引用格式：XIA Ying, LI Junyao, GUO Dongen. Semi-supervised Scene Classification of Remote Sensing Images Based on GAN[J]. Acta Photonica Sinica, 2022, 51(3):0310003
夏英,李骏垚,郭东恩. 基于GAN的半监督遥感图像场景分类[J].光子学报,2022,51(3):0310003

基于GAN的半监督遥感图像场景分类

夏英[1],李骏垚[1],郭东恩[1,2]
(1 重庆邮电大学 空间大数据智能技术重庆市工程研究中心,重庆 400065)
(2 南阳理工学院 计算机与软件学院,河南 南阳 473000)

摘　要:针对遥感图像背景复杂及有监督场景分类算法无法利用无标签数据的问题,提出一种基于生成对抗网络的半监督遥感图像场景分类方法。首先,引入谱归一化残差块代替传统生成对抗网络中的二维卷积,利用残差块的跳跃连接解决梯度消失问题;其次,引入特征融合思想,将浅层特征与深层特征进行融合,从而减少特征损失;最后,在生成对抗网络的判别器中加入结合门控的注意力模块,以增强特征判别能力。在EuroSAT和UC Merced数据集上的实验结果表明,该方法能够有效提取判别力更强的特征,提高半监督分类性能。

关键词:遥感图像;场景分类;半监督;生成对抗网络;注意力机制
中图分类号:TP391.4　　**文献标识码**:A　　**doi**:10.3788/gzxb20225103.0310003

0 引言

高分辨率遥感卫星技术迅速发展,产生了大量场景丰富的高分辨率遥感图像,如何充分利用不断增长的遥感图像变得尤为重要。近年来,智能解释遥感图像已成为重要研究内容,场景分类是活跃的研究领域之一。遥感图像场景分类主要利用语义信息,将图像的场景作为一个整体进行分类,被广泛应用在智能城市建设、灾情监测与评估、目标判读和土地资源利用等领域[1]。目前,基于卷积神经网络对遥感图像进行有监督分类,需要大量有标签数据,并且已经达到较高的分类精度。然而,遥感图像的标注需要丰富的工程技能和专家知识,在遥感应用中,大部分情况下仅存在少量的有标签遥感图像进行有监督训练,大量无标签图像无法得到充分利用。因此,通过学习少量标注数据,从大量未标注数据提取有效特征的半监督学习方法,成为解决这类问题的潜在途径。

生成对抗网络(Generative Adversarial Networks, GAN)[2]是近年来最具有潜力的半监督方法之一,通过生成对抗的方式训练模型。GAN在训练时,通过生成器产生大量样本扩充数据集,解决有标签样本少的问题。同时,对抗训练提高了判别器的泛化能力和抗干扰能力,进而增强特征提取能力。因此,针对遥感领域有标签样本量不足、人工标注困难以及难以提取判别力强的特征等问题,相关研究人员已经将GAN应用在遥感图像场景分类领域。

RADFORD A等[3]在生成对抗网络中加入卷积层和归一化层,优化了网络结构,提高特征提取能力;ODENA A等[4]将GAN应用在半监督分类中,用少量有标签数据和大量无标签数据训练模型;TAO Y等[5]将GAN应用于遥感图像场景分类,用来解决有标签遥感图像样本少的问题。但由于遥感图像背景复杂,场景类别繁多,上述基于GAN的算法存在训练不稳定、假样本质量低以及不能收敛等问题,限制了分类性能的提高。

基于上述问题,ROY S等[6]提出了Semantic Fusion Generation Adversarial Network(SFGAN)算法,引

基金项目:国家自然科学基金(No.41871226),河南省科技攻关项目(No. 212102210492),重庆市教委重点合作项目(No. HZ2021008)
第一作者(通讯作者):夏英(1972—),女,教授,博士,主要研究方向为时空大数据、跨媒体计算等。Email: xiaying@cqupt.edu.cn
收稿日期:2021-06-19;**录用日期**:2021-07-30

入语义融合方法，增强分类性能。MIYATO T等[7]针对生成对抗训练时出现的模式坍塌等问题，提出了Spectral Normalization Generation Adversarial Network（SNGAN）算法，可以增强GAN训练的稳定性。MAO X等[8]提出Least squares Generation Adversarial Network（LSGAN），缓解了生成图像质量差、多样性不足的问题。LECOUAT B等[9]提出Manifold Regularization Generation Adversarial Network（REG-GAN），通过流行正则化提高生成图像的质量。GUO D等[10]提出基于门控单元的自注意力Self-Attention Gating Generation Adversarial Network（SAGGAN），增强对鲁棒性强的特征的提取，提升模型收敛速度。

综上，为了进一步增强生成对抗训练的稳定性，充分利用大量无标签数据提取判别力更强的特征，以SFGAN算法为基础，提出一种残差注意力生成对抗网络（Residual Attention Generation Adversarial Network，RAGAN）。该方法具有以下特点：1）在网络结构中，引入谱归一化的残差块（Spectral Normalized Residual Block，SNRB），增强生成对抗训练的稳定性，同时解决梯度消失问题；2）将浅层特征和深层特征融合，更全面地反映场景信息，进一步增强特征表示能力；3）引入结合门控的注意力模块（Gate Attention Module，GAM），让判别器聚焦于鲁棒性好、判别力强的特征，为其赋予更高的权重，同时过滤干扰信息。

1 相关技术基础

1.1 生成对抗网络

GAN是一种基于博弈论的深度学习模型。GAN采用了一个生成网络G来生成对抗样本，同时采用一个判别网络D来判别样本是否真实。G的训练目标就是生成接近真实的假样本欺骗D，而D则是尽可能地区分真实样本和G生成的假样本，博弈到最后的解是达到纳什平衡。此时，判别器D的判别能力足够强，可以区分出真假样本，并且生成器G生成的样本足够真实，判别器D难以判断其真假。整个生成对抗的训练过程可表示为

$$\min_G \max_D V(G,D)=E_{x\sim p_{\text{data}}(x)}[\log D(x)]+E_{z\sim P_z(z)}[\log(1-D(G(z)))] \tag{1}$$

式中，G、D分别表示生成网络和判别网络，$V(G,D)$表示G与D的差异值，E表示数学期望，z表示随机噪声，$p_{\text{data}}(x)$表示真实样本的分布，$P_z(z)$表示生成器生成的假样本分布。

SALIMANS T等[11]扩展上述框架应用于半监督学习，每个类别对应一个神经元，最终将K个完整神经元添加到判别器D里。D的输入由未标注的样本、已标注的样本以及生成的假样本组成，输出由K个真实类和代表假样本的$K+1$类组成。因此，D的损失函数分为有监督的损失和无监督的损失，即

$$L_D=L_{\text{sup}}+L_{\text{un sup}} \tag{2}$$

其中有监督的损失函数为

$$L_{\text{sup}}=-E_{x,y\sim P_{\text{data}}(x,y)}[\log(p_D(y|x,y<K+1))] \tag{3}$$

无监督的损失函数为

$$\begin{aligned}L_{\text{un sup}}=&-E_{x\sim P_{\text{data}}(x)}[\log(1-p_D(y=K+1|x))]\\&-E_{z\sim p_z(z)}[\log(p_D(y=K+1|G(z)))]\end{aligned} \tag{4}$$

式中，$p_D(y=K+1|x)$代表G生成假样本的概率，$p_D(y|x,y<K+1)$代表真实样本的概率。

1.2 SFGAN网络结构

SFGAN用于半监督遥感图像场景分类模型，引入语义分支增强判别器的特征提取能力。其网络结构如图1所示，判别器将原始的64×64×3遥感图像x和语义信息$f(x)$作为输入，引入Inception V3网络在ImageNet数据集提取的语义信息$s(x)$。通过融合两种不同通道的语义信息丰富特征表示能力，从而提高分类的性能。

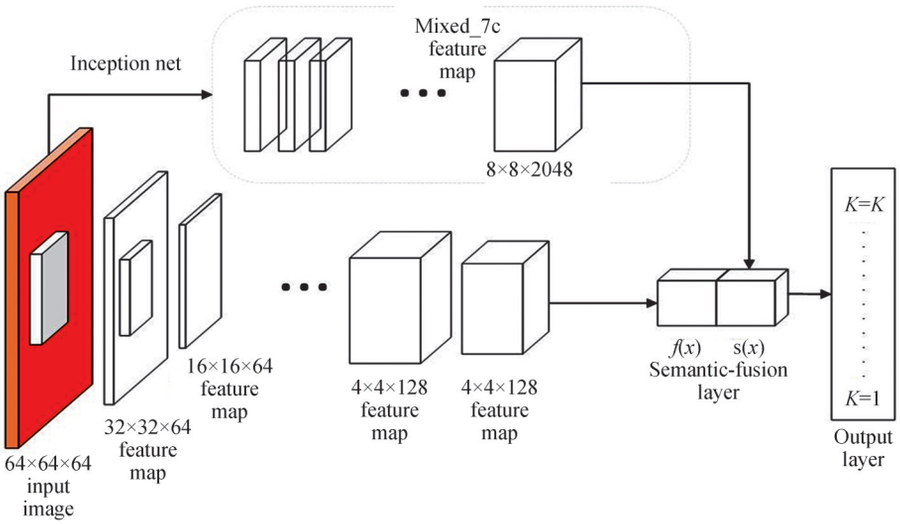


图1　SFGAN网络结构

Fig. 1　SFGAN network structure

2　方法

2.1　RAGAN网络结构

SFGAN算法虽然采用深度卷积生成对抗网络对遥感图像进行特征提取和场景分类，但网络层数的增加会导致梯度消失和特征损失的问题，无法更好地提取特征。

为解决这些问题，实现良好的分类性能，提出一种用于半监督的遥感图像场景分类方法，即一种残差注意力生成对抗网络RAGAN。该方法主要对SFGAN的判别器D做了以下三个方面的改进：1）采用谱归一化的残差块SNRB代替标准的二维卷积，每个残差块包含两层卷积，能够更充分地提取特征，解决梯度消失问题；2）将多层谱归一化残差块提取的深层特征和标准二维卷积提取的浅层特征进行融合，更全面地反映场景信息，同时减少训练造成的特征损失；3）引入结合门控的注意力模块GAM，让判别器充分提取融合后

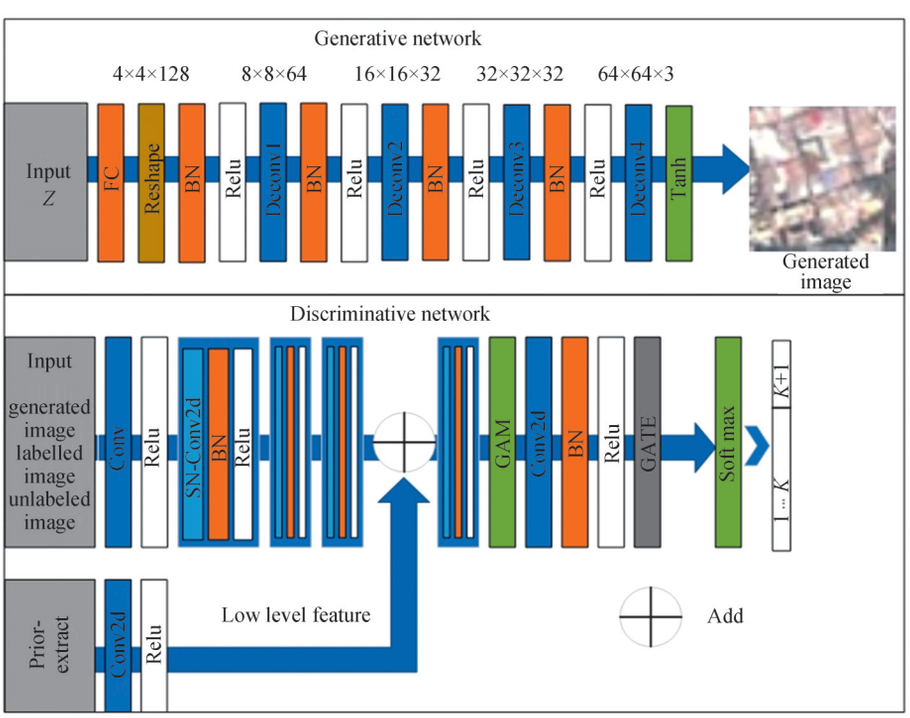


图2　RAGAN网络结构

Fig. 2　RAGAN network structure

的特征再进行权重分配，减少不相关信息的干扰。

为了让生成对抗网络更适合于图像的生成和处理，采取深度卷积生成对抗网络来构建模型[3]，它由全连接层（Fully Connected，FC）、反卷积层（Deconvolution，Deconv）、批归一化层（Batch Normalization，BN）以及激活函数ReLU与Tanh组成，RAGAN的网络结构如图2所示。

2.2 谱归一化残差块

受深度残差网络[12]的启发，深度神经网络引入跳跃结构形成残差模块，该模块由网络层、跳跃结构和激活函数Relu组成，如图3所示。x表示输入的数据，Relu表示线性激活函数，$F(x)$表示网络残差，$H(x)$表示学习到的特征，可表示为$H(x)=F(x)+x$。如果网络训练达到饱和的分类精度或下层的误差较大时，只需$F(x)=0$，使x的值近似等于$H(x)$，保证往后的网络层数不会造成精度下降，有效避免了退化现象[13]。

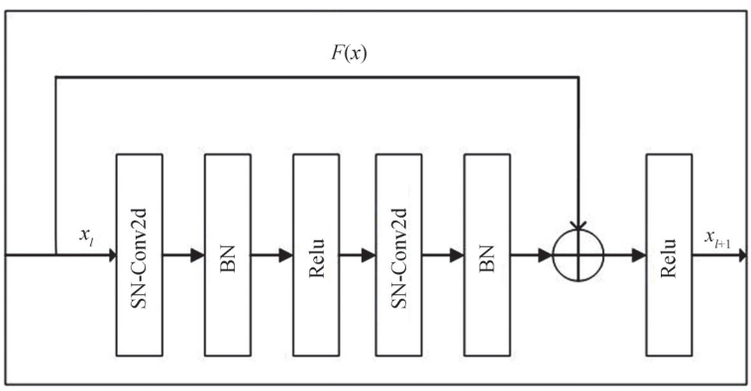


图3 残差神经网络的基本结构单元

Fig. 3 Basic structure of residual neural network

生成对抗网络的性能很大程度上依赖于判别器D的稳定性。判别器中C-Lipschitz的条件是

$$\frac{\|f(x)-f(x')\|}{\|x-x'\|}\leqslant C \tag{5}$$

式中，x和x'表示输入，$f(x)$和$f(x')$表示输出，Lipschitz常数C是f的最小常数，仅由谱范数决定。GAN的稳定性定理指出，当判别网络的输入输出满足1-Lipschitz连续时，GAN的训练就会稳定。因此，为增强训练的稳定性，在残差块的每一个卷积层引入谱归一化（Spectral Normalization，SN）[7]约束权重矩阵的谱范数。保证每一批次数据的输入和输出对权重的计算满足1-Lipschitz连续，使得生成对抗训练始终都保持平滑，提升训练的稳定性[14]。

由Lipschitz的复合函数可知，单个函数满足1-Lipschitz连续，那么它们所组成的复合函数同样满足1-Lipschitz连续。RAGAN的判别器由多层谱归一化残差块SNRB组成，每个残差块包括两个二维卷积和一个Relu激活函数，相当于一个复合函数。激活函数Relu满足1-Lipschitz连续，因此只需要将谱归一化应用在每一个卷积层，保证卷积层满足1-Lipschitz连续，那么整个判别网络就满足1-Lipschitz连续。卷积层的谱范数$\sigma(\boldsymbol{W})$计算方式为

$$\sigma(\boldsymbol{W})=\sup_{h\neq 0}\frac{\|\boldsymbol{W}h\|_2}{\|h\|_2}=\sup_{\|h\|_2\leqslant 1}\|\boldsymbol{W}h\|_2=C \tag{6}$$

式中，h表示输入，$\boldsymbol{W}$表示参数矩阵，$\sigma(\boldsymbol{W})$表示矩阵$\boldsymbol{W}$的谱范数，sup表示最小上界，卷积层的Lipschitz常数C等于该卷积层参数矩阵$\boldsymbol{W}$的谱范数$\sigma(\boldsymbol{W})$。谱归一化$\boldsymbol{W}_{\mathrm{SN}}(\boldsymbol{W})$计算方式为

$$\boldsymbol{W}_{\mathrm{SN}}(\boldsymbol{W})=\frac{\boldsymbol{W}}{\sigma(\boldsymbol{W})} \tag{7}$$

通过谱归一化使参数矩阵$\boldsymbol{W}$的谱范数$\sigma(\boldsymbol{W})$标准化，使其满足1-Lipschitz连续，即$\sigma(\boldsymbol{W}_{\mathrm{SN}}(\boldsymbol{W}))=1$。此时，卷积层满足1-Lipschitz连续，由Lipschitz的复合定理可知，整个判别网络也满足1-Lipschitz连续。

然而，谱范数的求解过程涉及矩阵奇异值分解，所需计算量较大。可采用幂迭代法近似求解，提高计算效率，实现谱归一化。随机初始化向量$\boldsymbol{m}$和$\boldsymbol{n}$，分别作为参数矩阵$\boldsymbol{W}$的左奇异值向量和右奇异值向量，即

$$\begin{cases} \boldsymbol{m}=\dfrac{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{n}}{\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{n}\|_2} \\ \boldsymbol{n}=\dfrac{\boldsymbol{W}\boldsymbol{m}}{\|\boldsymbol{W}\boldsymbol{m}\|_2} \end{cases} \tag{8}$$

经过式(8)多次迭代后，可估算出矩阵 $\boldsymbol{W}$ 的谱范数 $\sqrt{\lambda_1}$，即

$$\sqrt{\lambda_1}=\boldsymbol{n}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{m} \tag{9}$$

综上，为了缓解网络层数增加带来的梯度消失和训练不稳定问题，采用谱归一化残差块SNRB代替原本判别器 D 中的二维卷积，不仅保留全部的原始信息还减少网络参数，解决网络退化的问题，增强训练过程的稳定性。

2.3 特征融合与结合门控的注意力机制

图像中的特征包括浅层和深层的语义信息，底层卷积提取的浅层特征(Shallow Feature)包含更多位置和局部信息，但语义信息较弱。而随着网络层数的加深，提取到的深层特征(Deep Feature)包含较强的语义信息和全局信息，但对细节的感知能力较差。为了更全面地反映遥感图像场景信息，有必要对网络模型提取到的不同特征进行融合(Fusion)[15]。因此，在进入深层网络训练前，先进行一次普通的二维卷积提取浅层特征，然后与多层谱归一化残差块提取的深层特征进行融合，减少特征的损失，让模型学习到不同特征之间的互补关系，从而提升模型的表征能力。

由于遥感成像技术的进步，遥感图像的分辨率随之提高，类别也逐渐增多，导致图像背景复杂，神经网络模型很难聚焦到鲁棒性良好的特征。注意力机制(Attention)在很多计算机视觉任务中被证明可以有效提升网络性能，该方法模仿了人类视觉所特有的大脑信号处理过程，通过快速扫描全局图像，明确需要重点关注的区域，然后对这一区域投入更多的资源来获得充分的细节信息，从而过滤掉冗余无用的信息[16]。

传统的卷积是将通道信息和空间信息混合在一起提取信息特征，受WOO S[17]、LIU W [18]和GUO D等[19]的启发，卷积模块的注意力机制模块(Convolutional Block Attention Module，CBAM)是由通道注意力(Channel Attention)模块和空间注意力(Spatial Attention)模块两个部分组成，重点沿着通道和空间这两个维度分别进行特征聚焦，通道注意力关注什么样的特征以及空间注意力模块关注哪里的特征是有意义的。由于它是一个轻量级的通用模块，可以无缝地集成到任何神经网络模块中，因此将其引入到生成对抗网络的判别器 D 中，引导模型更有针对性地关注重要特征并抑制不必要的特征。同时，为了获得更强的特征表达能力，捕获特征之间的依赖关系，引入门控机制(Gate Block)，构造结合门控的注意力模块GAM，特征融合和GAM的结构如图4所示。

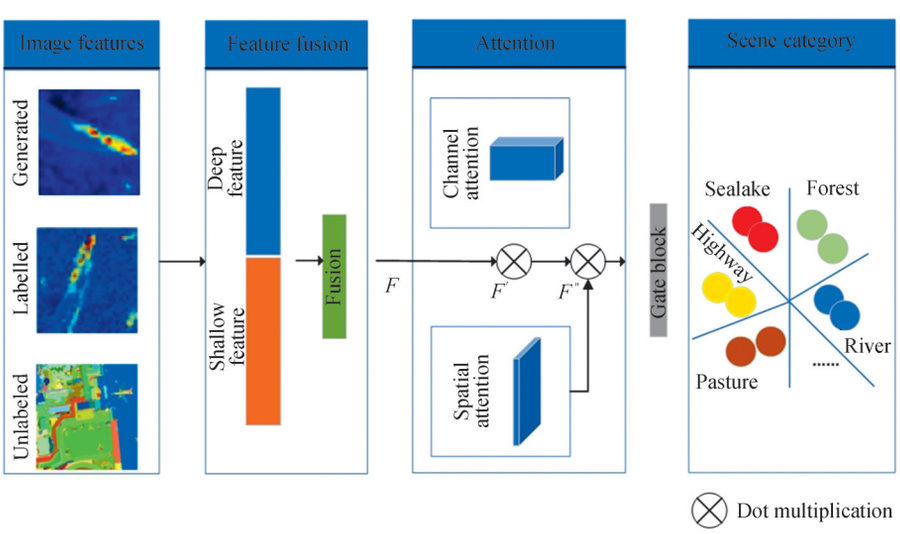


图4 特征融合和GAM结构

Fig. 4 Feature fusion and GAM structure

首先，对有标签图像（Labelled）、无标签图像（Unlabeled）以及生成的假样本（Generated）进行特征提取，并将浅层特征和深层特征融合；然后，通道注意力模块对融合的特征进行平均尺化和最大尺化产生两个空间元素，通过元素求和得到$M_C(F)$，最后，与输入的特征F进行点乘得到F'；同理，空间注意力模块以F'作为输入，求和得到$M_S(F')$，与F'点乘得到输出F''，可用公式表示为

$$F'=M_C(F)\otimes F \tag{10}$$

$$F''=M_S(F')\otimes F' \tag{11}$$

式中，F表示输入的特征，M_C表示通道注意力聚焦，F'表示通道注意力聚焦后的输出，M_S表示空间注意力聚焦，F''表示空间注意力聚焦后的输出，$\otimes$表示点乘。

为了增强注意力模块聚焦后特征之间的相关性，提高判别器D的表征能力，引入门控机制，特征图F''被输入到门控单元，并转换为内部关联性更强的新特征。最终，分类器通过关联性更强的特征分类出不同的场景类别（Scene Category），比如海洋湖泊（SeaLake）、高速公路（Highway）、牧场（Pasture）、河流（River）、森林（Forest）等。门控单元推导过程可表示为

$$f_{\text{gate}}(F'')=\sigma(\text{dense}(F'')) \tag{12}$$

$$F_{\text{gated}}=f_{\text{gate}}(F'')\otimes F'' \tag{13}$$

式中，$\sigma(x)$表示sigmoid激活函数，$f_{\text{gate}}(x)$表示门控机制，dense(x)表示完全连接操作。

3 实验分析

3.1 实验数据集

EuroSAT数据集[20]由Sentinel-2卫星获取的27 000幅带标记的卫星图像组成，覆盖了13个光谱带，将其分为10个不同的土地利用类别，如工业、住宅等，图像的分辨率为64×64。UC Merced Land-Use Dataset（UCM）[21]是一个用于研究的21级土地利用图像遥感数据集，其被用于美国各地的城市地区，图像的分辨率大小为256×256，包含21个类别的场景图像共计2 100张，其中每个类别有100张。数据集信息见表1。

表1 数据集信息

Table 1 Dataset information

Dataset	Images per category	Number of categories	Total images	Size
EuroSAT	2 000～3 000	10	27 000	64×64
UC Merced	100	21	2100	256×256

3.2 参数设置

针对经验风险最小化算法的过拟合的问题，采用简单交叉验证的方法，从EuroSAT和UC Merced数据集中随机选取80% 作为训练集，剩下的20% 作为测试集。然后，用训练集来训练模型，在测试集验证模型及参数。接着，再把样本打乱，重新选择训练集和测试集，继续训练数据和检验模型。通过反复交叉验证，用损失函数来度量得到模型的好坏，最终确立一个较好的模型。为了验证所提出的RAGAN方法的优越性，在EuroSAT数据集中，通过随机种子随机为图像进行标注，标记的样本数量M（Numbers of label M on EuroSAT）分别设置为100、1 000、2 000、21 600（全部训练集）；同理，在UC Merced数据集中，随机标注的样本数量M（Numbers of label M on UC Merced）设置为100、200、400、1 680（全部训练集）。实验在64位Ubuntu18.04操作系统下进行，框架采用TensorFlow-GPU 1.8.0，GPU为11GB的NVIDIA GeForce GTX 2080Ti。参数设置参考了SFGAN，即$\beta_1=0.5$，$\beta_2=0.9$，批处理大小bitch-size为128，训练周期epoch设置为30，初始学习率lr-rate设置为0.000 3，每次衰减设为0.9。

3.3 评价指标

在图像分类任务中，目前被学者广泛使用的评价指标是总体分类精度（Overall Accuracy，OA）和混淆矩阵（Confusion Matrix，CM）。

1）总体分类精度，即指被正确分类的类别像元数与总的类别个数的比值，计算公式为

$$\text{acc}(f;D)=\frac{1}{m}\sum_{i=1}^{m}\prod(f(x_i)=y_i) \tag{14}$$

2）混淆矩阵，也称误差矩阵，用n行n列的矩阵形式来表示，主要通过映射每个实测像元的位置和类别与分类图像中相应的位置和类别，来显示分类结果的准确性。

3.4 实验结果与分析

将RAGAN方法与其他几种具有代表性的图像分类方法在EuroSAT与UCM数据集进行性能比较，并通过总体分类精度OA和混淆矩阵CM来分析实验结果。CNN[6]作为传统的深度学习模型，是一种有监督的训练方法。Inception V3[6]采用了迁移学习，在Image Net自然图像数据集预训练了一个良好的模型。生成对抗网络作为最具潜力的半监督算法，在进行半监督遥感图像场景分类时，可以生成一定量的假样本，解决了样本数量不足的问题。FMGAN[2]、REG-GAN[9]、SFGAN[6]和SAGGAN[10]都基于生成对抗网络（GAN）的基础上改进，其中SFGAN和SAGGAN都沿用了Inception v3分支，增强了特征提取能力。各类方法的总体分类精度对比结果如表2所示。

表2 在EuroSAT和UCM数据集上的分类结果

Table 2 Classification results on EuroSAT and UCM datasets

Method	Numbers of label M on EuroSAT (10 class)				Time/h	Numbers of label M on Ucm (21 class)				Time/h
	100	1 000	2 000	21 600		100	200	400	1 680	
CNN[6]	29.3%	46.1%	59.0%	83.2%	25	18.5%	32.8%	43.6%	62.1%	1
Inception V3[6]	63.9%	84.6%	87.9%	91.5%	27	55.4%	71.1%	81.1%	85.4%	1.7
FMGAN[2]	63.0%	75.8%	78.3%	86.9%	30	43.6%	69.2%	74.5%	80.2%	1.5
REG-GAN[9]	64.7%	72.8%	76.4%	82.3%	28	40.4%	55.4%	63.6%	72.3%	1.3
SFGAN[6]	68.6%	86.1%	89.0%	93.2%	31.5	43.9%	52.1%	60.6%	79.5%	2
SAGGAN[10]	76.8%	88.1%	90.7%	94.3%	33	54.1%	69.7%	83.3%	90.5%	2
RAGAN(ours)	**71.5%**	**88.2%**	**93.3%**	**97.4%**	**37.5**	**55.2%**	**71.4%**	**85.7%**	**91.0%**	**2.25**

从表2可以看出：

1）Inception V3在Image Net自然图像数据集预训练了一个良好的模型，并且可以通过微调，迁移到遥感图像场景分类的应用上提高泛化性能。与没有引用Inception v3的FMGAN和REG-GAN相比，RAGAN、SFGAN和SAGGAN都沿用了Inception v3分支，以增强分类性能，当有标签的样本数量足够时，它们的总体分类精度均有小幅度提升。

2）由分类总体精度表可以看出，有标签的样本数量越多，分类精度越高。特别是在EuroSAT数据集中，可以看到M=1 000时（占总训练集的4.6%），RAGAN的总体分类精度可以达到88.2%。而当M=2 000（占总训练集的9.3%）和21 600（全部训练集）时，RAGAN的总体分类精度分别达到了93.3%和97.4%，相比半监督遥感图像分类算法SAGGAN提升了2.6%和3.1%。

3）当M=100时（占总训练集的0.46%），RAGAN相比SFGAN算法的总体分类精度提升了2.9%，但是和SAGGAN算法相比，精度下降了5.3%。经过分析，在有标签的样本数量低于1%时，RAGAN算法的优越性体现不出来，原因可能是样本数量太少，未能学习到判别性强的特征造成的。而当有标签的样本数量高于5%时，RAGAN方法的总体分类性能更有优越性。

4）从运行时间对比可知，RAGAN的运行时间相比其他基于GAN的半监督算法略长，其原因是：首先，RAGAN方法需要对每一个残差块的卷积层进行谱归一化，增加了运行成本；其次，卷积模块的注意力机制尽管是轻量级模型，但是经过不同层次的特征融合和门控单元的引入，RAGAN将注意力聚焦后的特征转化为内部关联性更强的新特征。因此，增加了计算资源的投入和运算时间。

实验同时生成了混淆矩阵图CM，进一步详细分析方法的效果，如图5和图6所示。在图5中，横纵坐标0～9代表的场景分别是“居民楼”、“河”、“高速公路”、“牧场”、“森林”、“庄稼作物”、“草本植被”、“工业建筑”、“永久性作物”、“海洋湖泊”。从混淆矩阵可以看出，RAGAN方法在6号草本植物场景中分类效果最好，在7号工业建筑场景中分类效果最差。原因是：飞机和卫星进行拍摄时，由于成像角度、云雾和光照辐射等因素的影响。居民楼、牧场和工业建筑等不同场景的相同对象例如房屋、道路和汽车等，出现深层语义重

叠，造成分类效果不明显。而RAGAN方法对草本植被、森林和永久性作物分类精度较高，表明RAGAN在类间相似性高的复杂场景中可以提取到判别力强的特征。在图6中，横纵坐标0～20代表的场景分别是“稀疏住宅区”、“飞机”、“高速公路”、“路口”、“河”、“网球场”、“密集住宅区”、“棒球场”、“立交桥”、“港口”、“储

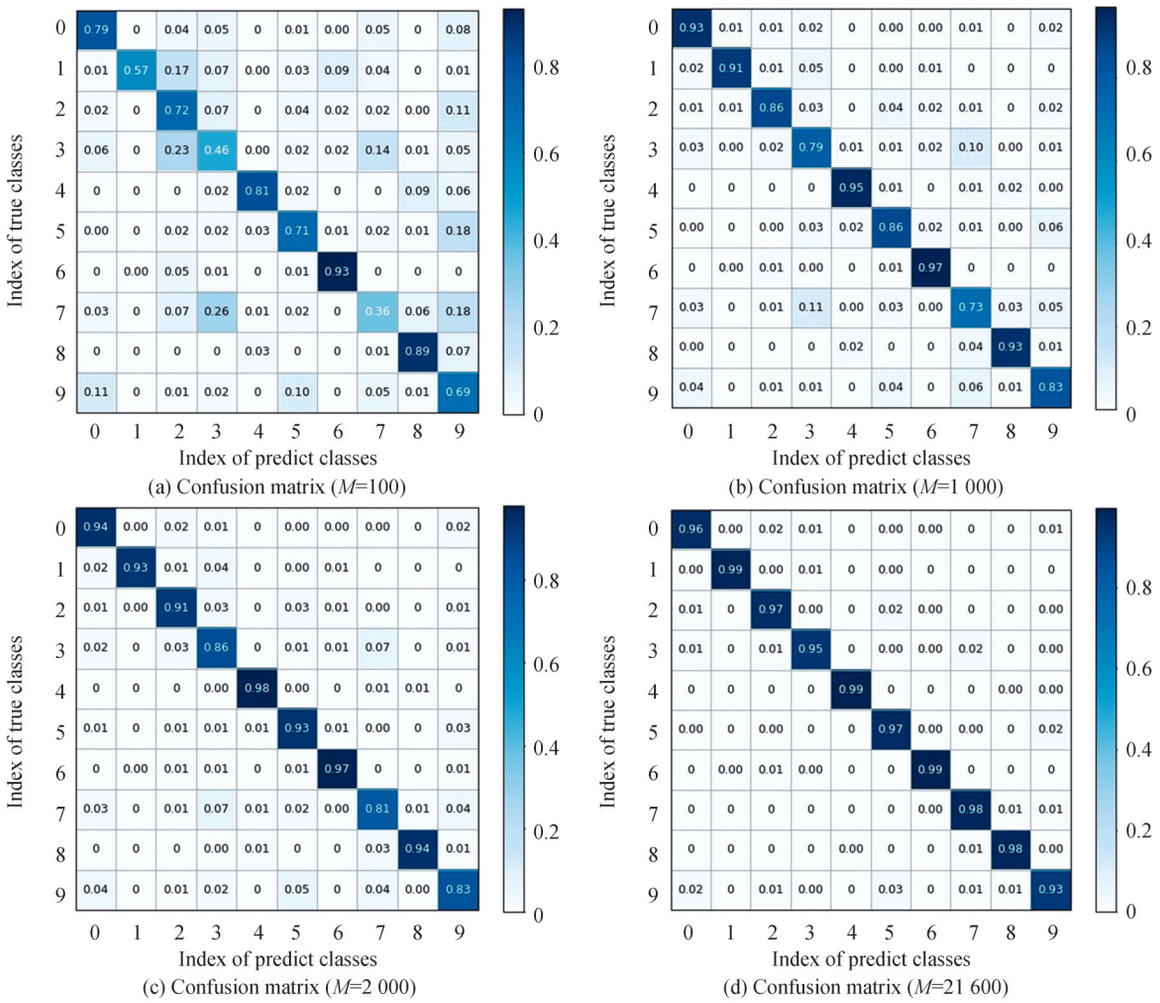
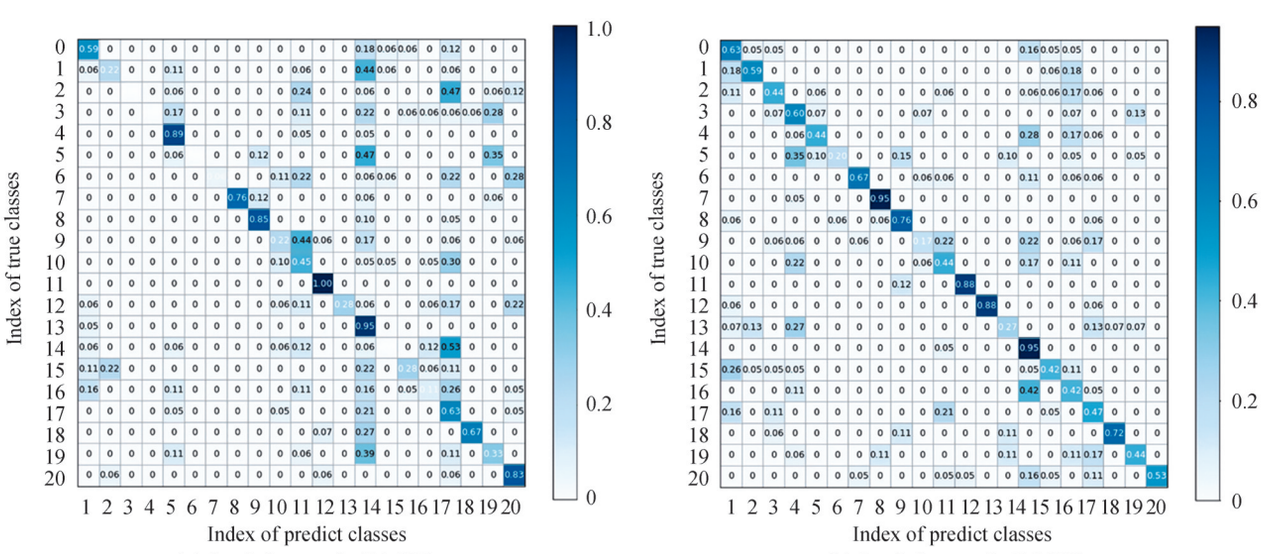


图5 EuroSAT数据集中不同标记数量的混淆矩阵

Fig. 5 Confusion matrix of different number of markers in EuroSAT dataset



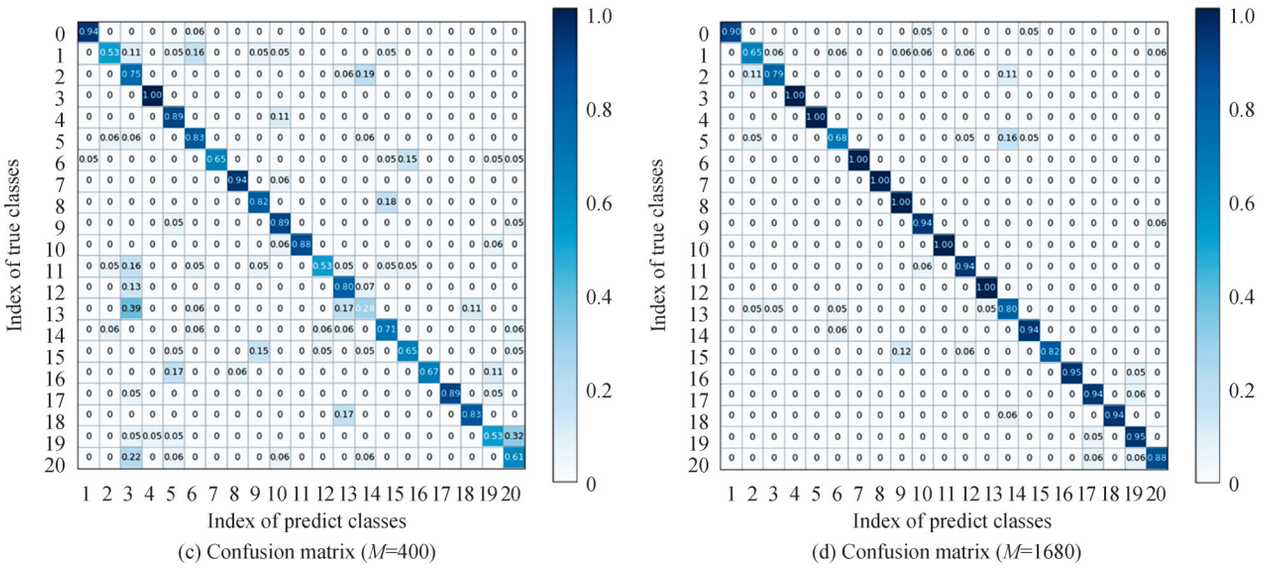


图6　UCM数据集中不同标记数量的混淆矩阵

Fig. 6　Confusion matrix of different number of markers in UCM dataset

油罐”、“农业”、“中型住宅”、“海滩”、“丛林”、“停车场”、“森林”、“移动家庭公园”、“跑道”、“高尔夫球场”、“建筑物”。同理，RAGAN方法在14号丛林、16号森林和11号的农业分类效果良好，体现该方法在复杂场景中分类的优越性。类间相似性高的场景，如0号稀疏住宅区、12号中型住宅和20号建筑物，分类效果较高，再次表明该方法能够提取判别力强的特征，具有更好的适应性和实际性。

3.5　RAGAN收敛性分析

在EuroSAT和UCM数据集中，通过不同标签数量的训练精度曲线和验证精度曲线，讨论该方法的收敛性，如图7所示。图7(a)是RAGAN方法在EuroSAT数据集里不同标记样本的训练精度曲线和验证精度曲线，可以看出RAGAN的训练精度曲线在epoch=8之前明显提高，在epoch=10以后逐渐收敛，并趋近于1。当有标记样本量M为100的时候，RAGAN的验证精度曲线在epoch=26以后逐渐收敛趋于1。其原因是因为有标签样本量过少，低于训练集的1%，RAGAN在较短的周期内没有充分学习深层特征，导致训练周期增长。除此之外，当epoch=18之前验证曲线逐步提高，并于epoch=20以后逐渐收敛。同理，7(b)是RAGAN方法在UCM数据集里不同标记样本的训练精度曲线和验证精度曲线，可以看出RAGAN的训练精度曲线在epoch=18以前明显提高，并于epoch=20以后逐渐收敛。而RAGAN的验证精度曲线在epoch=26以前逐渐提高，并于epoch=26以后逐渐收敛。RAGAN方法在UCM数据集收敛较慢的原因是：UCM数据集的总体样本量和有标记样本量较低，RAGAN无法充分训练，还学习了一部分冗余无用的特征。综上，引入谱归一化残差块后，生成对抗的训练精度曲线更加平滑稳定；不同层次的特征融合与结合门控的注意

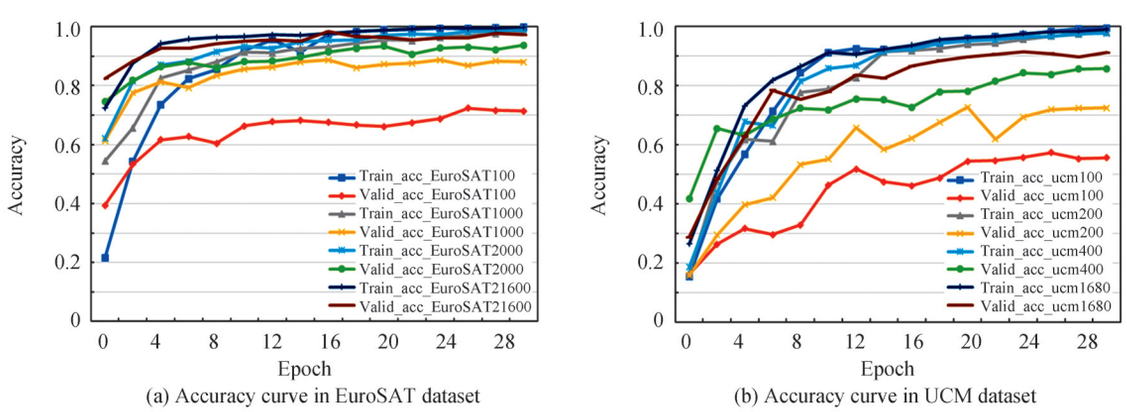


图7　EuroSAT和UCM数据集中不同标记数量的精度曲线

Fig. 7　Accuracy curves of different numbers of markers in EuroSAT and UCM datasets

力机制，可以在提取判别力强的特征同时让网络更快地收敛；特别是，当训练集的样本量充足且有标签的样本量超过总训练集的5%时，RAGAN只需要10个epoch即可实现稳定收敛。

3.6 消融实验验证

为了进一步验证谱归一化残差块、结合门控的注意力模块以及特征融合模块的贡献，在EuroSAT和UCM数据集验证各个模块的有效性。从表3的消融实验结果中，可以观察到谱归一化残差块、注意力模块、特征融合模块对分类精度的整体提升均有贡献。贡献最大的是谱归一化残差块，它对任何一个标签量的精度均有提升，原因是谱归一化相比其他方法稳定性更好，特别是应用在生成对抗网络里。其次是结合门控的注意力模块，特别是在有标签样本量大于10%时，分类效果提升比较大，因为样本数量充足能够学习到更全面的特征。贡献较小的是特征融合模块，分析得出是浅层特征的语义信息不够完整，没有深层特征丰富，因此和深层特征融合后，没有出现很好的分类效果。值得注意的是，在有标签样本量极少的情况下，使用了注意力模块和特征融合模块，分类精度和只使用谱归一化残差块相比有所降低。经过分析认为，由于样本量极少，没有充分训练和学习，提取了一部分冗余或无用的特征，造成分类精度降低。

表3 各模块对分类精度的影响

Table 3 Influence of each module on classification accuracy

Method	Numbers of label M on EuroSAT (10 class)				Time/h	Numbers of label M on UCM (21class)				Time/h
	100	1 000	2 000	21 600		100	200	400	1 680	
SFGAN	68.6%	86.1%	89.0%	93.2%	31.5	43.9%	52.1%	60.6%	79.5%	2
+SNRB	73.0%	88.1%	91.6%	95.9%	34.8	45.7%	52.9%	68.3%	82.4%	2.1
+GAM	70.7%	87.3%	92.6%	97.1%	33.5	54.8%	58.2%	77.8%	90.7%	2.1
+Fusion	65.4%	86.3%	90.4%	93.4%	32	41.9%	50.7%	65.6%	80.9%	2
+All	71.5%	88.2%	93.3%	97.4%	37.5	55.2%	71.4%	85.7%	91.0%	2.25

4 结论

针对有标签的高分辨率遥感图像样本较少、难以提取判别力强的特征的问题，提出了一种基于谱归一化残差块和门控注意力机制的半监督遥感图像场景分类方法RAGAN。该方法首先采用谱归一化残差块代替判别网络中的卷积层，增强了生成对抗的稳定性，同时每一个残差块包含两次卷积，可以更好地提取遥感图像特征。然后，融合不同层次的特征，更全面地反映场景信息。最后，引入结合门控的注意力模块，更好地聚焦于判别力强的特征，从而实现分类精度的提升。为了验证该方法的优越性，对EuroSAT和UC Merced两个高分辨率遥感图像数据集进行了实验，在EuroSAT数据集中，当有标签的数量M为2 000和21 600时，RAGAN有更好的分类效果，最高分类精度分别达到了93.3%和97.4%，相比半监督分类方法SAGGAN提高了2.6%和3.1%。同理，在UC Merced数据集中，当M为400和1 680时，分类精度分别达到了85.7%和91.0%，相比SAGGAN准确率提高了2.4%和0.5%。

参考文献

[1] YU Donghang, ZHANG Baoming, ZHAO Chuan, et al. Scene classification of remote sensing image using ensemble convolutional neural network[J]. Journal of Remote Sensing, 2020, 24(6): 717-727.
余东行, 张保明, 赵传, 等. 联合卷积神经网络与集成学习的遥感影像场景分类[J]. 遥感学报, 2020, 24(6): 717-727.

[2] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Advances in Neural Information Processing Systems, 2014, 3: 2672-2680.

[3] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. 2015, arXiv preprint arXiv: 1511.06434.

[4] ODENA A. Semi-supervised learning with generative adversarial networks[J]. 2016, arXiv preprint arXiv: 1606.01583.

[5] TAO Y, XU M, ZHONG Y, et al. GAN-assisted two-stream neural network for high-resolution remote sensing image classification[J]. Remote Sensing, 2017, 9(12): 1328.

[6] ROY S, SANGINETO E, SEBE N, et al. Semantic-fusion GANS for semi-supervised satellite image classification[C]. International Conference on Image Processing, 2018: 684-688.

[7] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral normalization for generative adversarial networks[J]. 2018,

arXiv preprint arXiv：1802.05957.
[8] MAO X，LI Q，XIE H，et al. Least squares generative adversarial networks[C]. International Conference on Computer Vision，2017：2794-2802.
[9] LECOUAT B，FOO C S，ZENATI H，et al. Manifold regularization with GANS for semi-supervised learning[J]. 2018，arXiv preprint arXiv：1807.04307.
[10] GUO D，XIA Y，LUO X. GAN-based semisupervised scene classification of remote sensing image[J]. IEEE Geoscience and Remote Sensing Letters，2020，(99)：1-5.
[11] SALIMANS T，GOODFELLOW I，ZAREMBA W，et al. Improved techniques for training gans[J]. 2016，arXiv preprint arXiv：1606.03498.
[12] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2016：770-778.
[13] ZHOU Wenhui，SHI Min，ZHU Dengming，et al. Seismic data super-resolution method based on residual attention network[J]. Computer Science，2021，48(8)：24-31.
周文辉，石敏，朱登明，等. 基于残差注意力网络的地震数据超分辨率方法[J]. 计算机科学，2021，48(8)：24-31.
[14] YU Yanjie，SUN Jiaqi，GE Sibo，et al. CycleGAN-SN：image stylization algorithm combining spectral normalization and cycleGAN[J]. Journal of Xi′an Jiaotong University，2020，54(5)：139-147.
余艳杰，孙嘉琪，葛思擘，等. CycleGAN-SN：结合谱归一化和CycleGAN的图像风格化算法[J]. 西安交通大学学报，2020，54(5)：139-147.
[15] ZHANG Tong ZHENG Enrang SHEN Junge，et al. Remote sensing image scene classification based on deep multi-branch feature fusion network[J]. Acta Photonica Sinica，2020，49(5)：0510002.
张桐，郑恩让，沈钧戈，等. 基于深度多分支特征融合网络的光学遥感场景分类[J]. 光子学报，2020，49(5)：0510002.
[16] GUO Fan，ZHANG Yongxiang，TANG Jin，et al. YOLOv3-A：a traffic sign detection network based on attention mechanism[J]. Journal on Communications，2021，42(1)：87-99.
郭璠，张泳祥，唐琎，等. YOLOv3-A：基于注意力机制的交通标志检测网络[J]. 通信学报，2021，42(1)：87-99.
[17] WOO S，PARK J，LEE J Y，et al. Cbam：convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision，2018：3-19.
[18] LIU W，HUANG X，CAO G，et al. Multi-modal sequence model with gated fully convolutional blocks for micro-video venue classification[J]. Multimedia Tools and Applications，2020，79(9)：6709-6726.
[19] GUO Dongen，XIA Ying，LUO Xiaobo，et al. Remote sensing image scene classification based on supervised contrastive learning[J]. Acta Photonica Sinica，2021，50(7)：0710002.
郭东恩，夏英，罗小波，等. 基于有监督对比学习的遥感图像场景分类[J]. 光子学报，2021，50(7)：0710002.
[20] HELBER P，BISCHKE B，DENGEL A，et al. Eurosat：a novel dataset and deep learning benchmark for land use and land cover classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing，2019，12(7)：2217-2226.
[21] YANG Y，NEWSAM S. Bag-of-visual-words and spatial extensions for land-use classification[C]. Proceedings of the 18th SIGSPATIAL，International Conference on Advances in Geographic Information Systems，2010：270-279.

Semi-supervised Scene Classification of Remote Sensing Images Based on GAN

XIA Ying[1]，LI Junyao[1]，GUO Dongen[1,2]
(1 *Chongqing University of Posts and Telecommunications，Chongqing Engineering Research Center of Spatial Big Data Intelligent Technology，Chongqing* 400065，*China*)
(2 *School of Computer and Software，Nanyang Institute of Technology，Nanyang，Henan* 473000，*China*)

Abstract：Remote sensing image scene classification is an important and challenging problem of remote sensing image interpretation. With the generation of a large number of scene-rich high-resolution remote sensing images，scene classification of remote sensing images is widely used in many fields such as smart city construction，natural disaster monitoring and land resource utilization. Due to the advancement of deep learning techniques and the establishment of large-scale scene classification datasets，scene classification methods have been significantly improved. Although the classification methods based on deep learning have achieved high classification accuracy，the supervised methods require a large number of training samples，

while the unsupervised classification methods are difficult to meet the practical needs and have low classification accuracy. Meanwhile, the annotation of remote sensing images requires rich engineering skills and expert knowledge, and in remote sensing applications, only a small amount of labeled remote sensing images exist for supervised training in most cases, and a large amount of unlabeled images cannot be fully utilized. Therefore, a semi-supervised learning method that extracts effective features from a large amount of unlabeled data by learning a small amount of labeled data becomes a potential way to solve such problems. To address the problems of complex background of remote sensing images and the inability of supervised scene classification algorithms to utilize unlabeled data, a semi-supervised remote sensing image scene classification method based on generative adversarial networks, namely, residual attention generative adversarial networks, is proposed. First, to enhance the stability of training, the residual blocks with jump structure are introduced in the deep neural network. At the same time, the spectral normalization constrains the spectral norm of the weight matrix in each convolutional layer of the residual block to ensure that the input and output of each batch of data satisfy the 1-Lipschitz continuity, which makes the generative adversarial training always smooth, not only improves the training stability, but also avoids network degradation. Secondly, since the shallow features extracted by the bottom convolution contain mostly local information and low semantics, while the deep features extracted by the top convolution contain more global information but lose part of the detail information. Therefore, the shallow features are fused with the deep features extracted from the multi-layer spectral normalized residual blocks to reduce the loss of features and allow the model to learn the complementary relationships between different features, thus improving the model's representational ability. Finally, to guide the model to focus more purposefully on important features and suppress unnecessary features, an attention module that mimics the signal processing of the human brain is used. Meanwhile, in order to obtain stronger feature representation ability and capture the dependency relationship between features, a gating mechanism is introduced to form an attention module combined with gating. To verify the superiority of the method, experiments were conducted on two high-resolution remote sensing image datasets, EuroSAT and UC Merced. In the EuroSAT dataset, the highest classification accuracy reached 93.3% and 97.4% when the number of labeled features was 2 000 and 21 600, respectively. In the UC Merced dataset, the classification accuracies reached 85.7% and 91.0% when the number of labeled was 400 and 1 680, respectively. To further validate the degree of contribution of each module, ablation experiments were also conducted in the EuroSAT and UCM public datasets, and it can be concluded from the validation that the spectral normalization residual module has the largest contribution, with improvement for different number of labeled samples. The reason is that the spectral normalization ensures that the gradient of the network is limited to a certain range during backpropagation, improving the stability of the generative adversarial network, and also does not destroy the network structure in the process. The next is the attention module combined with gating, especially when the labeled sample size is greater than 10%, the classification effect is improved more because the sample size is sufficient to learn more comprehensive features. The smallest contribution is the feature fusion module, because when the sample size is very small, the network is not sufficiently trained and learned, and a part of redundant or invalid features are extracted, resulting in lower classification accuracy. The above experimental results show that the proposed residual attention generation adversarial network classification method can effectively extract more discriminative features and improve the semi-supervised classification performance for the problem of small sample size of labeled high-resolution remote sensing images, which makes it difficult to extract discriminative features.

Key words: Remote sensing image; Scene classification; Semi-supervised; Generative Adversarial network; Attention mechanism

OCIS Codes: 100.3008; 100.4996; 150.1135

Foundation item: National Natural Science Foundation of China (No. 41871226), Special Research and Development and Promotion Project of Henan Province (No.212102210492), Key Cooperation Project of Chongqing Municipal Education Commission (No.HZ2021008)