 与 CA 注意力相比, 在 NWPU-RESISC45 数据集上的精度提高了 0.81%, EuroSAT 数据集上的精度提高了 0.97%, 10% 和 20% 训练占比的 VArcGIS 数据集上的精度分别提高了 0.98% 和 1.69%. 这验证了双分支注意力定位感兴趣区域和建立通道之间的依赖关系的能力. 综合来看, 同时加入十字型卷积和双分支注意力的模型与基础模型相比, 在 NWPU-RESISC45 数据集上的精度提高了 5.18%, 在 EuroSAT 数据集上的精度提高了 1.65%, 10% 和 20% 训练占比的 VArcGIS 数据集上的精度上分别提高了 6.94% 和 3.39%. 相比于单独使用一个方法提升更多, 由此说明两个方法均可以提升模型的性能.



(a) 准确率曲线



(b) 损失收敛曲线

图 13 VArcGIS 数据集上测试收敛性对比图

2.6 结果与分析

混淆矩阵常用于机器学习和统计学中用于评估分类模型性能的一种方法. 本文使用混淆矩阵来可视化最终的分类结果, 如图 14–图 17 所示, 分别表示模型在 EuroSAT 数据集、NWPU-RESISC45 数据集、10%

的训练占比的 VArcGIS 数据集和 20% 的训练占比的 VArcGIS 数据集的分类结果. 可以看出, 本文设计的模型分类性能较好, 在 3 个数据集上均取得了 95% 以上的准确率.

表 6　消融实验结果

Model	NWPU-RESISC45 (%)	EuroSAT (%)	VArcGIS (10%)(%)	VArcGIS (20%)(%)	推理时间 (ms/img)
①(baseline)	90.90	96.99	88.46	94.47	**1.6**
FasterNet	94.06	97.81	94.08	97.22	2.3
①+十字型卷积	94.63	97.81	94.02	97.08	1.9
①+双分支注意力	94.94	98.44	94.08	97.16	7.3
①+CA	94.13	97.47	93.10	95.47	5.1
①+ECA	94.06	97.22	92.18	95.18	4.0
Ours	**96.12**	**98.64**	**95.42**	**97.87**	7.1

图 14 是 EuroSAT 验证集的混淆矩阵, 数字 0–9 分别代表年作物、森林、草本植物、公路、工业区、牧场、农作物、住宅区、河流和湖泊. 由图 14 可知, 10 个类别的分类精度均达到了 95% 以上, 分类精度最低的类别是住宅区, 达到了 95.84%. 总体而言, 各个类别的分类精度比较平衡, 这是因为场景类别比较少.

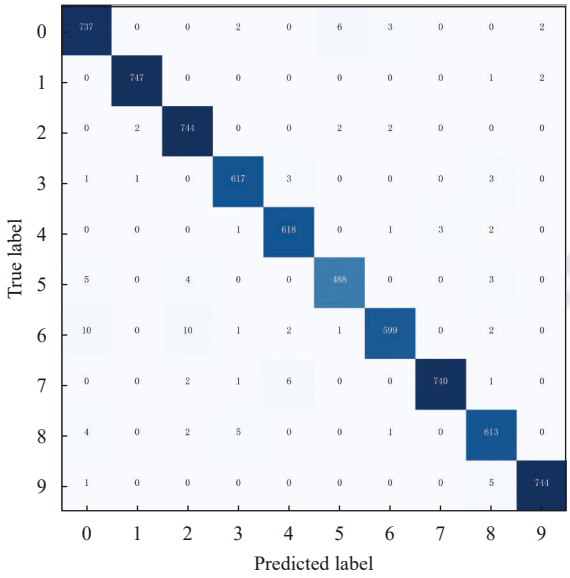


图 14　本文模型在 EuroSAT 验证集上的混淆矩阵

图 15 为 NWPU-RESISC45 验证集的混淆矩阵, 数字 0–44 分别代表飞机、机场、棒球场、篮球场、海滩、桥梁、灌木丛、教堂、圆形耕地、云、商业区、密度住宅区、沙漠、林地、公路、高尔夫球场、田径场、港口、工业区、交叉路口、岛屿、湖泊、草地、中型住宅区、活动房区、山脉、立交桥、宫殿、停车场、铁路、火车站、矩形耕地、河流、环岛、跑道、海冰、船舶、雪山、稀疏住宅区、体育场、储罐、网球场、梯田、火力发电站和湿地. 由图 5 可知, 36 个类别的分类精度达到了 95% 及以上, 42 个类别的分类精度达到了 90% 及以上; 其中, 教堂和宫殿是容易混淆的两个类别, 12.14% 的教堂被错分为宫殿, 7.14% 的宫殿被错分为教堂, 这是因为这两类具有结构特征相似度高, 建筑风格类似的特点.

图 16 和图 17 分别是不同训练比的 VArcGIS 验证集的混淆矩阵, 数字 0–37 分别代表飞机、棒球场、篮球场、海滩、桥、墓地、灌木丛、农场、封闭道路、海边豪宅、人行横道、密集住宅、客轮码头、足球场、森林、高速公路、高尔夫球场、港口、路口、移动家庭公园、疗养院、油气田、油井、立交桥、停车场、车位、铁路、河流、跑道、跑道标记、船场、太阳能板、稀疏住宅、储水箱、游泳池、网球场、变压器和污水处理厂.

由图 16 可知, 33 个类别的分类准确率达到了 90% 以上, 22 类场景的识别准确率达到了 95% 以上; 其中, 篮球场和网球场是容易混淆的两个类别, 5.43% 的篮球场被错分为网球场, 6.47% 的网球场被错分为篮球场, 这是因为这两类球场的背景特征极为相似, 球场建筑风格较为相同. 由图 17 可知, 38 个类别的分类精度达到了 90% 及以上, 34 个类别的分类精度达到了 95% 及以上; 分类精度最低的类别是篮球场, 达到了 93.07%; 由图 16 和图 17 可知, 扩大训练集的比例可以有效提高各个类别的分类精度, 有效缓解分类精度不平衡、类别被错分的问题.

3　结束语

本文针对遥感场景分类方法运行效率低和遥感场景图像类别多且相似度高导致的难分类问题, 本文从控制卷积提取特征过程中的运算量、模型定位感兴趣区域和计算不同通道处的权重关系出发, 提出残差模块优化部分卷积的策略提取特征以降低卷积运算量, 提高运行效率, 减少运行时间, 采用双分支注意力模块, 一条分支定位到重要的特征, 另一条分支建立通道间的依赖关系. 在 NWPU-RESISC45、EuroSAT、VArcGIS 这 3 种遥感场景数据集上开展对比实验以及一系列的消融实验, 分析实验结果, 证明了本文模型的有效性.

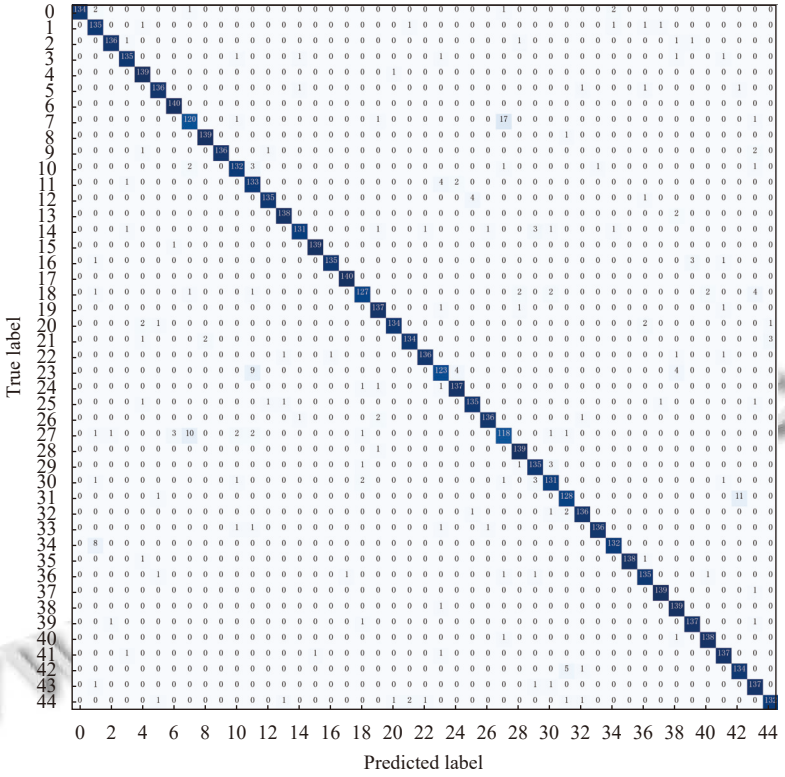


图 15　本文模型在 NWPU-RESISC45 验证集上的混淆矩阵

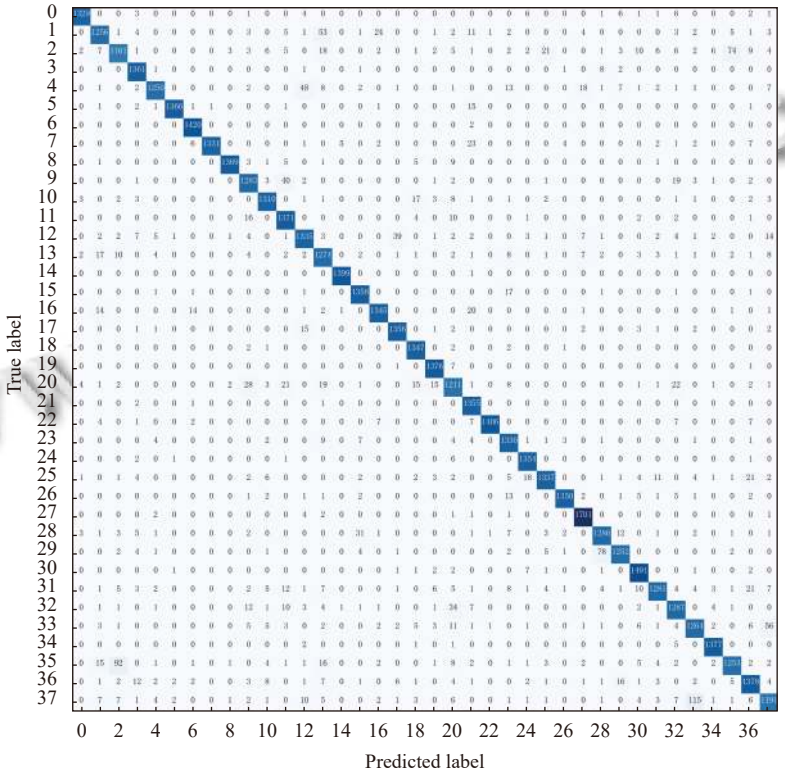


图 16　本文模型在 VArcGIS (10%) 验证集上的混淆矩阵

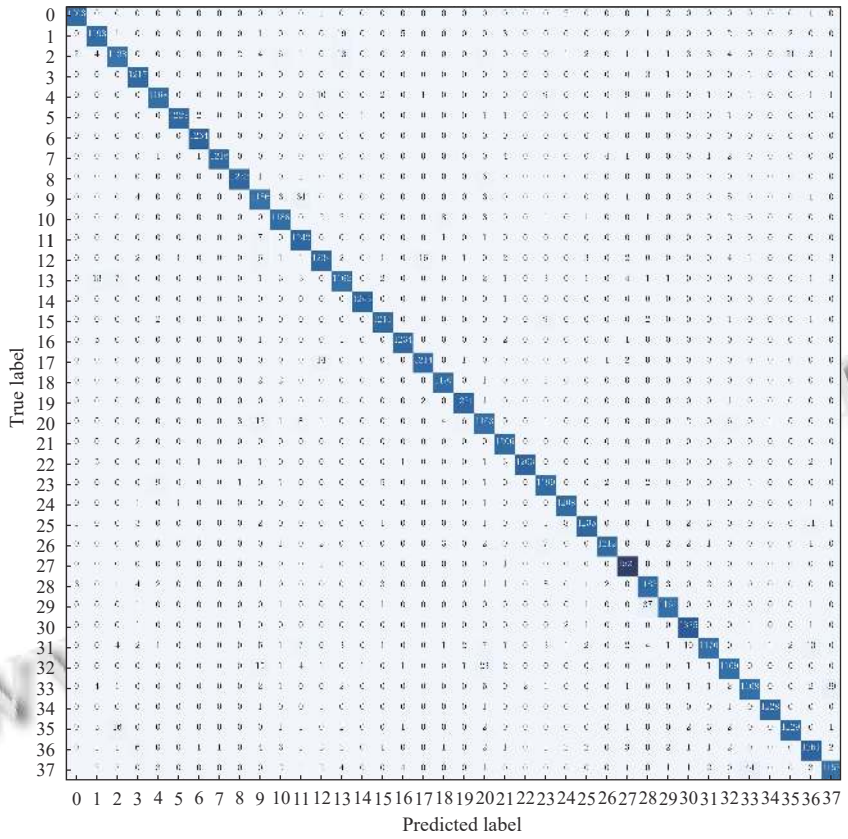


图 17 本文模型在 VArcGIS (20%) 验证集上的混淆矩阵

然而, 本文方法依然存在很多不足, 从消融实验可以得知, 虽然双分支注意力机制拥有一定的定位重要特征的能力, 但推理时间较长, 有很大的运行代价, 导致本模型优化运行速度的能力有限. 从热力图对比实验可知, 本文模型虽然能够定位到重要的特征, 但是不够集中, 依然无法有效区分相似高的特征的场景类. 因此, 后续将针对区分相似度高的场景类的问题和设计更快速地定位特征的模型是下一步重点研究方向.

参考文献

1 Máttyus G, Luo WJ, Urtasun R. DeepRoadMapper: Extracting road topology from aerial images. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 3458–3466.

2 Martha TR, Kerle N, van Westen CJ, *et al*. Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis. IEEE Transactions on Geoscience and Remote Sensing, 2011, 49(12): 4928–4943. [doi: 10.1109/TGRS.2011.2151866]

3 Longbotham N, Chaapel C, Bleiler L, *et al*. Very high resolution multiangle urban classification analysis. IEEE Transactions on Geoscience and Remote Sensing, 2012, 50(4): 1155–1170. [doi: 10.1109/TGRS.2011.2165548]

4 Kim M, Madden M, Warner TA. Forest type mapping using object-specific texture measures from multispectral Ikonos imagery: Segmentation quality and image classification issues. Photogrammetric Engineering & Remote Sensing, 2009, 75(7): 819–829.

5 Ghosh R, Jia XW, Kumar V. Land cover mapping in limited labels scenario: A survey. arXiv:2103.02429, 2021.

6 骆剑承, 王钦敏, 马江洪, 等. 遥感图像最大似然分类方法的 EM 改进算法. 测绘学报, 2002, 31(3): 234–239. [doi: 10.3321/j.issn:1001-1595.2002.03.010]

7 朱建华, 刘政凯, 俞能海. 一种多光谱遥感图象的自适应最小距离分类方法. 中国图象图形学报, 2000, 5(1): 21–24. [doi: 10.3969/j.issn.1006-8961.2000.01.005]

8 Foody GM, Mathur A. A relative evaluation of multiclass image classification by support vector machines. IEEE Transactions on Geoscience and Remote Sensing, 2004, 42(6): 1335–1343. [doi: 10.1109/TGRS.2004.827257]

9 Chen YS, Lin ZH, Zhao X, *et al*. Deep learning-based classification of hyperspectral data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014, 7(6): 2094–2107. [doi: 10.1109/JSTARS.2014.2329330]

10 Avramović A, Risojević V. Block-based semantic classification of high-resolution multispectral aerial images. Signal, Image and Video Processing, 2016, 10(1): 75–84.

[doi: 10.1007/s11760-014-0704-x]
11 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1–9.
12 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.
13 Hu W, Huang YY, Wei L, *et al*. Deep convolutional neural networks for hyperspectral image classification. Journal of Sensors, 2015, 2015: 258619.
14 Duan YP, Liu F, Jiao LC, *et al*. SAR image segmentation based on convolutional-wavelet neural network and Markov random field. Pattern Recognition, 2017, 64: 255–267. [doi: 10.1016/j.patcog.2016.11.015]
15 Wang GL, Fan B, Xiang SM, *et al*. Aggregating rich hierarchical features for scene classification in remote sensing imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(9): 4104–4115. [doi: 10.1109/JSTARS.2017.2705419]
16 Xie J, He NJ, Fang LY, *et al*. Scale-free convolutional neural network for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(9): 6916–6928. [doi: 10.1109/TGRS.2019.2909695]
17 Liu YF, Zhong YF, Qin QQ. Scene classification based on multiscale convolutional neural network. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(12): 7109–7121. [doi: 10.1109/TGRS.2018.2848473]
18 Roy SK, Deria A, Hong DF, *et al*. Multimodal fusion Transformer for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5515620.
19 金传, 童常青. 融合 CNN 与 Transformer 结构的遥感图像分类方法. 激光与光电子学进展, 2023, 60(20): 2028006.
20 Touvron H, Bojanowski P, Caron M, *et al*. ResMLP: Feedforward networks for image classification with data-efficient training. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 5314–5321.
21 Zhu HR, Chen BY, Yang C. Understanding why ViT trains badly on small datasets: An intuitive perspective. arXiv:2302.03751, 2023.
22 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.
23 Chen JR, Kao SH, He H, *et al*. Run, don't walk: Chasing higher FLOPs for faster neural networks. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 12021–12031.
24 Hou QB, Zhou DQ, Feng JS. Coordinate attention for efficient mobile network design. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13708–13717.
25 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical Vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002.
26 Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017.
27 Sandler M, Howard A, Zhu ML, *et al*. MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4510–4520.
28 Liu Z, Mao HZ, Wu CY, *et al*. A ConvNet for the 2020s. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11966–11976.
29 Ma NN, Zhang XY, Zheng HT, *et al*. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 122–138.
30 Zhang XY, Zhou XY, Lin MX, *et al*. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6848–6856.
31 Mehta S, Rastegari M. MobileViT: Light-weight, general-purpose, and mobile-friendly Vision Transformer. Proceedings of the 10th International Conference on Learning Representations. OpenReviw.net, 2022.
32 Helber P, Bischke B, Dengel A, *et al*. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(7): 2217–2226. [doi: 10.1109/JSTARS.2019.2918242]
33 Cheng G, Han JW, Lu XQ. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 2017, 105(10): 1865–1883. [doi: 10.1109/JPROC.2017.2675998]
34 Hou DY, Miao ZL, Xing HQ, *et al*. V-RSIR: An open access Web-based image annotation tool for remote sensing image retrieval. IEEE Access, 2019, 7: 83852–83862. [doi: 10.1109/ACCESS.2019.2924933]
35 Bazi Y, Bashmal L, Rahhal MMA, *et al*. Vision Transformers for remote sensing image classification. Remote Sensing, 2021, 13(3): 516. [doi: 10.3390/rs13030516]
36 宋冠武, 陈知明, 李建军. 基于 ResNet-50 的级联注意力遥感图像分类. 广西师范大学学报 (自然科学版), 2023, 41(6): 80–91.
37 郭东恩. 基于深度学习的遥感图像场景分类研究 [博士学位论文]. 重庆: 重庆邮电大学, 2021.

(校对责编: 张重毅)