引文格式: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.

高分辨率遥感影像场景的多尺度神经网络分类法

郑 卓[1,2],方 芳[1],刘袁缘[1],龚 希[1],郭明强[1],罗忠文[1]
1. 中国地质大学(武汉)信息工程学院,湖北 武汉 430074; 2. 武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉 430079

Joint Multi-scale Convolution Neural Network for Scene Classification of High Resolution Remote Sensing Imagery

ZHENG Zhuo[1,2],FANG Fang[1],LIU Yuanyuan[1],GONG Xi[1],GUO Mingqiang[1],LUO Zhongwen[1]
1. College of Information Engineering, China University of Geosciences, Wuhan 430074, China; 2. State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Abstract: High resolution remote sensing imagery scene classification is important for automatic complex scene recognition, which is the key technology for military and disaster relief, etc. In this paper, we propose a novel joint multi-scale convolution neural network (JMCNN) method using a limited amount of image data for high resolution remote sensing imagery scene classification. Different from traditional convolutional neural network, the proposed JMCNN is an end-to-end training model with joint enhanced high-level feature representation, which includes multi-channel feature extractor, joint multi-scale feature fusion and Softmax classifier. Multi-channel and scale convolutional extractors are used to extract scene middle features, firstly. Then, in order to achieve enhanced high-level feature representation in a limit dataset, joint multi-scale feature fusion is proposed to combine multi-channel and scale features using two feature fusions. Finally, enhanced high-level feature representation can be used for classification by Softmax. Experiments were conducted using two limit public UCM and SIRI datasets. Compared to state-of-the-art methods, the JMCNN achieved improved performance and great robustness with average accuracies of 89.3% and 88.3% on the two datasets.

Key words: high resolution remote sensing imagery; scene classification; joint multi-scale convolution neural network; enhanced high-level feature representation; limit datasets
Foundation support: The National Natural Science Foundation of China(Nos. 61602429;41701446),Chinese Geologic Survey Project (No. KZ17Z618)

摘 要:高分辨率遥感影像场景分类是实现复杂场景快速自动识别的基础,在军事、救灾等领域有十分重要的意义。为了在有限的遥感数据集上获得高识别精度,本文提出了一种基于联合多尺度卷积神经网络模型的高分辨率遥感影像场景分类方法。不同于传统的卷积神经网络模型,JMCNN 建立了一个具有 3 个不同尺度通道的端对端多尺度联合卷积网络模型,包括多通道特征提取器、多尺度特征联合和 Softmax 分类 3 个部分。首先,多通道特征提取器提取图像中、高层多尺度特征;然后,多尺度特征联合对多个通道的中、高层多尺度特征进行多次融合以增强特征表达;最后,Softmax 对高层特征进行分类。本文在 UC Merced 和 SIRI 遥感数据集进行测试,试验表明 JMCNN 模型在特征表达和计算速度方面均有显著提高,在小样本数据量下分别达到89.3%和 88.3%的识别精度。

关键词:高分辨率遥感影像;场景分类;联合多尺度卷积神经网络;高层特征增强表达;有限数据集

中图分类号：P237　　**文献标识码**：A　　**文章编号**：1001-1595(2018)05-0620-11
基金项目：国家自然科学基金(61602429；41701446)；中国地质调查项目(KZ17Z618)

随着 IKONOS、QuickBird 等高分遥感卫星的发射，高分辨率遥感影像比中、低分辨率的影像所包含的信息更加丰富。由于遥感影像场景中地物目标具有多样可变性、分布复杂性等特点，如何有效地对高分辨率遥感影像场景进行识别和语义提取成为了极具挑战的课题，已引起遥感学术界的广泛关注[1]。

为了解决遥感影像场景自动识别的问题，学者们先后提出了多种分类办法。文献[2]利用贝叶斯网络集成颜色特征、小波纹理特征和先验语义特征对室内、外场景影像进行分类。文献[3]利用金字塔表达方法提取底层特征，并利用 SVM (support vector machine) 和 KNN (k-nearest neighbor)完成场分类。文献[4]提出一种基于词袋的影像表达方法 SPMK (spatial pyramid matching kernel)，在 UC Merced(UCM)数据集上取得准确率为 74%的识别结果。文献[5]使用视觉词典，结合 BoVW(bag of visual words)，提出了一种空间共线核方法 SPCK++，相比 BoVW 和 SPMK 精度更高，取得 77.38%的准确率。文献[6]将概率主题模型 LDA (latent dirichlet allocation)用于场景分类，提出了 P-LDA 和 F-LDA，提高了 LDA 的分类精度。这些传统分类方法的关键在于分类器和人工特征提取。然而，在遥感场景影像中，复杂背景和尺度变化使得人工特征提取本身就是一个难点问题。

近年来，卷积神经网络 CNN(convolutional neural network)作为深度学习的一个模型，在大规模图像分类和识别中已经取得了巨大成功[7]。CNN 通过卷积层在大规模训练集中提取图像的中层特征，并通过反向传播算法[8]在全连接层中自动学习图像的高层特征表达，最后采用 Softmax 函数对目标分类。因此相比传统机器学习方法，CNN 具有权值共享，模型参数少，自动高层特征表达和易于训练的优点，已经开始应用于高分辨率遥感影像识别领域[9-11]。文献[9]利用显著性采样提取影像显著信息块，再利用卷积神经网络提取高层特征，最后使用 SVM 进行场景分类。文献[11]，利用 CaffeNet，在 UCM 遥感数据集上获得了 85.71%识别准确率。文献[11]等讨论了在数据增强的基础上，CNN 提取特征后直接分类的结果和在获得特征后做简单融合后的结果，识别准确率分别为 90.13%和 93.05%。

可见，遥感影像场景分类发展迅速，由人工提取图像底、中层特征，再到利用深度学习自动获取高层特征，已经取得了不错的分类结果。但是还存在一些难点和问题。一方面，人工提取特征只能解释一定信息量的数据，且受到环境、光照、遮挡等影响，对于信息量日益丰富的遥感影像数据的稳健性不高；另一方面，基于 CNN 的遥感影像场景分类研究中，良好的分类精度往往是依赖于大量的训练数据，而在小数据集上容易出现过拟合问题[12]。

为了解决 CNN 在有限数据集上的训练问题，增强高分遥感影像小数据集上的高层特征表达，本文提出基于 JMCNN 的高分遥感场景分类方法，如图 1 所示。每一个输入的遥感影像被提取 3 个尺度的随机子区域，并传入多通道卷积特征提取器，其提取得到的特征通过多个特征融合器进行融合，实现高层特征的联合增强表达，最后利用 Softmax 分类器对联合增强的特征进行分类。不同于现有 CNN 模型，本文提出的端对端的多尺度联合卷积神经网络模型，可以用更少的训练集实现高层特征的融合增强表达；其次 3 个尺度和通道的多输入模型，有效地解决了不同分辨率下的复杂图像分类，增强了模型的抗差性；第三，通过建立多个特征融合器对多通道多尺度特征融合，可实现高层特征的联合增强表达，提高网络效率。

1　JMCNN 网络结构

不同于以往分别训练多个 CNN[13]级联以增强特征表达的方式，JMCNN 建立了一个 3 个尺度、3 个通道的端对端训练模型，包括多通道特征提取器、多尺度特征联合、联合损失函数 3 个部分。JMCNN 的网络结构如图 2 所示。首先，将图像大小为 $\boldsymbol{N}\times\boldsymbol{N}$ 的遥感影像进行 3 个尺度随机子区域提取，获得图像子区域大小分别为 $\lfloor N/2\times N/2\rfloor$、$\lfloor N/4\times N/4\rfloor$ 和 $\lfloor N/8\times N/8\rfloor$，作为多通道卷积特征提取器输入；然后，通过建立多个特征融合器对多通道不同尺度特征进行融合，实现高层特征的联合增强表达，提高了网络的效率；最

后，Softmax 函数用来对场景的联合增强特征进行分类。它利用端对端的方式训练模型，对参数进行全局优化而不是对每一个 CNN 进行单独优化，且采用多通道子卷积网络提取不同尺度卷积特征，以及联合网络融合多尺度高层特征。最终 JMCNN 通过对多个通道的不同尺度的高层特征的联合增强表达，实现了在小样本训练集上的高精度分类。

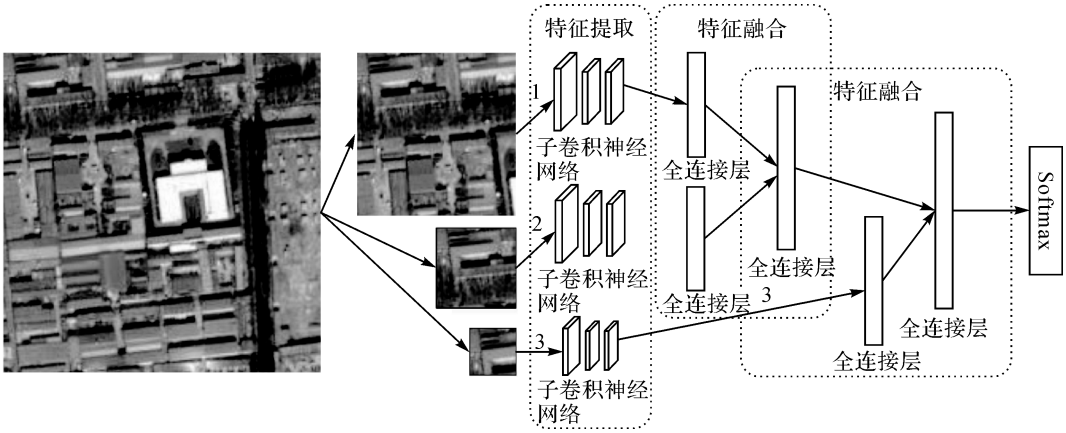

图 1 基于 JMCNN 的高分遥感影像场景分类流程

Fig.1 JMCNN framework for high resolution remote sensing image scene classification

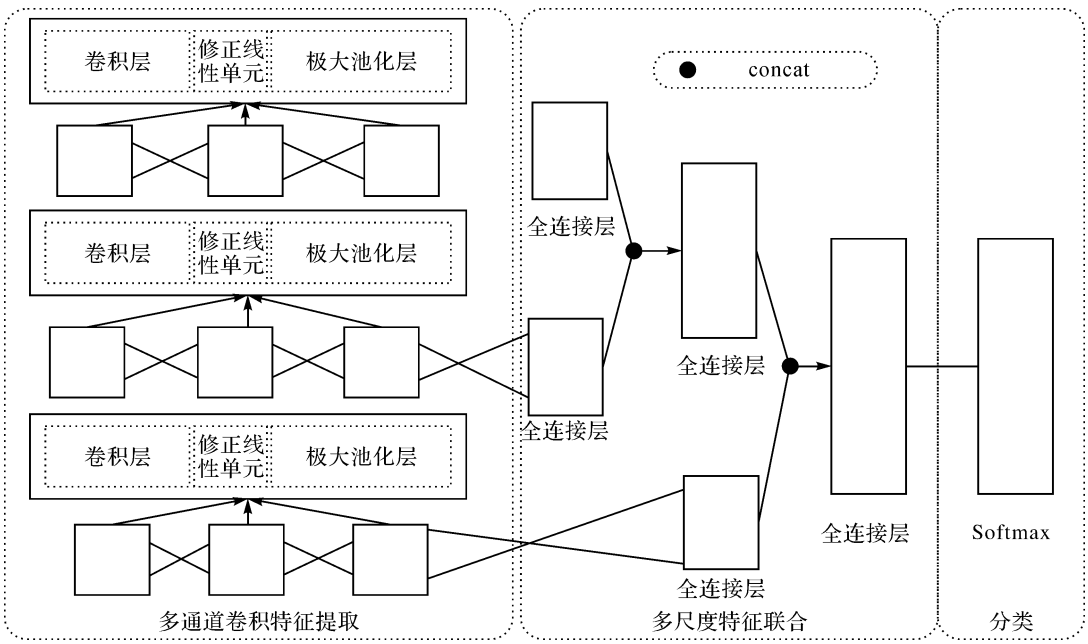

图 2 JMCNN 网络结构

Fig.2 The network architecture of JMCNN

1.1 多通道卷积特征提取

JMCNN 的多通道特征提取器是由 3 个单通道子卷积网络构成。每个单通道子卷积网络包括 3 个中间层，每个中间层分别由卷积层、ReLu[14] 激活函数和极大池化层构成，如图 3 所示。

首先，对输入大小为 $\boldsymbol{N}\times\boldsymbol{N}$ 的遥感影像，随机提取 3 个不同尺度（大小分别为 $\lfloor N/2\times N/2\rfloor$、$\lfloor N/4\times N/4\rfloor$、$\lfloor N/8\times N/8\rfloor$）不同位置的子影像，再利用 3 个中间层提取卷积特征矩阵。其中，ReLu 具有稀疏激活性，使得后面获得的高层特征矩阵稀疏度较大，利于后述的多尺度特征融合。

单通道子卷积网络的特征提取过程如下：

设输入影像为 $X\in R^{h\times w\times c}$，由宽卷积计算公式

$$Y^i=F\otimes X^i+b \tag{1}$$

其中，$\boldsymbol{h}$、$\boldsymbol{w}$、$\boldsymbol{c}$ 分别为影像的高、宽、颜色通道总数，$\boldsymbol{F}$ 为 5×5 的卷积核，$\boldsymbol{i}$ 为颜色通道号，$\boldsymbol{b}$ 为偏置项，⊗代表宽卷积运算。由于是宽卷积运算，输出的特征映射 $Y \in R^{h \times w \times c}$ 与 $\boldsymbol{X}$ 维度相同。然后，通过 ReLu 函数激活后和极大池化层计算特征映射 $M_k \in R^{h \times w \times c}$，其输出特征维度与 $\boldsymbol{Y}$ 相同，即为所提取的单通道卷积特征矩阵 $\boldsymbol{M}_k$，$\boldsymbol{k}$ 表示不同的特征通道。

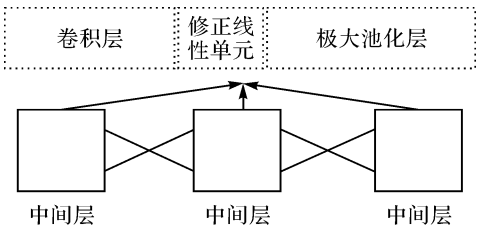

图 3　单个子卷积通道特征提取器

Fig.3　The single sub-convolutional feature extractor

在 JMCNN 中，3 个不同尺度的子影像分别通过 3 个单通道卷积子网络，则最终获得 3 个不同尺度的卷积特征矩阵 $\boldsymbol{M}\{\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3\}$。

1.2　多尺度特征联合

为了增强特征的表达能力，多尺度特征联合将对多通道卷积特征进行多尺度融合增强，获得高层特征增强表达。与 Inception 模块[15]相比，多尺度特征联合增强包括了两个级联的特征融合过程，从而减少了全连接层总连接数，提升了模型效率。第一个是将多通道特征提取器中输入图像大小为 $\lfloor N/2 \times N/2 \rfloor$ 和 $\lfloor N/4 \times N/4 \rfloor$ 的两个特征矩阵 $\boldsymbol{F}_{t_1}$、$\boldsymbol{F}_{t_2}$，利用特征融合器 f 进行联合，得到一个新的融合特征 TEM；第二个特征融合过程是将 TEM 与多通道特征提取器中输入影像大小为 $\lfloor N/8 \times N/8 \rfloor$ 的特征矩阵 $\boldsymbol{F}_{t_3}$ 再次使用 $\boldsymbol{f}$ 融合，最终获得高层增强联合特征表达 $\boldsymbol{FIN}$。多尺度特征联合过程的算法描述如下所示。

算法：多尺度特征联合。

输入：多通道特征矩阵 $\boldsymbol{F}_{t_1}$、$\boldsymbol{F}_{t_2}$、$\boldsymbol{F}_{t_3}$，特征融合器 $\boldsymbol{f}$。

输出：高层增强联合特征表达 FIN。

(1) 融合 $\boldsymbol{F}_{t_1}$ 和 $\boldsymbol{F}_{t_2}$，得到 $TMP = f(\boldsymbol{F}_{t_1}, \boldsymbol{F}_{t_2})$。

(2) 融合 TEM 和 $\boldsymbol{F}_{t_3}$，得到 $FIN = f(TEM, \boldsymbol{F}_{t_2})$。

(3) Return FIN。

图 4 为单个特征融合器 f 的结构图。特征融合器 $\boldsymbol{f}$ 的算法过程。假设任一个融合器输入的两个特征矩阵为 $\boldsymbol{M}_1$、$\boldsymbol{M}_2 \in R^{h \times w \times c}$，首先将 $\boldsymbol{M}_k (k=1,2)$ 以行、列、颜色通道的顺序展平为特征向量 $\boldsymbol{K}_i \in \boldsymbol{R}^{1 \times (h * w * c)}$，其中 * 代表数值乘法，× 代表笛卡儿积。然后将特征向量分别进入全连接层计算并使用 ReLu[13] 激活

$$\boldsymbol{V}_i = \boldsymbol{K}_i \boldsymbol{W} + b \tag{2}$$

$$\boldsymbol{T}_i = \mathrm{Re}Lu(\boldsymbol{V}_i) \tag{3}$$

其中 $\boldsymbol{T}_i$ 为 1024 维的特征向量；$\boldsymbol{W} \in R^{(h * w * c) \times 1024}$；$\boldsymbol{b}$ 为偏置项。$\boldsymbol{T}_1$ 和 $\boldsymbol{T}_2$ 通过“concat”变换成一个新的特征向量 $\boldsymbol{V}_3$，再将此向量通过一个全连接层计算得出最终的高层增强特征表达。其中，“concat”定义为两个特征向量的线性拼接，得到特征向量 $\boldsymbol{V}_3$ 空间维度为 $\boldsymbol{V}_3 \in R^{1 \times 2048}$。最后，式(3)再对 $\boldsymbol{V}_3$ 进行激活，得到融合的高层特征向量 $\boldsymbol{P} \in \boldsymbol{R}^{1 \times 512}$。

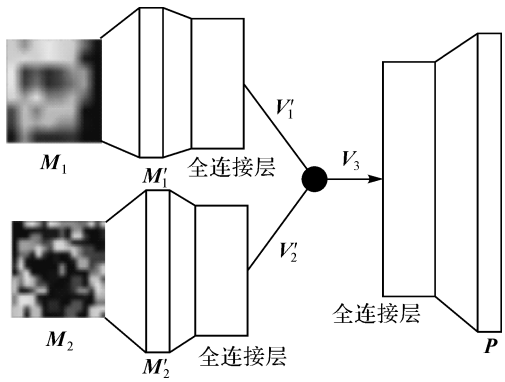

图 4　特征融合器

Fig.4　Multi-scale feature fusion

特征融合器 f 的算法过程如下：

算法：特征融合算法。

输入：特征矩阵 $\boldsymbol{M}_1$、$\boldsymbol{M}_2$，权重矩阵 $\boldsymbol{W}_1$、$\boldsymbol{W}_2$、$\boldsymbol{W}_3$，偏置向量 $\boldsymbol{b}_1$、$\boldsymbol{b}_2$、$\boldsymbol{b}_3$。

输出：融合特征向量 $\boldsymbol{P}$。

(1) 展平两个特征矩阵 $\boldsymbol{M}_1$、$\boldsymbol{M}_2$，得到 $\boldsymbol{K}_i = \mathrm{reshape}(\boldsymbol{M}_k), (k=1,2)$。

(2) 全连接层得到 $\boldsymbol{V}_i = \boldsymbol{K}_i \boldsymbol{W}_i + \boldsymbol{b}_i$。

(3) 激活得到 $\boldsymbol{T}_i = \mathrm{ReLu}(\boldsymbol{V}_i)$。

(4) 融合特征向量 $\boldsymbol{T}_1$ 和 $\boldsymbol{T}_2$，得到 $\boldsymbol{V}_3 = \mathrm{concat}(\boldsymbol{T}_1, \boldsymbol{T}_2)$。

(5) 计算全连接层计并 ReLu 激活，得到最终的特征向量 $\boldsymbol{P} = \mathrm{ReLu}(\boldsymbol{V}_3 \boldsymbol{W}_3 + \boldsymbol{b}_3)$。

(6) Return $\boldsymbol{P}$。

此外，为了防止过拟合问题，JMCNN 采用一种概率线性融合方式对两个特征进行融合。即在训练过程中每个全连接层后连接了一个 dropout[16] 层，即

每次随机保留一部分神经元参与训练。

1.3 Softmax 分类器及损失函数

JMCNN 模型采用 Softmax 分类器，因此本小节主要阐述模型的损失函数。JMCNN 的损失函数为交叉熵损失与正则化项之和，即在经验风险上加上表示模型复杂度的结构风险。

设 Softmax 函数输出的向量为 $\boldsymbol{Y}\in\boldsymbol{R}^{1\times n}$，$\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)$，式中 n 为样本类别数；y_i 表示向量中第 i 个元素的实数值。

损失函数可表示为

$$\left.\begin{aligned} L &= -\sum_i y'_i \log(y_i) + \frac{1}{2}\lambda \|W_i\|^2 \\ \lambda &= \prod_i \text{weight\_decay}_i \end{aligned}\right\} \quad (4)$$

其中，式(4)中前一项是交叉熵损失函数，后一项是权值的 L2 正则项；λ 为正则项系数，其由各权值的衰减系数乘积决定。式(4)引入了正则项的损失函数，其作为损失函数的一个惩罚项，平衡经验风险与模型复杂度，能有效防止过拟合现象。

2 基于 JMCNN 的高分遥感影像场景分类

本节主要描述了基于 JMCNN 的高分遥感影像场景分类过程。

2.1 遥感影像数据预处理

为了能更好地表达遥感影像中的场景信息，在开始训练 JMCNN 之前，需要对高分遥感影像进行数据预处理，以增加样本的多样性。首先，对于一张大小为 $N\times N$ 的影像，随机提取大小为 $\lfloor 0.875N\times 0.875N\rfloor$ 的图像区域。然后，通过图像归一化算法，调整影像的对比度和亮度，减少光照对场景分类的噪声影响。最后，在归一化的图像区域中，随机提取 $\lfloor N/2\times N/2\rfloor$、$\lfloor N/4\times N/4\rfloor$ 和 $\lfloor N/8\times N/8\rfloor$ 3 个不同尺度不同位置的子区域块，作为 JMCNN 的多尺度输入。

2.2 遥感影像场景分类

2.2.1 多通道卷积特征提取

JMCNN 由 3 个通道的卷积特征提取器组成，每个特征提取器由 3 个卷积层、ReLu 激活函数和池化层构成，每个卷积层的卷积图个数为 64。卷积层的卷积核的大小均为 5×5，步长为 1，权重衰弱系数为 0，即卷积层的权值的 L2 范数不加入正则项。池化层的卷积核的大小均为 3×3，步长为 2。其中卷积层、池化层中的卷积运算均采用宽卷积运算。

2.2.2 多尺度特征联合

多尺度特征联合过程由两个特征融合器构成。特征融合器中的全连接层的权值衰减系数均设置为 0.004，即全连接层的权值的 L2 范数均加入正则项。图 5 描述了多尺度特征联合中的特征向量融合过程。

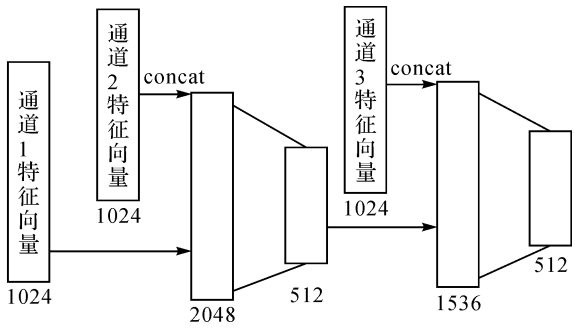

图 5 多尺度特征向量融合过程

Fig.5 Joint multi-scale feature vectors

如图 5 所示，第一个特征融合器的输入参数是两个不同尺度影像中根据特征矩阵的稀疏程度提取的特征向量，维度为 1024。“concat”后维度为 2048，再通过全连接层线性融合后得到特征向量，其输出维度为 512。

第二个特征融合器的输入参数分别为卷积特征通道所提取的特征向量(维度为 1024)和由第一个特征融合器融合的 512 维特征向量，通过“concat”，输出特征向量维度为 1536。接着进入全连接层线性融合后得到的特征向量的输出维度为 512。

每个特征融合器后面加入一个 dropout 层，从而在训练过程中可降低全连接层的复杂度，防止融合得到的特征产生过拟合现象。dropout 层会使得全连接层中的每个神经元以一定的概率“失活”，使得模型复杂度降低，计算量减少，模型收敛更快和泛化增强。参考 GoogleNet[15]，将第一个特征融合器的保留概率设置为 0.6，第二个则设置为 0.7。

2.2.3 基于 Softmax 的高层联合特征分类

Softmax 分类器用于对图 7 联合提取的 512 维的高层增强特征向量进行分类，获得最终的影像场景类别。

假设，输出一个维数与场景类别数 n 相同的一个向量 $\boldsymbol{Y}=\{\boldsymbol{Y}_i\}$，其中 $\boldsymbol{Y}_i(i=1,2,\cdots,n)$ 为该场景影像属于类别 i 的概率。Softmax 采用 $\boldsymbol{Y}_i$ 最

大概率判别该场别影像的类别 i，如图 6 所示。

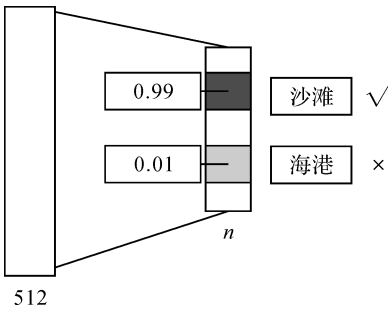

图 6　Softmax 分类

Fig.6　Classification using Softmax

3　基于 JMCNN 的场景分类试验和分析

为了有效地评估 JMCNN 模型在高分遥感影像场景分类，JMCNN 在 UCM 和 SIRI[17-19] 两个高分遥感影像数据集上分别进行了试验和分析，并与最新方法进行对比。试验均采用 5-折交叉验证，试验结果表明 JMCNN 可以在小数据集上实现较好的分类结果。

3.1　试验设置

试验环境：试验均在载有两块 NVIDA GeForce GTX1080 的显卡、Inter® core™ i7-6700K CPU@4.00 GHz、RAM：32.0 GB 的工作站上进行。本文的 JMCNN 与所使用的 CNN 模型均利用试验框架为 TensorFlow[20] 实现。

数据集：试验所采用的数据集为 UCM 和 SIRI 高分遥感数据集。UCM 数据集所包含的影像尺寸为 256×256，颜色通道为 RGB，空间分辨率为 0.3 m。该数据集影像总计为 2100 张，含 21 个场景类别，每类 100 张。类别包括(a)农田(b)机场(c)棒球场(d)沙滩(e)建筑(f)丛林(g)密集住宅区(h)森林(i)公路(j)高尔夫球场(k)海港(l)十字路口(m)中等住宅区(n)房车公园(o)天桥(p)停车场(q)河流(r)飞机跑道(s)稀疏住宅区(t)储油罐(u)网球场，如图 7 所示。

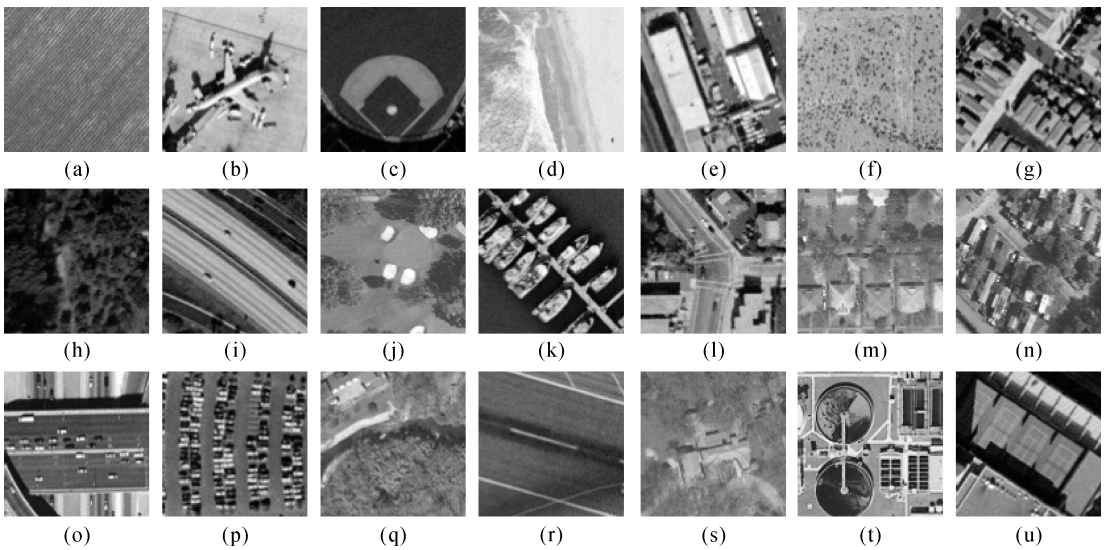

图 7　UCM 数据集图例

Fig.7　The samples of UCM dataset

SIRI 数据集为 Google Earth 上的影像数据，主要覆盖我国的城市及周边区域，由文献[20]的作者整理而成。影像尺寸为 200×200，颜色通道为 RGB，分辨率为 2 m。数据集总计 2400 张影像，有 12 类，每类 200 张。类别分别为(a)农田(b)商业区(c)港口(d)裸地(e)工业区(f)草地(g)交叉路口(h)公园(i)池塘(j)居民区(k)河流(l)水面，如图 8 所示。

试验验证方案：在试验中，均采用 5 折交叉验证方案，将数据集随机划分为 5 等份，每次利用其中 4 份作为样本集，余下 1 份即为测试集，轮流 5 次，取分类精度的平均值。

3.2　UCM 数据集场景识别试验

表 1 描述了使用不同网络结构和特征的场景分类时间和准确率的比较。表 1 第一行中“网络”、“Size”、“F”、“Acc”、“Kappa”，分别表示“网络结构”、“增强后数据集大小”、“单次前向计算耗时”、“识别准确率”、“Kappa 系数”。

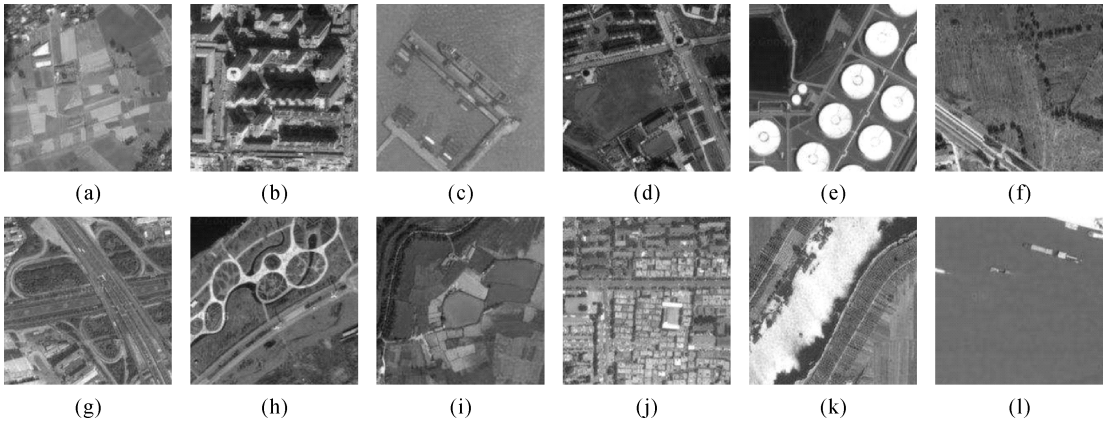

图 8 SIRI 数据集图例

Fig.8 The samples of SIRI dataset

试验中，CNN(6conv+2fc)代表卷积神经网络结构为 6 个卷积层(卷积核 5×5，步长为 1，卷积核数量分别为 60、50、64、128、256、512)且每个卷积层后接一个池化层(卷积核 3×3，步长为 2)和 2 个全连接层(输出维度分别为 1024、2048)和 Softmax 分类器。每个卷积层后均接极大池化层和 ReLu 激活函数，第一个全连接层的激活函数为 ReLu。CNN(5conv+2fc)的网络结构为 5 个卷积层(卷积核参数设置不变，卷积核数量分别为 60、50、64、256、512)。

表 1 结果显示，数据集大小同为 2100 张影像时，JMCNN 比 CNN(6conv+2fc)网络精度高出 25.03%，该结果说明在小数据集上利用融合后的多尺度特征的分类精度远远高于单一尺度的特征。同时，JMCNN 所用的卷积核数量远小于 CNN(6conv+2fc)和 CNN(5conv+2fc)，时间效率提高了 30%，一次前向计算时间减少到 145 ms。

表 1 使用不同网络和特征的时间和精度比较

Tab.1 Comparison of time and accuracy using different networks and features

网络	size	F/(ms)	Acc/(%)	Kappa
CNN(6conv+2fc)	2100	216	64.27	0.612 3
CNN(6conv+2fc)	2100×36	218	85.76	0.851 2
CNN(5conv+2fc)	2100×36	211	83.22	0.802 2
CNN(6conv+2fc)	2100×240	223	90.33	0.899 3
JMCNN	2100	153	89.30	0.874 2
JMCNN	2100×36	158	93.00	0.926 2
JMCNN	2100×240	145	98.30	0.967 7

同时，表 1 还对比了数据适当增广 36 倍和 240 倍后的分类效率和准确率。试验中所用的数据增广方法为在原影像上随机提取出 9 张影像，再令这 9 张影像顺时针旋转 0°、90°、180°、270°，从而获得了 36 倍增广数据集。240 倍增广数据是通过先保留图像的 60%、62%、64%、66%、68%、70%得到 6 个子影像，再在这 6 个子影像上随机提取出 10 张影像，然后按上述方法旋转 4 个角度，从而获得 6×10×4=240 倍的增广数据集。从试验结果可见，数据集大小为 2100×36 时，JMCNN 相比 CNN(5conv+2fc)要高出 9.78%，比 CNN(6conv+2fc)网络要高出 3.54%。而在相同网络结构之间，通过增强训练数据，JMCNN 精度最大提升了 9%，CNN 分类精度最大提升了 25.06%。数据表明，相对于传统的 CNN 网络，JMCNN 对大数据训练的依赖性更小，在小样本训练的情况下可以获得较强的高层特征。此外，Kappa 系数的结果表明，JMCNN 具有更好的分类一致性，其泛化能力更强。

图 9 为 JMCNN 在 UCM 训练数据无增广时的分类精度混淆矩阵。可见，JMCNN 对大部分场景的分类准确率高于 90%，对于极个别(13)中等住宅区(20)储油罐分类准确率低于 70%，相比于传统 CNN(6conv+2fc)的分类结果，JMCNN 在(b)机场(c)棒球场(p)停车场(q)河流等场景类别的识别准确率提升显著，最高提升了28.72%，总体提升了 25.03%，可见它对于尺度变化较大的场景类别识别更加准确。

为了进一步说明 JMCNN 在不同数据维度下的特征表达能力，图 10 描述了 JMCNN 和 CNN 在不同维度的训练样本数量下的分类准确率对比结果。图 10 表明，随着数据量增加，两种模型分类准确率均有提升，精度的变化率随数据量的增大而减小，并逐渐收敛。CNN 模型随着数据量的增加，准确率显著提升，表明其特征质量与训练样

本数据量相关程度较大，模型在数据量较小时特征表达不充分。JMCNN 的识别准确率随数据量的增加变化较为平缓，通过多通道多尺度高层特征的联合增强，能在小样本数据集上训练充分，获得较高的准确率。

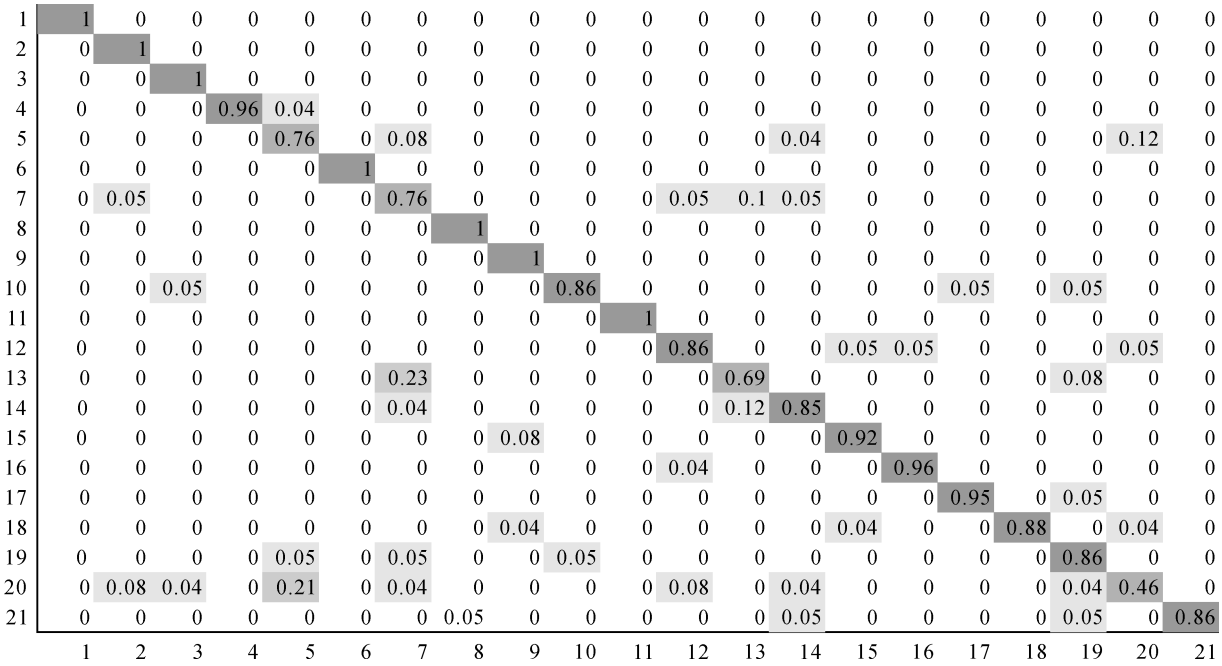

图 9　JMCNN 在 UCM 数据集上的分类混淆矩阵

Fig.9　Confusion matrix of JMCNN on the UCM dataset

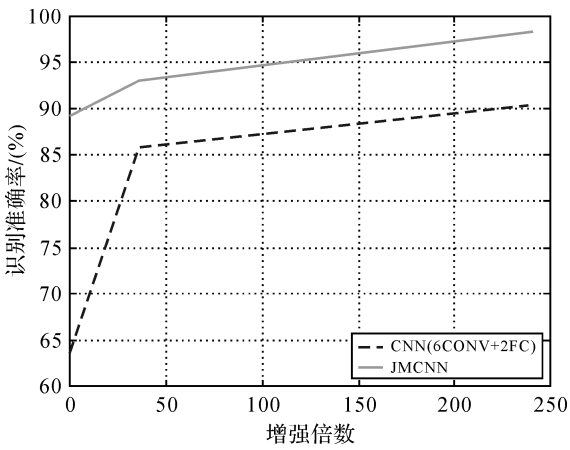

图 10　JMCNN 和 CNN 在不同数据量的识别准确率

Fig. 10　The accuracies comparison on different numbers of training images using JMCNN and CNN

表 2 显示了 JMCNN 与其他方法的对比结果。JMCNN 和 CNN 的样本大小均为 2100 * 0.8 张，输入数据为图像本身，均为高效的端对端网络模型。JMCNN 的识别率高于 CNN25.03%。SVM+LDA[21] 和 SAE[22] 均将数据增广了20 倍，其识别率结果比 JMCNN 分别低了 9.03% 和 6.58%。MeanStd-SIFT+LDA-H[17] 通过多种人工设计特征提取融合和聚类的方法，识别率提高到 84.98%，仍低于 JMCNN 约 5%。PSR[20] 结合 BOW 特征和金字塔空间关系模型获得了第二高的识别率 89.0%，然而其模型训练复杂度高，模型难以泛化使用。RF(Random forest，RF)[23] 采用随机森林对 SIFT 特征进行分类，在相同训练集下的识别率为 69.5%。对比结果表明，JMCNN 通过端对端训练模式，在不需要任何人工设计特征表达以及数据增广的情况下，识别率均高于其他方法。

表 2　JMCNN 与其他方法的识别率对比

Tab.2　Comparison of accuracy using different model

模型	准确率/(%)	Kappa
JMCNN	89.30	0.874 2
SVM+LDA[21]	80.33	-
SAE[22]	82.72	-
SPMK[4]	74.00	-
MeanStd-SIFT+LDA-H[16]	84.98	-
PSR[23]	89.00	-
RF+SIFT[24]	69.50	-

3.3　2SIRI 数据集场景分类试验

为了更好地验证本文提出模型的抗差性，

JMCNN 在 SIRI 数据集(总计 2400 张,200 * 200 的影像数据)上进行了试验分析。SIRI 数据集共有 12 类,每类 200 张。

表 3 描述了本文提出的 JMCNN 与 CNN(6conv + 2fc)、6conv + 2fc + SVM、SVM-LDA[21]、SPMK[4]、MeanStd-SIFI+LDA-H[16]方法的对比结果。JMCNN 在无数据增广的 SIRI 数据集上获得了 88.3%的分类精度,均高于 CNN 和传统机器学习方法。CNN(6conv+2fc)和 6conv+2fc+SVM 均采用 6 个卷积层、2 个全连接层提取高层特征,然后分别用 Softmax 与 SVM[25]分类器进行分类,其结果均低于 JMCNN 的识别率约 20%。同时,相比于 LDA-M[21]、SPM-SIFT[4]和 MeanStd-SIFI+LDA-H[15]方法的复杂特征设计和提取,JMCNN 模型不需要任何人工特征设计,采用端对端训练来统一优化参数,训练难度大大降低,特征表达能力更强,且分类准确率更高。Kappa 系数表示,JMCNN 与 CNN 及上述传统机器学习方法相比,具有更好的分类一致性。

图 11 为 JMCNN 在 SIRI 数据集上分类的混淆矩阵。结果显示,JMCNN 对(a)农田(b)商业区(e)工业区(j)居民区(k)水面等场景类别的识别准确率高于 95%,对于极个别(f)草地的识别率低于 70%,其余大部分在 85%左右。可见,该模型对于特征复杂的细粒度区域分类结果较好,而对于背景特征单一的区域分类结果需要进一步提升。

表 3 不同方法的对比

Tab.3 Comparison of different methods

模型	准确率/(%)	Kappa
JMCNN	88.30	0.859 5
SVM-LDA[23]	60.32	-
SPMK[4]	77.69	-
MeanStd-SIFI+LDA-H[17]	86.29	-
RF+SIFT	49.90	-
CNN(6conv+2fc)	68.81	0.652 1
6conv+2fc+SVM	67.29	0.637 7

3.4 3USGS 大幅影像场景标注试验

试验所用的大幅影像为 USGS 数据库中美国俄亥俄州蒙哥马利地区的影像,尺寸为 10 000×9000,空间分辨率为 0.6 m,如图 12(a)。在场景标注试验中,样本采样自上述大幅影像,每类样本包含 50 张图像大小为 150×150 的子影像,人工标注为 4 类,分别为住宅(图 13(a))、耕地(图 13(b))、森林(图 13(a))、停车场(图 13(d))。为了评估模型精度,样本以 80%、20%的比例分别划分为训练集和测试集。

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.978	0	0	0.022	0	0	0	0	0	0	0	0
2	0	0.978	0	0	0	0	0	0	0	0	0.022	0
3	0	0	0.789	0	0	0	0	0	0.079	0	0.026	0.105
4	0.025	0	0	0.85	0	0	0.075	0	0	0.025	0.025	0
5	0	0	0.021	0	0.957	0	0	0	0	0.021	0	0
6	0.098	0	0	0.073	0	0.634	0.024	0.146	0.024	0	0	0
7	0	0	0	0.02	0.082	0.02	0.878	0	0	0	0	0
8	0	0.02	0	0.041	0	0.082		0.857	0	0	0	0
9	0	0	0.028	0.056	0	0	0	0	0.861	0	0.056	0
10	0	0.019	0	0	0	0	0	0.019	0	0.962	0	0
11	0	0	0.091	0.023	0.023	0	0.023	0	0.091	0	0.75	0
12	0.04	0	0	0	0	0	0	0	0	0	0	0.96

图 11 JMCNN 在 SIRI 上的分类混淆矩阵

Fig.11 The confusion matrix of JMCNN on SIRI dataset

在该试验中,使用在 UCM 数据集上预训练的 JMCNN 模型并将其在该场景影像训练样本上微调。利用微调后的模型对整幅影像进行预测,如图 12(b),图 12(d)为某个预测类别为 forest 的区域。

通过观察局部细节(图 12(c)),JMCNN 在空间分布感知上具有一定优势,能较好地将房屋分布结构识别出。USGS 的场景分类准确率为98.5%,图 14 为场景分类混淆矩阵。可见,JMCNN 在 USGS 大幅影像上分类同样具有优势。

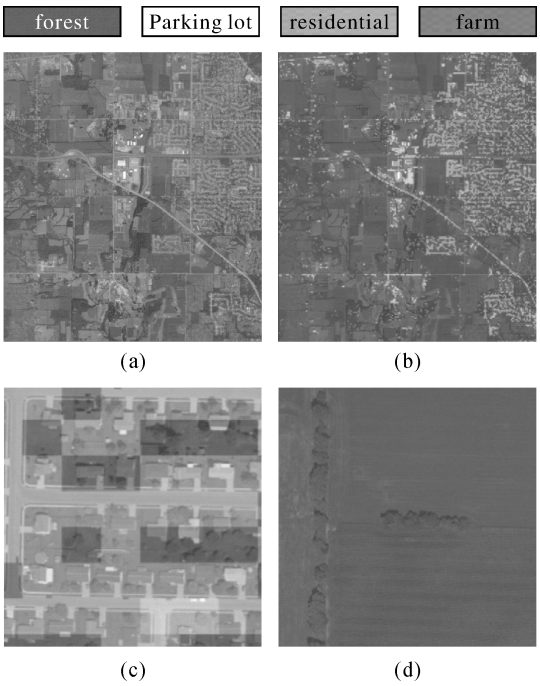

图 12　USGS 大幅遥感影像样本示例及分类结果

Fig.12　The result of classification on USGS large image

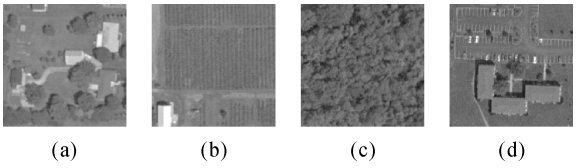

图 13　USGS 大幅遥感影像样本示例

Fig.13　Examples of USGS large image

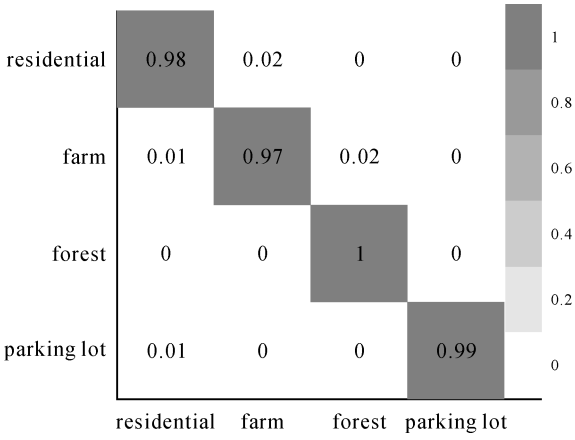

图 14　JMCNN 在 USGS 上分类的混淆矩阵

Fig.14　The confusion matrix of JMCNN on USGS large image

4　结　语

本文在 2400 张 UCM 和 2100 张 SIRI 的小数据集上进行试验，分别获得了 89.3% 和 88.3% 的识别准确率，均高于其他分类器的识别结果。然而，在多类别场景分类中，对于个别模糊场景的类别分类效果欠佳。未来，工作将从以下 3 个方面进行改进：①优化网络联合部分，使得联合特征更具抗差性，以提高 JMCNN 在模糊类别上的分类精度。②考虑调整多尺度特征提取器网络结构，使提取出的多尺度特征更为有效。③引入1×1 的卷积层用来减少参数量，进一步提高模型效率。同时还将探索基于 JMCNN 在不同视角下的遥感影像地物检测。

参考文献：

[1]　CHERIYADAT A M. Unsupervised Feature Learning for Aerial Scene Classification[J]. IEEE Transactions on Geoscience and Remote Sensing，2014，52(1)：439-451.

[2]　SERRANO N，SAVAKIS A E，LUO J B. Improved Scene Classification Using Efficient Low-Level Features and Semantic Cues[J]. Pattern Recognition，2004，37(9)：1773-1784.

[3]　殷慧，曹永锋，孙洪. 基于多维金字塔表达和 AdaBoost 的高分辨率 SAR 图像城区场景分类算法[J]. 自动化学报，2010，36(8)：1099-1106.
YIN Hui，CAO Yongfeng，SUN Hong. Urban Scene Classification Based on Multi-dimensional Pyramid Representation and AdaBoost Using High Resolution SAR Images[J]. Acta Automatica Sinica，2010，36(8)：1099-1106.

[4]　LAZEBNIK S，SCHMID C，PONCE J. Beyond Bags of Features：Spatial Pyramid Matching for Recognizing Natural Scene Categories[C]// Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York，NY：IEEE，2006：2169-2178.

[5]　YANG Yi，NEWSAM S. Spatial Pyramid Co-occurrence for Image Classification[C]// IEEE International Conference on Computer Vision. Barcelona，Spain：IEEE，2011：1465-1472.

[6]　ZHAO Bei，ZHONG Yanfei，ZHANG Liangpei. Scene Classification via Latent Dirichlet Allocation Using a Hybrid Generative/Discriminative Strategy for High Spatial Resolution Remote Sensing Imagery[J]. Remote Sensing Letters，2013，4(12)：1204-1213.

[7]　KRIZHEVSKY A，SUTSKEVER I，HINTON G E. ImageNet Classification with Deep Convolutional Neural Networks [C]// Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe，Nevada：ACM，2012：1097-1105.

[8]　HECHT-NIELSEN R. Theory of the Backpropagation Neural Network[C]// International Joint Conference on Neural Networks. Washington，DC：IEEE，1989(1)：593-605.

[9]　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.

[10] CASTELLUCCIO M, POGGI G, SANSONE C, et al. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks[J]. Acta Ecologica Sinica, 2015, 28(2): 627-635.

[11] PENATTI O A B, NOGUEIRA K, SANTOS J A D. Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains? [C]// IEEE Conference on Computer Vision and Pattern Recognition Workshops. Boston, MA: IEEE, 2015: 44-51.

[12] 李学龙，史建华，董永生，等. 场景图像分类技术综述[J]. 中国科学(信息科学)，2015，45(7)：827-848.
LI Xuelong, SHI Jianhua, DONG Yongsheng, et al. A Survey on Scene Image Classification[J]. Scientia Sinica (Informationis), 2015, 45(7): 827-848.

[13] LI Haoxiang, LIN Zhe, SHEN Xiaohui, et al. A Convolutional Neural Network Cascade for Face Detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: IEEE, 2015: 5325-5334.

[14] GLOROT X, BORDES A, BENGIO Y. Deep Sparse Rectifier Neural Networks[C]// Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, Florida: HAL, 2011: 315-323.

[15] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going Deeper with Convolutions[C]// IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: IEEE, 2015: 1-9.

[16] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[17] ZHAO Bei, ZHONG Yanfei, XIA Guisong, et al. Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(4): 2108-2123.

[18] ZHAO Bei, ZHONG Yanfei, ZHANG Liangpei, et al. The Fisher Kernel Coding Framework for High Spatial Resolution Scene Classification [J]. Remote Sensing, 2016, 8(2): 157.

[19] ZHU Qiqi, ZHONG Yanfei, ZHAO Bei, et al. Bag-of-Visual-Words Scene Classifier with Local and Global Features for High Spatial Resolution Remote Sensing Imagery [J]. IEEE Geoscience and Remote Sensing Letters, 2016, 13(6): 747-751.

[20] ABADI M, BARHAM P, CHEN Jianmin, et al. TensorFlow: A System for Large-scale Machine Learning[C]// Proceedings of the 12th Usenix Conference on Operating Systems Design and Implementation. Berkeley, CA: USENIX Association, 2016.

[21] ZHANG Fan, DU Bo, ZHANG Liangpei. Saliency-Guided Unsupervised Feature Learning for Scene Classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(4): 2175-2184.

[22] LIENOU M, MAITRE H, DATCU M. Semantic Annotation of Satellite Images Using Latent Dirichlet Allocation [J]. IEEE Geoscience and Remote Sensing Letters, 2010, 7(1): 28-32.

[23] CHEN Shizhi, TIAN YingLi. Pyramid of Spatial Relations for Scene-level Land Use Classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(4): 1947-1957.

[24] BREIMAN L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

[25] ADANKON M M, CHERIET M. Support Vector Machine [J]. Computer Science, 2002, 1(4): 1-28.

（责任编辑：张燕燕）

收稿日期：2017-04-17
修回日期：2018-02-09

第一作者简介：郑卓(1996—)，男，本科生，研究方向为深度学习，遥感影像解译。
First author: ZHENG Zhuo (1996—), male, undergraduate, majors in deep learning and Remote sensing image interpretation.
通信作者：刘袁缘
Corresponding author: LIU Yuanyuan
E-mail: liuyy@cug.edu.cn