引文格式:李彦胜,孔德宇,张永军,等.联合稳健跨域映射和渐进语义基准修正的零样本遥感影像场景分类[J].测绘学报,2020,49(12):1564-1574. DOI:10.11947/j.AGCS.2020.20200139.

LI Yansheng, KONG Deyu, ZHANG Yongjun, et al. Zero-shot remote sensing image scene classification based on robust cross-domain mapping and gradual refinement of semantic space[J]. Acta Geodaetica et Cartographica Sinica, 2020, 49(12): 1564-1574. DOI: 10.11947/j.AGCS.2020.20200139.

联合稳健跨域映射和渐进语义基准修正的零样本遥感影像场景分类

李彦胜,孔德宇,张永军,季 铮,肖 锐

武汉大学遥感信息工程学院,湖北 武汉 430079

Zero-shot remote sensing image scene classification based on robust cross-domain mapping and gradual refinement of semantic space

LI Yansheng, KONG Deyu, ZHANG Yongjun, JI Zheng, XIAO Rui

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

Abstract: Zero-shot classification technology aims to acquire the ability to identify categories that do not appear in the training stage (unseen classes) by learning some categories of the data set (seen classes), which has important practical significance in the era of remote sensing big data. Until now, the zero-shot classification methods in remote sensing field pay little attention to the semantic space optimization after mapping, which results in poor classification performance. Based on this consideration, this paper proposed a zero shot remote sensing image scene classification method based on cross-domain mapping with auto-encoder and collaborative representation learning. In the supervised learning module, based on the class semantic vector of seen class and the scene image sample, the depth feature extractor learning and robust mapping from visual space to semantic space are realized. In the unsupervised learning stage, based on the class semantic vectors of all classes and the unseen remote sensing image samples, collaborative representation learning and k-nearest neighbor algorithm are used to modify the semantic vectors of unseen classes, so as to alleviate the problem of the shift of seen class semantic space and unseen class semantic space one after another and unseen after self coding cross domain mapping model mapping the shift of class semantic space and unseen class semantic space after collaborative representation. In the testing phase, based on the depth feature extractor, self coding cross domain mapping model and modified unseen class semantic vector, the classification of unseen class remote sensing image scene can be realized. We integrate a number of open remote sensing image scene data sets and build a new remote sensing image scene data set, experiments were conducted using this dataset The experimental results show that the algorithm proposed in this paper were significantly better than the existing zero shot classification method in the case of a variety of seen and unseen classes.

Key words: zero-shot learning; remote sensing image scene classification; cross-domain mapping with auto-encoder; collaborative representation learning; natural language processing

Foundation support: The National Natural Science Foundation of China (Nos. 42030102; 41971284); The Foundation for Innovative Research Groups of the Hubei Natural Science Foundation (No. 2020CFA003)

摘 要:零样本影像分类技术旨在通过学习数据集的部分类别(可见类),获得识别在训练阶段未出现类别(不可见类)的能力。该技术在遥感大数据时代具有重要现实意义。目前,遥感领域的零样本场景分类方法对于映射后的语义空间优化关注很少,导致已有方法的整体分类性能较差。基于这一考虑,本文提出了一种基于稳健跨域映射和渐进语义基准修正的零样本遥感影像场景分类方法。在训练的有监督学习模块,基于可见类的类别语义向量和场景影像样本,实现深度特征提取器学习和视觉空间到语义空

间的稳健映射。在训练的无监督学习阶段，基于全体类别的类别语义向量和不可见类遥感影像样本，分别通过协同表示学习和 k 近邻算法来渐进修正不可见类类别的语义向量，从而缓解可见类语义空间与不可见类语义空间的漂移问题和自编码跨域映射模型映射后不可见类语义空间与协同表示后不可见类语义空间的偏移问题。在测试阶段，基于学习所得的深度特征提取器、自编码跨域映射模型和修正后的不见类语义向量，实现对不可见类遥感影像场景的分类。本文整合多个已有公开的遥感影像场景数据集，组建了一个新的遥感影像场景数据集，在此数据集上进行试验。试验结果表明本文提出的算法在多种不同的可见类与不可见类的划分情况下都明显优于已有公开零样本分类方法。

关键词：零样本学习；遥感影像场景分类；自编码跨域映射；协同表示学习；自然语言模型

中图分类号：P237　　**文献标识码**：A　　**文章编号**：1001-1595(2020)12-1564-11

基金项目：国家自然科学基金(42030102；41971284)；湖北省自然科学基金计划创新群体项目(2020CFA003)

进入 21 世纪之后，遥感技术发展越发迅猛，在土地资源调查、生态环境监测、灾害分析和预测等方面发挥着重要的作用[1]。随着遥感影像分辨率的提高[2]，基于像素和对象的分类方法广泛受到高分辨率遥感影像“同物异谱、同谱异物”现象的影响，无法满足高效稳定遥感影像解译的需求。基于这一考虑，遥感影像场景分类[3-5]受到国内外研究学者的广泛关注。遥感影像场景分类旨在通过挖掘遥感影像场景(影像块)内的视觉基元及视觉基元间的空间关系来预测影像块的语义类别，可以极大地降低像素级或对象级地物解译的混淆度，从而提高高分辨率遥感影像解译的稳定性及准确度，在基于内容的遥感影像检索[6-7]和遥感影像目标检测[8-10]等方面都有重要应用。

随着遥感影像场景数据集的不断开放，多领域研究人员提出了大量基于人工特征或深度学习的遥感影像场景分类方法[11-13]。随着遥感大数据时代的来临，遥感地物类别呈现爆炸式增长趋势，因此为所有类别都搜集充足的遥感影像样本是不现实的，但是现有的监督或半监督的遥感影像场景分类方法均需要依赖全部类别的遥感影像样本来学习分类模型，无法灵活应对出现的新的场景类别。如何将遥感领域的先验知识引入遥感影像场景理解过程，仅通过学习含有遥感影像的部分类别，就可以识别在训练阶段从未出现类别的遥感影像场景，一方面可以降低对场景类别样本的标注成本，另一方面将提高对新出现的场景的识别能力。因此，近年来零样本学习[14-15](zero-shot learning)的发展为遥感影像场景分类提供了新的思路。目前，零样本学习主要集中于计算机视觉领域，其在遥感影像场景分类中的研究还很少，需要大量研究工作来推进这一技术的发展。

大量心理学研究表明，人类可以识别大约 3 万种物体种类[16]，同时也可以对这些类中包含的子类进行分辨。人类可以从以往的学习中获得和积累先验知识，并将经验和知识运用到解决新的问题中，以此提高了人类的推理能力。零样本学习旨在模拟人类学习的过程，通过学习可见类(seen)样本中的知识，加以辅助信息(属性，词向量)的帮助来推断识别不可见类(unseen)中的样本，并且可见类与不可见类是不相交的，通常可见类样本用于训练，不可见类样本用于测试。在遥感影像中，不同场景类别可能包含相似或相同的对象，使得可以从一些已有场景中学习到各种对象，进行重新组合和演化得到新的场景类别。因此，零样本学习在遥感影像场景分类中具有广阔的发展前景。

近几年来，零样本学习在计算机视觉领域发展迅速。具体的，文献[17]提出的语义自编码(SAE)方法，采用基于编码-解码的架构，在自编码器进行编码和解码时，使用了原始数据作为约束，编码后的数据能够尽可能恢复为原来的数据，增强了对不可见类的泛化识别能力。文献[18]提出的双视觉语义映射(DMaP)方法，对语义嵌入空间进行迭代优化，希望能够使得视觉特征和语义特征在模型学习的过程中尽量对齐，得到样本最优的语义表示。这些方法在诸如 AWA2[19]、CUB[20]等数据集上都取得了较为理想的结果。然而在遥感领域，一方面遥感影像由于光照，拍摄角度和季节等原因具有类内差异性大和类间相似性高等现象；另一方面，对于类别的语义特征通常使用预训练的自然语言模型根据类别名称提取语义向量，但是广义的自然语言模型常常无法贴切地描述地学的地物类别。因此，现有的零样本学

习方法在遥感影像场景分类任务中都很难取得理想的结果。亟待提出适用于遥感领域的零样本分类方法。

基于上述考虑,本文提出了一种基于稳健跨域映射和渐进语义基准修正的零样本遥感影像场景分类方法。在训练的有监督学习模块,基于可见类的类别语义向量和遥感影像场景样本,联合场景类别分类和自编码跨域映射的多任务学习来实现深度特征提取器学习和遥感影像场景的视觉空间到类别语义空间的稳健映射。考虑到在学习映射矩阵的过程中,不可见类的语义特征没有参与映射模型的学习,而这通常会导致经过映射得到的不可见类语义向量出现域偏移问题,本文利用协同表示重构不可见类语义向量真值,增强了可见类的语义向量与不可见类的语义向量之间的联系。此外,由于视觉特征和语义特征的来源不同,这往往导致通过映射得到的语义向量与经过协同表示修正的语义向量之间的结构差异较大,对此,本文利用 k 近邻算法求经过协同表示重构的语义向量真值在映射得到的语义向量中的近邻向量并求其均值,以此对语义特征空间进行修正,使得映射得到的语义向量与协同表示修正的语义向量尽量对齐。考虑到已有的遥感影像场景数据集的类别数较少,不利于充分验证零样本分类技术的性能,本文基于公开的遥感数据集,筛选整合后得到类别更加多样、类内样本更加丰富的新遥感影像场景数据集,在此数据集上,使用 Word2Vec 和 Bert 两种语言模型分别提取语义向量。大量试验结果表明,本文方法相较于其他零样本分类方法在不同的可见类与不可见类划分比例上均具有更好的分类准确度。

1 零样本分类相关工作

本文主要从两个方面来综述讨论相关工作:计算机视觉领域的零样本分类;遥感领域的零样本分类。

1.1 计算机视觉领域的零样本分类

目前的零样本学习任务,很大一部分的研究思路是"视觉特征+语义特征+机器学习方法"。其中视觉特征通常由深度卷积网络提取得到。语义特征一般包括属性和词向量,在基于属性的零样本图像分类器模型中,属性需要考虑样本是否具有某一种属性,根据属性的有、无可以确定样本在属性空间的位置,进而确定样本的类别标签。属性一般需要人工对特定数据集进行标注,这需要标注人员具有一定的专业知识。例如文献[21]提出的直接属性预测模型 DAP(direct attribute prediction)和间接属性预测模型 IAP(indirect attribute prediction)。基于属性的方法存在的问题是:①建立一个合理有效的属性库十分困难;②类别属性向量的标注成本较大;③其扩展性较弱。

考虑到使用属性作为语义特征存在的种种局限,目前的研究普遍将词向量作为语义特征,词向量是指利用自然语言处理技术将词语映射到一个新的空间中,并以多维的连续实数向量表示叫作"Word Embedding"。例如文献[22]提出的 Word2Vec,文献[23]提出的 GloVe,以及 Bert 方法(arXiv preprint arXiv:1810.04805,2018.)。这些方法可以揭示词与词之间的关联性,使由自然语言转换得到的向量具有了语义上的信息,这些实数向量之间的距离可以很好地表述相应词之间的语义相似性。在得到视觉特征和语义特征后,利用机器学习方法完成分类任务。要给出适当总结与讨论。

随着生成对抗网络(generative adversarial network,GAN)[24]技术的快速发展,涌现出了一些利用生成对抗网络进行视觉样本生成的零样本分类方法。文献[25]借鉴了 Conditional GAN,以类别属性为输入生成类别对应的视觉特征,在判别器中加入了类别的分类损失;文献[26]在训练生成器 G 的过程中,引入了 hallucinated text,鼓励生成的视觉特征偏离可见类,希望生成的样本更具多样性。

目前零样本分类方法更多侧重计算机视觉任务。由于遥感影像场景的结构复杂性和充分描述场景类别的语义向量较难获取等问题,已有计算机视觉领域的零样本分类方法往往无法直接应用于遥感影像场景的零样本分类任务。

1.2 遥感领域的零样本分类

近几年来,零样本学习技术逐渐应用于遥感领域。文献[27]率先在遥感影像细粒度识别领域开展零样本分类研究,并建立了适用于细粒度识别任务的数据集。文献[28]将零样本学习应用于 SAR 影像的目标识别任务中。文献[29]在用于 PolSAR 土地覆盖分类的广义零样本学习框架进行了研究。面向遥感影像场景分类的零样本学习,也有一些学者开展了相关研究。具体的,文献[30]在零样本学习中引入了半监督 Sammon 嵌入算法。文献[31]提出一种标签传播算法,采用

基于稀疏学习的标签细化方法抑制分类结果中的噪声。文献[32]利用不同图像特征之间的互补性，提出基于图像特征融合的分类方法，减少冗余信息且保留各自图像特征自身的特点。文献[33]利用不同词向量之间的互补性，采用解析字典方法获取各语义词向量的稀疏系数，以减少冗余信息。目前，遥感领域的零样本分类方法通常通过融合多种语义特征或视觉特征，改善映射矩阵的方法提高分类精度。但是，对于映射后的语义空间很少关注。实际上，不可见类的语义向量重建与修正对于分类性能有较大影响，这也是本文的重要研究动机。

2　联合稳健跨域映射和渐进语义基准修正的零样本遥感影像场景分类方法

为了更清晰地描述零样本分类问题，本文首先给出零样本分类问题的符号定义。设 $\boldsymbol{D}=\{(x_i,y_i):i=1,2,\cdots,M\}$ 代表可见类遥感数据集，x_i 表示可见类中的第 i 张遥感影像场景，y_i 表示可见类中第 i 张影像的类别标签，M 为可见类遥感数据的样本总数；$\boldsymbol{D}^U=\{(x_i^U,y_i^U):i=1,2,\cdots,N\}$ 代表不可见类遥感数据集，x_i^U 表示不可见类中的第 i 张遥感影像场景，y_i^U 表示不可见类中第 i 张影像的类别标签，N 为不可见类数据的样本总数；$\boldsymbol{D}\bigcap\boldsymbol{D}^U=\varnothing$，也即可见类与不可见类的类别及数据是完全不重叠的。基于广义的语义基准（例如自然语言语料库），每个遥感场景类别都对应一个语义向量，令 $\boldsymbol{S}=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_p\}\in R^{d^s\times p}$ 表示可见类语义向量，$\boldsymbol{S}^U=\{\boldsymbol{s}_1^U,\boldsymbol{s}_1^U,\cdots,\boldsymbol{s}_q^U\}\in R^{d^s\times q}$ 表示不可见类语义向量集合，其中 p 和 q 分别表示可见类和不可见类的类别数，d^s 为语义向量维数。

本文提出的零样本遥感影像场景分类方法的整体框架如图 1 所示。本文算法的训练阶段包括有监督学习和无监督学习两个模块。有监督学习模块主要利用有类别标签的可见类遥感影像场景样本 $\boldsymbol{D}$ 和可见类类别的语义向量 $\boldsymbol{S}$，来完成深度特征提取器学习和语义映射模块学习，具体方法部分在 2.1 节讨论。无监督学习模块主要利用全体类别的类别语义向量 $\boldsymbol{S}$ 和 $\boldsymbol{S}^U$、无类别标签的不可见类的遥感影像场景样本库 $\boldsymbol{D}^U$ 来修正不可见类类别的语义向量，具体方法在 2.2 节描述。最后，2.3 节给出了具体测试的过程。

2.1　耦合深度卷积神经网络场景分类和自编码跨域映射的有监督稳健跨域映射

利用可见类场景分类约束训练特征提取器，使特征提取器更适用于遥感场景影像，另外由于视觉特征与语义特征在维度和结构方面具有较大差异，直接建立视觉特征到语义特征的映射往往会丢失重要信息，导致映射效果不佳。因此，建立场景分类与自编码跨域映射的多任务学习在映射过程中引入语义自编码，实现有监督下的映射矩阵学习。

2.1.1　基于可见类遥感影像场景样本的深度特征提取器学习

首先利用可见类遥感影像数据集 $\boldsymbol{D}$ 微调深度卷积网络，获取特征提取器用于描述遥感影像场景样本的视觉特征。如图 1 的有监督学习模块所示，令 $\boldsymbol{T}$ 表示深度卷积网络的卷积层超参数，$\boldsymbol{V}$ 为最后一个全连接层特征 $\boldsymbol{f}$ 与分类层 $\boldsymbol{c}$ 的映射超参数。给定一个可见类的遥感影像场景 x_i，其对应的全连接层特征可以表示为 $\boldsymbol{f}_i=Q(x_i;\boldsymbol{T})$，其中 $Q(.;.)$ 表示深度网络的非线性映射。那么，基于遥感影像场景数据集的网络优化损失函数为

$$\min_{\boldsymbol{T},\boldsymbol{V}}-\sum_{i=1}^{M}\sum_{j=1}^{p}y_i^j\log c_i^j \tag{1}$$

式中，$\boldsymbol{c}_i=\sigma(\boldsymbol{f}_i*\boldsymbol{V})$，$\sigma(\cdot)$ 表示 Softmax 映射。

通过后向传播算法优化式(1)，可以得到更新后的卷积层映射超参数 $\boldsymbol{T}$ 和全连接层映射超参数 $\boldsymbol{V}$。后续环节本文将 $\boldsymbol{T}$ 作为遥感影像场景的特征提取器。

2.1.2　基于自编码跨域映射的视觉特征空间到语义特征空间映射

对于可见类的每一张影像 x_i，提取其视觉特征 $\boldsymbol{f}_i=Q(x_i;\boldsymbol{T})$，$\boldsymbol{F}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_M]\in R^{d^f\times M}$ 为所有可见类影像视觉特征。其中 M 为可见类遥感数据的样本总数，d^f 为特征向量维数。为了减少错误传递等影响，固定上述获得特征提取器，然后学习视觉特征到语义特征的映射。考虑到遥感领域的语义向量不够精确的特点，本文在学习跨域映射模型时，引入自编码正则化损失。自编码器属于一种无监督的神经网络模型，对输入特征进行编码，然后对编码得到的特征进行解码以重构输入特征，希望重构得到的特征与初始输入特征的误差尽可能小。附加自编码损失的跨域映射的目标函数如式(2)所示

$$\min_{\boldsymbol{W}}\|\boldsymbol{W}\boldsymbol{f}-\boldsymbol{s}\|_F^2+\alpha\|\boldsymbol{f}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{f}\|_F^2 \tag{2}$$

式中，α 是自编码损失的正则化系数，来提高求解的稳定性。对式(1)优化求解得到

$$\boldsymbol{s}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{W}+\frac{1}{\alpha}\boldsymbol{W}\boldsymbol{f}\boldsymbol{f}^{\mathrm{T}}=\frac{1}{\alpha}\boldsymbol{s}\boldsymbol{f}^{\mathrm{T}}+\boldsymbol{s}\boldsymbol{f}^{\mathrm{T}} \tag{3}$$

令 $\boldsymbol{A}=\boldsymbol{s}\boldsymbol{s}^{\mathrm{T}}$，$\boldsymbol{B}=\frac{1}{\alpha}\boldsymbol{f}\boldsymbol{f}^{\mathrm{T}}$，$\boldsymbol{C}=\left(1+\frac{1}{\alpha}\right)\boldsymbol{s}\boldsymbol{f}^{\mathrm{T}}$，则式(3)最终变换为

$$\boldsymbol{A}\boldsymbol{W}+\boldsymbol{W}\boldsymbol{B}=\boldsymbol{C} \tag{4}$$

式(4)是一个 Sylvester 方程，可以利用 Bartels-Stewart 算法[34]求解得到映射矩阵 $\boldsymbol{W}$。在 python 中，利用 scipy 包中的 solve_sylvester 函数即可求解。

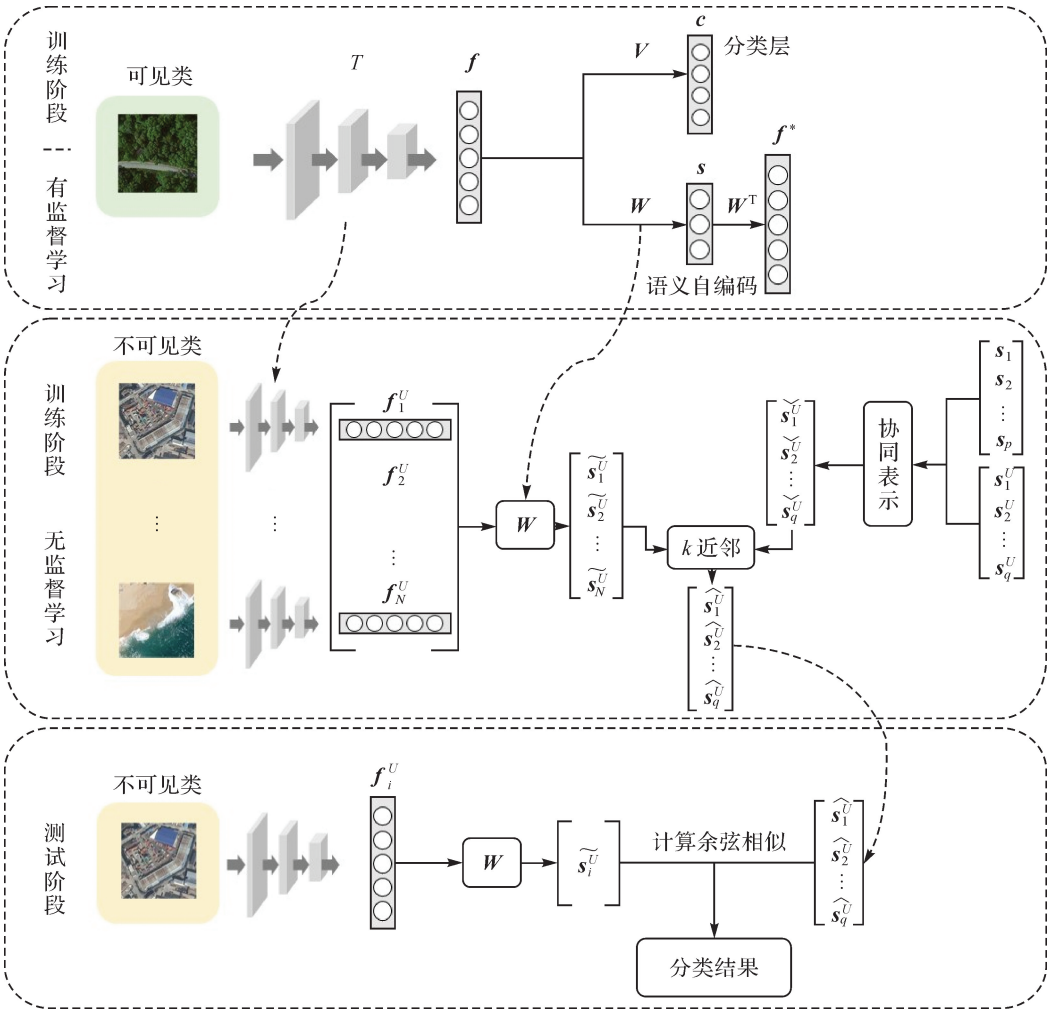


图 1 算法整体框架

Fig.1 Framework of the proposed approach

2.2 基于协同表示学习和 k 近邻算法的无监督渐进语义基准修正

经过深度特征提取器学习和自编码跨域映射，已经得到了图片特征提取器和视觉特征至语义特征的映射矩阵 $\boldsymbol{W}$。对于不可见类的每一张影像 x_i^U，提取其视觉特征 $\boldsymbol{f}_i^U=Q(x_i^U;\boldsymbol{T})$，$\boldsymbol{F}^U=[\boldsymbol{f}_1^U,\boldsymbol{f}_2^U,\cdots,\boldsymbol{f}_N^U]\in R^{d^f\times N}$ 为所有不可见类影像视觉特征。通过映射矩阵 $\boldsymbol{W}$ 对 $\boldsymbol{F}^U$ 映射得到语义矩阵 $\widetilde{\boldsymbol{S}^U}=\boldsymbol{F}^U\cdot\boldsymbol{W}$。其中 N 为可见类遥感数据的样本总数，d^f 为特征向量维数。

2.2.1 基于协同表示学习的不可见类语义向量重建

为了减小可见类语义空间与不可见类语义空间的漂移问题，无监督协同表示学习基于可见类语义向量来修正不可见类的语义向量。

考虑到可见类的语义特征 $\boldsymbol{S}$ 和不可见类的语义特征 $\boldsymbol{S}^U$ 共享同一个语义空间，因此 $\boldsymbol{S}$ 和 $\boldsymbol{S}^U$ 必然具有一定的局部相似性，即 $\boldsymbol{S}$ 一定程度上可以被 $\boldsymbol{S}^U$ 所表示。因此，在测试阶段引入了协同表示。协同表示(CR)即利用所有可见类的样本来共同表示不可见类的样本。计算展开系数的目标函数为

$$\min_{\boldsymbol{\rho}}\|\boldsymbol{S}^U-\boldsymbol{S}\cdot\boldsymbol{\rho}\|_F^2+\beta\cdot\|\boldsymbol{\rho}\|_F^2 \tag{5}$$

式中，β 为正则化常数。式(5)的闭式解为

$$\hat{\boldsymbol{\rho}}=(\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}+\beta\cdot\boldsymbol{I})\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}^U \tag{6}$$

式中，$\boldsymbol{I}$ 为判别矩阵。

利用式(7)求得的协同表示系数 $\hat{\boldsymbol{\rho}}$ 与 $\boldsymbol{S}$ 作矩阵运算即可以得到重建后的不可见类语义向量 $\breve{\boldsymbol{S}}^U$

$$\breve{\boldsymbol{S}}^U=\boldsymbol{S}\cdot\hat{\boldsymbol{\rho}} \tag{7}$$

2.2.2　基于 k 近邻算法的不可见类语义向量修正

为了缓和自编码跨域映射模型映射后不可见类语义空间与协同表示后不可见类语义空间的偏移，进一步利用无监督 k 近邻算法对不可见类语义向量进行修正。

虽然跨域映射算法通过在学习映射矩阵过程中加入自编码器约束，一定程度上提高了模型的泛化能力。然而，该方法并未充分利用已有的不可见类的语义信息。利用 k 近邻算法求 $\breve{\boldsymbol{S}}^U$ 在经过映射得到的语义向量 $\tilde{\boldsymbol{S}}^U$ 中的近邻向量并求其均值，得到更新后的不可见类语义特征 $\hat{\boldsymbol{S}}^U=[\hat{\boldsymbol{s}}_1^U,\hat{\boldsymbol{s}}_2^U,\cdots,\hat{\boldsymbol{s}}_q^U]\in R^{d^s\times p}$，如式(8)。以此对不可见类的语义特征进行进一步优化

$$\hat{\boldsymbol{s}}_j^U=\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{K}_j^i \tag{8}$$

式中，$\boldsymbol{K}_j^i(i=1\cdots m)$ 表示 $\breve{\boldsymbol{S}}^U$ 中第 j 类不可见类语义向量在 $\tilde{\boldsymbol{S}}^U$ 中寻找的 m 个近邻语义向量。

2.3　测试阶段

基于前面训练得到的特征提取器、跨域映射函数和修正后的不可见类语义向量集合，可以实现不可见类遥感影像场景的分类。具体的，给定一副测试遥感场景图像 x_i^U，遥感场景图像的视觉特征 $\boldsymbol{f}_i^U=\boldsymbol{Q}(x_i^U;\boldsymbol{T})$。进一步用矩阵 $\boldsymbol{W}$ 将其映射为语义向量 $\tilde{\boldsymbol{s}}_i^U=\boldsymbol{f}_i^U\boldsymbol{W}$，计算 $\tilde{\boldsymbol{s}}_i^U$ 与修正后不可见类语义向量集合 $\hat{\boldsymbol{S}}^U$ 之间的相似度量，从而得到其类别。余弦相似度量往往更注重向量之间在方向上的差异。通过语言模型提取出的语义向量正是通过向量之间的相似性来表示其代表的词或句子的相似度从而体现词句之间的联系。因此，余弦相似度适用于衡量语义向量之间的相似程度。基于以上考虑，本文采用余弦相似度作为相似度测度，其计算公式为

$$f(x_u^i)=\underset{j}{\operatorname{argmin}}\,d(\tilde{\boldsymbol{s}}_i^U,\hat{\boldsymbol{s}}_j^U) \tag{9}$$

式中，$f(x_u^i)$ 是场景图像 x_i^U 的预测标签；$d(\cdot)$ 是余弦距离方程。

3　试验及结果分析

3.1　数据集以试验设置

试验使用的数据集通过已有数据集整合而成，如图 2 所示。其结合了目前公开的 5 个遥感影像数据集 AID30[35]、NWPU-RESISC45[36]、PatternNet[37]、RSI-CB256[38]、UCM21[39]，对它们的类别进行筛选整理，最终形成具有 70 类遥感影像场景，每类包括 800 张影像，像素大小为 256×256 的数据集。

在本文试验中，采用 Resnet-50[40] 为深度网络的骨架，用可见类遥感影像数据集对网络模型微调，在试验中对 Resnet-50 网络的最后 1 层卷积层和全连接层采取微调(其余层次冻结)。利用深度网络计算得到每张遥感影像对应的 2048 维特征向量作为视觉特征。对于语义特征：①采用 Word2Vec[41] 模型将每个类别的类别名称映射为 300 维的词向量作为类别的语义特征；②对每个类别添加一段语义描述，利用 Bert 模型将语义描述语句映射为 1024 维向量作为类别的语义特征。本文的计算硬件资源为 Inter I7 3.2 GHz CPU、32 GB RAM 和 GTX2070 显卡。

后续试验中，分别选取 60 类可见类与 10 类不可见类；50 类可见类与 20 类不可见类；40 类可见类与 30 类不可见类这 3 种划分方式。为了客观评估方法的性能，采用总体分类准确率(OA)作为评价指标。

3.2　关键参数分析

首先讨论深度网络的微调次数对于方法的整体性能影响情况。以 Word2Vec 作为语义向量，随机选取 50 类作为可见类与 20 类作为不可见类，深度网络模型分别进行 1、3、5、7、9 个轮次的微调。基于这一试验设置，表 1 分别统计了不同迭代次数下深度网络微调的时间消耗及方法的整体精度水平。考虑到分类精度与计算时间的平衡，本文试验统一将深度网络的微调次数设置为 5 次。

表 1　不同迭代次数下方法准确率和时间消耗

Tab. 1　Overall accuracy and time consumption in different epochs

迭代次数	1	3	5	7	9
准确率(OA)/(%)	18.88	20.93	22.08	19.55	21.38
耗时/h	1.2	3.8	6.4	8.9	11.5

接下来，分析本文方法涉及的参数敏感性情况。本文提出的方法包含有 3 个参数：语义自编码器中的权重参数 α，协同表示中的参数 β 和计算最近邻向量的个数 m。后续参数分析试验中，

每次试验都对类别进行20次随机划分,然后统计方法在20次随机划分试验上分类精度的均值和标准差。

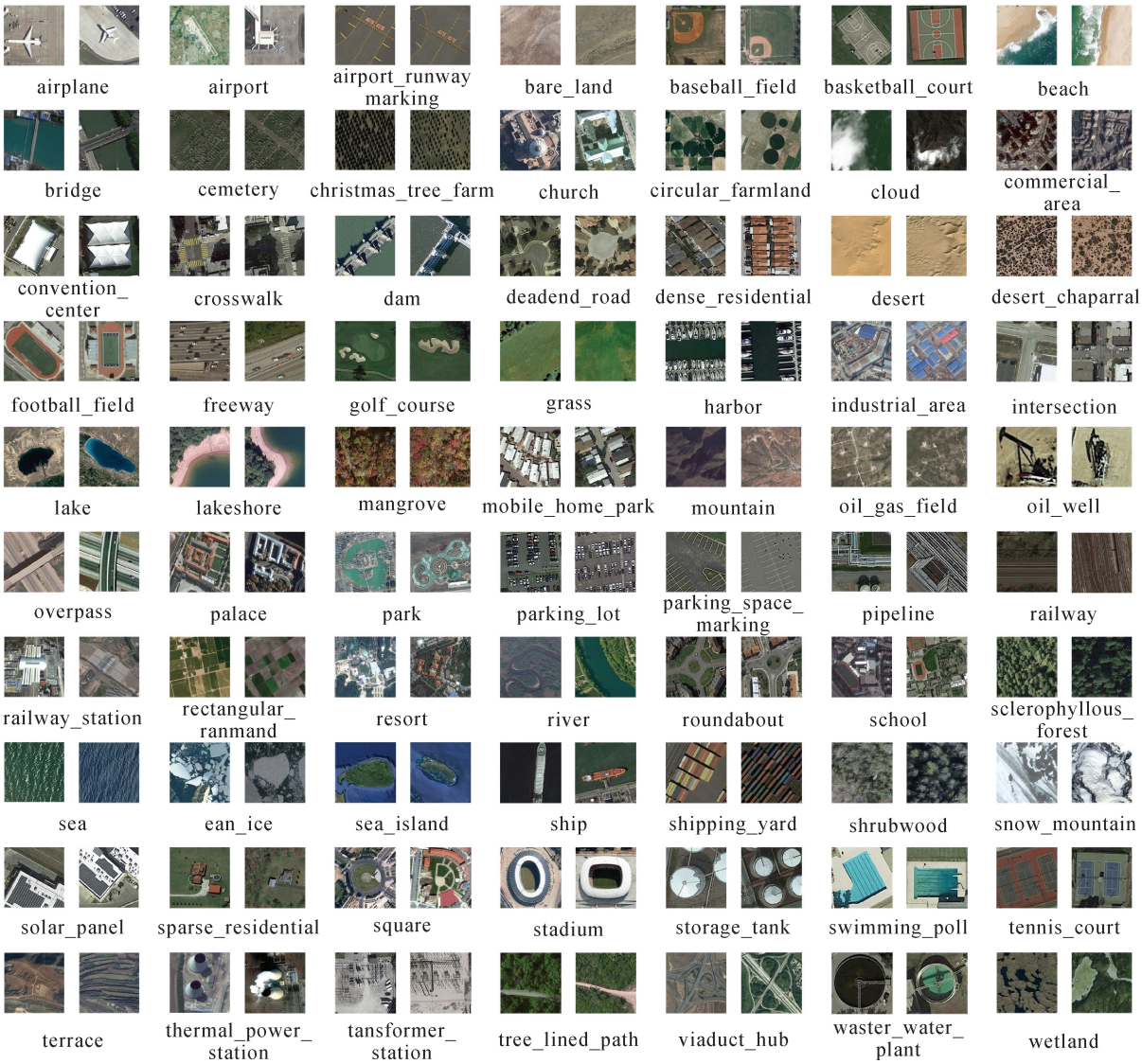


图2 遥感影像场景数据集

Fig.2 Remote sensing image scene dataset

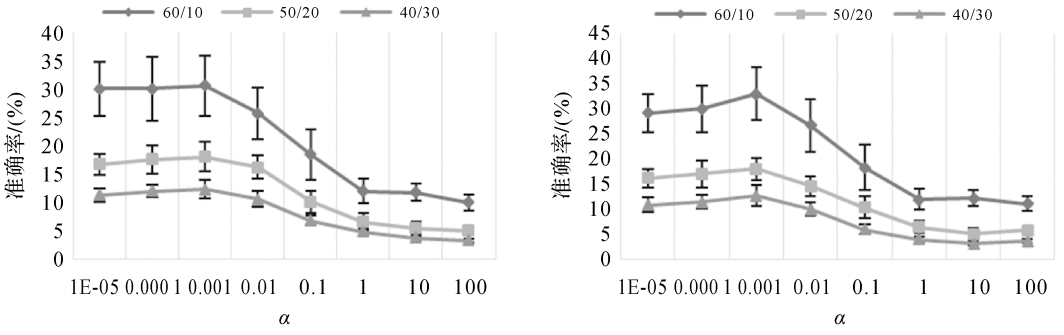
本文首先固定协同表示参数$\beta=1$,最近邻向量个数$m=200$,分析语义自编码器参数α的最佳取值。图3给出了使用两种不同语义向量情况下准确率随α变化的准确率曲线,分别取$\alpha=\{0.000\,01,0.0001,0.001,0.01,0.1,1,10,100\}$。由图4可以看出,由Word2Vec和Bert提取的语义向量均在α为0.001时,准确率最高。

接下来固定语义自编码器参数α为0.001,最近邻向量个数$m=200$,分析协同表示参数β的最佳取值。图4给出了使用两种不同语义向量情况下准确率随β变化的准确率曲线,分别取$\beta=\{0.000\,1,0.001,0.01,0.1,1,10,100,1000\}$。由图5可以看出,由Word2Vec和Bert提取的语义向量均在β为0.01时,准确率最高。

接下来固定语义自编码器参数α为1000,协同表示参数β为0.01,分析最近邻向量个数m的最佳取值。图5给出了使用两种不同语义向量情况下准确率随m变化的准确率曲线,分别取$m=\{100,200,300,400,500,600,700,800\}$。

由图5可以看出,针对不同的划分方式,参数m的最佳值也有所差别,在由Word2Vec提取语义向量的情况下,对于60/10的划分方式,$m=$

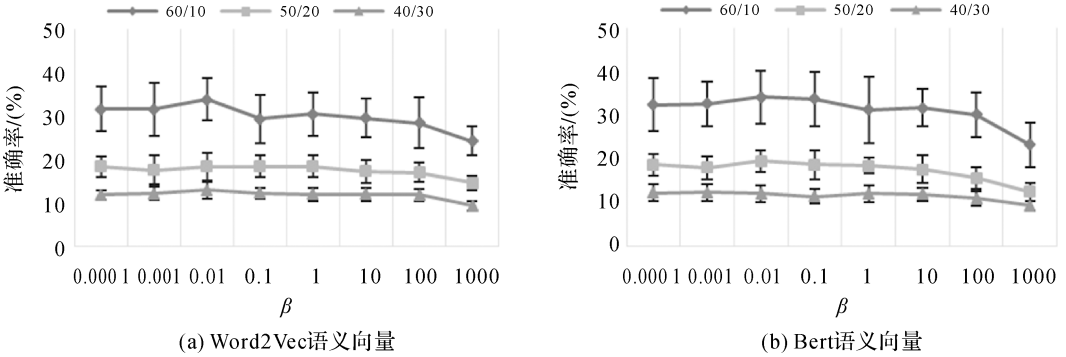
300 时准确率最高为 33.1%；对于 50/20 的划分方式，$m=500$ 时准确率最高为 19%；对于 40/30 的划分方式，$m=700$ 时准确率最高为 12.5%；在由 Bert 提取语义向量的情况下，对于 60/10 的划分方式，$m=200$ 时准确率最高为 35.8%；对于 50/20 的划分方式，$m=500$ 时准确率最高为 19.6%；对于 40/30 的划分方式，$m=500$ 时准确率最高为 12.7%。



(a) Word2Vec语义向量　　(b) Bert语义向量

图 3　α 参数分析

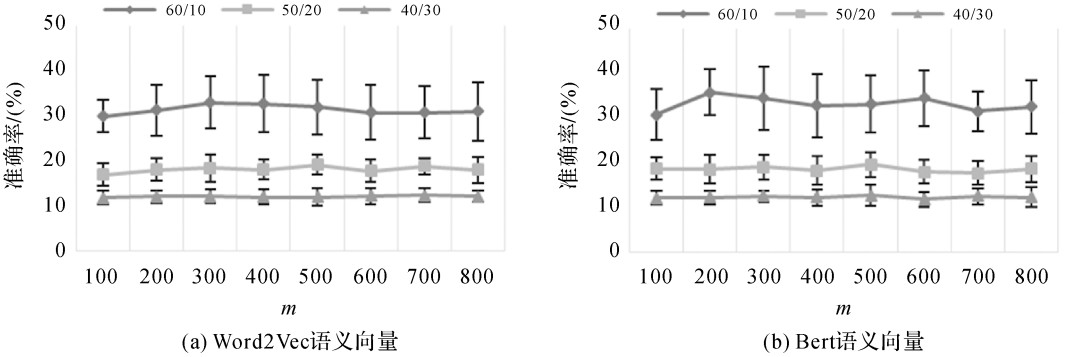
Fig.3　α parameter analysis



(a) Word2Vec语义向量　　(b) Bert语义向量

图 4　β 参数分析

Fig.4　β parameter analysis



(a) Word2Vec语义向量　　(b) Bert语义向量

图 5　m 参数分析

Fig.5　m parameter analysis

3.3　与已有方法的对比分析

为了验证本文方法的有效性，将本文方法与已有的方法进行了试验对比。为了提高方法的可比性，所有参与对比方法都对类别进行 20 次随机划分，然后统计方法在 20 次随机划分试验上分类精度的均值和标准差。对比方法包括学习映射矩

阵方法和利用生成对抗网络进行视觉样本生成方法。

对于学习映射矩阵方法,本文主要测试的方法有:DMaP[18]:对语义嵌入空间进行迭代优化,使得视觉特征和语义特征在模型学习的过程中尽量对齐,得到样本最优的语义表示。Ridge_regression[42]:岭回归是在平方误差的基础上增加正则项,用于控制与最小二乘估计相关的方差膨胀性和产生的不稳定性。SPLE[43]:引入了语义保持位置嵌入的思想,通过保留类内数据的位置来实现视觉特征与语义特征之间更好的匹配。

对于利用生成对抗网络进行视觉样本生成的零样本分类方法,本文主要测试的方法为:CIZSL[26]:在训练生成器 G 的过程中,引入了 hallucinated text,鼓励生成的视觉特征偏离可见类,从而使得生成的样本更具多样性。

表 2 不同划分方式下不同方法的准确率对比

Tab.2 comparison of accuracy of different methods under different classification methods (%)

语义向量	Word2Vec			Bert		
可见类/不可见类	40/30	50/20	60/10	40/30	50/20	60/10
SAE[17]	9.6±1.4	13.7±1.7	23.5±4.2	8.8±1.3	12.4±1.9	22.0±1.7
DMaP[18]	10.4±0.9	16.7±2.2	26.0±3.6	10.0±0.8	15.6±1.9	16.4±1.9
Ridge_regression[42]	7.3±1.2	10.9±2.2	19.2±3.4	8.8.±0.6	12.5±1.3	22.8±2.0
SPLE[43]	9.8±1.4	13.2±1.9	20.1±3.7	8.3±2.0	13.2±2.6	19.0±3.8
CIZSL[26]	6.0±1.2	10.6±3.7	20.6±0.4	6.2±2.1	10.3±1.9	20.4±4.1
本文方法	**12.5±1.5**	**19±1.8**	**33.1±5.8**	**12.7±2.3**	**19.6±2.7**	**35.8±5.1**

表 2 给出了不同方法在本文数据集上的结果对比。可以看出,在测试方法中,综合 Word2Vec 和 Bert 两种语义向量条件下,学习映射矩阵的 SAE 方法具有相对较好的结果,利用生成对抗网络的 CIZSL 方法结果较差,这可能是因为遥感场景包含多种对象,内容较为复杂,因此生成对抗网络还不能较好地生成高质量的样本,导致分类结果不理想。本文方法对不可见类的语义向量做了修正处理,使得经过映射过后的语义空间和协同表示后的语义空间更加一致,针对 3 种不同的划分方式的分类结果都明显优于其他方法。由结果可以看出,使用 Bert 提取语义向量的结果对比 Word2Vec 在 3 种不同划分方式上均更优,本文分析这是由于遥感影像场景复杂多样,对于不同的场景却可能包含几乎相同的地物目标,又或者是相同的场景中却包含不同的地物目标,因此 Bert 使用对场景的描述语句提取的语义向量相比 Word2Vec 单纯使用场景类别名称提取的语义向量包含了更多深层的语义信息,从而取得了更好的结果。

4 结 论

本文着眼于遥感场景零样本分类任务中的稳健跨域映射和语义基准修正,在有监督学习阶段,基于可见类的类别语义向量和遥感影像场景样本,联合场景类别分类和自编码跨域映射的多任务学习来实现深度特征提取器学习和遥感影像场景的视觉空间到类别语义空间的稳健映射。针对可见类语义空间与不可见类语义空间的偏移问题和自编码跨域映射模型映射后不可见类语义空间与协同表示后不可见类语义空间的偏移问题,本文基于全体类别的类别语义向量和不可见类遥感影像样本,分别通过无监督协同表示学习和无监督 k 近邻算法来渐进修正不可见类类别的语义向量,从而实现稳定的不可见类遥感影像场景识别任务。在整合的数据集上分别对 Word2Vec 和 Bert 两种模型提取的语义向量做了对比分析试验,验证了 Bert 模型提取的语义向量在零样本学习任务中的优越性。

参考文献:

[1] 李德仁, 张良培, 夏桂松. 遥感大数据自动分析与数据挖掘[J]. 测绘学报, 2014, 43(12): 1211-1216. DOI: 10.13485/j.cnki.11-2089.2014.0187.
LI Deren, ZHANG Liangpei, XIA Guisong. Automatic analysis and mining of remote sensing big data[J]. Acta Geodaetica et Cartographica Sinica, 2014, 43(12): 1211-1216. DOI: 10.13485/j.cnki.11-2089.2014.0187.

[2] 张鑫龙, 陈秀万, 李飞, 等. 高分辨率遥感影像的深度学习变化检测方法[J]. 测绘学报, 2017, 46(8): 999-1008. DOI: 10.11947/j.AGCS.2017.20170036.
ZHANG Xinlong, CHEN Xiuwan, LI Fei, et al. Change detection method for high resolution remote sensing images using deep learning[J]. Acta Geodaetica et Carto-

graphica Sinica, 2017, 46(8): 999-1008. DOI: 10.11947/j.AGCS.2017.20170036.

[3] LI Yansheng, TAO Chao, TAN Yihua, et al. Unsupervised multilayer feature learning for satellite image scene classification[J]. IEEE Geoscience and Remote Sensing Letters, 2016, 13(2): 157-161.

[4] 许夙晖，慕晓冬，赵鹏，等. 利用多尺度特征与深度网络对遥感影像进行场景分类[J]. 测绘学报，2016，45(7)：834-840. DOI：10.11947/j.AGCS.2016.20150623.
XU Suhui, MU Xiaodong, ZHAO Peng, et al. Scene classification of remote sensing image based on multi-scale feature and deep neural network[J]. Acta Geodaetica et Cartographica Sinica, 2016, 45(7): 834-840. DOI: 10.11947/j.AGCS.2016.20150623.

[5] 郑卓，方芳，刘袁缘，等. 高分辨率遥感影像场景的多尺度神经网络分类法[J]. 测绘学报，2018，47(5)：620-630. DOI：10.11947/j.AGCS.2018.20170191.
ZHENG Zhuo, FANG Fang, LIU Yuanyuan, et al. Joint multi-scale convolution neural network for scene classification of high resolution remote sensing imagery[J]. Acta Geodaetica et Cartographica Sinica, 2018, 47(5): 620-630. DOI: 10.11947/j.AGCS.2018.20170191.

[6] LI Yansheng, ZHANG Yongjun, HUANG Xin, et al. Large-scale remote sensing image retrieval by deep hashing neural networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(2): 950-965.

[7] LI Yansheng, ZHANG Yongjun, HUANG Xin, et al. Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(11): 6521-6536.

[8] LI Yansheng, ZHANG Yongjun, HUANG Xin, et al. Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 146: 182-196.

[9] DAI Yuchao, ZHANG Jing, HE Mingyi, et al. Salient object detection from multi-spectral remote sensing images with deep residual network[J]. Journal of Geodesy and Geoinformation Science, 2019, 2(2): 101-110.

[10] LI Yansheng, CHEN Wei, ZHANG Yongjun, et al. Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning[J]. Remote Sensing of Environment, 2020, 250: 112045.

[11] 何小飞，邹峥嵘，陶超，等. 联合显著性和多层卷积神经网络的高分影像场景分类[J]. 测绘学报，2016，45(9)：1073-1080. DOI：10.11947/j.AGCS.2016.20150612.
HE Xiaofei, ZOU Zhengrong, TAO Chao, et al. Combined saliency with multi-convolutional neural network for high resolution remote sensing scene classification[J]. Acta Geodaetica et Cartographica Sinica, 2016, 45(9): 1073-1080. DOI: 10.11947/j.AGCS.2016.20150612.

[12] ZHANG Fan, DU Bo, ZHANG Liangpei. Scene classification via a gradient boosting random convolutional network framework[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(3): 1793-1802.

[13] LI Yansheng, ZHANG Yongjun, ZHU Zhihui. Error-tolerant deep learning for remote sensing image scene classification[J]. IEEE Transactions on Cybernetics, 2020. DOI: 10.1109/TCYB.2020.2989241.

[14] LAROCHELLE H, ERHAN D, BENGIO Y. Zero-data learning of new tasks[C]// Proceedings of the 23rd AAAI Conference on Artificial Intelligence. Chicago, IL: AAAI, 2008: 3.

[15] PALATUCCI M, POMERLEAU D, HINTON G, et al. Zero-shot learning with semantic output codes[C]// Proceedings of the 22nd International Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada: NIPS, 2009: 1410-1418.

[16] BIEDERMAN I. Recognition-by-components: a theory of human image understanding[J]. Psychological Review, 1987, 94(2): 115-147.

[17] KODIROV E, XIANG Tao, GONG Shaogang. Semantic autoencoder for zero-shot learning[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI: IEEE, 2017: 4447-4456.

[18] LI Yanan, WANG Donghui, HU Huanhang, et al. Zero-shot recognition using dual visual-semantic mapping paths[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI: IEEE, 2017: 5207-5215.

[19] XIAN Yongqin, LAMPERT C H, SCHIELE B, et al. Zero-shot learning: a comprehensive evaluation of the good, the bad and the ugly[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2251-2265.

[20] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD birds-200-2011 dataset[R]. Pasadena: California Institute of Technology, 2011.

[21] LAMPERT C H, NICKISCH H, HARMELING S. Learning to detect unseen object classes by between-class attribute transfer[C]// Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE, 2009: 951-958.

[22] MIKOLOV T, SUTSKEVER I, CHEN Kai, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NE: NIPS, 2013: 3111-3119.

[23] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]// Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: EMNLP, 2014: 1532-1543.

[24] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Proceedings of the

27th International Conference on Neural Information Processing Systems. Montreal, Quebec, Canada: NIPS, 2014: 2672-2680.

[25] XIAN Yongqin, LORENZ T, SCHIELE B, et al. Feature generating networks for zero-shot learning [C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018: 5542-5551.

[26] ELHOSEINY M, ELFEKI M. Creativity inspired zero-shot learning [C]// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 5783-5792.

[27] SUMBUL G, CINBIS R G, AKSOY S. Fine-grained object recognition and zero-shot learning in remote sensing imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(2): 770-779.

[28] SONG Qian, XU Feng. Zero-shot learning of SAR target feature space with deep generative neural networks[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(12): 2245-2249.

[29] GUI Rong, XU Xin, WANG Lei, et al. A generalized zero-shot learning framework for PolSAR land cover classification[J]. Remote Sensing, 2018, 10(8): 1307.

[30] QUAN Jicheng, WU Chen, WANG Hongwei, et al. Structural alignment based zero-shot classification for remote sensing scenes[C]// Proceedings of 2018 IEEE International Conference on Electronics and Communication Engineering. Xi'an, China: IEEE, 2018: 17-21.

[31] LI Aoxue, LU Zhiwu, WANG Liwei, et al. Zero-shot scene classification for high spatial resolution remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 4157-4167.

[32] 吴晨, 王宏伟, 袁昱纬, 等. 基于图像特征融合的遥感场景零样本分类算法[J]. 光学学报, 2019, 39(6): 61-68.
WU Chen, WANG Hongwei, YUAN Yuwei, et al. Image feature fusion based remote sensing scene zero-shot classification algorithm[J]. Acta Optica Sinica, 2019, 39(6): 61-68.

[33] 吴晨, 袁昱纬, 王宏伟, 等. 基于词向量融合的遥感场景零样本分类算法[J]. 计算机科学, 2019, 46(12): 286-291.
WU Chen, YUAN Yuwei, WANG Hongwei, et al. Word vectors fusion based remote sensing scenes zero-shot classification algorithm[J]. Computer Science, 2019, 46(12): 286-291.

[34] BARTELS R H, STEWART G W. Solution of the matrix equation AX+ XB= C[F4][J]. Communications of the ACM, 1972, 15(9): 820-826.

[35] XIA Guisong S, HU Jingwen, HU Fan, et al. AID: a benchmark data set for performance evaluation of aerial scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 3965-3981.

[36] CHENG Gong, HAN Junwei, LU Xiaoqiang. Remote sensing image scene classification: benchmark and state of the art [J]. Proceedings of the IEEE, 2017, 105(10): 1865-1883.

[37] YANG Yi, 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. San Jose, CA: GIS, 2010: 270-279.

[38] ZHOU Weixun, NEWSAM S, LI Congmin, et al. Pattern net: a benchmark dataset for performance evaluation of remote sensing image retrieval[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 145: 197-209.

[39] LI Haifeng, DOU Xin, TAO Chao, et al. RSI-CB: a large-scale remote sensing image classification benchmark using crowdsourced data[J]. Sensors, 2020, 20(6): 1594.

[40] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV: IEEE, 2016: 770-778.

[41] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

[42] HOERL A E, KENNARD R W. Ridge regression: biased estimation for nonorthogonal problems[J]. Technometrics, 1970, 12(1): 55-67.

[43] TAO S Y, YEH Y R, WANG Y C F. Semantics-preserving locality embedding for zero-shot learning[C]// Proceedings of British Machine Vision Conference. London, UK: BMVC, 2017: 2017.

(责任编辑:张艳玲)

收稿日期: 2020-04-14
修回日期: 2020-11-02

第一作者简介: 李彦胜(1987—),男,博士,副教授,研究方向为遥感大数据处理与知识挖掘、人工智能与深度学习。
First author: LI Yansheng (1987—), male, PhD, associate professor, majors in remote sensing big data processing and knowledge discovery, and artificial intelligence and deep learning.
E-mail: yansheng.li@whu.edu.cn
通信作者: 张永军
Corresponding author: ZHANG Yongjun
E-mail: zhangyj@whu.edu.cn