弱监督尺度自适应增强的高分辨率遥感影像场景分类

王梨名[1]，祁昆仑[1,2]，杨超[1,2]，吴华意[3]

1. 中国地质大学(武汉) 地理与信息工程学院, 武汉 430074;
2. 国家地理信息系统工程技术研究中心, 武汉 430074;
3. 武汉大学 测绘遥感信息工程国家重点实验室, 武汉 430079

摘 要：遥感图像中同一种地物可能对应不同大小尺寸，而卷积核感受野大小固定严重影响了卷积神经网络在遥感场景分类中的性能。针对上述尺度效应问题，本文提出了一种面向高分辨率遥感影像场景分类的弱监督尺度自适应增强网络WSADAN（Weakly-supervised Scale Adaptation Data Augmentation Network），主要包括尺度生成和尺度融合两个模块。尺度生成模块利用卷积神经网络提取的原图像高层特征学习出适合于不同样本实例的最佳尺度参数；而尺度融合模块通过融合原尺度图像和最佳尺度图像的高层特征进行精化去除冗余，挖掘出不同尺度下特征间的关联信息。最后，联合多尺度特征表达输入到全连接层实现场景类别的预测。本文采用RSSCN7、AID和NWPU这3个遥感场景分类数据集验证方法的有效性，结果表明所提出的网络模型优于传统卷积神经网络，尤其对于尺度变化较大的类别性能提升最为明显。

关键词：遥感，场景分类，深度学习，卷积神经网络，弱监督，多尺度，数据增强

中图分类号：TP75/P2

引用格式：王梨名,祁昆仑,杨超,吴华意.2023.弱监督尺度自适应增强的高分辨率遥感影像场景分类.遥感学报,27(12):2815-2830

Wang L M, Qi K L, Yang C and Wu H Y. 2023. Weakly supervised scale adaptation data augmentation for scene classification of high-resolution remote sensing images. National Remote Sensing Bulletin, 27(12):2815-2830[DOI: 10.11834/jrs.20221481]

1 引 言

随着卫星遥感技术的发展，能够获取的高分辨率遥感卫星影像越来越多，如何快速解译海量卫星遥感数据成为亟需解决的问题。遥感影像场景分类根据遥感数据的内容，赋予遥感图像一个或多个有意义的具有高层次语义信息的场景类别标签（Zhong等，2015）。遥感影像场景分类在土地利用和土地覆盖分类（Zhu等，2016；Zhao和Du，2016；Chen等，2018）、自然灾害检测（Martha等，2011；Cheng等，2013）、城市规划（Lv等，2016）等领域有着广泛的应用。

遥感场景分类方法根据提取的特征分为：基于低层视觉特征、中层视觉特征和高层视觉特征的遥感场景分类方法（Xia等，2017）。基于低层视觉特征的遥感场景分类方法一般采用一种或几种低层特征，如光谱、纹理和结构等描述遥感图像（Yang和Newsam，2008；Luo等，2013）。基于低层视觉特征的遥感场景分类方法计算相对简单，耗费时间少，适合于结构和空间分布相对简单的场景类别，如森林、农田等；但是对于地物类型繁杂或空间分布多变的复杂场景，低层视觉特征表达能力有限。中层视觉特征采用图像局部特征，如SIFT、LBP和颜色直方图等，将其映射到字典空间中从而获得更具有判别力的特征表示，如视觉词袋模型和主题模型等（Yang和Newsam，2010；Chen等，2011）。基于中层视觉特征的场景分类方法可以建立场景局部特征到全局特征之间的语义

收稿日期：2021-07-20；**预印本**：2022-02-04

基金项目：湖北省科技厅重大专项(编号:2020AAA004)

第一作者简介：王梨名,研究方向为高分遥感影像解译。E-mail：limingwang@cug.edu.cn

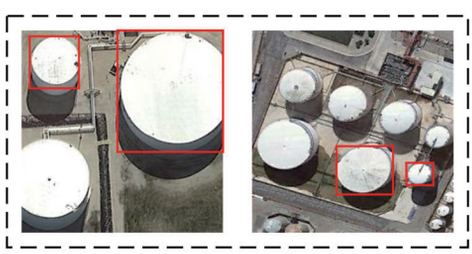
通信作者简介：杨超,研究方向为遥感影像解译和时空大数据。E-mail：yangchao@cug.edu.cn

关联，从而获得更具有判别力的场景特征；该方法依赖于低层特征的描述能力，限制了中层特征的场景表达性能。高层视觉特征通过卷积神经网络自适应地学习特征表达，从而解决低层局部特征在场景表达时的瓶颈问题（Gu等，2019）。近年来，以卷积神经网络为代表的深度学习方法在遥感场景分类任务上取得了突破性的进展，在许多极具挑战性的场景数据集上，性能均大幅度地超越了传统方法，展现出强大的特征表达能力（Han等，2017；钱晓亮 等，2018；余东行 等，2020）。

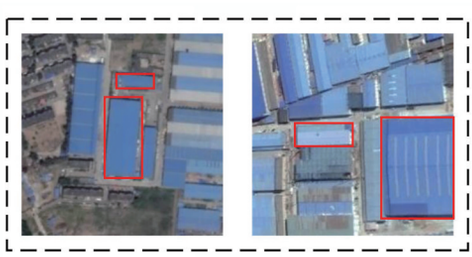
卷积神经网络一般采用固定的卷积核大小，其感受野有限，而遥感图像中蕴含丰富的空间特征和尺度效应，从不同的尺度能够获取不同层次的地物和空间关系特征（张书瑜，2020）。如图1所示，从上到下依次为储油罐、工业区和岛屿场景类别的影像，在同一幅影像相同图像分辨率情况下，储罐、工业区建筑物和岛屿等都具有多种不同尺度大小，其所需要的感受野也不一样。

如何解决遥感图像中的尺度效应问题，提升卷积神经网络在不同尺度下遥感图像场景的辨识力，是遥感影像场景分类中的热点问题。针对遥感图像的尺度效应问题，常用的方法包括：多尺度池化和多尺度特征融合两种方法。Liu等（2016）提出自适应深度金字塔匹配模型，将输入图像扭曲成不同的尺度输入空间金字塔网络提取多尺度深度特征。Zheng等（2019）提出一种深度场景表示方法，通过多尺度池化和费希尔向量对提取的深度特征构建场景的局部描述符。姚艳清等（2021）通过特征金字塔提取多尺度特征并在其后嵌入多分辨率特征提取网络，促使网络学习目标在不同分辨率下的特征。咎露洋等（2021）采用特征金字塔结构和任意四边形检测头，提高了对复杂尺度目标和不规则目标的处理能力。上述多尺度池化的方法一般都是对输入图像不同尺度的下采样获取图像更加宏观的概括性信息，无法对图像上采样得到图像更微观上的细节信息。此外，多尺度池化层一般置于网络的最后几层，对于前面卷积层不同尺度表达能力的提升有限。

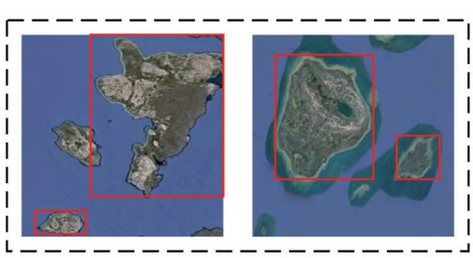
多尺度融合方法可以通过对输入图像进行不同尺度的上下采样，有助于辅助网络的卷积层学习到图像的多尺度信息。Qi等（2017）提出一种基于多尺度深度描述关联子的场景分类方法，将不同尺度的外观信息和空间信息联合起来提升场景分类的性能。Alhichri等（2018）提出一个适应各种图像大小的多尺度深度卷积神经网络架构，使用3个并行网络接收不同尺度的图像，从而提升模型的尺度表达能力。马欣悦等（2021）提出了一种基于多尺度循环注意力网络的遥感影像场景分类方法，采用注意力机制得到影像不同尺度下的关注区域，融合原始影像不同尺度及其关注区域的影像特征。朱祺琪等（2021）提出了一种全局局部细节感知条件随机场框架，在有效利用多尺度建筑物信息的同时保留局部结构信息，解决了传统条件随机场一元势能丢失边界信息的问题。上述多尺度特征融合方法融合了图像不同尺度的信息，但一般只能采用固定的几种图像尺度，而各场景类别中不同图像的最佳尺度大小会有所不同。



（a）储油罐
（a）Storage_tank



（b）工业区
（b）Industrial_area



（c）岛屿
（c）Island

表示地物目标尺度大小

图1 遥感图像的尺度效应示例

Fig. 1 Examples of scale effect in remote sensing scene image

为了提升卷积神经网络在不同尺度下的表达能力，解决多尺度特征融合中图像大小固定的问题，本文提出了一种弱监督尺度自适应增强网络WSADAN（Weakly-supervised Scale Adaptation Data Augmentation Network），该网络采用尺度生成模块SGM（Scale Generation Module）学习不同影像的最佳尺度大小，通过尺度融合模块SFM（Scale Fusion Module）实现原始图像和尺度变换后图像中高层特征的融合，从而提升遥感影像场景分类的精度。

2　方　法

2.1　弱监督尺度自适应增强网络

针对遥感图像尺度效应问题，本文提出了一种弱监督尺度自适应增强网络，如图2所示。（1）输入图像$\boldsymbol{X}$通过卷积神经网络得到输入图像的高层特征$\boldsymbol{y}$，$\boldsymbol{y}$经过尺度生成模块学习到该图像的最佳尺度参数u；（2）输入图像$\boldsymbol{X}$根据得到的最佳尺度参数u重采样后得到尺度变换图像$\boldsymbol{X}'$，尺度变换图像$\boldsymbol{X}'$输入到卷积神经网络得到尺度变换图像高层特征$\boldsymbol{y}'$；（3）将输入图像高层特征$\boldsymbol{y}$和尺度变换图像高层特征$\boldsymbol{y}'$进行拼接后输入到尺度融合模块，最终通过softmax激活函数得到遥感图像的场景分类结果。

由于随机数据增强方法效率低，并且容易产生很多不受控制的噪声，因此，本文提出的WSADAN网络采用一种尺度弱监督数据增强策略，即将原始图像根据尺度生成模块的最佳尺度参数，通过双线性重采样得到尺度变换图像，然后将尺度变换图像重新输入到预训练的基础网络模型中提取其特征信息。相比于随机数据增强，本文的尺度弱监督数据增强策略能够根据场景内容选择数据增广的参数。

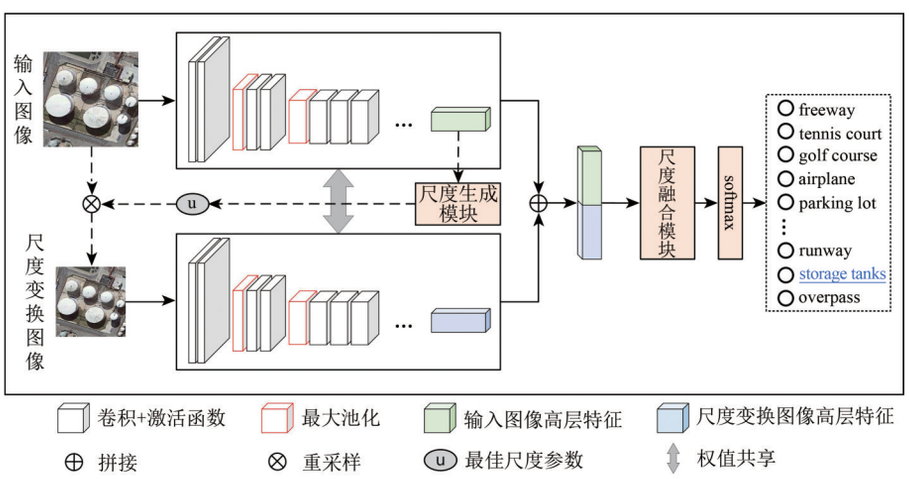


图2　弱监督尺度自适应增强网络(WSADAN)结构图

Fig. 2　Architecture of Weakly-supervised Scale Adaptation Data Augmentation Network (WSADAN)

在双线性重采样中有上采样和下采样两种情况。式（1）所示为双线性下采样过程，如果对一幅图像进行s倍的下采样，则原始图像中$s\times s$窗口大小的像元变为一个像元，像元值为该窗口内像元的均值，图像尺寸会由原始的$M\times N$变为$(M/s)\times(N/s)$。用公式可以表示为

$$P_k=\frac{\sum_{i\in \boldsymbol{win}(s*s)}\boldsymbol{I}_i}{s^2} \tag{1}$$

式中，P_k表示双线性下采样后生成的像元值，$\boldsymbol{win}(\cdot)$表示像元窗口，$\boldsymbol{I}_i$表示像元i的值，s表示下采样倍数。

式（2）所示为双线性上采样过程，通过在像元点之间插入新的像元使图像变大，插入的新像元值根据下列式（2）得到：

$$f(x,y)=\frac{\begin{bmatrix}x_2-x,x-x_1\end{bmatrix}\begin{bmatrix}Q_{11} & Q_{12}\\ Q_{21} & Q_{22}\end{bmatrix}\begin{bmatrix}y_2-y\\ y-y_1\end{bmatrix}}{(x_2-x_1)(y_2-y_1)} \tag{2}$$

式中，$(x_1,\ y_1)$、$(x_1,\ y_2)$、$(x_2,\ y_1)$、$(x_2,\ y_2)$为已知像元值的4个角像元坐标，Q_{11}、Q_{12}、Q_{21}、Q_{22}分别为4个角的像元值，$(x,\ y)$表示双线性下采样新插入的像元坐标，$f(\cdot)$表示该点像元值大小。

2.2 尺度生成模块

在高分辨率遥感图像中，不同场景图像分类的最佳尺度大小不同。同一幅遥感图像在不同尺度下能够得到不同的空间特征信息，综合不同尺度的特征信息有助于提高遥感场景分类的精度。目前学者们对于多尺度图像生成方法主要有3种：构建金字塔的方式获取图像多个尺度的特征（例如，Zhao和Du，2016）；不同上下文输入大小的多个神经网络模型获取图像多尺度的特征（例如，Längkvist等，2016）；尺度参数估计的方式得到最佳尺度组合（Chen等，2019）。上述3种获取多尺度特征的方式一般对整个数据集，挑选同一种最优的尺度组合方案，没有考虑到不同类别场景图像对于不同尺度的适应性，而且不同数据集的最佳尺度组合方案需要精心设计和大量的实验选择最优方案。

针对上述不同场景图像的尺度差异问题，本文设计了一个尺度生成模块，结构如表1所示。该模块通过对原始图像高层特征进行处理，提取出适合该图像的最佳尺度变换参数。尺度生成模块由一个全局平均池化层、两个全连接层和缩放函数组成，采用ReLU激活函数对第一个全连接层特征非线性变换，采用Sigmoid激活函数将第二个全连接层特征映射到0到1空间。通过全局平均池化层对原始图像高层特征每个通道值平均，由原来的$7 \times 7 \times C$压缩成为$1 \times 1 \times C$。压缩后的高层特征输入到全连接层中进行特征提取，特征提取过程为

$$\boldsymbol{t} = z(\boldsymbol{max}(0, z(\boldsymbol{y}))) \tag{3}$$

式中，$\boldsymbol{y}$表示输入图像高层特征，$z(\cdot)$表示为全连接层，$\boldsymbol{max}(0, \cdot)$相当于ReLU激活函数，$\boldsymbol{t}$表示为特征提取的结果。该特征提取过程可以描述为高层特征经过第一个全连接层后使用ReLU函数进行激活，然后再输入到第二个全连接层中。特征提取后通过激活函数Sigmoid映射到0到1的区间：

$$v = \frac{1}{1 + \mathrm{e}^{-t}} \tag{4}$$

式中，v表示生成的尺度变换参数。采用缩放函数把尺度变换参数线性缩放到对应的尺度，得到适合的尺度参数。这里采用缩放范围为[0.5，2]，故缩放函数为

$$u = av + b \tag{5}$$

式中，a为1.5，b为0.5，u表示生成的尺度参数。

表1 尺度生成模块(SGM)详细结构(括号中为输入输出大小)

Table 1 Detailed structure of Scale Generation Module (SGM) (The values in brackets denote sizes of input or output)

结构组成	输入(数据大小)	输出(数据大小)
全局平均池化	卷积深度特征(512×7×7)	卷积特征代表(512×1×1)
压缩	卷积特征代表(512×1×1)	压缩卷积特征代表(512)
全连接层1	压缩卷积特征代表(512)	全连接层1特征(128)
ReLU激活函数	全连接层1特征(128)	激活全连接层1特征(128)
全连接层2	激活全连接层1特征(128)	全连接层2特征(1)
Sigmoid激活函数	全连接层2特征(1)	尺度变换参数(1)
缩放函数:$u=av+b$	尺度变换参数(1)	尺度参数(1)

2.3 尺度融合模块

传统方法在多尺度特征融合时，一般将不同尺度卷积特征简单串联或相加，然后输入到分类层中，这种方式虽然易于操作，但是缺乏对多尺度特征的深层次信息挖掘。

为充分利用图像多个尺度的特征，本文设计了一个尺度融合模块，用于提升多尺度特征融合的表达能力。如图3所示，该模块主要包括4个部分：批量标准化BN（Batch Normalization）层、通道注意力层、1维卷积层和全局平均池化GAP（Global Average Pooling）层。按通道拼接原始图像高层特征$\boldsymbol{y}$（大小为$N \times C \times H \times W$）与尺度变换图像高层特征$\boldsymbol{y}'$（大小为$N \times C \times H \times W$）得到组合特征$\boldsymbol{Y}$（大小为$N \times 2C \times H \times W$），并输入至BN层，将来自不同尺度的特征变换到同一基准下。采用通道注意力机制对不同尺度所有特征通道赋予不同权重，权重值越大代表该特征相关度越高，进一步筛选出不同尺度中最重要的特征。通道注意力主要有挤压、激励和注意3个过程。挤压是把每个通道特征值都压缩成一个值，挤压函数如下所示：

$$\boldsymbol{F}_{sq}(\boldsymbol{u}_c) = \sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{u}_c(i,j) \Big/ H \times W \tag{6}$$

式中，H表示输入图像特征的高，W表示输入图像特征的宽，$\boldsymbol{u}_c(\cdot)$表示通道c像元的特征值，$\boldsymbol{F}_{sq}$表示压缩后的特征值。激励是通过全连接层和激活函数来得到每个通道的权重，激励函数如下所示：

$$\boldsymbol{F}_{ex}\left(\boldsymbol{F}_{sq},\boldsymbol{W}\right)=\boldsymbol{\sigma}\left(\boldsymbol{W}_2\boldsymbol{\delta}\left(\boldsymbol{W}_1\boldsymbol{F}_{sq}+\boldsymbol{b}_1\right)+\boldsymbol{b}_2\right) \quad (7)$$

式中，$\boldsymbol{\sigma}$表示Sigmoid激活函数，$\boldsymbol{\delta}$表示ReLU激活函数，$\boldsymbol{W}_1$和$\boldsymbol{W}_2$表示两层全连接层的权重系数，$\boldsymbol{b}_1$和$\boldsymbol{b}_2$表示两层全连接层的偏置系数，$\boldsymbol{F}_{ex}$表示通道权重系数。利用特征的每一个通道与通道权重相乘计算不同通道的注意力，注意函数如下所示：

$$\boldsymbol{F}_{at}\left(\boldsymbol{u}_c,\boldsymbol{F}_{exc}\right)=\boldsymbol{F}_{exc}\cdot\boldsymbol{u}_c \quad (8)$$

式中，$\boldsymbol{F}_{exc}$表示每个通道的权重系数，$\boldsymbol{u}_c$每个通道的特征值。

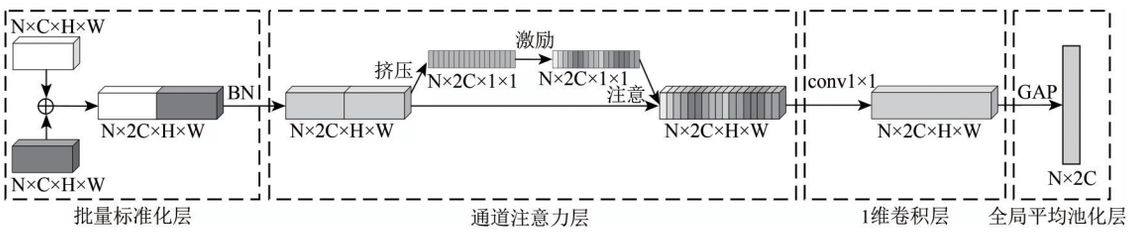


图3 尺度融合模块(SFM)

Fig. 3 Scale fusion module (SFM)

通过1维卷积联合所有通道特征，选择各尺度特征并加权融合，进一步挖掘多尺度特征之间的关联信息。最后，通过GAP层输入到全连接层进行场景类别的预测。

3 实验与分析

3.1 实验设置

3.1.1 基础模型

本文采用遥感影像场景分类中应用最为广泛的基础模型VGG16（Simonyan和Zisserman，2015）和ResNet50（He等，2016）验证本文方法的有效性，这两种模型均曾获得ImageNet挑战赛的冠军。VGG16网络结构设计简洁，整个网络通过重复堆叠很多基础模块组成；ResNet50提出了一种残差模块，有效抑制了深层网络梯度消失问题，有效提升了特征的表达能力。由于深度卷积神经网络模型参数量较大，现有的遥感场景分类数据集相对较少不足以完全训练，所以本文采用迁移学习的方法（He等，2020；Zeng等，2018），基于在ImageNet数据集上预训练的VGG16和ResNet50网络对模型参数进行微调，这种方式可以缓解数据量小而导致的过拟合现象（Zheng，2015）。

3.1.2 实验数据

为验证模型的有效性，本文采用具有代表性的RSSCN7（Zou等，2015）、AID（Aerial Image Dataset）（Xia等，2017）和NWPU（NWPU-RESISC45）（Cheng等，2017）3种遥感场景数据集进行试验。

RSSCN7数据集包含2800张遥感场景图像，包括草地、森林、农田、停车场、居住区、工业区、河流湖泊等7个典型场景类别。每个场景类别有400张图像，分别在4个不同尺度上采样100张，每张图像大小为400×400像素。由于季节变化和不同天气条件以及采样尺度的影响，造成该数据集下场景图像的复杂性和多样性。本文实验中训练数据比率分别设置为20%和50%。

AID（Aerial Image Dataset）数据集是武汉大学2017年发布的大型场景数据集。数据集包含30个场景类别，每一个场景类别分别含有220—420张图像，每幅图像大小600×600，共计10000幅场景图像。AID数据集中的图像来自于不同传感器，数据集中图像的像素分辨率多种多样，大约在0.5—8 m。本文实验中该数据集训练数据比率设置为20%和50%。

NWPU（NWPU-RESISC45）数据集是包含45个场景类别，每个类别由700幅图像组成，每张图像像素大小为256×256，共计31500幅场景图像。数据集中的图像来自100多个国家和地区，除了一些空间分辨率较低的特定类（如岛、湖、山、冰山），大多数场景类别的像素分辨率大约在0.2—30 m。本文实验中该数据集训练数据比率设置为10%和20%。

表2 数据集相关信息

Table 2 Dataset information

数据集	场景类别	数据量	像素大小	训练集比例/%
RSSCN7	7	2800	400×400	20、50
AID	30	10000	600×600	20、50
NWPU	45	31500	256×256	10、20

3.1.3 评价标准

全局精度OA（Overall Accuracy）是模型在所有测试集数据上预测正确的个数与全部数据总数的比值。全局精度是衡量一个深度学习模型的性能核心指标。全局精度的定义如下：

$$\mathrm{OA}=\sum_{i=1}^{C}M_i\Big/\sum_{i=1}^{C}N_i \tag{9}$$

式中，C为数据类别数量，M_i是第i类预测正确的样本数，N_i是第i类所有的样本数。

标准差STD（Standard Deviation）是所有数值与他们平均值之间差异的一种反映。通常情况下每次训练的网络模型性能都会有所差异，得到的全局精度和全局平均值精度也会不一样，所以要科学评估一个算法模型的可靠性需要多次实验，然后取其平均值并计算标准差。标准差的定义如下：

$$\mathrm{STD}=\sqrt{\frac{\sum_{i=1}^{N}(\mathrm{OA}_i-\overline{\mathrm{OA}})^2}{N}} \tag{10}$$

式中，N表示实验的次数，OA_i表示第i次实验的总体精度，$\overline{\mathrm{OA}}$表示N次实验的平均总体精度。

混淆矩阵CM（Confusion Matrix）可以详细的描述各类别的精度及类别间的误分状况。混淆矩阵中，元素a_{ij}表示i类别的图像被分为j类别的概率。

3.1.4 实验环境

本文实验采用Windows10服务器，配置Intel（R）Xeon（R）W-2133处理器、64 G运行内存、NVIDIA GeForce RTX 2080 Ti显卡，基于Pytorch深度学习框架实现算法模型。优化器采用Adam（Kingma和Ba，2014），学习率（learning rate）设置为1E-4，权重衰减（weight decay）系数设置为1E-5，批量大小（batch size）设置为8。

3.2 消融实验分析

为验证设计的尺度生成模块和尺度融合模块的有效性，本文在RSSCN7、AID和NWPU 3个数据集上分别进行实验，以原始VGG16模型（Baseline-VGG16）为基准评估各模块的作用，表3为不同模块在不同数据下的总体精度。

表3 不同模块的消融实验

Table 3 Ablation experiments of different modules

数据集	训练数据占比/%	Baseline-VGG16	$\mathrm{WSADAN}_{\mathrm{SGM}}$	$\mathrm{WSADAN}_{\mathrm{SGM+SFM}}$
RSSCN7	20	89.06±0.58	90.00±0.69	91.65±0.62
	50	91.86±0.51	92.75±0.53	94.07±0.49
AID	20	89.86±0.19	91.13±0.15	92.78±0.16
	50	93.05±0.42	93.92±0.39	95.18±0.37
NWPU	10	84.17±0.32	85.26±0.23	87.01±0.26
	20	88.31±0.48	89.02±0.37	90.44±0.40

本文方法只添加尺度生成模块（$\mathrm{WSADAN}_{\mathrm{SGM}}$）情况下，各数据集的总体精度都有提升。RSSCN7数据集20%和50%训练数据比率下精度分别提升了0.94%和0.89%，AID数据集20%和50%训练数据比率下精度分别提升了1.27%和0.87%，NWPU数据集10%和20%训练数据比率下精度分别提升了1.09%和0.71%。尺度生成模块通过生成最佳尺度图像对卷积神经网络进行数据增强从而提升了模型的鲁棒性。

本文方法在同时添加尺度生成模块和尺度融合模块情况下，$\mathrm{WSADAN}_{\mathrm{SGM+SFM}}$模型的总体精度提升效果明显。相比于$\mathrm{WSADAN}_{\mathrm{SGM}}$，RSSCN7数据集20%和50%训练数据比率下精度分别提升了1.65%和1.32%，AID数据集20%和50%训练数据比率下精度分别提升了1.65%和1.26%，NWPU数据集10%和20%训练数据比率下精度分别提升了1.75%和1.42%。添加尺度融合模块后，通道注意力对多尺度融合后特征进行过滤和筛选去除特征中包含的一些噪声，1维卷积对多尺度的特征进行深度融合并提取尺度间的特征信息，使得模型的总体精度明显提升。

对比不同训练集比例的总体精度，本文方法在训练数据较少的情况下精度提升更加明显。RSSCN7数据集中20%训练数据的总体精度提升了2.59%，50%训练数据的总体精度提升了2.21%；AID数据集中20%训练数据的总体精度提升了2.92%，50%训练数据的总体精度提升了2.13%；NWPU数据集中20%训练数据的总体精度提升了2.84%，50%训练数据的总体精度提升了2.13%。采用较少的训练

数据精度提升更加明显可能原因是本文中采用的是弱监督数据增强的方式，在训练数据较少的情况下更加有效。

3.3 场景尺度变化分析

为更准确的验证本文方法对场景尺度变化的学习能力和适应能力，本文在RSSCN7数据集20%训练比例上进行对比实验，以原始VGG16模型（Baseline-VGG16）为基准，与本文提出的基于VGG16模型的方法WSADAN-VGG16进行对比分析。图4展示了256尺度大小图像上训练出的模型在图像尺度变换到［0.5，2］倍时，测试模型总体精度变化。由图4可知，WSADAN-VGG16在图像所有尺度大小的总体精度均高于Baseline-VGG16，展现出更强的尺度适应能力。WSADAN-VGG16和Baseline-VGG16的精度随着图像尺度先增加后降低，在图像尺度为256时同时达到最大值，这是因为训练模型时采用的图像尺度为256×256，模型对于256尺度或接近256尺度的图像表现出更好的性能。

为更好的分析不同场景类型的尺寸变化特点和规律，图5展示了RSSCN7测试数据中不同场景类别的尺度参数箱形图。其中，各场景类别的尺度参数都大于1且小于1.9，分布范围比较广泛，说明同一场景类别下不同影像的最佳尺度大小差距较大。除草地和牧场两个场景类别外，其他场景类别图像的尺度参数多数集中在1.3—1.6。图6展示为草地和牧场两个类别的图像，草地的场景图像特征比较复杂，包含很多道路、树木和建筑物等非草地目标的地物类别，因此需要较大尺度参数以获取更加准确的细节特征。牧场场景图像地物类别比较简单，不需要获取太多细节信息，因此该类场景得到的尺度参数相对较小。

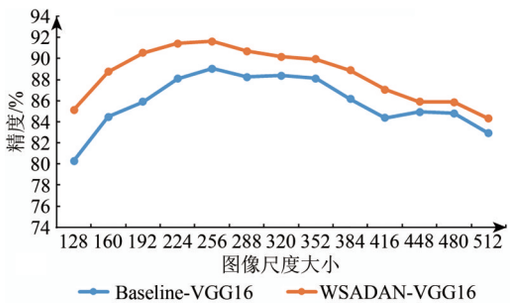


图4 Baseline-VGG16和WSADAN-VGG16随着图像尺度变化分类总体精度对比图

Fig. 4 Comparison of classification overall accuracy between Baseline-VGG16 and WSADAN-VGG16 with different image scale

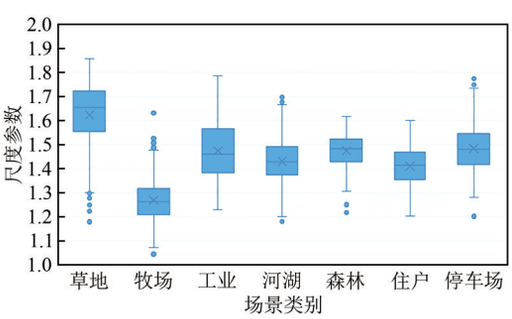
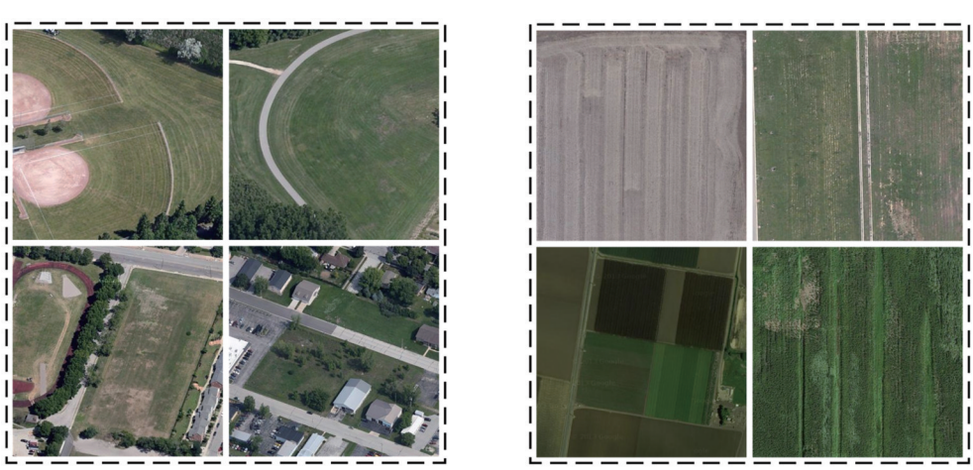


图5 RSSCN7数据集下本文方法测试模型得到的尺度参数图

Fig.5 The scale parameter boxplot obtained from our method on RSSCN7 testing dataset



（a）草地 （b）牧场

（a）Grass （b）Field

图6 RSSCN7数据集草地和牧场场景图片示例图

Fig.6 Examples of grass and pasture scene categories on RSSCN7 dataset

3.4 遥感影像分类算法对比

3.4.1 RSSCN7数据集实验

由于RSSCN7数据集从4个不同的尺度和角度采集的遥感场景图像，因此该数据集适合于评定本文提出方法的有效性。本文在RSSCN7数据集上对几种先进的方法进行了比较，实验结果如表4所示。在20%和50%的训练数据集下，本文提出的WSADAN-VGG16模型总体精度分别为91.65%和94.07%，基于ResNet50的WSADAN-ResNet50模型总体精度分别达到92.69%和94.82%。本文方法与传统的CaffeNet、VGG-VD-16和GoogLeNet相比性能有大幅提升，主要是因为本文设计的弱监督尺度自适应增强方法可以缓解遥感图像中的尺度效应问题。Attention GANs（Yu等，2020）采用多尺度池化的方式获取影像上下文信息，并设计了基于上下文聚合的特征融合体系结构，但是缺乏对融合特征筛选的过程。Deep Filter Banks（Wu等，2016）、Two-stage deep feature fusion（Liu等，2018）和Dual Attention-Aware Network（Gao等，2020）都采用了多尺度特征融合的方法。Deep Filter Banks结合多列堆叠去噪稀疏自编码（SDSAE）和Fisher向量（FV），以分层的方式自动学习具有代表性和鉴别性的特征。Two-stage deep feature fusion的方法结合来自中间层和FC层的激活，生成一个新的具有有向无环图拓扑CNN，并将两个这样的拓扑CNN融合在一起。Dual Attention-Aware Network采用通道注意力和空间注意力机制，分别从通道和空间维度探讨语境依赖关系，最后将两个注意模块的输出集成为注意感知特征表示。本文方法相较于上述研究性能较优的原因可能是因为本文中设计的尺度融合模块，可以对不同尺度的特征进行筛选和深度融合。

表4 RSSCN7数据集不同方法的总体精度对比

Table 4 Comparison of overall accuracy for different methods on the RSSCN7 dataset

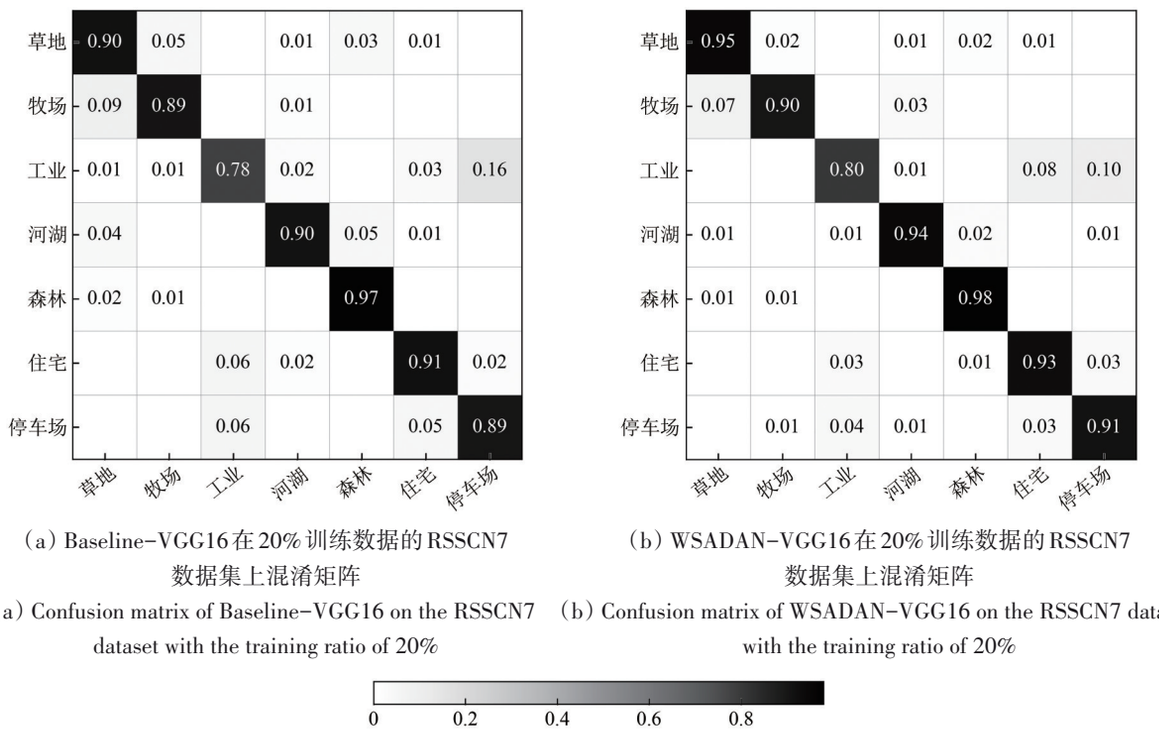
方法	不同训练集比例下的总体精度/%	
	20	50
CaffeNet(Xia等,2017)	85.57±0.95	88.25±0.62
VGG-VD-16(Xia等,2017)	83.98±0.87	87.18±0.94
GoogLeNet(Xia等,2017)	82.55±1.11	85.84±0.92
Attention GANs(Yu等,2020)	83.47±0.63	87.32±0.54
Deep Filter Banks(Wu等,2016)	—	90.4±0.6
Two-stage deep feature fusion（Liu等,2018）	—	92.37±0.72
Dual Attention-Aware Network(Gao等,2020)	91.07±0.65	93.25±0.28
WSADAN-VGG16（本文方法）	91.65±0.62	94.07±0.49
WSADAN-ResNet50（本文方法）	92.69±0.48	94.82±0.45

图7展示了RSSCN7数据集20%训练比率下不同模型得到的混淆矩阵，其中图7（a）和图7（b）分别表示Baseline-VGG16和WSADAN-VGG16模型的混淆矩阵。WSADAN-VGG16在每一个场景类别中精度都比Baseline-VGG16要高，WSADAN-VGG16除工业这一个场景类别外，其他所有的场景类别精度都达到了90%以上。最容易混淆的草地、牧场和森林3个场景类别精度都分别达到了95%、90%和98%。

3.4.2 AID数据集实验

表5对比了AID数据集上本文方法与多种先进方法的总体精度。在20%和50%训练数据情况下，本文提出的基于VGG16模型的方法WSADAN-VGG16总体精度分别为92.78%和95.18%，基于ResNet50模型的方法WSADAN-ResNet50总体精度分别为93.73%和95.88%。本文方法与传统的网络模型CaffeNet、VGG-VD-16、GoogLeNet和ResNet-50相比性能提升明显。DTDCNN（施慧慧等，2021）、LCPP（Sun等，2021）、MICP（Qi等，2021）采用多尺度池化的方式获取影像多尺度信息。DTDCNN通过可变形卷积对不同尺度目标感受野自适应调整使得网络模型对遥感场景的几何形变更具有稳健性。LCPP采用多层次卷积金字塔语义融合框架，融合BOVW模型和卷积神经网络模型提取的多层次语义特征。MICP采用多级改进循环池方法，提取并融合生成不同层次下的环形池化

特征。SAFF（Cao等，2021）和MDFR（Zheng等，2019）采用多尺度特征融合方法，其中，SAFF模型采用一种强调复杂对象权重的基于自注意深度特征融合方法，MDFR采用基于区域特征选择和表示的多尺度深度特征表示方法。本文相较于上述方法性能表现更好的原因可能是弱监督数据增强的方式使得模型具有更强的鲁棒性。



（a）Baseline-VGG16在20%训练数据的RSSCN7数据集上混淆矩阵

（a）Confusion matrix of Baseline-VGG16 on the RSSCN7 dataset with the training ratio of 20%

（b）WSADAN-VGG16在20%训练数据的RSSCN7数据集上混淆矩阵

（b）Confusion matrix of WSADAN-VGG16 on the RSSCN7 dataset with the training ratio of 20%

图7 RSSCN7数据集上WSADAN与Baseline混淆矩阵对比图

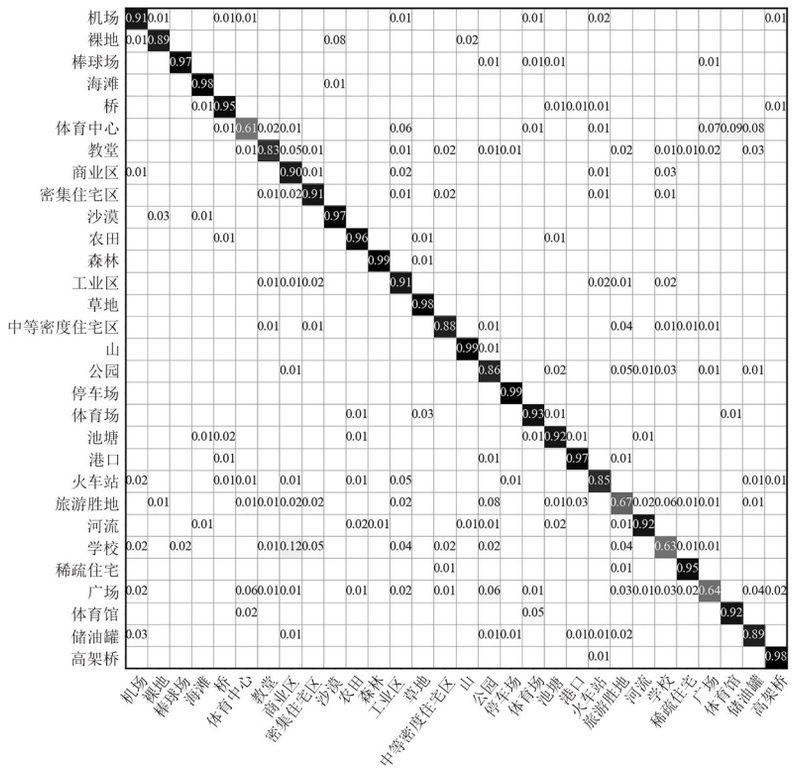
Fig. 7 WSADAN vs. Baseline confusion matrix on the RSSCN7 dataset

表5 AID数据集不同方法的总体精度对比

Table 5 Comparison of overall accuracy for different methods on the AID dataset

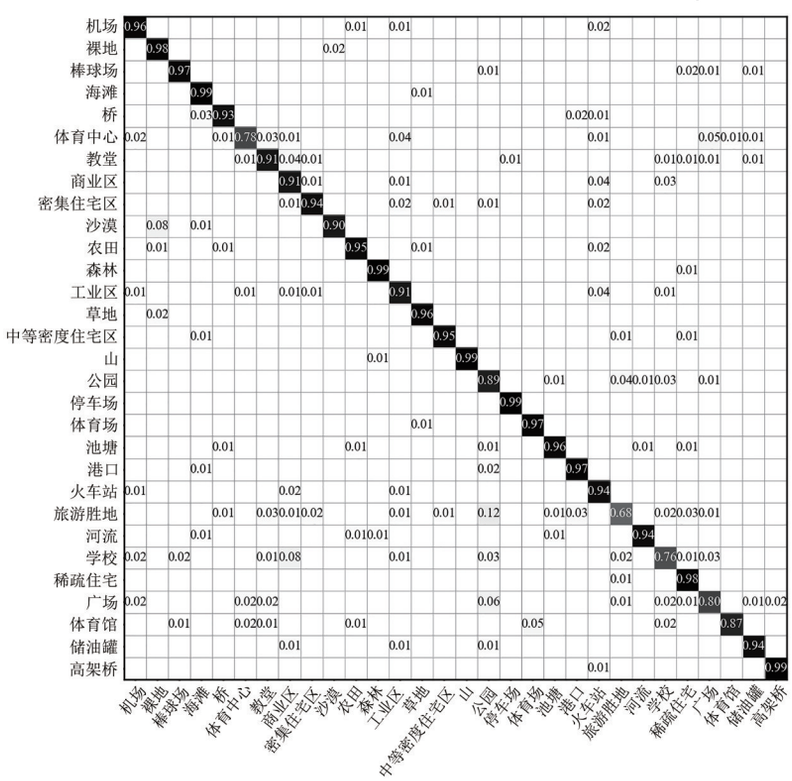
方法	不同训练集比例下的总体精度/%	
	20	50
CaffeNet(Xia等,2017)	86.86±0.47	89.53±0.31
VGG-VD-16(Xia等,2017)	86.59±0.29	89.64±0.36
GoogLeNet(Xia等,2017)	83.44±0.40	86.39±0.55
ResNet-50(钱晓亮 等,2018)	88.23±0.70	91.31±0.58
DTDCNN(施慧慧 等,2021)	—	89.26
LCPP(Sun等,2021)	90.96±0.33	93.12±0.28
MICP(Qi等,2021)	92.45±0.46	94.94±0.34
SAFF(Cao等,2021)	90.25±0.29	93.83±0.28
MDFR(Zhang等,2019)	90.62±0.27	93.37±0.29
WSADAN-VGG16（本文方法）	92.78±0.16	95.18±0.37
WSADAN-ResNet50（本文方法）	93.73±0.23	95.88±0.24

图8展示了AID数据集20%训练比率下不同模型得到的混淆矩阵，其中图8（a）和图8（b）分别表示Baseline-VGG16和WSADAN-VGG16得到的混淆矩阵。WSADAN-VGG16对于大部分类别的精度都达到了90%以上，其中提升较为明显的场景类别如体育中心、学校和广场类别。沙漠和体育馆的精度有所下降，原因是本文方法融合其他尺度特征时对尺度不敏感图像的特征表达可能引入无效信息，导致分类精度有所下降。沙漠最易混淆的类别是裸地，体育馆最易混淆的类别是体育场。图9展示了上述几种易混淆类别的图像及其热力图，沙漠和裸地、体育馆和体育场图像非常相似，尺度信息无法提升模型对于这几种类别的判别力。



(a) Baseline-VGG16在20%训练数据的AID数据集上混淆矩阵

(a) Confusion matrix of Baseline-VGG16 on the AID dataset with the training ratio of 20%



(b) WSADAN-VGG16在20%训练数据的AID数据集上混淆矩阵

(b) Confusion matrix of WSADAN-VGG16 on the AID dataset with the training ratio of 20%

图8 AID数据集上WSADAN与Baseline混淆矩阵对比图

Fig. 8 WSADAN vs. Baseline confusion matrix on the AID dataset

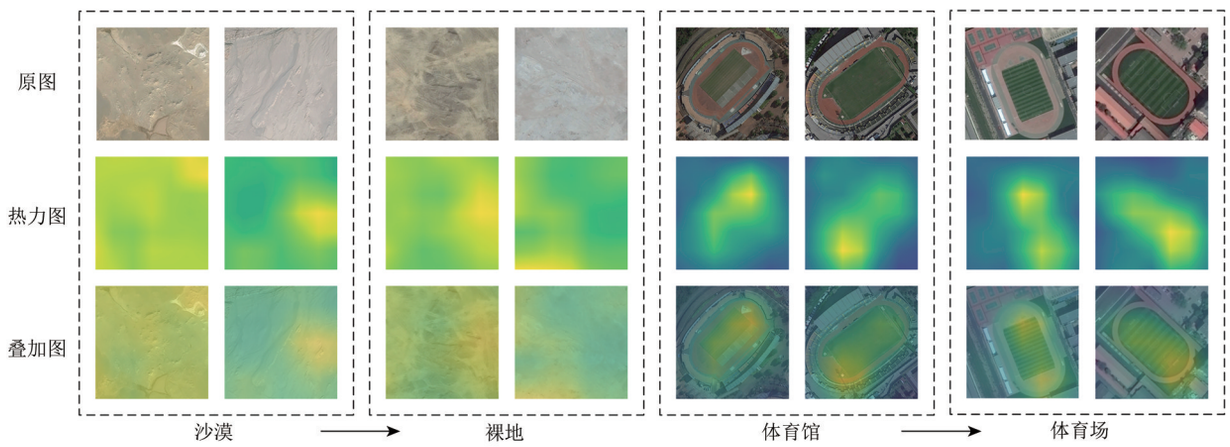


图9 AID数据集上错误分类样本

Fig.9 The misclassified samples of the AID dataset

3.4.3 NWPU数据集实验

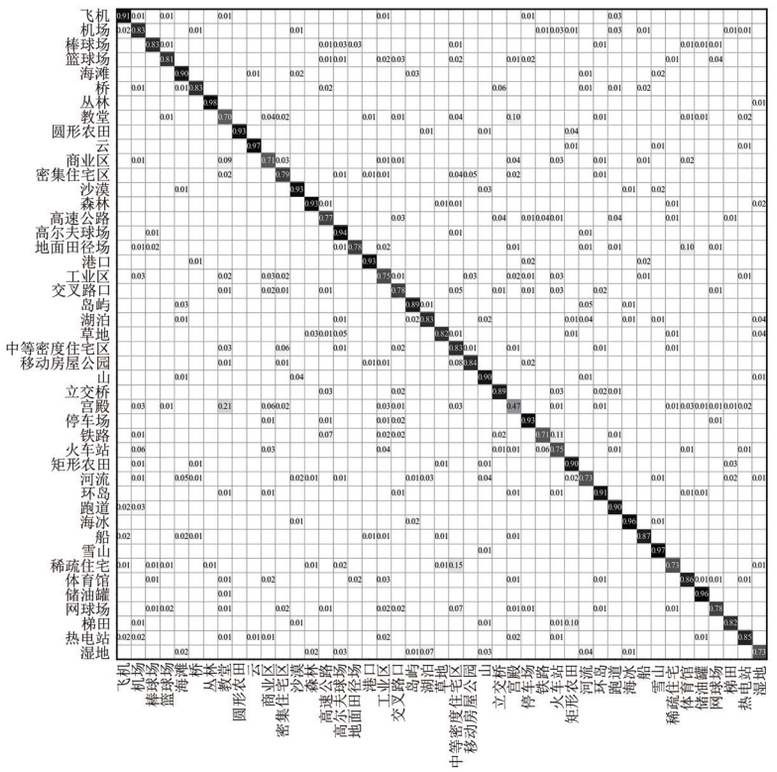
本文在NWPU数据集上进行了评估实验，表6展示为总体精度对比结果表。本文基于VGG16和基于ResNet50的方法WSADAN-VGG16、WSADAN-ResNet50在NWPU数据集10%训练数据情况下的总体精度分别为87.01%和90.71%，在20%训练数据情况下的总体精度分别为90.44%和92.63%。与传统的网络模型CaffeNet、VGG-VD-16、GoogLeNet、ResNet-50相比，本文方法网络性能表现较优。DTDCNN（施慧慧 等，2021）和MICP（Qi等，2021）采用多尺度池化的方式获取影像多尺度特征，SAFF（Cao等，2021）和MDFR（Zheng等，2019）采用多尺度特征融合的方式获取影像多尺度特征。本文方法相较于DTDCNN、SAFF和MDFR总体精度更高，模型表现更好，与MICP相比本文的WSADAN-VGG16方法总体精度相当，这可能是因为MICP方法中融合了尺度信息和旋转不变性信息。

表6 NWPU数据集不同方法的总体精度对比

Table 6 Comparison of overall accuracy for different methods on the NWPU dataset

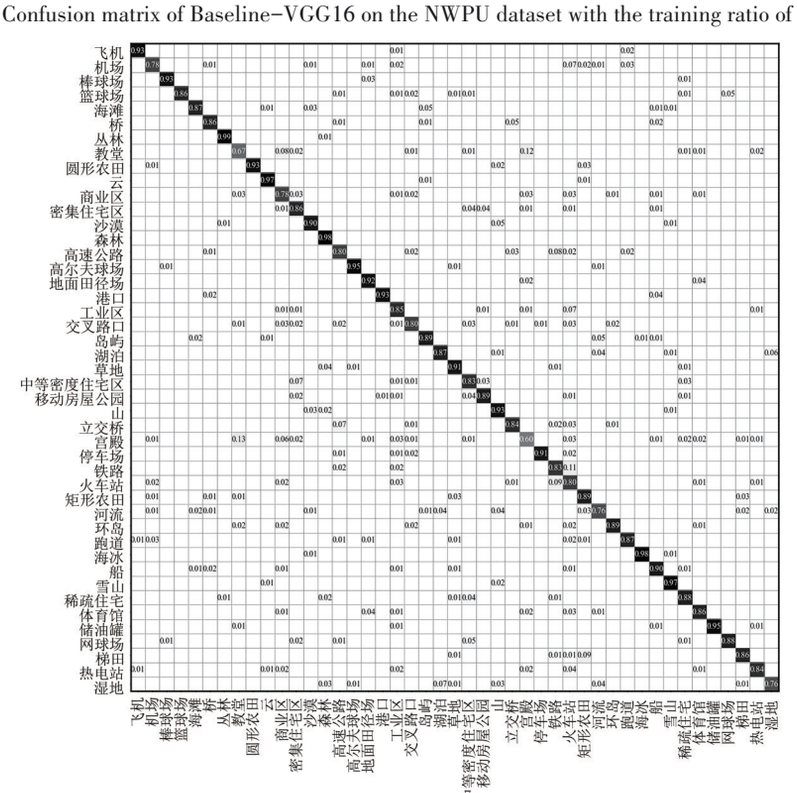
方法	不同训练集比例下的总体精度/%	
	10	20
CaffeNet(Xia等,2017)	78.01±0.27	81.08±0.21
VGG-VD-16(Xia等,2017)	76.47±0.18	79.79±0.15
GoogLeNet(Xia等,2017)	76.19±0.38	78.48±0.26
ResNet-50(钱晓亮 等,2018)	—	84.39±0.29
DTDCNN(施慧慧 等,2021)	—	84.12
MICP(Qi等,2021)	87.54±0.31	90.49±0.28
SAFF(Cao等,2021)	84.38±0.19	87.86±0.14
MDFR(Zhang等,2019)	83.37±0.26	86.89±0.17
WSADAN-VGG16（本文方法）	87.01±0.26	90.44±0.40
WSADAN-ResNet50（本文方法）	90.71±0.23	92.63±0.19

图10展示了NWPU数据集10%训练比率下不同模型得到的混淆矩阵，其中图10（a）和图10（b）分别表示Baseline-VGG16和WSADAN-VGG16的混淆矩阵。WSADAN-VGG16较Baseline-VGG16精度提升明显的类别包括：棒球场、地面田径场、工业区、宫殿、铁路和稀疏住宅等场景类别。机场和立交桥分类总体精度有所下降，可能是因为在考虑到多尺度宏观特征时忽略了细节上的注意力信息，导致分类错误。机场最易混淆的类别是火车站，立交桥最易混淆的类别是高速公路。图11展示了上述几种易混淆类别的图像及其热力图，可以看到机场和火车站建筑类型、颜色和纹理等比较相似，立交桥和高速公路具有一样的道路、车辆等地物信息。



（a）Baseline-VGG16在10%训练数据的NWPU数据集上混淆矩阵

（a）Confusion matrix of Baseline-VGG16 on the NWPU dataset with the training ratio of 10%



（b）WSADAN-VGG16在10%训练数据的NWPU数据集上混淆矩阵

（b）Confusion matrix of WSADAN-VGG16 on the NWPU dataset with the training ratio of 10%

图10 NWPU数据集上WSADAN与Baseline混淆矩阵对比图

Fig. 10 WSADAN vs. Baseline confusion matrix on the NWPU dataset

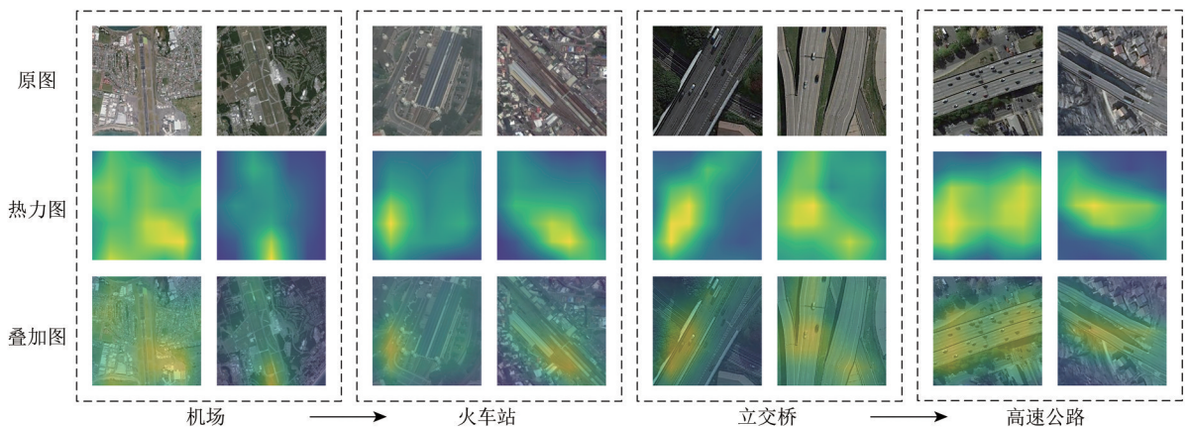


图11 NWPU数据集上错误分类样本

Fig. 11 Misclassified samples of the NWPU dataset

4 结 论

传统卷积神经网络具有固定感受野，很难获取遥感图像中不同尺度地物特征，进而影响遥感图像场景分类的精度。因此，本文提出了一种弱监督尺度自适应增强网络，该网络通过弱监督方式提升卷积特征对于不同尺度地物的表达能力。通过消融实验验证了本文提出的尺度生成模块和尺度融合模块的有效性，通过场景尺度变化分析验证了本文方法对于场景尺度变化具有一定的学习能力和适应能力。本文在RSSCN7、AID和NWPU这3个数据集上的实验结果证明了提出模型性能优于其他解决尺度效应问题的遥感场景分类方法。本文方法相比多尺度池化和多尺度特征融合方法泛化能力更强，并且无需针对特定数据集做大量尺度探索实验。

本文提出的方法对于尺度不敏感地物和依赖于细节特征的场景类别效果不理想。在未来研究中，将探讨在提取多尺度特征的同时融合图像注意力机制进一步提升模型表达能力。

参考文献（References）

Alhichri H, Alajlan N, Bazi Y and Rabczuk T. 2018. Multi-scale convolutional neural network for remote sensing scene classification// Proceedings of the 2018 IEEE International Conference on Electro/Information Technology (EIT). Rochester: IEEE: 1-5 [DOI: 10.1109/EIT.2018.8500107]

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]

Chen L J, Yang W, Xu K and Xu T. 2011. Evaluation of local features for scene classification using VHR satellite images//Proceedings of 2011 Joint Urban Remote Sensing Event. Munich: IEEE: 385-388 [DOI: 10.1109/JURSE.2011.5764800]

Chen W T, Li X J, He H X and Wang L Z. 2018. Assessing different feature sets' effects on land cover classification in complex surface-mined landscapes by ZiYuan-3 satellite imagery. Remote Sensing, 10(1): 23 [DOI: 10.3390/rs10010023]

Chen Y Y, Ming D P and Lv X W. 2019. Superpixel based land cover classification of VHR satellite image combining multi-scale CNN and scale parameter estimation. Earth Science Informatics, 12(3): 341-363 [DOI: 10.1007/s12145-019-00383-2]

Cheng G, Guo L, Zhao T Y, Han J W, Li H H and Fang J. 2013. Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. International Journal of Remote Sensing, 34(1): 45-59 [DOI: 10.1080/01431161.2012.705443]

Cheng G, Han J W and Lu X Q. 2017. 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]

Gao Y, Shi J, Li J and Wang R Y. 2020. Remote sensing scene classification with dual attention-aware network//Proceedings of the IEEE 5th International Conference on Image, Vision and Computing (ICIVC). Beijing: IEEE: 171-175 [DOI: 10.1109/ICIVC50857.2020.9177460]

Gu Y T, Wang Y T and Li Y S. 2019. A survey on deep learning-driven remote sensing image scene understanding: scene classification, scene retrieval and scene-guided object detection. Applied Sciences, 9(10): 2110 [DOI: 10.3390/app9102110]

Han X B, Zhong Y F, Cao L Q and Zhang L P. 2017. Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. Remote Sensing, 9(8): 848 [DOI: 10.3390/rs9080848]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

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]

Kingma D P and Ba J. 2014. Adam: a method for stochastic optimization//Proceedings of the 3rd International Conference on Learning Representations. San Diego: [s.n.]

Längkvist M, Kiselev A, Alirezaie M and Loutfi A. 2016. Classification and segmentation of satellite orthoimagery using convolutional neural networks. Remote Sensing, 8(4): 329 [DOI: 10.3390/rs8040329]

Liu Q S, Hang R L, Song H H, Zhu F P, Plaza J and Plaza A. 2016. Adaptive deep pyramid matching for remote sensing scene classification. arXiv:1611.03589 [DOI: 10.48550/arXiv.1611.03589]

Liu Y S, Liu Y B and Ding L W. 2018. Scene classification based on two-stage deep feature fusion. IEEE Geoscience and Remote Sensing Letters, 15(2): 183-186 [DOI: 10.1109/LGRS.2017.2779469]

Luo B, Jiang S J and Zhang L P. 2013. Indexing of remote sensing images with different resolutions by multiple features. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6(4): 1899-1912 [DOI: 10.1109/JSTARS.2012.2228254]

Lv Z H, Li X M, Zhang B Y, Wang W X, Zhu Y Y, Hu J X and Feng S Z. 2016. Managing big city information based on WebVRGIS. IEEE Access, 4: 407-415 [DOI: 10.1109/ACCESS.2016.2517076]

Ma X Y, Wang L M, Qi K L and Zheng G Z. 2021. Remote sensing image scene classification method based on multi-scale cyclic attention network. Earth Science, 46(10): 3740-3752 (马欣悦, 王梨名, 祁昆仑, 郑贵洲. 2021. 基于多尺度循环注意力网络的遥感影像场景分类方法. 地球科学, 46(10): 3740-3752) [DOI: 10.3799/dqkx.2020.365]

Martha T R, Kerle N, Van Westen C J, Jetten V and Kumar K V. 2011. Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis. IEEE Transactions on Geoscience and Remote Sensing, 49(12): 4928-4943 [DOI: 10.1109/TGRS.2011.2151866]

Qi K L, Yang C, Guan Q F, Wu H Y and Gong J Y. 2017. A multiscale deeply described correlatons-based model for land-use scene classification. Remote Sensing, 9(9): 917 [DOI: 10.3390/rs9090917]

Qi K L, Yang C, Hu C L, Zhai H, Guan Q F and Shen S Y. 2021. A multi-level improved circle pooling for scene classification of high-resolution remote sensing imagery. Neurocomputing, 462: 506-522 [DOI: 10.1016/j.neucom.2021.08.022]

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, 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]

Shi H H, Xu Y N, Teng W X and Wang N. 2021. Scene classification of high-resolution remote sensing imagery based on deep transfer deformable convolutional neural networks. Acta Geodaetica et Cartographica Sinica. 50(05): 652-663 (施慧慧, 徐雁南, 滕文秀, 王妮. 2021. 高分辨率遥感影像深度迁移可变形卷积的场景分类法. 测绘学报, 50(5): 652-663) [DOI: 10.11947/j.AGCS.2021.20200190]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego: IEEE: 7-9

Sun X L, Zhu Q Q and Qin Q Q. 2021. A multi-level convolution pyramid semantic fusion framework for high-resolution remote sensing image scene classification and annotation. IEEE Access, 9: 18195-18208 [DOI: 10.1109/ACCESS.2021.3052977]

Wu H, Liu B Z, Su W H, Zhang W C and Sun J G. 2016. Deep filter banks for land-use scene classification. IEEE Geoscience and Remote Sensing Letters, 13(12): 1895-1899 [DOI: 10.1109/LGRS.2016.2616440]

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]

Yang Y and Newsam S. 2008. Comparing SIFT descriptors and gabor texture features for classification of remote sensed imagery//Proceedings of the 15th IEEE International Conference on Image Processing. San Diego: IEEE: 1852-1855 [DOI: 10.1109/ICIP.2008.4712139]

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]

Yao Y Q, Cheng G, Xie X X and Han J W. 2021. Optical remote sensing image object detection based on multi-resolution feature fusion. National Remote Sensing Bulletin, 25(5): 1124-1137 (姚艳清, 程塨, 谢星星, 韩军伟. 2021. 多分辨率特征融合的光学遥感图像目标检测. 遥感学报, 25(5): 1124-1137) [DOI: 10.11834/jrs.20210505]

Yu D H, Zhang B M, Zhao C, Guo H T and Lu J. 2020. Scene classification of remote sensing image using ensemble convolutional neural network. Journal of Remote Sensing, 24(6): 717-727 (余东行, 张保明, 赵传, 郭海涛, 卢俊. 2020. 联合卷积神经网络与集成学习的遥感影像场景分类. 遥感学报, 24(6): 717-727) [DOI: 10.11834/jrs.20208273]

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]

Zan L Y, Li B P, Lu K X, Chen Z C and Zhang B. 2021. Intelligent detection of chemical plant based on poly FPN neural network model. National Remote Sensing Bulletin (咎露洋, 李柏鹏, 卢凯旋, 陈正超, 张兵. 2021. 基于Poly-FPN神经网络模型的化工厂智能检测. 遥感学报) [DOI: 10.11834/jrs.20210005]

Zeng D, Chen S J, Chen B Y and Li S Y. 2018. Improving remote sensing scene classification by integrating global-context and local-object features. Remote Sensing, 10(5): 734 [DOI: 10.3390/rs10050734]

Zhang J, Zhang M, Shi L K, Yan W J and Pan B. 2019. A multi-scale approach for remote sensing scene classification based on feature maps selection and region representation. Remote Sensing, 11(21): 2504 [DOI: 10.3390/rs11212504]

Zhang S Y. 2020. High-Resolution Remote Sensing Image Land Cover Classification Based on Deep Learning and Multi-Scale and Multi-Feature Fusion. Hangzhou: Zhejiang University (张书瑜. 2020. 基于深度学习和多尺度多特征融合的高分辨率遥感地表覆盖分类研究. 杭州: 浙江大学) [DOI: 10.27461/d.cnki.gzjdx.2020.001324]

Zhao L J, Tang P and Huo L Z. 2016. Feature significance-based multi-bag-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 W Z and Du S H. 2016. Learning multiscale and deep representations for classifying remotely sensed imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 113: 155-165 [DOI: 10.1016/j.isprsjprs.2016.01.004]

Zheng X T, Yuan Y and Lu X Q. 2019. A deep scene representation for aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing, 57(7): 4799-4809 [DOI: 10.1109/TGRS.2019.2893115]

Zheng Y. 2015. Methodologies for cross-domain data fusion: an overview. IEEE Transactions on Big Data, 1(1): 16-34 [DOI: 10.1109/TBDATA.2015.2465959]

Zhong Y F, Zhu Q Q and Zhang L P. 2015. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 53(11): 6207-6222 [DOI: 10.1109/TGRS.2015.2435801]

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 Q Q, Li Z, Zhang Y N, Li J L, Du Y Q, Guan Q F and Li D R. 2021. Global-Local-Aware conditional random fields based building extraction for high spatial resolution remote sensing images. National Remote Sensing Bulletin, 25(7): 1422-1433 (朱祺琪, 李真, 张亚男, 李佳伦, 杜禹强, 关庆锋, 李德仁. 2021. 全局局部细节感知条件随机场的高分辨率遥感影像建筑物提取. 遥感学报, 25(7): 1422-1433) [DOI: 10.11834/jrs.20210360]

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]

Weakly supervised scale adaptation data augmentation for scene classification of high−resolution remote sensing images

WANG Liming[1], QI Kunlun[1,2], YANG Chao[1,2], WU Huayi[3]

1.School of Geography and Information Engineering, China University of Geosciences (Wuhan), Wuhan 430074, China;
2.National Engineering Research Center of Geographic Information System, Wuhan 430074, China;
3.State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Abstract: Scene classification of remote sensing images aims to assign a meaningful label to a given image. In recent years, Convolutional Neural Networks (CNNs)-based methods make a breakthrough and substantially outperform traditional methods in scene classification tasks of remote sensing images. However, obtaining features under different scales in remote sensing images is difficult due to the fixed receptive field of CNNs. This complexity seriously affects the performance of CNNs in scene classification of remote sensing images. This study proposes a method to learn the optimal scales for different scene image instances in a weakly supervised manner.

A Weakly Supervised Scale Adaptive Data Augmentation Network (WSADAN) is proposed to capture feature information at different scales of remote sensing scenes, and a scale generation module and a scale fusion module are designed to improve the robustness. The scale generation module learns the optimal scale parameters based on the CNN features of the original image. The scale fusion module filters the CNN features of images with original and optimal scales to remove the noise and then deeply fuses them to exploit the correlation between

features at different scales. The deeply fused multi-scale features are input into a fully connected layer to predict categories of scene images.

The effectiveness of the scale generation and scale fusion modules is verified by ablation experiments. The accuracy of WSADANSGM compared with the baseline improves by 0.94% and 0.89% for the 20% and 50% training data ratios of RSSCN7 dataset, 1.27% and 0.87% for the 20% and 50% training data ratios of AID dataset, and 1.09% and 0.71% for the 10% and 20% training data ratios of NWPU dataset, respectively. Compared with WSADANSGM, WSADANSGM+SFM improves by 1.65% and 1.32% for the RSSCN7 dataset at 20% and 50% training data ratios, 1.65% and 1.26% for the AID dataset at 20% and 50% training data ratios, and 1.75% and 1.42% for the NWPU dataset at 10% and 20% training data ratios, respectively. In the experiment for scene scale change analysis, the classification accuracy of our method is higher than the baseline at any scale of image, which proves that our method can learn certain image scale information and has strong scale adaptation ability. We use three datasets for remote sensing scene classification, namely, RSSCN7, AID, and NWPU, for the experiments. On the RSSCN7 dataset, the overall accuracies are 91.65% and 94.07% with the training ratios of 20% and 50% for WSADAN-VGG16. For WSADAN-ResNet50, the corresponding accuracies are 92.69% and 94.82%. On the AID dataset, the overall accuracies are 92.78% and 95.18% with the training ratios of 20% and 50% for WSADAN-VGG16. For WSADAN-ResNet50, the corresponding accuracies are 93.73% and 95.88%. On the NWPU dataset, the overall accuracies are 87.01% and 90.44% with the training ratios of 10% and 20% for WSADAN-VGG16. For WSADAN-ResNet50, the corresponding accuracies are 90.71% and 92.63%.

The proposed method can learn CNN features at a wider range of scales without manual multi-scale selection for different datasets. The performance of the proposed method is better than that of traditional CNNs, especially for the scene categories containing objects with large-scale variations.

key word: remote sensing, scene classification, deep learning, convolutional neural networks, weakly supervision, multi-scale, data augmentation

Supported by Hubei Key Research and Development Program in China (No. 2020AAA004)