高分辨率遥感图像场景分类研究进展

李智[1,2]，高连如[1]，郑珂[3]，倪丽[1]

1. 中国科学院空天信息创新研究院 计算光学成像技术重点实验室, 北京 100094;
2. 中国科学院大学 资源与环境学院, 北京 100049;
3. 聊城大学 地理与环境学院, 聊城 252000

摘 要：高分辨率遥感图像场景分类作为遥感图像智能解译中的基本任务，在土地监测、环境保护等诸多领域都拥有着广泛且重要的作用。随着深度学习技术、大数据、大模型的快速发展，遥感图像场景分类取得了一系列全新的成果，并面临新的机遇与挑战。本文对遥感图像场景分类领域中的深度学习方法进行系统性研究，包括卷积神经网络、Vision Transformer、生成对抗网络等模型架构，并总结了从场景分类概念提出以来至今的具有代表性的24个数据集，基于其中的基准数据集评估了一系列经典的场景分类方法，最后讨论了遥感图像场景分类面临的主要挑战和技术发展趋势。

关键词：高分辨率遥感图像，图像分类，场景分类，深度学习

中图分类号：TP701/P2

引用格式：李智,高连如,郑珂,倪丽.2024.高分辨率遥感图像场景分类研究进展.遥感学报,28(11):2739-2760

Li Z, Gao L R, Zheng K and Ni L. 2024. Research progress of high-resolution remote sensing image scene classification. National Remote Sensing Bulletin, 28(11):2739-2760[DOI:10.11834/jrs.20243519]

1 引 言

随着遥感技术的快速发展，遥感卫星的分辨率越来越高、谱段越来越多、重访周期越来越短，人们可以从遥感图像中获得更多有用的数据和信息。遥感大数据、遥感基础模型、智慧城市等概念在近些年相继提出，海量遥感数据应用对遥感图像信息智能提取技术提出了更高的要求（张兵，2018）。遥感图像分类方法作为遥感图像信息智能提取技术中的重要一环，在土地利用土地覆盖、国土资源调查、自然灾害观测、农业估产、林业保护等领域都拥有着重要的实际应用意义（赵理君和唐娉，2016；Cheng等，2017a）。

在较早的研究中，遥感图像的分类方法主要是基于像素和基于对象块进行处理的。随着近年来遥感图像空间分辨率的不断提高，遥感图像中可能蕴含着不同的对象类信息，小区域的分割不足以正确地将图像的完整语义信息表征出来，因此为了对遥感图像进行正确的分类，有必要了解遥感图像的全局语义信息。

在这种背景下，遥感图像场景分类被提出。遥感图像中的场景分类的目的是从整体上对每个给定的遥感图像进行语义类别的判定，对提取到的特征信息进行高层的语义汇总和分析，将感兴趣的场景依特征赋予不同类别的标签（Xia等，2017）。与自然图像相比，虽然包含颜色、纹理和形状等特征信息，但由于其俯视视角成像带来的复杂场景内容以及低分辨率带来的弱纹理和颜色信息使得遥感场景分类的困难远大于自然图像识别。作为遥感应用的技术手段之一，遥感图像场景分类技术对实际应用技术的发展具有重要意义。

遥感图像场景分类的传统方法主要是针对中低级特征，即提取图像的颜色、纹理和形状等信息。而深度学习的分类方法主要针对图像的高级语义信息，即提取了图像的高级抽象信息（钱晓亮 等，2018）。仅使用自然图像处理方法，难以进

收稿日期：2023-12-11；**预印本：**2024-02-29

基金项目：国家重点研发计划(编号：2021YFB3900502)

第一作者简介：李智,研究方向为高光谱图像处理和遥感信息智能提取。E-mail：lizhi21@mails.ucas.ac.cn

通信作者简介：高连如,研究方向为高光谱图像处理。E-mail：gaolr@aircas.ac.cn

一步提高遥感图像场景分类的准确率，如何能够高效地提高场景分类精度是一个具有重要意义的科学问题。

随着深度学习技术的兴起，从2015年开始，深度学习方法逐渐被引入到遥感图像场景分类任务中。卷积神经网络（CNN）结构由于其强大的特征提取能力，以及在自然图像分类中良好的性能，受到了遥感领域的关注（Wang等，2023a，2023d；Gao等，2023）。从对经典的网络结构如AlexNet、GoogLeNet、VGGNet等进行迁移，到如今各种网络结构层出不穷，CNN在遥感图像场景分类中的效果提升迅速，并逐步发展出了适用于遥感场景的系列分类模型方法。而近些年引入ViT（Transformer和Vision Transformer）结构概念进行遥感图像场景分类，同样为这一领域构建了新的方法框架。

由于监督学习需要充足的训练样本支撑，在应用场景上往往会受到一定的限制。而无监督学习可以在没有先验类别知识的情况下，根据图像本身的特点特征来让网络模型学习。当无监督分类与深度学习结合，自编码器方法和生成对抗网络（GAN）先后应用于场景分类任务，同样表现出来了良好的性能。

由于深度学习方法需要大量数据进行驱动，遥感图像场景分类方法的不断发展为其数据集提出了更大的挑战。在近些年中，数据集的样本数量、分类体系和制作方法取得了长足的进步。如今的遥感图像场景分类数据集逐步完善，且越来越贴近于真实应用场景中所需要的类别（欧阳淑冰 等，2022）。数据集也从单纯的检验分类方法精度向提供大规模预训练样本库方向发展。

经过13年的发展，国内外已经有多篇关于遥感图像场景分类的研究综述（Xia等，2017；Cheng等，2017a，2020）。然而，随着近些年来遥感大数据的兴起，场景分类面临新的形势：

（1）随着深度学习技术的不断演进，尤其是卷积神经网络和Transformer的广泛运用，遥感图像场景分类取得了显著的进展。本文有针对性总结神经网络及Transformer模型在场景分类中的相关进展，并对国内外研究现状进行分析与比较，针对目前行业内可行性方向给出发展趋势的总结与展望。

（2）自监督学习作为一种不依赖标注数据的学习方法，在遥感图像场景分类领域显得尤为关键。基于自监督学习的遥感基础大模型已经成功应用于场景分类，为该领域提供了创新性的解决方案。

（3）随着遥感数据量的不断增加，遥感图像场景分类的数据集规模也在迅速扩大，相应的分类任务变得日益复杂。遥感图像场景分类数据集正朝着多源、多标签、大规模样本的方向迅速发展。

本文结合当前文献的调研结果，从分类方法和数据集两个方面进行归纳、分析和总结，并阐明未来的发展趋势。

2 文献分析

本节选择Web of Science作为主要的分析工具。文献分析基于“Remote Sensing”和“Scene Classification”两个关键词展开，共1043个结果。

本文将自2015年至2023年间的共1016篇文章绘制成图1。图1表明，近些年，人们对这一领域的关注越来越多。特别的，从2017年深度学习被广泛应用于该领域，AID和NWPU-RESISC45这两个大规模场景分类数据集公开之后，公开发表的论文数量有一个明显的跃升。

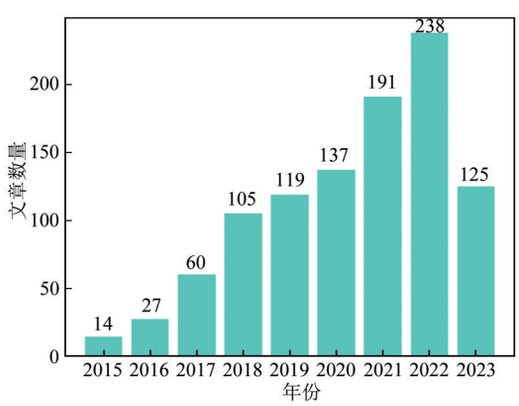

图1 2015年—2023年遥感图像场景分类领域发表论文数量

Fig. 1 Number of papers published in the field of remote sensing image scene classification from 2015 to 2023

本文对这部分论文的研究学者国籍分布进行了统计如图2所示。如图2所示，中国学者在这一领域做出了最大的贡献，远超其他国籍的学者。美国、德国、印度等紧随其后，也有一定数量的论文发表。而在世界范围内，发文量最大的机构分布为武汉大学、中国科学院、中国地质大学、西北工业大学、西安电子科技大学、中南大学等，排名在前列的以中国的高校和研究机构为主。

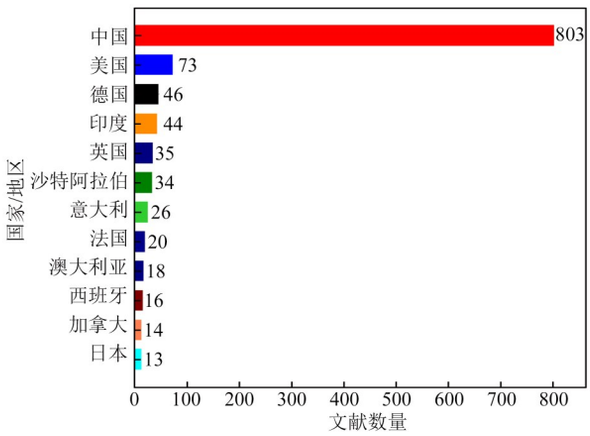

图2 各个国家/地区在遥感图像场景分类领域发表论文数量情况

Fig. 2 Number of papers published in the field of remote sensing image scene classification by country/region

我们对823篇刊登在期刊上的文章进行了统计如图3所示。其中，IEEE Transactions on Geoscience and Remote Sensing、Remote Sensing、IEEE Geoscience and Remote Sensing Letters和IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing这4个期刊包含的文章数量超过了总出版物的一半。这些期刊大多数是当下遥感领域的热门期刊。由此可见，遥感图像场景分类是遥感领域中典型的、热门的科学问题。

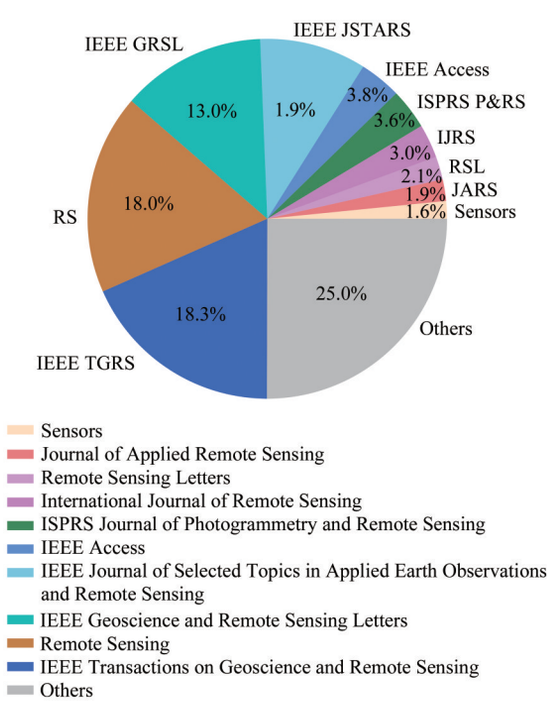

图3 遥感图像场景分类的期刊论文在不同期刊上的刊登数量情况

Fig. 3 The number of journal papers published in different journals on remote sensing image scene classification

3 遥感图像场景分类方法研究现状

本节将从5个方向阐述当前的场景分类方法，包括基于手工特征提取的分类模型、基于卷积神经网络的分类模型、基于ViT的分类模型和基于生成对抗网络的分类模型、基于遥感基础模型的方法。其中，基于CNN的模型包括以CNN为特征提取器的方法、利用CNN端到端的分类方法和将注意力机制引入CNN的分类方法。

3.1 基于手工特征提取的方法

伴随着遥感图像场景分类的数据集逐渐丰富和计算机视觉方法的不断革新进步，场景分类方法研究也不断迭代深入。从传统模式识别和机器学习方法出发，分类问题是一个从数据中提取描述数据的特征，再依据特征构造分类器进行分类的过程。而对于遥感图像场景分类而言，遥感图像所表现出来的场景语义特征是遥感图像场景分类的依据，相同类别的图像在统计意义上具有相同的整体视觉特征，因此，遥感图像的特征提取便成为了影响场景分类效果的至关重要的环节。大多数早期的遥感图像场景分类方法依赖于手工特征描述算子，例如颜色直方图（Swain和Ballard，1991）、纹理描述符（Haralick等，1973）、GIST（Oliva和Torralba，2001）、尺度不变特征变换（Lowe，2004）、定向梯度直方图（Dalal和Triggs，2005）等方法。

其中，颜色直方图、纹理描述符、GIST方法是利用图像的颜色、纹理、空间结构信息等图像特征信息描述整个图像场景的的全局特征，因此这些具有统计意义的特征可以描述图像在场景级别的语义信息，可以直接输入到分类器应用于场景分类。然而，尺度不变特征变换和定向梯度直方图特征是局部特征，用于表示局部结构和形状信息。为了描述图像的场景语义信息，需要通过某些编码方法对提取到的特征进行特征编码。常见的方法有改进Fisher核（Perronnin等，2010），局部聚合描述符（Jégou等，2012），空间金字塔匹配（Lazebnik等，2006），概率主题模型（Lienou等，2010）和视觉词袋模型（Yang和Newsam，2010），由于这些特征编码简单有效，在遥感图像场景分类领域得到了广泛的应用（Zhang等，2013）。

在现实场景中，图像的场景语义信息往往包含光谱、颜色、纹理、形状等多种特征，单一类型的特征往往不足以描述整个图像的场景语义信息，因此，将联合多个具有互补性的特征用于场景分类任务是一个可以提高分类性能的有效的策略。Zhao等（2016a）提出了一种狄利克雷衍生的多主题模型，在主题级别结合3种类型的特征进行场景分类。Zhu等（2016）提出了基于局部—全局特征的视觉词袋场景分类方法，该方法融合了基于形状的全局纹理特征、局部光谱特征和局部密集尺度不变特征。虽然组合多个互补特征可以改善提升分类结果，但是如何有效地融合不同类型的特征仍然没有定论，并且受限于手工特征提取算子的局限性，当场景图像变得更加复杂时，这些特征的描述能力会表现出性能上的局限性。

3.2 利用卷积神经网络作为特征提取器的方法

使用深度神经网络结构，同样可以起到特征提取的效果。对于一些经典的网络结构，例如AlexNet、VGGNet、GoogLeNet等，由于是在ImageNet自然图像数据集上进行的训练提取特征的方法，应用这种方法提取出的深层特征是普遍存在于自然图像中的。而自然图像和遥感图像在一定程度上存在较大的相似性，因此在自然图像中提取的现有的深层特征，应用在遥感图像上，存在一定的合理性。通过前人的实验结果表明，直接应用这一特征，也可以得到很好的效果（Penatti等，2015）。

Hu等（2015）将CNN视为特征提取器，并研究了如何充分利用预训练的CNN进行场景分类。Chaib等（2017）融合了使用VGGNet的深度特征，用来增强场景分类性能。Li等（2017）融合了预训练的CNN特征，表现出了比原始CNN特征更好的效果。Cheng等（2017b）通过现成的CNN特征代替了传统的局部描述符，设计了用一种卷积特征描述符（BoCF）。Yuan等（2019）重新排列了已训练的VGGNet-19的特征。He等（2018）提出了一种用于场景分类的新型多层堆叠协方差池化算法（MSCP）。上述方法都使用预训练的CNN作为特征提取器，然后融合或组合现有CNN提取的特征。值得注意的是，使用现成的CNN作为特征提取器的策略在小规模数据集上简单有效。

除了直接用CNN作为特征提取器以外，将CNN作为特征提取器与传统的特征提取方法进行融合表示同样受到关注。Zhu等（2018）提出了一种自适应深度稀疏语义建模框架，利用了基于均值和标准差的光谱特征，基于小波变换的纹理特征，以及SIFT特征，使用k-means聚类完成FSTM建模，作为图像的中级语义特征，再使用CaffeNet作为高级语义特征提取器，通过对这两种语义的特征融合，作为图像的特征表示，最终用SVM进行分类。

3.3 端到端的卷积神经网络的方法

在经典的机器学习中，需要对原始数据进行初步的处理，输入的不是直接的原始数据，而是在原始数据中人工参与提取的一些关键特征。由于遥感场景图像像素量大，数据维数高，这种处理方法尤为常见，需要提取图像中的一些关键信息。无论是利用传统手工提取的图像特征信息，还是基于深度学习网络结构得到的高级语义特征，都是基于这一基本框架。

然而，随着深度神经网络的发展，基于端对端的方法可以让网络自己学习如何提取更好的特征。其原理是把特征提取的任务也交给模型去做，而在这一过程中的具体细节不需要人工进行干预。通过这种方式，缩减了人工预处理和后续的处理工作，尽可能使模型从原始输入到最终的输出，给模型更多可以根据数据自动调节的空间，增加了模型的整体契合度。

深度神经网络包含大量的参数，往往需要大量的数据进行驱动。然而，对于场景分类任务而言，训练样本的数量往往不足以从头开始训练一个新的CNN。因此，更普遍的做法是先对网络进行预训练，再在目标数据集上进行微调。除此之外，采用数据增强、迁移学习等策略，同样可以有效的缓解样本量不足的问题。

尽管经典的网络如AlexNet，VGGNet，GoogLeNet在场景分类任务中表现出了较好的结果，但是，不同任务仍然需要一个更复杂、具有更强建模能力、更契合任务本身的网络结构。Sun等（2020）提出了一种将分层特征聚合和消除干扰信息集成到CNN的门控双向网络，旨在解决场景分类存在的忽略多层卷积特征的层次结构和其中冗余、互斥信息的问题。考虑到遥感图像的具体场景方向，Wang等（2018）提出了基于改进的定向响应网络

来处理场景分类中的定向问题。Liu等（2018）提出一种多尺度CNN框架来解决遥感图像中的物体尺度变化问题。

也有学者结合其他网络结构发展场景分类方法。Zhang等（2019b）充分利用了CNN和胶囊网络两种模型的有点，提出了一种场景分类架构。Hu等（2020）将CNN作为基础学习器，与AdaBoost技术相结合，构建了集成框架。Peng等（2022）采用了CNN和GCN联合学习的训练策略，通过网络的信息传播机制，利用空间关系和拓扑关系，使得参数同步更新。

高度的类内相似性和类间多样性是遥感图像场景分类的两大问题，从视觉的角度看属于多类分类问题。而在多类分类中，CNN往往与交叉熵损失函数相结合，修改损失函数以更契合任务从而提升效果成为了一种发展方向。Cheng等（2018）在传统CNN模型的目标函数上嵌入一个度量学习正则化项来训练D-CNN模型。Liu等（2019）基于Wasserstein距离等理论构造了一个新的损失函数，克服了交叉熵损失函数的缺点，并提出HW-CNNs模型。类间关系的先验知识被嵌入到模型中，来自多个CNN的信息可以在训练单个HW-CNN的过程中提供信息指导。

在图像处理领域，频率域特征也经常作为图像的特征进行使用。Fang等（2019）不仅利用了空谱特征，还利用了频谱特征。将傅里叶变换得到频谱图像应用于场景分类的方法这一领域目前较为稀缺，而这一部分的特征并不具备空间域低级特征的可视性，反而可能提供了对于特征描述的另一个可行的角度。

尽管基于预训练的方法可以得到很好的性能，但依赖于预训练的神经网络并不能学习到完全适应场景分类任务的特征。例如，大部分场景分类方法通常需要将输入图像调整为固定大小，以生成预训练CNN模型中完全连接层所需大小的输入向量。这种调整大小的过程往往会丢弃场景中的关键信息，从而降低分类性能。为了解决这一问题，Xie等（2019）引入了一种用于遥感场景分类的无尺度CNN（SF-CNN），不仅允许输入图像具有任意大小，而且还保留了使用传统的基于滑动窗口的策略提取判别特征的能力。

同样的，为了应对场景分类中样本量不足的问题，设计网络结构来处理小样本问题受到学者关注。将场景分类设计成小样本问题时，由于有限的样本难以描述数据的分布，且不一定具备代表性，因此会使得模型缺乏广义特征，以及模型的学习依赖于边界的样本偏差问题。针对这两个问题，Cui等（2022）将参数线性分类器集成到元学习框架，通过组合元核策略核拉伸损失来应对。Tang等（2022）提出了一种新的训练算法，可以在少量标记样本的情况下顺利工作，以解决样本数量少，网络性能减弱的问题。

3.4 基于注意力机制的卷积神经网络方法

尽管CNN具有很强的拟合数据的能力，但是由于优化算法和计算能力的限制，在实践中处理规模较大的数据、实现复杂任务时，模型依然存在一定瓶颈。深度学习中的注意力机制作为一种即插即用的网络模块，可以在不过多增加模型复杂度的同时提高模型的表达能力，在2017年SENet提出并刷新在ImageNet的分类精度记录之后，更是受到了广泛的关注（Hu等，2018）。

将注意力机制应用到遥感图像场景分类中，是近些年来备受关注的且效果显著的研究方向。Wang等（2019）设计了一种新的循环注意力结构ARCNet，将高级语义和空间特征压缩到几个向量中，通过自适应选择注意区域进行学习，是首个将注意力机制引入遥感图像场景分类中的模型。Cao等（2021），提出了基于自注意力的深度特征融合分类方法（SAFF），对CNN的最后3层卷积特征图进行拼接，再依次基于通道注意力和空间注意力提取特征，最后使用SVM进行分类，文章主要测试了基于AlexNet作为主干网络的SAFF模型和基于VGG-16作为主干网络的SAFF模型的表现能力。Fan等（2019）提出了一种注意力模型，称之为基于注意力的残差网络，其在ResNet50后连接了一个注意力模块，该模块由一个主干模块和一个掩码模块构成，实现了一个简单的注意力结构。

部分学者提出一些改进的注意力机制模型，其比较普遍的方法是使用一个较为成熟的网络结构作为主干，再通过提出的注意力机制来提升主干网络的性能。而ResNet结构是其中较为简单和热门的结构。Zhu等（2019）融合了ResNet18和一种空间特征转换器（SFT）模型进行综合判别的框架，称之为ADFF。Guo等（2020）提出了

SDAResNet模型，将通道注意力和空间注意力混合使用到了ResNet101之中，具体表现为交替连接在Bottleneck模块之后。除此之外，文章使用了Mixup、余弦权重衰减（CosLR）、MultiStepLR、Xavier参数初始化方法、Warmup、No bias decay、标签平滑和数据增强中的随机擦除等技巧，并系统进行了相关优化方式的消融实验。

DenseNet增加了ResNet的连接数量，具有减轻梯度消失，加强特征传递，鼓励特征重用和较少的参数数量的特点，同样适合选取为主干网络。Bi等（2020b）提出的RADC-Net在每个DenseBlock后面加入注意力，而在APDC-Net中，Bi等（2020c）在最后3个DenseBlock的输出并列通过注意力，直接进入全连接层分类，均取得了较好的效果。Tong等（2020）提出了CAD模型，将SE Block引入DenseNet121，与SENet在ResNet中的插入位置不同，选择在每个Dense Block前插入了注意力机制，并且选择了标签平滑的交叉熵损失函数来替代传统的交叉熵损失函数。Tian等（2021）提出了一种多尺度注意力网络（SEMSDNet），将多尺度信息结构引入DenseNet121作为基本框架，再使用SE Block实现进一步特征增强。Chen等（2022）提出了GCSANet模型，选用了DenseNet121作为主干网络，在每个Dense Block前连接了注意力机制进行注意力增强，并且使用Mixup的数据增强方法，提升了网络的效果。Shen等（2022）分别使用ResNet50和DenseNet121的最后4层卷积提取尺度信息，再利用提出的ACGLNet框架分别对尺度信息进行通道注意力，最后对两个分支使用Count-Sketch做深度特征融合，通过快速傅里叶变换与逆变换实现。

另外，Tang等（2021），基于孪生网络结构提出了一种注意力一致网络（ACNet），它将图像转置之后，通过共享权重的网络分别训练，再分别使用通道注意力和空间注意力强化特征，再使用MSE进行对比。Zhang等（2019a）使用MobileNet V2作为基础网络，并引入膨胀卷积和通道注意力来提取判别特征。为了进一步提高CNN的性能，还提出了一个多尺度池化模块来提取多尺度特征。Bai等（2022）基于MobileNet-v2提出了一种轻量级多尺度网络（ESPA-MSDWNet），文章使用了DW卷积，这也是MobileNet系列的基本构造。并且参考Res2Net的想法建立多尺度信息卷积核，即MSDWNet，用来替代Bottleneck中的卷积。再每个卷积后面连接一个注意力机制进行特征提取。Alhichri等（2021）在EfficientNetB3中引入了注意力机制，构建了EfficientNetB3-Attn-2。Shi等（2022a）提出了一种基于分组混合注意力的轻量级CNN结构（LCNN-GWHA）。该模型串联了4个分组混合注意力模块（GWHAM），每个模块对通道注意力和空间注意力进行了多层级，多并联的混合。为了解决场景分类中的小物体识别，Zhao等（2021）设计了一个增强的注意力模块，提高深度网络的特征提取和泛化能力，使其能够学习更多的判别特征。

在过去的研究中，研究人员将注意力机制引入CNN应用于遥感图像场景分类研究开展了深入且丰富的工作。例如，SE Block，GC Block（Cao等，2019），EPSA Block（Zhang等，2023）等注意力模块引入了遥感分类模型中，取得了很好的效果。还有一些设计了独特的注意力机制模块，同样满足轻量化，即插即用的便捷特性，同样取得了良好的效果。在这部分研究中，除了提高模型精度分类精度外，逐渐聚焦于改善网络结构提取多尺度信息，与实现网络轻量化，从而更贴近于现实应用需求。

3.5 基于Vision Transformer的方法

Vaswani等（2017）在自然语言处理领域首次提出了Transformer模型，Transformer使用一个完整的注意力机制来直接考虑局部特征之间的交互。之后，Dosovitskiy等（2021）基于Transformer中的Encoder部分，提出了Vision Transformer模型，Transformer模型逐渐应用于计算机视觉领域中。简单而言，ViT模型由3个模块组成：嵌入层，Transformer中的编码部分，和一个多层感知器用于分类。一般来说，模型首先需要将图像分割成若干个token，然后对每个token考虑位置编码进行线性嵌入。接下来，将token序列输入到Transformer中的编码器部分，通过多头自注意力机制考虑token之间的交互特征。最后，通过多层感知机额对提取到的特征进行分类。ViT模型可以直接考虑图像中所包含的上下文信息和物体的空间分布，具有一定的优势。

在计算机视觉领域取得了一定进步之后，研究人员开始专注于将这一架构引入遥感图像场景

分类中，将几个常用的ViT模型在几个场景分类数据集上做了测试，并且测试了几种数据增强方法，和通过只是蒸馏和压缩之后的模型版本。ViT相较CNN模型表现出一定的优势（Bazi等，2021；Bashmal等，2021）。并且，仅仅将预训练过的ViT模型作为特征提取器也是可行的。Chaib等（2022）将通过预训练的ViT模型作为一个特征提取器，将ViT提取的特征合并成一个信号数据集，再针对特征设计了一种特征和图像选择算法。

但是，Transformer结构在视觉任务中存在一定的缺陷，例如patch编码限制了模型学习图像整体特征的能力，对局部信息的学习能力有限，具有较大的计算复杂度等。通过增加模块的方式，可以有效的减少模型本身所带来的缺陷。为了解决普通的ViT模型仅仅简单地将图像分割成固定大小的patch作为token的做法，Lv等（2022）提出了一种空间通道特征保持模型。首先使用逐步聚合相邻重叠patch生成token，提取图像局部结构特征，再采用多头自注意力机制进行建模，然后使用一种轻量级通道注意力机制来考虑不同通道的重要权重，最后使用多层感知机进行分类。实验结果证实了ViT模型在场景分类中的潜力。Swin Transformer是Transformer在计算机视觉领域的又一次碰撞。Wang等（2022）结合Swin Transformer，提出了一种多级别融合Swin Transformer模型，该方法集成了多级特征融合模块和自适应特征压缩模块来进一步提高场景分类性能。

将一些框架与ViT结合，同样可以起到很好的效果。Sha和Li（2022）通过将多实例学习与ViT结合，以应对ViT忽略关键局部特征的缺陷。使用多实例学习的方法，可以有助于突出遥感场景关键局部区域的特征。Bi等（2023）将监督对比学习与ViT结合，并且设计了一种新的结合了CE损失和SupCon损失的联合损失函数已替代传统的交叉损失函数，通过大量的实验验证了其优越性能。

CNN结构在过去的研究中始终占据主导位置，并且具有强大的表现能力，如何将CNN与ViT之间的特点进行扬弃，成为一个研究热点。Zhang等（2021）使用了一种新颖的方式，将自注意力机制集成到了ResNet中，提出了TRS模型，它是一种纯卷积-卷积+Transformer-纯Transformer的结构。与注意力机制模型和ViT模型相比均表现了更好的效果。Deng等（2022）结合了ViT和CNN提出了CTNet，使用双流网络的框架，将ViT设定为T-stream来获取语义特征，CNN设定为C-stream来获取结构特征，最后定义了一个联合损失函数来整体优化。Li等（2022a）同样使用双流网络来结合CNN和Transformer。文章提出了一种双向特征交互模块，该模块不仅可以有效地融合基于CNN的局部特征和基于Transformer的全局特征，还可以从遥感场景中提取多尺度信息。Tang等（2022）也是一种结合CNN与Transformer的模型，利用多层特征提取模块从场景中获取全局视觉特征和多层卷积特征，提出语义信息提取模块，从多层特征中获取丰富的语义信息，并设计了一个跨层次注意力模块来聚合特征之间的相关性，最后设计了评分融合模块，来整合各个级别特征的贡献。Yu等（2022）构建了一种跨语义，跨尺度的胶囊ViT结构，在几个标准数据集上同样取得了很好的效果。

知识蒸馏是一种经典的模型压缩方法，核心思想是通过引导轻量化的学生模型模仿性能更好、结构更复杂的教师模型，在不改变学生模型结构的情况下提高其性能。通过知识蒸馏的思想，将CNN与Transformer结合也是一个有效的方法。Xu等（2022）提出了ET-GSNet，利用ViT模型作为教师网络，选用ResNet作为学生网络进行训练，该模型通过知识蒸馏综合两种模型的优点，而不增加计算复杂度，可以顺利的将教师模型中的暗知识转移到学生模型中。Nabi等（2022）则是将CNN作为教师网络，而将ViT模型设置为学生网络，同样证明了方法的有效性。

从2021年ViT开始在遥感图像场景分类中受到关注以来，相较于传统的CNN而言，研究成果相对较少，依然存在着巨大的研究潜力。

3.6 基于生成对抗网络的方法

在深度学习在遥感领域兴起之前，视觉特征在遥感图像场景分类中起着重要的作用。从本质上说，深度CNN模型在特征学习中属于监督学习方法，而在无监督学习方向方法中，研究人员最初主要应用自动编码器作为基本模型（Zhang等，2015；Cheng等，2015；Du等，2017）。然而，基于自编码器的方法大多没有充分利用场景类信息，无法学习到更有效表征场景信息的语义特征。

基于CNN的方法需要使用大量的标记样本来训练模型，然而对于海量的遥感图像而言，注释

样本是一项劳动密集型的工作。而GAN可以生成伪样本，以实现数据增广的目的（Goodfellow等，2014）。因此一些研究人员开始用GAN来生成场景分类样本。Xu等（2018）为了生成用于场景分类的高质量遥感图像，将缩放的指数线性单元（SELU）添加到GAN。Ma等（2019）设计了Sifting GAN，它可以生成大量真实的带注释样本用于场景分类。该方法扩展了传统的GAN框架，包括用于样本生成、模型筛选和样本筛选等多种方法。Han等（2020）提出了一种新的基于GAN的遥感图像生成方法，可以用于创建用于场景分类的高分辨率注释样本。Ma等（2022a）提出了一种有监督的渐进式GAN，在样本有限的情况下显著的提高了分类精度。文章主要提出了标签样本的条件生成框架，并介绍了一种渐进生长样本生成方法。

生成对抗网络的提出是利用噪声通过训练来生成伪样本，在视觉领域为了将其应用在图像处理上，将GAN和CNN结合形成了DCGAN模型（Radford等，2016）。近年来随着GAN应用的兴起，GAN同样被应用在了场景分类中。Lin等（2017）提出了一种用于场景分类任务的多层特征匹配生成对抗网络（MARTA GANs）。CNN模型的训练需要数百万个参数，但是样本集能提供的样本量远远没有达到这个量级，MARTA GANs模型的提出着重解决的是样本量不足的问题。Duan等（2018）提出了一种用于遥感图像分类的具有非局部空间信息的生成对抗网络（GAN-NL）。具体来说，将非局部层合并到GAN中以进行无监督表示学习。然后，设计一个分类网络来推断图像的标签。Han等（2018）提出了称之为SSGA-E的生成框架来进行场景分类。该框架将深度学习特征、自标签技术和判别评估方法结合起来，完成场景分类和数据集标注任务。Yu等（2020）设计了一种用于场景分类的注意力GAN，即Attention GAN，它通过增强判别器的表示能力来实现更好的场景分类性能。Attention-GAN模型在GAN的基础上进一步加入了注意力机制。模型将判别器设计成了CNN的结构，并且在每一组卷积层中间加入了注意力层来进行实现，最终在无监督学习处理场景分类领域中得到了很好的效果。在近年的研究中，Guo等（2021）将自注意力的门控单元引入Inception V3网络，提出了SAG模块和SAGGAN网络。Pan等（2020）提出了Diversity-GAN模型，充分利用了GAN训练过程采用循序渐进的方式，和训练进度可控的优势，不仅保证了生成样本的多样性，还可以在训练阶段通过少量迭代实现场景图像结构的多样性。Wei等（2020）提出了一种改进的无监督表示学习模型，可以从无标记样本中提取用于场景分类的特征信息。Guo等（2021）提出了一种基于相似自监督门控自注意力GAN的场景分类方法。

尽管针对生成对抗实例有了大量的方法，但是如何提高对未知攻击的防御能力仍然需要突破。Cheng等（2022）提出了一种用于场景分类的有效防御框架，通过引入图像重建过程中生成的实例，以及纯净的和对抗的实例，来训练分类器。

在遥感图像场景分类中，跨域分类是一个值得关注的方向。GAN为跨域场景分类提供了一种新的思路和方向。Bashmal等（2018）提供了一种用于学习来自两个不同域图像的不变特征方法，用来处理跨域条件下的遥感图像场景分类问题。Teng等（2020）提出了一种用于跨域半监督场景分类的分类器约束对抗网络。文章采用采用深度CNN构建特征表示来描述场景的语义内容，然后进行自适应。然后，使用对抗域自适应来对齐源和目标的特征分布，并生成器在分类器约束下创建远离原始土地覆盖类边界的鲁棒可转移特征。

在遥感图像场景分类领域，大多数基于GAN的方法通常以对抗的方式使用GAN进行样本生成或特征学习。与基于CNN的场景分类方法相比，目前关于基于GAN的场景分类方法的文章较少，并且基于GAN的场景分类性能不如基于CNN的方法。除此之外，因为基于GAN的场景分类方法通常需要标签来训练额外的分类器，导致大多数方法无法进行端到端的训练。然而，GAN强大的自监督特征学习能力为场景分类的未来提供了一个很有前途的方向。

3.7 基于遥感基础模型的方法

基础模型是指作为各种下游任务的基本构建块的一种深度学习模型（Yuan，2023）。其旨在从大规模数据集中学习到一般的特征表示，然后根据特定的应用进行微调和延展，以期达到更好的效果（He等，2024）。在自然图像领域，基础模型的研究已经取得了巨大的进展，从自然语言处理到计算机视觉领域，基础模型层出不穷，典型的方

法有：BERT（Devlin等，2019）、iBOT（Zhou等，2021）、T5（Raffel等，2020）、ChatGPT（Dehouche，2021）、CodeGeeX（Zheng等，2023）、DALL-E（Ramesh等，2021）、CLIP（Li等，2022b）、DINO（Caron等，2020）等。

然而与自然图像相比，遥感图像受到传感器影响，通常具有多个空间分辨率，由于尺度和角度的多样性，同一物体在遥感图像中具有不同的特征。另外，遥感图像中具有许多小而密集的目标，在大而宽的遥感图像中，这些小而密集的目标影响了解译的精度。除此以外，除了目标信息外，遥感图像包含了大量的背景信息，导致图像的信噪比较低。遥感图像容易受到天气、光线、云、雾等自然因素的干扰，影响成像质量。基于上述自然图像与遥感图像的差异，用自然图像训练的基础模型在遥感图像上的表现较差，因此，有必要设计合适遥感数据的基础模型（Sun等，2023）。

遥感图像的基础大模型研究已经取得了一定进展。Wang等（2023b）对遥感预训练数据进行了实验，结果表明预训练方法可以有效缓解数据差异，但由于下游任务需要与场景识别任务不同的表征，因此仍可能受到任务差异的影响。Sun等（2023）收集了200万张遥感图像，构建了一个覆盖全球不同场景和对象的大规模数据集进行预训练，其提出的空天·灵眸基础模型，为遥感领域多行业应用提供了一套通用便捷、性能优良的解决方案。Wang等（2023c）训练了一个具有1亿参数的视觉遥感transformer模型，并引入了一种新的旋转可变大小的注意力方法，以适应遥感图像中的目标密集特点。针对遥感目标尺度差异性较大的问题，Reed等（2023）提出了scale-mae方法，该方法学习了不同已知尺度数据之间的关系。此外，Cong等（2022）建立了一个基于mae的多光谱卫星图像预训练框架，将该范式扩展到多光谱图像和时序维度上。Mai等（2023）利用图像中的地理空间信息构建了一个用于对比学习的预训练框架。Mendieta等（2023）构建了一个紧凑而多样的数据集，称之为GeoPile，从而增加预训练数据中的信息量。

场景分类作为遥感图像信息处理的任务，是各个遥感基础大模型的主要下游任务之一。基于遥感基础大模型的场景分类方法，往往具有比单纯的进行图像处理算法设计更优的结果。利用遥感基础大模型统一场景分类、目标检测、语义分割等多种下游任务已经成为当今处理高分辨率遥感图像的新的热点方向。

对于已经在遥感图像场景分类领域有一定背景知识的研究人员来说，开源代码可以为他们提供巨大的帮助。我们在表1中总结了部分在GitHub上的该领域的开源代码，以便对不同方法进行比较。这些开源代码既有基于MATLAB开发的，又有使用Python语言，基于Caffe，TensorFlow，PyTorch框架进行开发的。其中，由Meta开发的PyTorch框架使用频次较高，更获得当前开发者们的青睐。值得注意是，Google在近期推出的JAX框架，暂时还没有应用到该领域的开发中。

表1 部分遥感图像场景分类深度学习方法的公开代码汇总

Table 1 Summary of the open-source codes of some remote sensing image scene classification deep learning methods

模型名称	语言/框架	代码链接
MSCP(He等,2018)	MATLAB	https://github.com/henanjun/code_MSCP
D-CNNs(Cheng等,2018)	C/C++	https://github.com/limbo0000/PairLoss
IORN(Wang等,2018)	PyTorch	https://github.com/wdczs/ImprovedORN
FACNN(Lu等,2019)	PyTorch	https://github.com/Hua-YS/Multi-Scene-Recognition
Hydra(Minetto等,2019)	TensorFlow	https://github.com/maups/hydra-fmow
SCcov(He等,2020)	MATLAB	https://github.com/henanjun/SccovNet
SF-CNN(Xie等,2019)	Caffe	https://github.com/Aaromxj/SF-CNN
EAM(Zhao等,2021)	PyTorch	https://github.com/williamzhao95/Pay-More-Attention
CPGL(Tang等,2022)	PyTorch	https://github.com/TangXu-Group/Remote-Sensing-Images-Classification/tree/main/CPGL
MF2CNet(Bai等,2022)	PyTorch	https://github.com/liuqingxin-chd/MF2CNet
ARCNet(Wang等,2019)	PyTorch	https://github.com/laserbox/ARCNet
SAFF(Cao等,2021)	PyTorch	https://github.com/zh-hike/SAFF

续表

模型名称	语言/框架	代码链接
GCSANet(Chen等,2022)	PyTorch	https://github.com/ShubingOuyangcug/GCSANet
SCDAE(Du等,2017)	PyTorch	https://github.com/296769150/SCDAE
MARTA GANs(Lin等,2017)	TensorFlow	https://github.com/BUPTLdy/MARTA-GAN
Simaese-GANs(Bashmal等,2018)	TensorFlow	https://github.com/LailaMB/Siamese-GANs
SSGA-E(Han等,2018)	Caffe	https://github.com/weihancug/SSGF-for-HRRS-scene-classification
SiftingGAN(Ma等,2019)	TensorFlow	https://github.com/MDAooo/SiftingGAN
PSGAN(Cheng等,2022)	PyTorch	https://github.com/xuxiangsun/PSGAN

4 遥感图像场景分类数据集与方法性能比较

4.1 数据集研究现状

为了方便开展遥感图像场景分类这一领域的研究，近十年来，研究人员公布了大量用于遥感图像场景分类的数据集。数据集可以为方法提供评估标准，以检验方法的有效性。在早期，场景分类领域最具代表性的数据集是UC Merced数据集（Yang和Newsam，2010）。该数据由2100个场景图像组成，分为21个场景类别。每个类别由100个256×256像素的航空遥感图像组成，共有3个波段通道，空间分辨率为0.3 m。该数据的数据源是美国地质调查局（United States Geological Survey）国家地图下载的航空正射影像，涵盖了美国20座城市。该数据集包含高度重叠的土地利用类别，例如密集住宅区，中等密度的住宅区和稀疏住宅区，它们的主要区别在于结构的密度，通过这样的方式令数据集更加丰富和具有挑战性。

除此之外，为了适应概率主题模型的方法评估，涌现了一批优质的，被广泛使用的样本量在1000—3000之间，类别数不超过21个的较小规模的数据集，例如：WHU-RS19（Xia等，2010；Sheng等，2012），RSSCN7（Zou等，2015），Brazilian Coffee Scene（Penatti等，2015），RSC11（Zhao等，2016b），SIRI-WHU（Zhu等，2016）。另外，SAT-4和SAT-6数据集则是建立了包含500000/405000个图像斑块的数据集，斑块的大小仅为28×28（Basu等，2015）。伴随着深度学习方法在场景分类中得到密切的关注，数据驱动和深度学习方法开始逐渐替代传统的人工解译方法，这为遥感图像的自动解译和分析提供了广阔的前景。然而，受限于场景分类数据集规模较小，深度网络的性能往往会受到数据的限制，进一步致使智能方法在实际应用中应用受限。因此，开发一个更大规模的数据集作为分类方法的基准成为趋势。

AID数据集（Xia等，2017）是用于航空场景分类的相对大规模的数据集。数据集由30个场景类别组成，每个场景类别包含220到420张图像，共包含了10000张图像。图像从Google Earth中裁剪得来，并固定在600×600大小。由于AID使用的航拍图像是用不同的传感器获取的，因此该数据集是多源的，数据集提供了传感器参数信息。此外，该数据集还是多分辨率的，每个场景类别的空间分辨率从大约8 m到大约0.5 m不等。相对于之前的数据集将城市区域内的建筑笼统的归类为建筑物，该数据集对具体的城区职能进行了进一步细分。典型的包括：教堂、商业区、工业区、港口、火车站等，这些类别表示了城市中一些具体的场景，进一步提升了分类的难度。

NWPU-RESISC45数据集（Cheng等，2017a）在很长一段时间内都是最大的场景分类数据集。它由45个场景类别组成，每个类别包含700张图像。它从Google Earth中获得，大小为256×256像素。该数据集总共包含31500张场景图像，选择100多个国家和地区。除了一些空间分辨率较低的特定类别外，大多数场景类别的空间分辨率在30 m到0.2 m之间变化。该数据集将同属道路的多种场景做出了进一步拓展，如：桥梁、高速公路、交叉口、立交桥、铁路、环形交通枢纽、跑道等。相对而言，NWPU-RESISC45数据集包含的类别十分丰富，且包含城市中的地物类别、多种典型地貌、水体等难以区分的类别。

自此之后，更多的具有挑战性和不同特点的数据集被提出，如RSI-CB256，RSI-CB128（Li等，2020a），MASATI（Gallego等，2018），RSD46-WHU

(Xiao 等, 2017), PatternNet (Zhou 等, 2017), Optimal-31 (Wang 等, 2019), CLRS (Li 等, 2020b), EuroSAT (Helber 等, 2019) 等。其中, fMoW 包含多达 62 个类别, 132716 张图像, 且各个类别中存在不平衡的问题 (Christie 等, 2018)。而 WH-MAVS 数据集的图像分别属于 2014 年和 2016 年两个年份, 给场景分类任务带来了新的挑战 (Yuan 等, 2022)。

除了期刊会议论文发布数据集外, 遥感图像场景分类也备受比赛关注。例如 Kaggle 在 2017 年和 2019 年分布发布了 Planet-Understanding the Amazon from Space 数据集和 WiDS Datathon 2019 数据集。

即便是一个普通的深度 CNN 模型, 往往拥有数百万个参数, 会过度拟合训练集中的数万个训练样本。因此, 利用现有的场景分类数据集完全训练深度分类模型几乎是不可行的。当下的场景分类方法普遍采用迁移学习的方法使用在 ImageNet 预训练过的模型, 但是自然图像和遥感图像依然存在着特征表示上的差距。虽然迁移方法在类型和样本有限的目标数据集上表现得相当好, 但与完全训练深度 CNN 模型相比, 它们并不是最优的解决方案, 因为从头开始训练的模型能够在训练样本足够大的情况下提取更具体的特征, 以适应目标域。

随着遥感影像数据发布平台的发展, 对于遥感影像从原始传感器数据到可发布使用的数据的流程将日趋成熟, 遥感影像数据也会在数据存储, 数据管理, 数据应用上不断提升, 这将为场景分类数据集的扩充逐步创建条件。Million-AID 数据集包含 51 个类别, 每个类别大约包括 20000 张图像, 共 1000000 张图像。空间分辨率在 0.5 m 到 153 m 之间, 图像大小为 256×256 和 512×512。数据来自于 Google Earth 中的全球各个地方。数据集制作中分类方法采用了交互式制作, 并标注了地理位置, 为网络模型在遥感图像上进行预训练提供了可能 (Long 等, 2021)。

随着类别数量的增多, 在过去的数据集设计中完全独立的各个类别之间内在的联系开始受到关注。Million-AID 数据集将总共 51 个类别划分为农业用地、商业用地、工业用地、公共服务用地、住宅用地、交通用地、未利用土地和水体区域 8 个大类别, 其中较为复杂的农业用地、工业用地、公共服务用地和交通用地 4 个较复杂的类别做出了进一步细分, 每个子类别包含 1—6 个具体的类别。通过这种方式, Million-AID 对现今分类中出现频次较高的各种类别建立了一个普适性很强的类别架构体系。

遥感图像场景分类问题的提出和发展的内在动力是遥感图像空间分辨率的提升, 因此场景分类的研究一般主要基于高分辨率遥感图像。而高分辨率遥感图像可表示的信息量也仅存在于目标地物所呈现出来的可见光光谱信息, 而这部分信息若在完整的描述多种多样的遥感场景类型, 依然存在着一定的困难。出于这个原因, 可以依靠更多模态的数据来对扩充可用的信息量来进行综合判断, 从而进一步提升任务效果。BigEarthNet 数据集不仅将样本量扩展到 590326 张。数据集包含 43 个类别, 每个类别包含 328—217119 张图像, 共 590326 张图像。空间分辨率为 10 m, 20 m 和 60 m。图像大小为 20×20, 60×60 和 120×120, 提供了地理位置、成像时间和传感器参数信息 (Sumbul 等, 2019)。值得注意的是, BigEarthNet 中包含许多不同类别的森林都关联了很多的样本量, 其中混交林关联了 217119 个样本、针叶林关联了 211703 个样本、过渡林灌木关联 173506 个样本、阔叶林关联 150944 个样本, 这从侧面也反映出了该数据集对自然地物地貌的分类有所侧重。

BigEarthNet 在 2021 年进行了补充, 在 Sentinel-2 的基础上引入了 Sentinel-1 数据, 即 BigEarthNet-MM 数据集, 构建了多模态数据集, 从而支持了对多模态遥感和深度学习的研究 (Sumbul 等, 2021)。另外, 文章还证明了在 BigEarthNet-MM 上从头训练的深度学习模型优于在 ImageNet 上预训练的模型, 尤其是在一些包括农业和其他植被自然环境的复杂类别方面。

在现实工程中, 当很多遥感图像并不是具有单一场景类别的, 而当下的研究更多的是将遥感影像依据场景特征硬分类为某一种特定类别。随着数据集的发展, 以及场景分类方法在越来越多的领域上的发展, 对遥感影像进行多标签分类具有重要的意义, 可以有效的将模棱两可的分类结果呈现给遥感解译技术员做出进一步的判断。MultiScene 数据集的提出创建了多标签遥感影像场景分类数据集, 相较于传统的数据集一张影像仅用一个标签进行表示, 该数据集将标签量扩展到至多 13 个 (Hua 等, 2022)。该数据集相较于一般

的单标签分类数据集视野更大，例如该数据集中某一个样本便包含了桥梁、停车场、河流、环形交叉路和密集建筑物5种类别，将多种场景类别融合在一个样本中，可以有效的检验方法的场景识别能力。

在已提出的数据集中，UC Merced，AID和NWPU-RESISC45最经常被用来作为检验方法效果的标准数据集。如今，大多数的分类方法选择UC Merced，AID和NWPU-RESISC45数据集作为实验数据集，以证明方法的有效性。遥感图像场景分类常用的数据集如表2所示。

表2 遥感图像场景分类数据集汇总表

Table 2 Summarization of remote sensing image scene classification datasets

数据集名称	类别数	样本量	空间分辨率/m	图像大小	数据链接
UC Merced（Yang和Newsam，2010）	21	2100	0.3	256×256	http://weegee.vision.ucmerced.edu/datasets/landuse.html
WHU-RS19（Xia等，2010）	19	1013	最高0.5	600×600	http://www.lmars.whu.edu.cn/prof_web/zhongyanfei/e-code.html
RSSCN7（Zou等，2015）	7	2800		400×400	https://sites.google.com/site/qinzoucn/documents
SAT-4（Basu等，2015）	4	500000	1—6	28×28	https://www.kaggle.com/datasets/crawford/deepsat-sat4
SAT-6（Basu等，2015）	6	405000	1—6	28×28	https://www.kaggle.com/datasets/crawford/deepsat-sat6
Brazilian Coffee Scenes（Penatti等，2015）	2	2876		600×600	http://patreo.dcc.ufmg.br/2017/11/12/brazilian-coffee-scenes-dataset/
RSC11（Zhao等，2016b）	11	1232	约0.2	512×512	https://www.researchgate.net/publication/271647282_RS_C11_Database
SIRI-WHU（Zhu等，2016）	12	2400	2	200×200	http://www.lmars.whu.edu.cn/prof_web/zhongyanfei/e-code.html
NWPU-RESISC45（Cheng等，2017a）	45	31500	0.2—30	256×256	https://hyper.ai/datasets/5449
AID（Xia等，2017）	30	10000	0.5—8	600×600	https://pan.baidu.com/s/1mifOBv6?_at_=1695288119618#list/path=/
RSD46-WHU（Xiao等，2017）	46	40480	0.5—2	256×256	https://github.com/RSIA-LIESMARS-WHU/RSD46-WHU
MASATI（Gallego等，2018）	7	7389		512×512	https://www.iuii.ua.es/datasets/masati/
EuroSAT（Helber等，2019）	10	27000	10	64×64	https://github.com/phelber/eurosat#
PatternNet（Zhou等，2017）	38	30400	0.062—4.693	256×256	https://sites.google.com/view/zhouwx/dataset
fMoW（Christie等，2018）	62	132716	0.5	74×58—16184×16288	https://github.com/fMoW/dataset
Optimal-31（Wang等，2019）	31	1860		256×256	https://drive.google.com/open?id=1Fk9a0DW8UyyQsR8dP2Qdakmr69NVBhq9
BigEarthNet（Sumbul等，2019）	43	590326	10—60	20×20，60×60，120×120	https://bigearth.net/
CLRS（Li等，2020b）	25	15000	0.26—8.85	256×256	https://github.com/lehaifeng/CLRS
RSI-CB128（Li等，2020a）	45	36000	0.3—3	128×128	https://github.com/lehaifeng/RSI-CB
RSI-CB256（Li等，2020a）	35	24000	0.3—3	256×256	https://github.com/lehaifeng/RSI-CB
MLRSN（Qi等，2020）	46	109161	0.1—10	256×256	https://github.com/cugbrs/MLRSNet
Million-AID（Long等，2021）	51	1000000	0.5—153	256×256或512×512	https://jin-pu.github.io/Million-AID/2018/12/12/Million-AID/
MultiScene（Hua等，2022）	36	100000	0.3—0.6	512×512	https://multiscene.github.io/
WH-MAVS（Yuan等，2022）	14	47137	1.2	200×200	http://sigma.whu.edu.cn/newspage.php?q=2021_06_27_eng

4.2 评估方法

在场景分类中，较为单薄的使用总体分类精度（OA）和混淆矩阵（Confusion Matrix）来进行精度评价。

混淆矩阵是一个用于表示预测值和真实值的数量统计情况的矩阵。通常来说，矩阵中的列，代表真实数据，行代表由遥感数据分类得到的预测数据。通过混淆矩阵，可以比较清晰地观察到在多分类过程中，特定的某一种类别的分类效果，分析模型在各个类别上的精度。由此根据具体的分类情况来进行模型的调整。混淆矩阵能够很直观的表现出模型的较为具体的分类情况。

依据混淆矩阵，应用总体分类精度来综合评价分类结果的优劣，有公式：

$$\mathrm{OA}=\frac{1}{N}\sum_{i=0}^{r-1}x_{ii} \tag{1}$$

式中，r是混淆矩阵的行列数，由于混淆矩阵是方阵，因此行列数一致。N是样本总数。x是混淆矩阵对应位置的值。通过应用上述两种精度评价方法，可以有效的应用在各个模型中，作为分析模型效果的工具和调整模型的依据。

Kappa系数用于一致性检验，同样也是一种衡量分类精度的指标，在遥感图像分类中被广泛使用。

$$\kappa=\frac{N\cdot\sum_{i=0}^{r-1}x_{ii}-\sum_{i=0}^{r-1}\left(\sum_{j=0}^{r-1}x_{ij}\cdot\sum_{j=0}^{r-1}x_{ji}\right)}{N^2-\sum_{i=0}^{r-1}\left(\sum_{j=0}^{r-1}x_{ij}\cdot\sum_{j=0}^{r-1}x_{ji}\right)} \tag{2}$$

和总体精度一样，Kappa系数也可以从数值上直观的反应分类模型的性能。

对于将注意力机制引入CNN的方法，可以使用CAM，Grad-CAM和Grad-CAM++来更直观清晰的展示加入注意力机制前后的效果（Zhou等，2016；Selvaraju等，2017；Chattopadhyay等，2018）。该方法的通过热力图展示，直观的展示出CNN学习到的特征，在场景分类中得到了广泛的应用（Zhu等，2019；Tong等，2020；Bai等，2022；Shi等，2022b）。

4.3 性能比较

近年来，各种场景分类算法相继问世。本节使用总体分类精度指标作为评估标准，在UC Merced，AID和NWPU-RESISC45这3个最常用的数据集上进行比较。对于UC Merced数据集，每类选择抽取80%或50%样本作为训练样本。对于AID数据集，每类选取50%和20%样本作为训练样本，而对于NWPU-RESISC45数据集，每类选取20%或10%样本进行训练样本。

首先，我们列举了基于CNN的方法，各个方法在数据集上表现的性能如表3所示。随着时间的发展，UC Merced数据集的在80%样本量训练精度上逐渐趋近于100.00%。在这样的实验设置下，共有420张图片作为测试样本，当仅有1张测试样本被误判时，准确率即为419/420≈99.76%，与而利用CNN的许多网络在平均值上已经达到了这一水平。2017年，随着AID和NWPU-RESISC45数据集的提出，为各个方法给出了更具挑战性的基准。而随着时间发展，在这两个数据集上进行性能检验的总体精度也逐渐突破了90%。CNN结构在遥感图像场景分类中表现出了强大的性能优势。从性能上分析，仅仅使用CNN作为特征提取器并没有发挥CNN的潜力，而注意力机制可以进一步增强了CNN的特征表示能力。

表3 基于卷积神经网络的遥感图像场景分类模型在标准数据集上的分类精度

Table 3 Classification accuracy of remote sensing image scene classification model based on convolutional neural network on standard dataset

/%

模型名称	UCM80	AID50	AID20	NWPU20	NWPU10
D-CNNs(Cheng等,2018)	98.93±0.10	96.89±0.10	90.82±0.16	91.89±0.22	89.22±0.50
SF-CNN(Xie等,2019)	99.05±0.27	96.66±0.11	93.60±0.12	92.55±0.14	89.89±0.16
SCCov(He等,2020)	99.05±0.25	96.10±0.16	93.12±0.25	92.10±0.25	89.30±0.35
ADFF(Zhu等,2019)	97.53±0.63	94.75±0.25	93.68±0.29	91.91±0.23	90.58±0.19
CNN-CapsNet(Zhang等,2019b)	99.05±0.24	96.32±0.12	93.79±0.13	89.03±0.21	89.03±0.21
DCCNN(Bi等,2019)	96.21±0.67	91.49±0.22	87.37±0.41	85.63±0.18	83.97±0.19
RADC-Net(Bi等,2020b)	97.05±0.48	92.35±0.19	88.12±0.43	87.63±0.28	85.72±0.25

续表

模型名称	UCM80	AID50	AID20	NWPU20	NWPU10
SE-MDPMNet(Zhang等,2019a)	98.95±0.12	97.14±0.15	94.68±0.17	94.11±0.03	91.80±0.07
APDC-Net(Bi等,2020c)	97.05±0.43	92.15±0.29	88.56±0.29	87.84±0.26	85.94±0.22
MIDC-Net(Bi等,2020a)	97.40±0.48	92.95±0.17	88.51±0.41	87.99±0.18	86.12±0.29
CAD(Tong等,2020)	99.66±0.27	97.16±0.26	95.73±0.22	94.58±0.26	92.70±0.32
SAFF(Cao等,2021)	97.02±0.78	93.83±0.28	90.25±0.29	87.86±0.14	84.38±0.19
ACNet(Tang等,2021)	99.76±0.10	95.38±0.29	93.33±0.29	92.42±0.16	91.09±0.13
SEMSDNet(Tian等,2021)	99.41±0.14	97.64±0.51	94.23±0.63	93.89±0.63	91.68±0.39
ResNet101-EAM(Zhao等,2021)	99.21±0.26	97.06±0.19	94.26±0.11	94.29±0.09	91.91±0.22
VGG-MS2AP(Bi等,2021)	99.45±0.32	96.86±0.20	95.42±0.28	93.91±0.15	92.27±0.21
GCSANet(Chen等,2022)	99.31±0.56	97.53±0.32	95.96±0.38	94.95±0.36	93.39±0.39
LCNN-GWHA(Shi等,2022a)	99.76±0.25	97.64±0.28	93.85±0.16	94.26±0.25	92.24±0.12
HHTL(Ma等,2022b)	99.48±0.28	96.88±0.21	95.62±0.13	94.21±0.09	92.07±0.44
MF2CNet(Bai等,2022)	99.52±0.25	97.02±0.28	95.54±0.17	93.85±0.27	92.07±0.22
SCCNN(Shi等,2022b)	99.76±0.05	97.31±0.10	93.15±0.25	94.39±0.16	92.02±0.50

注：UCM80为使用UC Merced数据集使用80%/20%的样本进行训练/测试；AID50为使用AID数据集使用50%/50%的样本进行训练/测试；AID20为使用AID数据集使用20%/80%的样本进行训练/测试；NWPU20为使用NWPU-RESISC45数据集使用20%/80%的样本进行训练/测试；NWPU10为使用NWPU-RESISC45数据集使用10%/90%的样本进行训练/测试。

随着Vision Transformer近两年在场景分类的兴起，本节列举了一部分改进Transformer的效果如表4所示。可见，对于这部分方法而言，其效果与CNN中的较优的性能网络类似。尽管没有极为明显的提升，但是依然依靠其较高的平均总体精度，成为后续研究的热点方向。

表4 基于Vision Transformer的遥感图像场景分类模型在标准数据集上的分类精度

Table 4 Classification accuracy of remote sensing image scene classification model based on vision transformer on standard dataset

/%

模型名称	UCM80	AID50	AID20	NWPU20	NWPU10
TRS(Zhang等,2021)	99.52±0.17	98.48±0.06	95.54±0.18	95.56±0.20	93.06±0.11
EMTCAL(Tang等,2022)	99.57±0.28	96.41±0.23	94.69±0.14	93.65±0.12	91.63±0.19
MFST(Wang等,2022)	—	97.38±0.08	96.23±0.16	94.90±0.06	92.64±0.08
Resformer(Li等,2022a)	—	—	96.01±0.21	—	92.68±0.14
ET-GSNet(Xu等,2022)	99.29±0.34	96.88±0.19	95.58±0.18	94.50±0.18	92.72±0.18
SCViT(Lv等,2022)	99.57±0.31	96.98±0.16	95.56±0.17	94.66±0.10	92.72±0.04
C^2-CapsViT(Yu等,2022)	99.76±0.12	97.50±0.15	96.05±0.11	95.28±0.08	93.32±0.05
CTNet-ResNet34(Deng等,2022)	—	97.56±0.20	96.35±0.13	95.49±0.12	93.86±0.22
CTNet-MobileNet_v2(Deng等,2022)	—	97.70±0.11	96.25±0.10	95.40±0.15	93.90±0.14
MITformer(Sha和Li,2022)	99.83±0.24	97.96±0.18	96.04±0.21	95.93±0.17	94.09±0.15
ViT-CL(Bi等,2023)	99.76±1.06	97.31±3.09	95.39±4.68	94.69±4.38	92.85±5.38

注：UCM80为使用UC Merced数据集使用80%/20%的样本进行训练/测试；AID50为使用AID数据集使用50%/50%的样本进行训练/测试；AID20为使用AID数据集使用20%/80%的样本进行训练/测试；NWPU20为使用NWPU-RESISC45数据集使用20%/80%的样本进行训练/测试；NWPU10为使用NWPU-RESISC45数据集使用10%/90%的样本进行训练/测试。

5 结 语

场景分类是遥感图像解译中一个重要且具有挑战性的问题。因其广泛的应用，引起了许多研究者的关注。由于深度学习技术的应用和大规模场景分类数据集的建立，场景分类有了快速的发展。尽管在过去几年中取得了惊人的成绩，但目前机器的理解水平与人类的表现水平之间仍然存

在巨大的差距，因此，在场景分类领域还有很多工作要做。本文通过对现有场景分类算法和现有数据集的研究，探讨了遥感图像场景分类的几个潜在发展方向。

（1）一个成熟的、理想化的场景分类系统，应该包含一个完善丰富的类别划分体系，一套足以支撑训练海量参数模型的数据集和一系列高性能的分类方法。然而，人类对于地面场景本身存在着一定的认知局限性，不同自然地理条件和人文地理条件都会对类别的划分判定带来一定程度的限制。因此在标准数据集中训练的模型标签，在面对实际应用中，存在目标类别数量和类型与模型不一致的情况。现行的经常被使用的数据集包含了几十个场景类别，远远少于人类可以判定的场景类别数量。随着遥感图像数据发布平台的发展，今后对于遥感图像从原始传感器数据到可发布使用的数据的流程将日趋成熟，遥感图像数据也会在数据存储、数据管理及数据应用上不断提升，这将为场景分类数据集的扩充逐步创建条件，从而使得大规模场景分类数据集得到充分的应用。

（2）在过去的几十年里，人们对单标签图像分类进行了广泛的研究。然而，在现实工程中，遥感图像并不是具有单一场景类别，而当下的研究更多地是将遥感图像依据场景特征硬分类为某一种特定类别。因此，单标签遥感图像场景分类或许无法深入理解遥感图像错综复杂的内容。随着数据集和场景分类方法在越来越多的领域上的应用，遥感图像的多标签分类具有重要的意义，其可以有效地将更多的标签分类结果呈现给遥感解译技术员做出进一步的判断。

（3）遥感图像场景分类问题的提出和发展的内在动力是遥感图像空间分辨率的提升，因此场景分类研究一般主要针对高分辨率遥感图像进行处理。然而，高分辨率遥感图像可表示的信息量也仅存在于目标地物所呈现出来的可见光光谱信息，而仅利用可见光光谱无法完整地描述多样的遥感场景类型。这就需要结合高空间分辨率、多光谱/高光谱及合成孔径雷达等多种对地观测手段进行联合观测，以扩充信息量并进行综合判断，从而进一步提升任务效果。

（4）在最近的研究中，深度学习方法逐渐成为场景分类方法的主要方法，从CNN到ViT，为了提升模型的精度，模型参数量也在爆炸性增长。尽管许多研究专注于网络模型的剪枝和轻量化处理，但是将当前的算法部署在机载或星载嵌入式系统中进行实时处理依然存在着一定的难度。

（5）针对遥感图像场景分类的实际应用，为了训练一个高精度的分类模型，往往需要大量的可标注样本，然而，海量的标注样本需要消耗大量的人力和物力投入。而且，在一些特定的应用中，受限于观测条件限制，往往导致目标样本数据量不足。因此，在目标样本数据较少的情况下，如何让场景分类模型做出更准确的判断，对实际应用有着巨大的价值。当前，在小样本学习、无监督学习和跨域场景分类都有相关研究成果，这些方向都可以提升在不同应用场景下的模型分类精度，从而缓解标注样本不足带来的问题。

（6）当前的场景分类方法很多都是受到自然图像分类方法的启发，然而对于遥感图像场景分类而言，在图像中存在着空间相关性与丰富的地学信息。从传感器的角度而言，遥感图像的成像过程也包含着丰富的物理机制。相比于自然图像，如何突出遥感图像场景分类的特点，有针对性地融入地学专家知识，是一个值得探讨的方向。

参考文献（References）

Alhichri H, Alswayed A S, Bazi Y, Ammour N and Alajlan N A. 2021. Classification of remote sensing images using EfficientNet-B3 CNN model with attention. IEEE Access, 9: 14078-14094 [DOI: 10.1109/ACCESS.2021.3051085]

Bai L, Liu Q X, Li C L, Zhu C L, Ye Z and Xi M. 2022. A lightweight and multiscale network for remote sensing image scene classification. IEEE Geoscience and Remote Sensing Letters, 19: 8012605 [DOI: 10.1109/LGRS.2021.3078518]

Bashmal L, Bazi Y, AlHichri H, AlRahhal M M, Ammour N and Alajlan N. 2018. Siamese-GAN: learning invariant representations for aerial vehicle image categorization. Remote Sensing, 10(2): 351 [DOI: 10.3390/rs10020351]

Bashmal L, Bazi Y and Rahhal M A. 2021. Deep vision transformers for remote sensing scene classification//2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. Brussels: IEEE: 2815-2818 [DOI: 10.1109/IGARSS47720.2021.9553684]

Basu S, Ganguly S, Mukhopadhyay S, DiBiano R, Karki M and Nemani R. 2015. DeepSat: a learning framework for satellite imagery//Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. Seattle: ACM: 37 [DOI: 10.1145/2820783.2820816]

Bazi Y, Bashmal L, Rahhal M M A, Dayil R A and Ajlan N A. 2021.

Vision transformers for remote sensing image classification. Remote Sensing, 13(3): 516 [DOI: 10.3390/rs13030516]

Bi M Q, Wang M H, Li Z and Hong D F. 2023. Vision transformer with contrastive learning for remote sensing image scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16: 738-749 [DOI: 10.1109/JSTARS.2022.3230835]

Bi Q, Qin K, Li Z L, Zhang H and Xu K. 2019. Multiple instance dense connected convolution neural network for aerial image scene classification//2019 IEEE International Conference on Image Processing (ICIP). Taipei, China: IEEE: 2501-2505 [DOI: 10.1109/ICIP.2019.8803322]

Bi Q, Qin K, Li Z L, Zhang H, Xu K and Xia G S. 2020a. A multiple-instance densely-connected ConvNet for aerial scene classification. IEEE Transactions on Image Processing, 29: 4911-4926 [DOI: 10.1109/TIP.2020.2975718]

Bi Q, Qin K, Zhang H, Li Z L and Xu K. 2020b. RADC-Net: a residual attention based convolution network for aerial scene classification. Neurocomputing, 377: 345-359 [DOI: 10.1016/j. neucom.2019.11.068]

Bi Q, Qin K, Zhang H, Xie J F, Li Z L and Xu K. 2020c. APDC-Net: attention pooling-based convolutional network for aerial scene classification. IEEE Geoscience and Remote Sensing Letters, 17(9): 1603-1607 [DOI: 10.1109/LGRS.2019.2949930]

Bi Q, Zhang H and Qin K. 2021. Multi-scale stacking attention pooling for remote sensing scene classification. Neurocomputing, 436: 147-161 [DOI: 10.1016/j.neucom.2021.01.038]

Cao R, Fang L Y, Lu T and He N J. 2021. Self-attention-based deep feature fusion for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters, 18(1): 43-47 [DOI: 10.1109/LGRS.2020.2968550]

Cao Y, Xu J R, Lin S, Wei F Y and Hu H. 2019. GCNet: non-local networks meet squeeze-excitation networks and beyond//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul: IEEE: 1971-1980 [DOI: 10.1109/ICCVW. 2019.00246]

Caron M, Misra I, Mairal J, Goyal P, Bojanowski P and Joulin A. 2020. Unsupervised learning of visual features by contrasting cluster assignments//Proceedings of the 34th International Conference on Neural Information Processing System. Vancouver: Curran Associates Inc.: 831

Chaib S, Liu H, Gu Y F and Yao H X. 2017. Deep feature fusion for VHR remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 55(8): 4775-4784 [DOI: 10.1109/TGRS.2017.2700322]

Chaib S, Mansouri D E K, Omara I, Hagag A, Dhelim S and Bensaber D A. 2022. On the co-selection of vision transformer features and images for very high-resolution image scene classification. Remote Sensing, 14(22): 5817 [DOI: 10.3390/rs14225817]

Chattopadhyay A, Sarkar A, Howlader P and Balasubramanian V N. 2018. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe: IEEE: 839-847 [DOI: 10.1109/WACV.2018.00097]

Chen W T, Ouyang S B, Tong W, Li X J, Zheng X W and Wang L Z. 2022. GCSANet: a global context spatial attention deep learning network for remote sensing scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15: 1150-1162 [DOI: 10.1109/JSTARS.2022.3141826]

Cheng G, Han J W, Guo L, Liu Z B, Bu S H and Ren J C. 2015. Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 53(8): 4238-4249 [DOI: 10.1109/TGRS.2015.2393857]

Cheng G, Han J W and Lu X Q. 2017a. Remote sensing image scene classification: benchmark and state of the art. Proceedings of the IEEE, 105(10): 1865-1883 [DOI: 10.1109/JPROC.2017.2675998]

Cheng G, Li Z P, Yao X W, Guo L and Wei Z L. 2017b. Remote sensing image scene classification using bag of convolutional features. IEEE Geoscience and Remote Sensing Letters, 14(10): 1735-1739 [DOI: 10.1109/LGRS.2017.2731997]

Cheng G, Sun X X, Li K, Guo L and Han J W. 2022. Perturbation-seeking generative adversarial networks: a defense framework for remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 60: 5605111 [DOI: 10.1109/TGRS.2021.3081421]

Cheng G, Xie X X, Han J W, Guo L and Xia G S. 2020. Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13: 3735-3756 [DOI: 10.1109/JSTARS.2020.3005403]

Cheng G, Yang C Y, Yao X W, Guo L and Han J W. 2018. When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs. IEEE Transactions on Geoscience and Remote Sensing, 56(5): 2811-2821 [DOI: 10.1109/TGRS.2017.2783902]

Christie G, Fendley N, Wilson J and Mukherjee R. 2018. Functional map of the world//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 6172-6180 [DOI: 10.1109/CVPR.2018.00646]

Cong Y Z, Khanna S, Meng C L, Liu P, Rozi E, He Y T, Burke M, Lobell D B and Ermon S. 2022. SatMAE: pre-training transformers for temporal and multi-spectral satellite imagery//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc.: 15

Cui Z Q, Yang W, Chen L and Li H F. 2022. MKN: metakernel networks for few shot remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 60: 4705611 [DOI: 10.1109/TGRS.2022.3153679]

Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego: IEEE:: 886-893 [DOI: 10.1109/CVPR.2005.177]

Dehouche N. 2021. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). Ethics in Science and Environmental Politics, 21: 17-23 [DOI: 10.3354/esep00195]

Deng P F, Xu K J and Huang H. 2022. When CNNs meet vision transformer: a joint framework for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters, 19: 8020305 [DOI: 10.1109/LGRS.2021.3109061]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16x16 words: transformers for image recognition at scale//9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Du B, Xiong W, Wu J, Zhang L F, Zhang L P and Tao D C. 2017. Stacked convolutional denoising auto-encoders for feature representation. IEEE Transactions on Cybernetics, 47(4): 1017-1027 [DOI: 10.1109/TCYB.2016.2536638]

Duan Y P, Tao X M, Xu M, Han C Y and Lu J H. 2018. GAN-NL: unsupervised representation learning for remote sensing image classification//2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Anaheim: IEEE: 375-379 [DOI: 10.1109/GlobalSIP.2018.8646414]

Fan R Y, Wang L Z, Feng R Y and Zhu Y Q. 2019. Attention based residual network for high-resolution remote sensing imagery scene classification//2019 IEEE International Geoscience and Remote Sensing Symposium. Yokohama: IEEE: 1346-1349 [DOI: 10.1109/IGARSS.2019.8900199]

Fang J, Yuan Y, Lu X Q and Feng Y C. 2019. Robust space–frequency joint representation for remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 57 (10): 7492-7502 [DOI: 10.1109/TGRS.2019.2913816]

Gallego A J, Pertusa A and Gil P. 2018. Automatic ship classification from optical aerial images with convolutional neural networks. Remote Sensing, 10(4): 511 [DOI: 10.3390/rs10040511]

Gao L R, Wang D G, Zhuang L N, Sun X, Huang M and Plaza A. 2023. BS^{3}LNet: a new blind-spot self-supervised learning network for hyperspectral anomaly detection. IEEE Transactions on Geoscience and Remote Sensing, 61: 5504218 [DOI: 10.1109/TGRS.2023.3246565]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Advances in neural information processing systems. [DOI: 10.48550/arXiv.1406.2661]

Guo D E, Xia Y and Luo X B. 2020. Scene classification of remote sensing images based on saliency dual attention residual network. IEEE Access, 8: 6344-6357 [DOI: 10.1109/ACCESS. 2019. 2963769]

Guo D E, Xia Y and Luo X B. 2021. GAN-based semisupervised scene classification of remote sensing image. IEEE Geoscience and Remote Sensing Letters, 18(12): 2067-2071 [DOI: 10.1109/LGRS.2020.3014108]

Han W, Feng R Y, Wang L Z and Cheng Y F. 2018. A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification. ISPRS Journal of Photogrammetry and Remote Sensing, 145: 23-43 [DOI: 10.1016/j.isprsjprs.2017.11.004]

Han W, Wang L Z, Feng R Y, Gao L, Chen X D, Deng Z, Chen J and Liu P. 2020. Sample generation based on a supervised Wasserstein Generative Adversarial Network for high-resolution remote-sensing scene classification. Information Sciences, 539: 177-194 [DOI: 10.1016/j.ins.2020.06.018]

Haralick R M, Shanmugam K and Dinstein I H. 1973. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3(6): 610-621 [DOI: 10.1109/TSMC.1973.4309314]

He N J, Fang L Y, Li S T, Plaza A and Plaza J. 2018. Remote sensing scene classification using multilayer stacked covariance pooling. IEEE Transactions on Geoscience and Remote Sensing, 56(12): 6899-6910 [DOI: 10.1109/TGRS.2018.2845668]

He N J, Fang L Y, Li S T, Plaza J and Plaza A. 2020. Skip-connected covariance network for remote sensing scene classification. IEEE Transactions on Neural Networks and Learning Systems, 31(5): 1461-1474 [DOI: 10.1109/TNNLS.2019.2920374]

He X, Chen Y S, Huang L B, Hong D F and Du Q. 2024. Foundation model-based multimodal remote sensing data classification. IEEE Transactions on Geoscience and Remote Sensing, 62: 5502117 [DOI: 10.1109/TGRS.2023.3344698]

Helber P, Bischke B, Dengel A and Borth D. 2019. 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, 12(7): 2217-2226 [DOI: 10.1109/JSTARS.2019.2918242]

Hu F, Xia G S, Hu J W and Zhang L P. 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sensing, 7(11): 14680-14707 [DOI: 10.3390/rs71114680]

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Hu X D, Zhang P L and Zhang Q. 2020. A novel framework of CNN integrated with adaboost for remote sensing scene classification//2020 IEEE International Geoscience and Remote Sensing Symposium. Waikoloa: IEEE: 2643-2646 [DOI: 10.1109/IGARSS39084.2020.9324261]

Hua Y, Mou L, Jin P and Zhu X X. 2021. MultiScene: A large-scale dataset and benchmark for multiscene recognition in single aerial images. IEEE Transactions on Geoscience and Remote Sensing, 60:1-13 [DOI: 10.1109/TGRS.2021.3110314]

Jégou H, Perronnin F, Douze M, Sánchez J, Pérez P and Schmid C. 2012. Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(9): 1704-1716 [DOI: 10.1109/TPAMI.2011.235]

Lazebnik S, Schmid C and Ponce J. 2006. Beyond bags of features: spatial pyramid matching for recognizing natural scene catego-

ries//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). New York: IEEE: 2169-2178 [DOI: 10.1109/CVPR.2006.68]

Li E Z, Xia J S, Du P J, Lin C and Samat A. 2017. Integrating multilayer features of convolutional neural networks for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 55(10): 5653-5665 [DOI: 10.1109/TGRS.2017.2711275]

Li H F, Dou X, Tao C, Wu Z X, Chen J, Peng J, Deng M and Zhao L. 2020a. RSI-CB: a large-scale remote sensing image classification benchmark using crowdsourced data. Sensors, 20(6): 1594 [DOI: 10.3390/s20061594]

Li H F, Jiang H, Gu X, Peng J, Li W B, Hong L and Tao C. 2020b. CLRS: continual learning benchmark for remote sensing image scene classification. Sensors, 20(4): 1226 [DOI: 10.3390/s20041226]

Li M T, Ma J J, Tang X, Han X, Zhu C and Jiao L C. 2022a. Resformer: bridging residual network and transformer for remote sensing scene classification//2022 IEEE International Geoscience and Remote Sensing Symposium. Kuala Lumpur: IEEE: 3147-3150 [DOI: 10.1109/IGARSS46834.2022.9883041]

Li Y G, Liang F, Zhao L C, Cui Y F, Ouyang W L, Shao J, Yu F W and Yan J J. 2022b. Supervision exists everywhere: a data efficient contrastive language-image pre-training paradigm//10th International Conference on Learning Representations. [s. l.]: OpenReview.net

Lienou M, Maitre H and Datcu M. 2010. Semantic annotation of satellite images using latent dirichlet allocation. IEEE Geoscience and Remote Sensing Letters, 7(1): 28-32 [DOI: 10.1109/LGRS.2009.2023536]

Lin D Y, Fu K, Wang Y, Xu G L and Sun X. 2017. MARTA GANs: unsupervised representation learning for remote sensing image classification. IEEE Geoscience and Remote Sensing Letters, 14(11): 2092-2096 [DOI: 10.1109/LGRS.2017.2752750]

Liu S T, Wang Q and Li X L. 2018. Attention based network for remote sensing scene classification//2018 IEEE International Geoscience and Remote Sensing Symposium. Valencia: IEEE: 4740-4743 [DOI: 10.1109/IGARSS.2018.8519232]

Liu Y S, Suen C Y, Liu Y B and Ding L W. 2019. Scene classification using hierarchical wasserstein CNN. IEEE Transactions on Geoscience and Remote Sensing, 57(5): 2494-2509 [DOI: 10.1109/TGRS.2018.2873966]

Long Y, Xia G S, Li S Y, Yang W, Yang M Y, Zhu X X, Zhang L P and Li D R. 2021. On creating benchmark dataset for aerial image interpretation: reviews, guidances, and million-AID. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 4205-4230 [DOI: 10.1109/JSTARS.2021.3070368]

Lowe D G. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2): 91-110 [DOI: 10.1023/B:VISI.0000029664.99615.94]

Lu X, Sun H and Zheng X. 2019. A Feature Aggregation Convolutional Neural Network for Remote Sensing Scene Classification. IEEE Transactions on Geoscience and Remote Sensing, 57(10): 7894-7906 [DOI: 10.1109/TGRS.2019.2917161]

Lv P Y, Wu W J, Zhong Y F, Du F and Zhang L P. 2022. SCViT: a spatial-channel feature preserving vision transformer for remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 60: 4409512 [DOI: 10.1109/TGRS.2022.3157671]

Ma A L, Yu N, Zheng Z, Zhong Y F and Zhang L P. 2022a. A supervised progressive growing generative adversarial network for remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 60: 5618818 [DOI: 10.1109/TGRS.2022.3151405]

Ma D A, Tang P and Zhao L J. 2019. SiftingGAN: generating and sifting labeled samples to improve the remote sensing image scene classification baseline *in vitro*. IEEE Geoscience and Remote Sensing Letters, 16(7): 1046-1050 [DOI: 10.1109/LGRS.2018.2890413]

Ma J J, Li M T, Tang X, Zhang X R, Liu F and Jiao L C. 2022b. Homo – heterogenous transformer learning framework for RS scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15: 2223-2239 [DOI: 10.1109/JSTARS.2022.3155665]

Mai G C, Lao N, He Y T, Song J M and Ermon S. 2023. CSP: self-supervised contrastive spatial pre-training for geospatial-visual representations//International Conference on Machine Learning. Honolulu: PMLR: 23498-23515

Mendieta M, Han B R, Shi X J, Zhu Y and Chen C. 2023. Towards geospatial foundation models via continual pretraining//IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE: 16760-16770 [DOI: 10.1109/ICCV51070.2023.01541]

Minetto R, Segundo M and Sarker S. 2019. Hydra: An ensemble of convolutional neural networks for geospatial land classification. IEEE Transactions on Geoscience and Remote Sensing, 57(9): 6530-6541 [DOI: 10.1109/TGRS.2019.2906883]

Nabi M, Maggiolo L, Moser G and Serpico S B. 2022. A CNN-transformer knowledge distillation for remote sensing scene classification//2022 IEEE International Geoscience and Remote Sensing Symposium. Kuala Lumpur: IEEE: 663-666 [DOI: 10.1109/IGARSS46834.2022.9884099]

Oliva A and Torralba A. 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision, 42(3): 145-175 [DOI: 10.1023/A:1011139631724]

Ouyang S B, Chen W T, Li X J, Dong Y S and Wang L Z. 2022. Geomorphological scene classification dataset of high-resolution remote sensing imagery in vegetation-covered areas. National Remote Sensing Bulletin, 26(4): 606-619 (欧阳淑冰, 陈伟涛, 李显巨, 董玉森, 王力哲. 2022. 植被覆盖区高精度遥感地貌场景分类数据集. 遥感学报, 26(4): 606-619) [DOI: 10.11834/jrs.20221385]

Pan X, Zhao J and Xu J. A scene images diversity improvement generative adversarial network for remote sensing image scene classification. 2019. IEEE Geoscience and Remote Sensing Letters, 17(10): 1692-1696. [DOI: 10.1109/LGRS.2019.2953192]

Penatti O A B, Nogueira K and dos Santos J A. 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?//2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Boston: IEEE: 44-

51 [DOI: 10.1109/CVPRW.2015.7301382]

Peng F, Lu W, Tan W, Qi K, Zhang X and Zhu Q. 2022. Multi-output network combining GNN and CNN for remote sensing scene classification. Remote Sensing, 14(6):1478 [DOI: 10.3390/rs14061478]

Perronnin F, Sánchez J and Mensink T. 2010. Improving the fisher kernel for large-scale image classification//11th European Conference on Computer Vision. Heraklion: Springer: 143-156 [DOI: 10.1007/978-3-642-15561-1_11]

Qi X M, Zhu P P, Wang Y B, Zhang L Q, Peng J H, Wu M F, Chen J L, Zhao X D, Zang N and Mathiopoulos P T. 2020. MLRSNet: a multi-label high spatial resolution remote sensing dataset for semantic scene understanding. ISPRS Journal of Photogrammetry and Remote Sensing, 169: 337-350 [DOI: 10.1016/j.isprsjprs.2020.09.020]

Qian X L, Li J, Cheng G, Yao X W, Zhao S N, Chen Y B and Jiang L Y. 2018. Evaluation of the effect of feature extraction strategy on the performance of high-resolution remote sensing image scene classification. Journal of Remote Sensing (in Chinese), 22(5): 758-776 (钱晓亮, 李佳, 程塨, 姚西文, 赵素娜, 陈宜滨, 姜利英. 2018. 特征提取策略对高分辨率遥感图像场景分类性能影响的评估. 遥感学报, 22(5): 758-776) [DOI: 10.11834/jrs.20188015]

Radford A, Metz L and Chintala S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks//4th International Conference on Learning Representations. San Juan: [s.n.]

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y Q, Li W and Liu P J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1): 140

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M and Sutskever I. 2021. Zero-shot text-to-image generation//38th International Conference on Machine Learning. [s.l.]: PMLR: 8821-8831

Reed C J, Gupta R, Li S F, Brockman S, Funk C, Clipp B, Keutzer K, Candido S, Uyttendaele M and Darrell T. 2023. Scale-MAE: a scale-aware masked autoencoder for multiscale geospatial representation learning//IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE: 16760-16770 [DOI: 10.1109/ICCV51070.2023.00378]

Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization//2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE: 618-626 [DOI: 10.1109/ICCV.2017.74]

Sha Z Y and Li J F. 2022. MITformer: a multiinstance vision transformer for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters, 19: 6510305 [DOI: 10.1109/LGRS.2022.3176499]

Shen J G, Yu T W, Yang H P, Wang R X and Wang Q. 2022. An attention cascade Global-Local network for remote sensing scene classification. Remote Sensing, 14(9): 2042 [DOI: 10.3390/rs14092042]

Sheng G F, Yang W, Xu T and Sun H. 2012. High-resolution satellite scene classification using a sparse coding based multiple feature combination. International Journal of Remote Sensing, 33(8): 2395-2412 [DOI: 10.1080/01431161.2011.608740]

Shi C P, Zhang X L, Sun J W and Wang L G. 2022a. A lightweight convolutional neural network based on group-wise hybrid attention for remote sensing scene classification. Remote Sensing, 14(1): 161 [DOI: 10.3390/rs14010161]

Shi C P, Zhang X L, Sun J W and Wang L G. 2022b. Remote sensing scene image classification based on self-compensating convolution neural network. Remote Sensing, 14(3): 545 [DOI: 10.3390/rs14030545]

Sumbul G, Charfuelan M, Demir B and Markl V. 2019. Bigearthnet: a large-scale benchmark archive for remote sensing image understanding//2019 IEEE International Geoscience and Remote Sensing Symposium. Yokohama: IEEE: 5901-5904 [DOI: 10.1109/IGARSS.2019.8900532]

Sumbul G, de Wall A, Kreuziger T, Marcelino F, Costa H, Benevides P, Caetano M, Demir B and Markl V. 2021. BigEarthNet-MM: a large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [Software and Data Sets]. IEEE Geoscience and Remote Sensing Magazine, 9(3): 174-180 [DOI: 10.1109/MGRS.2021.3089174]

Sun H, Li S Y, Zheng X T and Lu X Q. 2020. Remote sensing scene classification by gated bidirectional network. IEEE Transactions on Geoscience and Remote Sensing, 58(1): 82-96 [DOI: 10.1109/TGRS.2019.2931801]

Sun X, Wang P J, Lu W X, Zhu Z C, Lu X N, He Q B, Li J X, Rong X E, Yang Z J, Chang H, He Q L, Yang G, Wang R P, Lu J W and Fu K. 2023. RingMo: a remote sensing foundation model with masked image modeling. IEEE Transactions on Geoscience and Remote Sensing, 61: 5612822 [DOI: 10.1109/TGRS. 2022. 3194732]

Swain M J and Ballard D H. 1991. Color indexing. International Journal of Computer Vision, 7(1): 11-32 [DOI: 10.1007/BF00130487]

Tang X, Li M T, Ma J J, Zhang X R, Liu F and Jiao L C. 2022. EMTCAL: efficient multiscale transformer and cross-level attention learning for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 60: 5626915 [DOI: 10.1109/TGRS.2022.3194505]

Tang X, Ma Q S, Zhang X R, Liu F, Ma J J and Jiao L C. 2021. Attention consistent network for remote sensing scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 2030-2045 [DOI: 10.1109/JSTARS.2021.3051569]

Teng W X, Wang N, Shi H H, Liu Y C and Wang J. 2020. Classifier-constrained deep adversarial domain adaptation for cross-domain semisupervised classification in remote sensing images. IEEE Geoscience and Remote Sensing Letters, 17(5): 789-793 [DOI: 10.1109/LGRS.2019.2931305]

Tian T, Li L L, Chen W T and Zhou H B. 2021. SEMSDNet: a multi-scale dense network with attention for remote sensing scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 5501-5514 [DOI: 10.1109/JSTARS.2021.3074508]

Tong W, Chen W T, Han W, Li X J and Wang L Z. 2020. Channel-at-

tention-based densenet network for remote sensing image scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13: 4121-4132 [DOI: 10.1109/JSTARS.2020.3009352]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc.: 6000-6010

Wang D, Zhang J, Du B, Xia G S and Tao D C. 2023b. An empirical study of remote sensing pretraining. IEEE Transactions on Geoscience and Remote Sensing, 61: 5608020 [DOI: 10.1109/TGRS.2022.3176603]

Wang D, Zhang Q M, Xu Y F, Zhang J, Du B, Tao D C and Zhang L P. 2023c. Advancing plain vision transformer toward remote sensing foundation model. IEEE Transactions on Geoscience and Remote Sensing, 61: 5607315 [DOI: 10.1109/TGRS.2022.3222818]

Wang D G, Gao L R, Qu Y, Sun X and Liao W Z. 2023a. Frequency-to-spectrum mapping GAN for semisupervised hyperspectral anomaly detection. CAAI Transactions on Intelligence Technology, 8(4): 1258-1273 [DOI: 10.1049/cit2.12154]

Wang D G, Zhuang L N, Gao L R, Sun X, Huang M and Plaza A J. 2023. PDBSNet: pixel-shuffle downsampling blind-spot reconstruction network for hyperspectral anomaly detection. IEEE Transactions on Geoscience and Remote Sensing, 61: 5511914 [DOI: 10.1109/TGRS.2023.3276175]

Wang G Q, Zhang N, Liu W C, Chen H and Xie Y Z. 2022. MFST: a multi-level fusion network for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters, 19: 6516005 [DOI: 10.1109/LGRS.2022.3205417]

Wang J, Liu W C, Ma L, Chen H and Chen L. 2018. IORN: an effective remote sensing image scene classification framework. IEEE Geoscience and Remote Sensing Letters, 15(11): 1695-1699 [DOI: 10.1109/LGRS.2018.2859024]

Wang Q, Liu S T, Chanussot J and Li X L. 2019. Scene classification with recurrent attention of VHR remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 57(2): 1155-1167 [DOI: 10.1109/TGRS.2018.2864987]

Wei Y, Luo X, Hu L, Peng Y and Feng J. 2020. An improved unsupervised representation learning generative adversarial network for remote sensing image scene classification. Remote Sensing Letters, 11(6) :598-607 [DOI: 10.1080/2150704X.2020.1746854]

Xia G S, Hu J W, Hu F, Shi B G, Bai X, Zhong Y F, Zhang L P and Lu X Q. 2017. AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing, 55(7): 3965-3981 [DOI: 10.1109/TGRS.2017.2685945]

Xia G S, Yang W, Delon J, Gousseau Y, Sun H and Maître H. 2010. Structural high-resolution satellite image indexing[EB/OL]. [2024-02-28]. https://hal.science/file/index/docid/458685/filename/structural_satellite_indexing_XYDG.pdf

Xiao Z, Long Y, Li D, Wei C, Tang G and Liu J. 2017. High-resolution remote sensing image retrieval based on CNNs from a dimensional perspective. Remote Sensing, 9(7): 725 [DOI: 10.3390/rs9070725]

Xie J, He N J, Fang L Y and Plaza A. 2019. Scale-free convolutional neural network for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 57(9): 6916-6928 [DOI: 10.1109/TGRS.2019.2909695]

Xu K J, Deng P F and Huang H. 2022. Vision transformer: an excellent teacher for guiding small networks in remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 60: 5618715 [DOI: 10.1109/TGRS.2022.3152566]

Xu S H, Mu X D, Chai D and Zhang X M. 2018. Remote sensing image scene classification based on generative adversarial networks. Remote Sensing Letters, 9(7): 617-626 [DOI: 10.1080/2150704X.2018.1453173]

Yang Y and Newsam S. 2010. Bag-of-visual-words and spatial extensions for land-use classification//Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. San Jose: ACM: 270-279 [DOI: 10.1145/1869790.1869829]

Yu Y L, Li X Z and Liu F X. 2020. Attention GANs: unsupervised deep feature learning for aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing, 58(1): 519-531 [DOI: 10.1109/TGRS.2019.2937830]

Yu Y T, Li Y Y, Wang J, Guan H Y, Li F F, Xiao S Z, Tang E and Ding X W. 2022. C^2-CapsViT: cross-context and cross-scale capsule vision transformers for remote sensing image scene classification. IEEE Geoscience and Remote Sensing Letters, 19: 6512005 [DOI: 10.1109/LGRS.2022.3185454]

Yuan J W, Ru L X, Wang S G and Wu C. 2022. WH-MAVS: a novel dataset and deep learning benchmark for multiple land use and land cover applications. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15: 1575-1590 [DOI: 10.1109/JSTARS.2022.3142898]

Yuan Y. 2023. On the power of foundation models//40th International Conference on Machine Learning. Honolulu: PMLR: 40519-40530

Yuan Y, Fang J, Lu X Q and Feng Y C. 2019. Remote sensing image scene classification using rearranged local features. IEEE Transactions on Geoscience and Remote Sensing, 57(3): 1779-1792 [DOI: 10.1109/TGRS.2018.2869101]

Zhang B. 2018. Remotely Sensed big data era and intelligent information extraction. Geomatics and Information Science of Wuhan University, 43(12): 1861-1871 (张兵. 2018. 遥感大数据时代与智能信息提取. 武汉大学学报(信息科学版), 43(12): 1861-1871) [DOI: 10.13203/j.whugis20180172]

Zhang B, Zhang Y J and Wang S G. 2019a. A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(8): 2636-2653 [DOI: 10.1109/JSTARS.2019.2919317]

Zhang F, Du B and Zhang L P. 2015. Saliency-guided unsupervised feature learning for scene classification. IEEE Transactions on Geoscience and Remote Sensing, 53(4): 2175-2184 [DOI: 10.1109/TGRS.2014.2357078]

Zhang H, Zu K K, Lu J, Zou Y R and Meng D Y. 2023. EPSANet: an

efficient pyramid squeeze attention block on convolutional neural network//16th Asian Conference on Computer Vision. Macao, China: Springer: 541-557 [DOI: 10.1007/978-3-031-26313-2_33]

Zhang J R, Zhao H W and Li J. 2021. TRS: transformers for remote sensing scene classification. Remote Sensing, 13(20): 4143 [DOI: 10.3390/rs13204143]

Zhang W, Tang P and Zhao L J. 2019b. Remote sensing image scene classification using CNN-CapsNet. Remote Sensing, 11(5): 494 [DOI: 10.3390/rs11050494]

Zhang Y S, Sun X, Wang H Q and Fu K. 2013. High-resolution remote-sensing image classification via an approximate earth mover's distance-based bag-of-features model. IEEE Geoscience and Remote Sensing Letters, 10(5): 1055-1059 [DOI: 10.1109/LGRS. 2012. 2228625]

Zhao B, Zhong Y F, Xia G S and Zhang L P. 2016a. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 54(4): 2108-2123 [DOI: 10.1109/TGRS. 2015.2496185]

Zhao L J and Tang P. 2016. Scalability analysis of typical remote sensing data classification methods: a case of remote sensing image scene. Journal of Remote Sensing (in Chinese), 20(2): 157-171 (赵理君, 唐娉. 2016. 典型遥感数据分类方法的适用性分析——以遥感图像场景分类为例. 遥感学报, 20(2): 157-171) [DOI: 10.11834/jrs.20164279]

Zhao L J, Tang P and Huo L Z. 2016b. Feature significance-based multibag-of-visual-words model for remote sensing image scene classification. Journal of Applied Remote Sensing, 10(3): 035004 [DOI: 10.1117/1.JRS.10.035004]

Zhao Z C, Li J Q, Luo Z, Li J and Chen C. 2021. Remote sensing image scene classification based on an enhanced attention module. IEEE Geoscience and Remote Sensing Letters, 18(11): 1926-1930 [DOI: 10.1109/LGRS.2020.3011405]

Zheng Q K, Xia X, Zou X, Dong Y X, Wang S, Xue Y F, Wang Z H, Shen L, Wang A D, Li Y, Su T, Yang Z L and Tang J. 2023. CodeGeeX: a pre-trained model for code generation with multilingual benchmarking on HumanEval-X[EB/OL]. [2024-02-28]. https://arxiv.org/abs/2303.17568

Zhou B L, Khosla A, Lapedriza A, Oliva A and Torralba A. 2016. Learning deep features for discriminative localization//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE: 2921-2929 [DOI: 10.1109/CVPR.2016.319]

Zhou J H, Wei C, Wang H Y, Shen W, Xie C H, Yuille A and Kong T. 2021. iBOT: image BERT pre-training with online tokenizer[EB/OL]. [2024-02-28]. https://arxiv.org/abs/2111.07832

Zhou W X, Newsam S, Li C M and Shao Z F. 2017. PatternNet: a benchmark dataset for performance evaluation of remote sensing image retrieval[EB/OL]. [2024-02-28]. https://arxiv.org/abs/1706.03424

Zhu Q Q, Zhong Y F, Zhang L P and Li D R. 2018. Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 56(10): 6180-6195 [DOI: 10.1109/TGRS. 2018. 2833293]

Zhu Q Q, Zhong Y F, Zhao B, Xia G S and Zhang L P. 2016. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. IEEE Geoscience and Remote Sensing Letters, 13(6): 747-751 [DOI: 10.1109/LGRS.2015.2513443]

Zhu R X, Yan L, Mo N and Liu Y. 2019. RETRACTED: attention-based deep feature fusion for the scene classification of high-resolution remote sensing images. Remote Sensing, 11(17): 1996 [DOI: 10.3390/rs11171996]

Zou Q, Ni L H, Zhang T and Wang Q. 2015. Deep learning based feature selection for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters, 12(11): 2321-2325 [DOI: 10.1109/LGRS.2015.2475299]

Research progress of high−resolution remote sensing image scene classification

LI Zhi[1,2], GAO Lianru[1], ZHENG Ke[3], NI Li[1]

1. Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China;
2. College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China;
3. School of Geography and Environment, Liaocheng University, Liaocheng 252000, China

Abstract: With the rapid advancement of remote sensing technology, the resolution of remote sensing satellites is improving, the number of spectral bands is increasing, and revisit periods are contracting. This progression empowers researchers to access more valuable data and information from remote sensing images. Concepts, such as remote sensing big data, remote sensing foundation models, and smart cities, have successively emerged in recent years, imposing increased demands on the intelligent extraction technology of massive remote sensing data, particularly regarding remote sensing image information.

As an indispensable element of intelligent information extraction technology applied in fields, such as land use and cover, national land

resource surveys, natural disaster observation, agricultural yield estimation, and forestry protection, remote sensing image classification exhibits substantial practical importance. Remote sensing image scene classification has been introduced in this context. The objective of scene classification in remote sensing images is to comprehensively and semantically categorize each given remote sensing image. This task entails summarizing and analyzing the extracted feature information at a high level and assigning different labels to areas of interest based on their features.

In contrast with natural images, although they contain features, such as color, texture, and shape, remote sensing images encounter more challenges in classification due to the intricate scene content resulting from the overhead perspective, weak texture, and color information caused by low resolution. Nevertheless, as one of the technical means in remote sensing applications, remote sensing image scene classification technology plays a pivotal role in the development of practical application technologies.

After years of development, numerous comprehensive review studies on remote sensing image scene classification have been conducted locally and abroad. However, the recent surge in remote sensing big data has introduced new challenges into scene classification. The ongoing evolution of deep learning technology, particularly the widespread application of Convolutional Neural Networks (CNNs) and transformers, has resulted in significant advancements in remote sensing image scene classification. In this context, self-supervised learning, as a method that is independent of annotated data, has become indispensable in the field of remote sensing image scene classification. Foundation models based on self-supervised learning have been successfully implemented in scene classification, presenting innovative solutions to this field. As the volume of remote sensing data continues to increase, the dataset scale for remote sensing image scene classification is expanding rapidly, giving rise to increasingly intricate classification tasks. Remote sensing image scene classification datasets are swiftly progressing toward the integration of multiple sources, the incorporation of multiple labels, and the inclusion of large-scale samples.

Drawing from the findings of the current literature survey, this study systematically compiles a summary of deep learning methods within the domain of remote sensing image scene classification. Encompassing CNNs, visual transformers, and generative adversarial networks, this overview also introduces representative datasets and foundation models since the inception of scene classification. Several classical scene classification methods have undergone evaluation across various benchmark datasets. In addition, this study delves into primary challenges and prospects, paving the way for further research in the classification of scenes in remote sensing images.

Key words: high-resolution remote sensing image, image classification, scene classification, deep learning

Supported by National Key Research and Development Program of China (No. 2021YFB3900502)