基于区域注意力机制的遥感图像检索

彭晏飞**, 梅金业*, 王恺欣, 訾玲玲, 桑雨
辽宁工程技术大学电子与信息工程学院, 辽宁 葫芦岛 125105

摘要 遥感图像存在大量语义对象,相同的语义对象视觉差异较大,针对卷积神经网络(CNN)提取的全局特征不能准确描述图像内容的问题,提出了一种使用区域注意力机制的遥感图像检索方法。首先去除 CNN 的全连接层,将高层特征作为区域注意力网络的输入;然后在遥感图像数据集上分别训练 CNN 和区域注意力网络,提取具有区域关注度的图像特征;最后构建了一种多距离相似性度量矩阵并采用扩展查询以提高检索性能。实验结果表明,相比基于全局特征的遥感图像检索方法,本方法能有效抑制遥感图像背景和不相关的图像区域,在两大遥感实验数据集上的检索性能更好。

关键词 遥感图像检索; 卷积神经网络; 区域注意力机制; 多距离矩阵; 扩展查询
中图分类号 TP753 **文献标志码** A **doi**: 10.3788/LOP57.101017

Remote Sensing Image Retrieval Based on Regional Attention Mechanism

Peng Yanfei**, Mei Jinye*, Wang Kaixin, Zi Lingling, Sang Yu
School of Electronic and Information Engineering, Liaoning Technical University, Huludao, Liaoning 125105, *China*

Abstract Remote sensing images have a large number of semantic objects, and the visual differences of the same semantic objects are large. Aiming at the problem that the global features extracted by convolutional neural network (CNN) cannot accurately describe the image content, a remote sensing image retrieval method based on regional attention mechanism is proposed. First, the fully connected layer of the CNN is removed, and the deep features are used as the input of regional attention network. Then, the CNN and regional attention network are trained respectively on remote sensing image dataset. After that, local image features with attention can be extracted. Finally, a multi-distance similarity metric matrix is constructed, and extended query is used to improve retrieval performance. Experimental results show that, compared with remote sensing image retrieval method based on global features, this method can effectively suppress the background of remote sensing images and unrelated image regions, and the retrieval performance is better on the two large remote sensing experimental data sets.

Key words remote sensing image retrieval; convolutional neural network; regional attention mechanism; multi-distance matrix; query expansion
OCIS codes 100.2000; 100.4996; 100.3008

1 引 言

随着遥感技术的快速发展,遥感图像数据库呈爆炸式增长,为了高效管理遥感图像数据库,基于内容的图像检索(CBIR)系统成为国内外研究的热点。CBIR 有两个主要步骤:图像的特征提取和图像的相似性匹配。早期的 CBIR 根据图像的纹理、颜色、形状等视觉特征进行提取,例如尺度不变特征转换(SIFT)算法[1]和方向梯度直方图(HOG)算法[2],但这种低层的全局特征易受视角、光照、遮挡等影响。

随着卷积神经网络(CNN)的提出,CNN 在计算机视觉领域的应用越来越广泛,如图像分类[3]、目标检测[4-5]、图像检索[6]领域。Babenko 等[7]提出了基于 CNN 的图像检索方法,根据目标数据集训练增强 CNN 的图像分类识别能力,相比传统手工提

收稿日期: 2019-09-12; **修回日期**: 2019-10-11; **录用日期**: 2019-10-22
基金项目: 国家自然科学基金(61702241,61602226)、辽宁省教育厅高等学校基本科研项目(LJ2017FBL004)
* **E-mail**: 1113417696@qq.com; ** **E-mail**: pengyf75@126.com

取方法，CNN 提取的深层特征包含了更丰富的图像信息，提高了检索性能。文献[8-10]指出，CNN 经分类训练和微调后，从中间层提取的特征包含了空间信息和语义信息，更有利于图像检索。文献[11-12]采用注意力机制提取图像的深层局部特征，在图像背景内容复杂的情况下，得到了较好的检索效果。文献[13]在 CNN 深层特征图上用不同的尺度进行均匀采样，提取图像的局部特征，实验结果表明，该特征能更好表达图像内容。文献[14]基于文献[11]的方法提取了遥感图像的深层局部特征，再通过 VLAD（Vector of locally aggregated descriptors）方法[15]将这些局部特征进行组合，结果表明，该方法的检索性能优于 CNN。

上述方法中，文献[7]提取的全局特征不能准确地描述图像内容，对检索性能提升有限。文献[11]、[12]、[14]使用注意力机制提取图像的局部特征，虽然在图像背景内容复杂的情况下检索效果较好，但耗费时间长，不满足遥感图像检索的高效性要求。在传统遥感图像检索方法中普遍使用了单一的距离度量公式，研究表明，当数据集出现异常值时，计算样本间的相似度会导致结果不稳定。

本文提出了一种基于区域注意力机制和区域卷积最大激活（R-MAC）算法[16]的深层网络模型，该网络考虑了遥感图像的局部特征和全局特征，并将两者聚合为最终的遥感图像特征。构建了多距离公式矩阵进行遥感图像匹配，采用扩展查询进一步提高检索精度。在两大遥感数据集进行实验，结果表明，本方法的检索性能明显优于基于全局特征的图像检索方法。

2 实验方法

实验使用的检索系统框架如图 1 所示，先由 CNN 提取遥感图像的深层特征，再输入注意力网络提取局部特征，并与全局特征相结合。在检索阶段使用多距离矩阵以及扩展查询进行图像的相似性匹配，返回前 n 个相似内容作为检索结果。

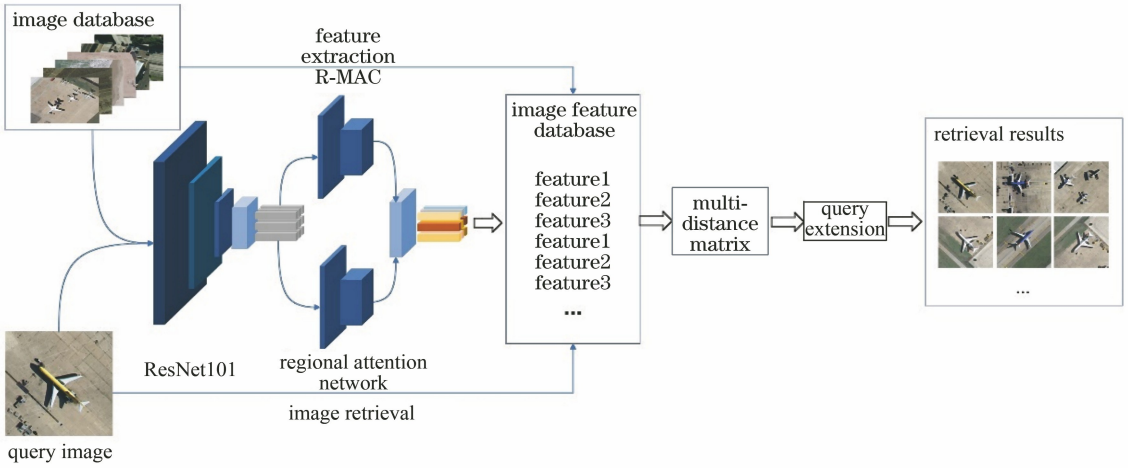

图 1 检索系统框架

Fig. 1 Framework of the retrieval system

2.1 卷积区域最大激活

R-MAC 是一种以固定尺寸窗口在特征图上滑动采样的方法，具体步骤：

1）用特定尺度 $S=1\times1,2\times2,3\times3\cdots$ 的特征图 $\boldsymbol{R}_S$ 进行滑动采样，相邻的两个采样区域的重叠率为 40%。如图 2 所示，在尺度 S 下，产生的区域尺寸为 $2\ \min(W,H)(S+1)$，其中 W 和 H 分别为特征图的宽度和高度。

2）获得当前尺度下的区域后，计算每个区域的特征向量。

3）然后对其依次进行 L2 标准化[17]、主成分分析与白化[18]、L2 标准化。

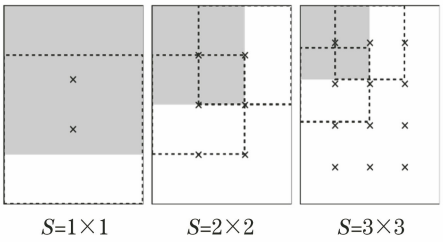

图 2 R-MAC 区域尺度

Fig. 2 Regional scale of R-MAC

4）图像 $\boldsymbol{f}$ 由各尺度的区域特征向量加权求和得到，可表示为

$$\boldsymbol{f}=\sum \boldsymbol{f}_{\boldsymbol{R}_S}, \tag{1}$$

式中，$\boldsymbol{f}_{\boldsymbol{R}_S}$为图像尺度 S 下的区域特征向量。

2.2 区域注意力网络

提取准确描述图像内容的深层特征是提高遥感图像检索性能的关键，以深度残差网络[19](ResNet)模型为主体提取遥感图像的深层特征，能更好地描述图像内容，提高系统的检索性能。ResNet 模型如图 3 所示，其中 FLOPs 表示浮点运算次数，用来衡量模型的复杂度。ResNet50，ResNet101，ResNet152 为三种常用的网络模型。本方法去掉了卷积层 conv5_x 之后的全连接层，使用 ResNet101 conv5_x 的高层输出作为注意力网络的输入。

Layer name	Output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

图 3 ResNet 框架

Fig. 3 Framework of ResNet

区域注意力网络由 R-MAC 和注意力机制构成，采样器先对 ResNet101 conv5_x 得到的特征图进行采样，获取每个区域的特征向量，然后通过区域特征向量的加权平均值获得图像 $\boldsymbol{f}$ 的全局特征向量 $\boldsymbol{f}_i$，可表示为

$$\boldsymbol{f}_i=[f_i,1,\cdots,f_i,n]^{\mathrm{T}}=\frac{\sum_{\boldsymbol{R}\in\boldsymbol{\Omega}}\phi[k]P[M(\boldsymbol{R})]}{|\boldsymbol{\Omega}|},\tag{2}$$

式中，$\boldsymbol{\Omega}$ 为 R-MAC 在 conv5_x 提取的特征图组成的一组区域特征图，$\boldsymbol{R}$ 为 $\boldsymbol{\Omega}$ 中的一个区域特征图，$P[M(\boldsymbol{R})]$为区域特征向量经过最大池化以及后处理操作，k 为对区域特征向量 $\boldsymbol{R}$ 进行平均池化操作，ϕ 为注意力机制函数，用来获得注意力权重，可由两个线性变换矩阵计算得到

$$\phi(k)=X_{\mathrm{softplus}}[\boldsymbol{W}_{\mathrm{c}}\pi(k)+\boldsymbol{b}_{\mathrm{c}}],\tag{3}$$

$$\pi(k)=X_{\tanh}(\boldsymbol{W}_{\mathrm{r}}k+\boldsymbol{b}_{\mathrm{r}}),\tag{4}$$

式中，X_{softplus}、$X_{\tanh}$为激活函数，$\boldsymbol{W}_{\mathrm{r}}\in\boldsymbol{R}^{d\times n}$和 $\boldsymbol{W}_{\mathrm{c}}\in\boldsymbol{R}^{1\times d}$是线性变换矩阵，$\boldsymbol{b}_{\mathrm{r}}\in\boldsymbol{R}^{d\times n}$和 $\boldsymbol{b}_{\mathrm{c}}\in\boldsymbol{R}^{1\times d}$分别为偏置向量和标量。从(1)式～(4)式可以发现注意力机制减小了遥感图像背景与不重要区域的权重，增加了检索目标的权重，实现了对遥感图像局部特征的关注。与传统遥感图像检索方法提取局部特征不同，该方法将图像的全局特征向量和局部特征串联起来，作为遥感图像的特征向量 $\boldsymbol{f}_{\mathrm{I}}$，可表示为

$$\boldsymbol{f}_{\mathrm{I}}=[\boldsymbol{f}_{\mathrm{I}},1,\cdots\boldsymbol{f}_{\mathrm{I}},n]^{\mathrm{T}}=\frac{\sum_{\boldsymbol{R}\in\boldsymbol{\Omega}}\phi[k\oplus J(\boldsymbol{V}_{\mathrm{I}})]P[M(\boldsymbol{R})]}{|\boldsymbol{\Omega}|},\tag{5}$$

式中，$\boldsymbol{V}_{\mathrm{I}}$为基于 ResNet101 的 conv5_x 产生的全局特征，$J(\boldsymbol{V}_{\mathrm{I}})$为平均池化，$\oplus$为通道空间中向量的串联。本方法结合了 ResNet101 和区域注意力网络用于提取图像特征，网络框架如图 4 所示。

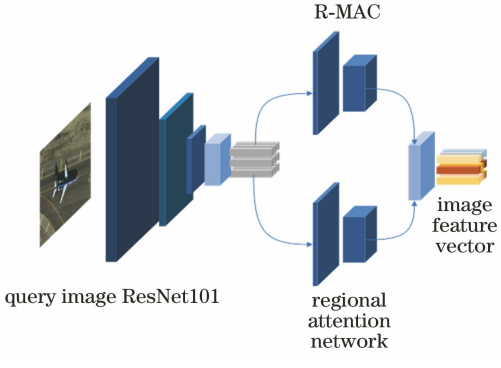

图 4 本网络的框架图

Fig. 4 Framework diagram of our network

2.3 多距离相似性度量矩阵

传统遥感图像检索方法普遍采用欧氏距离计算不同图像的相似性，但对于不同类别的遥感图像或图像出现异常时，检索结果与实际相差较大。针对

单一的距离计算公式不能满足遥感图像检索稳定性的要求，构建一个多种距离相似性度量矩阵进行遥感图像的相似性匹配。

对于空间中的 n 维特征向量 $\boldsymbol{X}(x_1,x_2,\cdots,x_n)$ 和 $\boldsymbol{Y}(y_1,y_2,\cdots,y_n)$，两者间的欧氏距离为

$$D_1(\boldsymbol{X},\boldsymbol{Y})=\sqrt{\sum_i^n (x_i-y_i)^2}, \tag{6}$$

切比雪夫距离为

$$D_2(\boldsymbol{X},\boldsymbol{Y})=\lim_{p\to\infty}\left(\sum_{i=1}^n |x_i-y_i|^p\right)^{\frac{1}{p}}, \tag{7}$$

余弦距离为

$$D_3(\boldsymbol{X},\boldsymbol{Y})=\frac{\sum_{p=1}^n x_i y_i}{\sqrt{\sum_{p=1}^n x_i^2}\sqrt{\sum_{p=1}^n y_i^2}}, \tag{8}$$

构建的多距离相似性度量矩阵为

$$\boldsymbol{D}_{\mathrm{M}}=[D_1 \quad D_2 \quad D_3]。 \tag{9}$$

将 $\boldsymbol{D}_{\mathrm{M}}$ 进行 L2 归一化，并通过实验对比了使用单一距离公式和多距离相似性度量矩阵的性能。

2.4 扩展查询

扩展查询可提高图像检索的查全率，具体步骤：

1）首先将待查询图像作为检索输入，得到 N 幅内容相似的遥感图像；

2）对第一次查询返回的前 N 幅遥感图像包括待查询图像的特征向量求取平均值；

3）将步骤 2）得到的特征向量平均值作为检索输入再次进行查询。

采用扩展查询，在两个遥感数据集上进行实验，结果表明，图像检索查准率相比未使用扩展查询提升了 2%～3%。

3 实验结果与分析

3.1 实验设置

3.1.1 数据集

UC Merced Land-Use[20]（UCM）数据集如图 5 所示，该数据集中包含农田类、飞机类等 21 类图像，每类有 100 幅尺寸为 256 pixel×256 pixel 的遥感图像。使用数据增强方法，将该数据集扩充至 4200 幅。实验中，每类随机抽取 160 幅图像作为训练集，剩余 40 幅作为测试集。

SIRI-WHU[21-23]（SIRI）遥感数据集如图 6 所示，该数据集包含 12 大类，每类有 200 幅尺寸为 200 pixel×200 pixel 的遥感图像。同样每类随机抽取 40 幅作为测试集，剩余作为训练集。

图 5 UCM 数据集

Fig. 5 UCM dataset

图 6 SIRI 数据集

Fig. 6 SIRI dataset

3.1.2 网络训练

基于 ResNet101 CNN 训练时，使用学习率为 0.001，batch_size 为 32，动量为 0.9 的随机梯度下降优化算法进行训练。当 ResNet101 CNN 完成训练后，添加区域注意力网络，固定并冻结 ResNet101 CNN 的参数，更新区域注意力网络参数。在特征降维时，使用 ResNet101 CNN 预训练网络在两个数据集上进行主成分分析(PCA)学习，最后进行遥感图像的特征提取。

3.1.3 对比方法

为了验证本方法的有效性，选取基于预训练的 VGG16[24] CNN 模型、DenseNet121[25] CNN 模型及未添加区域注意力网络的 ResNet101 CNN 模型作为对比进行实验。

3.1.4 评价标准

图像检索的常用评价标准为查准率(precision)、查全率(recall)、以及平均检索精度(mAP)。查准率、查全率可表示为

$$\begin{cases} X_{\text{precision}} = \dfrac{m}{N} \times 100\% \\ X_{\text{recall}} = \dfrac{m}{M} \times 100\% \end{cases}, \tag{10}$$

式中，N 为检索系统返回最相似的图像数量，m 为真正与待查询图像相似的正确候选项数量，M 为遥感数据库中与待查询图像相似的图像数量。mAP 可表示为

$$X_{\text{mAP}} = \frac{1}{C}\sum_{i=1}^{C} A_i, \tag{11}$$

$$A_i = \frac{1}{m}\sum_{i=1}^{N} P_i R_i, \tag{12}$$

式中，C 为查询次数，m 为检索返回 N 幅图像时真正与待查询图像相似的个数，A_i 为检索精度，

$$R_i = \begin{cases} 1, & \text{if } \ i \text{ similar to query image} \\ 0, & \text{other} \end{cases}, \tag{13}$$

$$P_i = \frac{N_i}{i}, \tag{14}$$

式中，N_i 为检索结果中真正与待查询图像相似的排序。

3.2 实验结果分析

返回 20 幅、40 幅检索图像的检索效果如图 7、图 8 所示，可以发现当检索返回 20 幅图像时，遥感图像的查准率为 100%；当检索返回 40 幅图像时，本方法的检索性能依然较好，原因是使用了区域注意力网络。

图 7 返回 20 幅图像的检索效果

Fig. 7 Retrieval effect when return 20 images

图 8 返回 40 幅图像的检索效果

Fig. 8 Retrieval effect when return 40 images

3.2.1 平均检索精度分析

当返回 40 幅检索图像时，不同方法的 mAP 如表 1 所示，可以发现，本方法的检索性能明显优于使用预训练的 VGG16、DenseNet121、ResNet101 的 CNN 模型。在 UCM 数据集上本方法的 mAP 为 96.8%，在 SIRI 数据集上的 mAP 为 88.6%。

表 1 不同方法的 mAP 对比

Table 1 Comparison of mAP of different methods

unit：%

Method	VGG16	ResNet101	DenseNet121	Our method
UCM	71.2	75.9	78.9	96.8
SIRI	70.0	75.7	74.4	88.6
Average	70.6	75.8	76.7	92.7

在遥感图像的特征提取阶段，输入的尺寸不同，检索精度也不同。同一数据集中输入不同尺寸遥感图像的 mAP 如表 2 所示，其中 1.0 表示图像为原始尺寸，其他尺寸为图像的缩放倍数。可以发现，输入原始图像尺寸时，检索精度最高；将原始图像缩小一半时，特征提取速度有明显提升，但检索精度大幅度下降，即输入遥感图像的尺寸会影响检索精度。

表 2 不同尺寸图像的 mAP 对比

Table 2 Comparison of mAP of different image sizes

Image size	1.0	0.9	0.8	0.7	0.6	0.5
mAP /%	88.6	85.3	84.7	82.1	77.3	70.9

3.2.2 平均查全率分析

四种方法在两个数据集上的平均查全率如图 9 所示，可以发现，在 UCM 数据集上，本方法在返回 100 幅检索图像时的平均查全率为 100%。这表明测试集中检索任意一幅遥感图像，在返回 100 幅图像时即可被全部检索出来，而其他方法，最高平均查全率仅为 76.7%，全部检索到相似遥感图像需要返回 500 幅到 600 幅。在 SIRI 数据集上，本方法同样优于其他方法，在返回图像数更少的情况下，正确候选项更多，这表明了本方法的优越性。

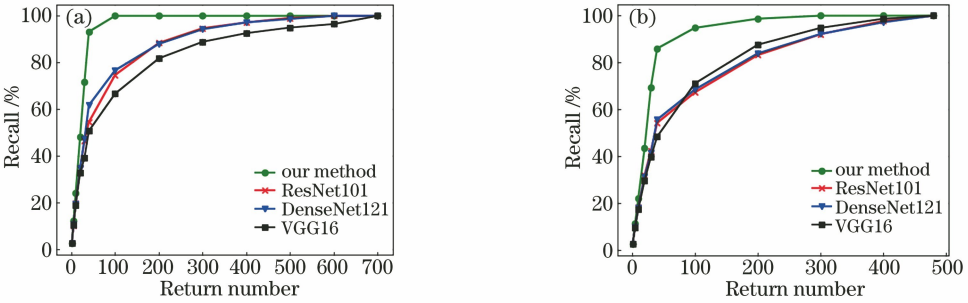

图 9 平均查全率对比。(a) UCM 数据集；(b) SIRI 数据集

Fig. 9 Comparison of average recalls. (a) UCM dataset; (b) SIRI dataset

3.2.3 平均查准率分析

四种方法对两个数据集中每个类别的遥感图像平均查准率如表 3、表 4 所示。其中 RAN 表示使用 ResNet101 CNN 并添加了区域注意力网络，RAN+MD表示在 RAN 的基础上采用多距离相似性度量矩阵，RAN+MD+QE 表示在 RAN+MD 的基础上添加了扩展查询。

实验结果表明，在 UCM 数据集中，某些遥感图像类别，如建筑物、密集住宅类、中型住宅区、稀疏住宅区的检索效果与其他类差距较大。原因是这三类遥感图像都含有树木、建筑物等目标，在视觉上，中型住宅区类和密集住宅区类较为相似，导致区域注意力网络没有学习到如何正确提取这些目标特征。同样在 SIRI 数据集中，海港类、工业类、牧场类、池塘类检索效果较差，但总体上，本方法比其他方法检索性能更高。

表 3 不同方法在 UCM 数据集上的平均查准率

Table 3 Average accuracy of different methods on the UCM dataset unit: %

Method	VGG16	DenseNet121	ResNet101	Our method		
				RAN	RAN+MD	RAN+MD+QE
Agricultural	92	86	81	**97**	**97**	**97**
Airplane	69	83	84	95	**96**	95
Baseball diamond	43	38	46	**100**	**100**	**100**
Beach	81	85	90	98	99	**100**
Building	23	44	46	69	**73**	65
Chaparral	92	99	95	**100**	**100**	**100**
Dense residential	29	46	27	77	**85**	82
Forest	80	92	84	**100**	**100**	**100**
Freeway	43	44	30	88	89	**92**
Golf course	33	41	51	94	95	**97**
Harbor	43	75	75	98	98	**100**
Intersection	26	45	40	96	95	**98**
Medium residential	37	78	55	80	**82**	80
Mobile home park	69	80	39	94	**95**	**95**
Overpass	48	64	43	94	**95**	92
Parking lot	62	77	60	**99**	93	95
River	25	42	50	88	85	**93**
Runway	57	65	49	87	83	**96**
Sparse residential	49	45	41	79	**85**	**85**
Storage tanks	25	31	25	92	86	**95**
Tennis court	38	36	33	98	99	**100**
Average	50.7	61.7	54.5	91.6	91.9	**93.2**

表 4 不同方法在 SIRI 数据集上的平均查准率

Table 4 Average accuracy of different methods on the SIRI dataset unit: %

Method	VGG16	DenseNet121	ResNet101	Our method		
				RAN	RAN+MD	RAN+MD+QE
Agriculture	59	73	29	97	95	**99**
Commercial	51	59	67	87	89	**94**
Harbor	50	54	65	60	62	**67**
Idle land	27	43	39	85	85	**88**
Industrial	42	49	66	**78**	73	75
Meadow	29	32	27	64	**69**	**69**
Overpass	71	81	89	97	97	**98**
Park	35	43	44	77	79	**83**
Pond	39	49	46	66	62	**76**
Residential	51	60	48	87	87	**90**
River	34	30	35	89	90	**95**
Water	93	95	96	97	**98**	**98**
Average	48.4	55.7	54.3	82.0	82.2	**86.0**

从表 3、表 4 中可以发现，本方法平均查准率比 VGG16 提高了 40.1 个百分点，比 DenseNet121 提高了30.9个百分点，比 ResNet101 提高了 35.2 个百分点。采用 RAN 方法比直接使用 ResNet101 CNN 的平均查准率更高，这证明了本方法的检索性能比提取全局特征的方法更好；相比 RAN 方法，使用 RAN+MD 方法提高了遥感图像检索的平均查准率，但在某些类别，如储油罐类、河流类检索性能有所下降，在 SIRI 数据集中也有此类现象。原因是单一的距离公式在某些类别的遥感图像上会产生较大的性能差异，而本方法使用多距离相似性度量矩阵，结合了多种样本的相似性计算方法，使网络的检索性能更稳定，平均查准率得到提升。

在两个数据集中的大部分图像类别上，相比 RAN+MD 方法，采用扩展查询的 RAN+MD+QE 方法检索性能有明显提升，特别是公路类、停车场类、河流类、储油罐类。原因是扩展查询将正确候选项的前 k 个遥感图像特征向量与待查询图像特征向量求和并取平均值，反映了这一类遥感图像的大致特征，这表明本方法能有效提升网络的图像检索性能。

本方法在 UCM 数据集上的检索性能优于 SIRI 数据集。原因是 SIRI 数据集中遥感图像数量较少，只有 12 类，且每类只有 200 幅遥感图像，CNN 没有正确学习到相关特征，可提前终止训练防止网络过拟合解决该问题。此外，SIRI 数据集的遥感图像比较模糊，如在池塘类中的两幅图像，视觉差异性较大，且有些与农田类的图像视觉相似性较大。而 UCM 数据集中，同类别图像的视觉差异较小，且不同类别图像的区分度较大。

3.2.4 区域注意力尺度

区域尺度 S 是影响检索性能的重要因素，为避免多距离相似性度量矩阵和扩展查询对实验结果的影响，将上述两个方法剔除，只考虑不同尺度下的平均查准率，结果如图 10 所示，横坐标为区域尺度 S 的值，纵坐标为查准率。

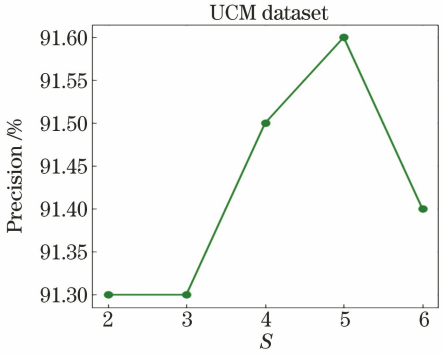

图 10 不同尺度下的平均查准率

Fig. 10 Average precision at different scales

从图 10 可以发现，在 UCM 数据集上，当 $S<5$ 时，遥感图像检索平均查准率随 S 的增加而增加，当 $S=5$ 时，平均查准率达到饱和，随后下降。尺度参数过大会加大系统计算量，因此设置合适的尺度参数能有效提升遥感图像检索性能。

经一系列实验表明，采用基于区域注意力机制提取遥感图像特征结合多距离相似性度量矩阵的图像匹配方法，可有效提升遥感图像的检索性能，同时满足遥感图像检索的高效性要求。

4 结　论

针对遥感图像的特殊性，即具有相同语义的图像，视觉上存在巨大差异，提出一种基于区域注意力机制的遥感图像检索方法，该方法关注遥感图像不同尺度下的区域特征，有效抑制了背景和不重要的遥感图像区域。使用两个公共数据集训练 CNN，随后冻结网络参数，更新区域注意力网络参数，并将该网络用于遥感图像的特征提取，采用多距离相似性度量矩阵和扩展查询进行图像检索。实验结果表明，本方法能显著提高遥感图像的检索性能，与基于全局特征的遥感图像检索方法相比，对视觉上相似而语义信息不同的两幅遥感图像区分性更强。

参 考 文 献

[1] Lowe D G. Object recognition from local scale-invariant features[C] // Proceedings of the Seventh IEEE International Conference on Computer Vision, September 20-27, 1999, Kerkyra, Greece. New York: IEEE, 1999, 2: 1150-1157.

[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C] // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 20-25, 2005, San Diego, CA, USA, New York: IEEE, 2005, 1: 886-893.

[3] Zhao X H, Yin L F, Zhu Y N, et al. Improved image classification algorithm based on principal component analysis network [J]. Laser & Optoelectronics Progress, 2019, 56(2): 021004.
赵小虎, 尹良飞, 朱亚楠, 等. 基于主成分分析网络的改进图像分类算法[J]. 激光与光电子学进展, 2019, 56(2): 021004.

[4] Ou P, Zhang Z, Lu K, et al. Object detection of remote sensing images based on convolutional neural networks[J]. Laser & Optoelectronics Progress, 2019, 56(5): 051002.
欧攀, 张正, 路奎, 等. 基于卷积神经网络的遥感图像目标检测[J]. 激光与光电子学进展, 2019, 56(5): 051002.

[5] Wang J Q, Li J S, Zhou X W, et al. Improved SSD algorithm and its performance analysis of small target detection in remote sensing images[J]. Acta Optica Sinica, 2019, 39(6): 0628005.
王俊强, 李建胜, 周学文, 等. 改进的 SSD 算法及其对遥感影像小目标检测性能的分析[J]. 光学学报, 2019, 39(6): 0628005.

[6] Peng Y F, Song X N, Zi L L, et al. Remote sensing image retrieval based on convolutional neural network and modified fuzzy C-means [J]. Laser & Optoelectronics Progress, 2018, 55(9): 091008.
彭晏飞, 宋晓男, 訾玲玲, 等. 基于卷积神经网络和改进模糊 C 均值的遥感图像检索[J]. 激光与光电子学进展, 2018, 55(9): 091008.

[7] Babenko A, Slesarev A, Chigorin A, et al. Neural codes for image retrieval[M] // Fleet D, Pajdla T, Schiele B, et al. Computer Vision-ECCV 2014, Lecture Notes in Computer Science. Cham: Springer, 2014, 8689: 584-599.

[8] Gordo A, Almazán J, Revaud J, et al. End-to-end learning of deep visual representations for image retrieval[J]. International Journal of Computer Vision, 2017, 124(2): 237-254.

[9] Zhou W G, Li H Q, Sun J, et al. Collaborative index embedding for image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1154-1166.

[10] Hao J D, Dong J, Wang W, et al. What is the best practice for CNNs applied to visual instance retrieval? [EB/OL]. [2019-09-02]. https: // arxiv. org/abs/1611.01640.

[11] Noh H, Araujo A, Sim J, et al. Large-scale image retrieval with attentive deep local features[C] // 2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice Italy. New York: IEEE, 2017: 3456-3465.

[12] Teichmann M, Araujo A, Zhu M L, et al. Detect-to-retrieve: efficient regional aggregation for image search [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 5109-5118.

[13] Kim J, Yoon S E. Regional Attention Based Deep Feature for Image Retrieval[C] // Proceedings of British Machine Vision Conference , School of Computing, Korea Advanced Institute of Science and Technology, 2018: 209. [2019-09-02]. http: // sglab.kaist.ac.kr/RegionalAttention/.

[14] Imbriaco R, Sebastian C, Bondarev E, et al. Aggregated deep local features for remote sensing image retrieval[J]. Remote Sensing, 2019, 11(5): 493.

[15] Jégou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation [C] // 2010 IEEE Computer Society Conference on Computer Vision & Pattern Recognition, June 13-18,

2010, San Francisco, CA, USA. New York: IEEE, 2010: 3304-3311.

[16] Tolias G, Sicre R, Jégou H. Particular object retrieval with integral max-pooling of CNN activations[EB/OL]. [2019-08-25]. https://arxiv.org/abs/1511.05879.

[17] Hoffer E, Banner R, Golan I, et al. Norm matters: efficient and accurate normalization schemes in deep networks[EB/OL]. [2019-08-30]. https://arxiv.org/abs/1803.01814.

[18] Jégou H, Chum O. Negative evidences andCo-occurences in image retrieval: the benefit of PCA and whitening[M]// Fitzgibbon A, Lazebnik S, Perona P, et al. Computer Vision-ECCV 2012, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2012, 7535: 774-787.

[19] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[20] Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification[C]// Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, November 2-5, 2010, San Jose, California. New York, USA: ACM Press, 2010: 270-279.

[21] Zhao B, Zhong Y F, Xia G S, 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.

[22] Zhao B, Zhong Y F, Zhang L P, et al. The fisher kernel coding framework for high spatial resolution scene classification[J]. Remote Sensing, 2016, 8(2): 157.

[23] Zhu Q Q, Zhong Y F, Zhao B, 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.

[24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2019-08-28]. https://arxiv.org/abs/1409.1556

[25] Huang G, Liu Z, van der Maaten, L, et al. Densely connected convolutional networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 4700-4708.