Computer Science and Application 计算机科学与应用, 2021, 11(3), 515-524
Published Online March 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.113052

Hans汉斯

一种用于遥感图像检索的双重注意力深度神经网络

陈光明[1]，王卓薇[1]，陈立宜[1]，邱俊豪[2]，何俊霖[1]

[1]广东工业大学计算机学院，广东 广州
[2]广东工业大学机电工程学院，广东 广州
Email: chenguangmingfe@foxmail.com

收稿日期：2021年2月15日；录用日期：2021年3月9日；发布日期：2021年3月16日

摘　要

因为遥感图像背景复杂，所以提取判别性强特征是遥感图像检索的一个核心技术。本文引入双重自注意力模块，利用空间和通道上的长距离上下文信息，编码局部特征，从而增强特征的表达能力。本文分别在3个典型的数据集上做了实验，在UC Merced Land Use、Satellite Remote Sensing Image Database、NWPU-RESISC45的平局检索精度分别为0.92、0.90和0.89。实验表明，双重自注意力深度学习网络对遥感图像检索性能的提升有显著的作用。

关键词

遥感图像检索，注意力机制，CNN，深度学习

A Dual Attention Deep Neural Network for Remote Sensing Image Retrieval

Guangming Chen[1], Zhuowei Wang[1], Liyi Chen[1], Junhao Qiu[2], Junlin He[1]

[1]School of Computer Science, Guangdong University of Technology, Guangzhou Guangdong
[2]School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou Guangdong
Email: chenguangmingfe@foxmail.com

Received: Feb. 15th, 2021; accepted: Mar. 9th, 2021; published: Mar. 16th, 2021

Abstract

Extracting discriminative features is a core technology for remote sensing image retrieval due to

文章引用: 陈光明, 王卓薇, 陈立宜, 邱俊豪, 何俊霖. 一种用于遥感图像检索的双重注意力深度神经网络[J]. 计算机科学与应用, 2021, 11(3): 515-524. DOI: 10.12677/csa.2021.113052

the complex background of the remote sensing image. In order to enhance the expressive ability of the features, the paper introduces dual attention module to encode the long-distance length information on the spatial and the channel dimensions into local features. Experiments were carried out on three typical datasets. We have conducted experiments on three typical datasets to ascertain the effectiveness of our method. The retrieval precisions on UC Merced Land Use, Satellite Remote Sensing Image Database, and NWPU-RESISC45 are 0.92, 0.90 and 0.89. The experiment shows the self-attention deep learning network gets a significant effect on the improvement of remote sensing image retrieval performance.

Keywords

Remote Sensing Image Retrieval, Attention Mechanism, CNN, Deep Learning

1. 引言

随着遥感图像技术的高速发展，遥感图像的数量急剧增加。如何在大型遥感数据库中有效地组织，管理和检索遥感图像，已经成为遥感图像应用中的紧迫而迫切的问题。其中，基于内容的遥感图像检索(CBRSIR) [1] [2]是遥感应用中最关键的技术。CBRSIR 可以概括为两个步骤：特征提取和相似性度量。CBRSIR 的性能通常取决于从遥感图像中提取的判别特征[3]。因此，作为 CBRSIR 的最关键步骤，特征提取是大多数 CBRSIR 研究的重点[4]。

特征提取主要有两种方法：基于手工特征的方法和基于学习特征的方法[5]。基于手工特征包括颜色、纹理、形状等全局特征和基于 SIFT [6]和 SURF [7]的局部特征。此外，词袋模型(BOW) [8] [9]和局部聚集描述符的向量(VLAD) [10]用于编码局部特征，可以进一步增强特征的表达能力。无论是全局特征还是局部特征，它们都不能很精确地表达图像，所以在高级语义和低级语义之间存在“语义鸿沟”。随着深度学习的发展，卷积神经网络(CNN)，在计算机视觉领域，例如分类[11] [12] [13]、检测[14] [15] [16]、分割[17] [18]等方面，展现了优异的性能优势，CNN 已经广泛应用于图像特征提取。CNN 可以通过大量的卷积层堆叠来提取高级语义特征。GE 等人[19]将在 ImageNet 上训练得到预训练模型应用到遥感图像数据集上，表明 CNN 的特征明显优于传统的手工特征。

然而，在现有的 CNN 特征提取方法中存在一个问题。与其他图像相比，遥感图像具有几个特殊特征。例如，即使在同一类别中，不同图像中的目标也可能具有不同的大小，颜色和角度。更重要的是，目标区域周围的其他材料和背景可能会导致较高的组内差异和较低的组间差异。因此，在现有的 CNN 特征提取方法中存在一个问题，即提取到的特征空间中的图像表示可能无法准确反映其真实类别信息。准确检索具有相似视觉内容的图像需要提取足够描述性和鲁棒性的特征。

针对上述问题，在 Fu 的研究[20]的启发下，本文提出了一种双重注意力模型。本文的主要贡献分为两个部分：

1) 本文设计了一种双重注意力深度学习网络，通过捕获空间和通道的特征依赖关系，提取具有复杂背景的遥感图像的显著性特征，以准确反映真实类别信息。

2) 本文引入的双重注意力模块包括空间注意力模块和通道注意力模块。对于空间注意力模块，使用自注意力机制来捕获特征图上任意两个位置上的依赖关系。对于通道注意力模块，我们引入自注意力机

制来捕获特征图上任意两个通道的依赖关系。

本文的结构如下：第二章介绍遥感图像检索的相关工作，第三章介绍提出的方法，第四章展示实验，第五章总结本文。

2. 相关工作

2.1. 基于学习的特征提取

在特征提取领域，CNN 逐渐取代了传统的方法，变得越来越流行了。CNN 通过深层网络结构中的非线性函数，从训练数据中学习参数权重。但是，遥感图像数据集数据量小，导致了不能从零开始训练 CNN 模型。即使大型基准数据集与遥感图像数据集有很大差异，但是使用大型基准数据集上训练得到的预训练模型可以在一定程度上解决因为遥感图像数据集数据量不足而带来的问题。一些研究[21] [22] [23] [24]已经比较了在不同网络和不同层之间提取的特征的性能。Mnih 等人[21]使用预训练模型的特征和简单的聚合特征，提高了检索性能。Wang 等人[24]提出了对遥感图像数据集上微调预训练模型的方法，同时提出了一种基于三层感知器的 CNN 网络结构。该感知器不仅参数较少，而且可以学习底层局部特征。Shao 等人[25]研究了在多标签遥感图像检索框架下研究了不同深度学习架构的有效性，并获得较好的检索效果。Roy 等人[26]提出三元组深度度量学习深度卷积神经网络，利用三元组损失函数，使得在语义空间中，来自同一类别的图像彼此接近，而来自不同类别的图像则彼此远离。

2.2. 注意力机制

注意力机制通过学习不同区域的权重分布，来为不同区域分配不同的“关注度”。注意力机制的有效性已在许多任务中得到证明，包括机器翻译和文本等基于序列的任务以及分类和分割等计算机视觉任务。一些研究[27] [28] [29]将学习到的权重应用于原始图像，Yuan 等人[30]将权重学习应用于特征图。Huang 等人[31]考虑了特征通道之间的关系，在特征通道上加入了注意力机制。Fu 等人[20]结合了特征通道和特征空间两个维度的注意力机制。Wang 等[32]提出多头注意机制，引入额外的特征映射和实现了自注意力机制。所有这些工作都被用于自然图像处理方面，其中在分类，检测等方面表现出了出色的性能。目前比较少应用于遥感图像处理的注意力模型。Du 等人[33]将结合了特征通道和特征空间两个维度的注意力机制应用于遥感图像处理。Maxim 等人[34]采用注意力机制提取遥感图像的深层局部特征，在图像背景内容复杂情况下，依然实现较好的检索性能。

3. 本文方法

本节介绍双重注意力深度学习网络的具体细节。在检索具有复杂背景的遥感图像时，关键是提取遥感图像关键特征。因此，我们引入双重注意力机制来，模拟跨越图像区域的长距离、多层的依赖关系，有效地对上下文进行建模，编码为局部特征，从而增强特征的表达能力。该模型的总体结构见图 1。

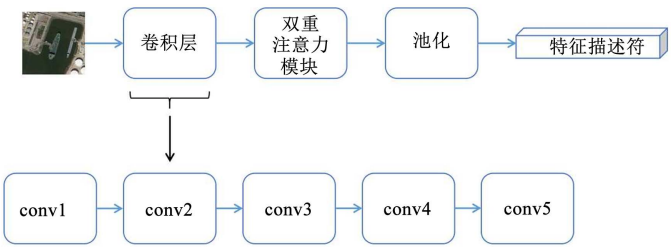

Figure 1. The overall architecture of the model
图 1. 模型的整体结构

3.1. 网络结构与池化

我们使用 ResNet-50 [13]作为模型的骨架。ResNet-50 包括五个卷积层，每个卷积层包括一个卷积操作、一个修正线性单元和最大池化操作。输入一个图像，我们只需要 ResNet-50 最后的一个卷积层的输出的特征图，不需要全连接层的输出。我们从最后一个卷积层得到一个张量 $\chi \in R^{W\times H\times N}$，其中 N 表示通道数，W 表示特征图的宽度，H 表示特征图的高度。

SPoC [35] 使用的是平均池化操作，，如公式(1)所示。MAC [36]使用的是最大池化操作，如公式(2)所示。这两种池化方法已经在标准数据库中已经取得较好的结果。

$$\left[F_1^{(\mathrm{SPoC})},\cdots,F_M^{(\mathrm{SPoC})}\right]^{\mathrm{T}},F_m^{(\mathrm{SPoC})}=\frac{1}{|O_m|}\sum\nolimits_{o\in O_m}o \tag{1}$$

$$\left[F_1^{(\mathrm{MAC})},\cdots,F_M^{(\mathrm{MAC})}\right]^{\mathrm{T}},F_m^{(\mathrm{MAC})}=\max\nolimits_{o\in O_m}o \tag{2}$$

相比这两种池化方法，GeM [37]使用的是一种广义平均池化操作，，如公式(3)所示。GeM 性能更好，可以提升检索精度。所以，我们使用 GeM 来聚合特征，获得更紧凑的特征。假设 χ_k 表示 χ 的第 k 个特征图。

$$\left[F_1^{(\mathrm{GeM})},\cdots,F_M^{(\mathrm{GeM})}\right]^{\mathrm{T}},F_m^{(\mathrm{GeM})}=\left(\frac{1}{|O_m|}\sum\nolimits_{o\in O_m}o^{p_m}\right)^{\frac{1}{p_m}} \tag{3}$$

其中 O 是双重注意力模块输出的特征。SPoC 和 MAC 是 GeM 特殊情况。在公式中，当 $p_m\to\infty$时，公式(3)会转化为公式(2)，即最大池化操作；当 $p_m\to 1$时，公式(3)会转化为公式(1)，即平均池化操作。最后得到的特征图的维数等于 M。在我们的模型中，M 等于 2048。最后，对特征图进行 l^2 归一化操作。

3.2. 双重注意力模块

双重注意力模块的结构见图 2。双重注意力模块包括空间注意力模块和通道注意力模块。我们将从 ResNet-50 最后一层卷积层得到的特征图，分别输入到空间注意力模块和通道注意力模块中。下一步，将两个注意力模块的结果通过加操作融合在一起。最后通过一个卷积层输出结果。

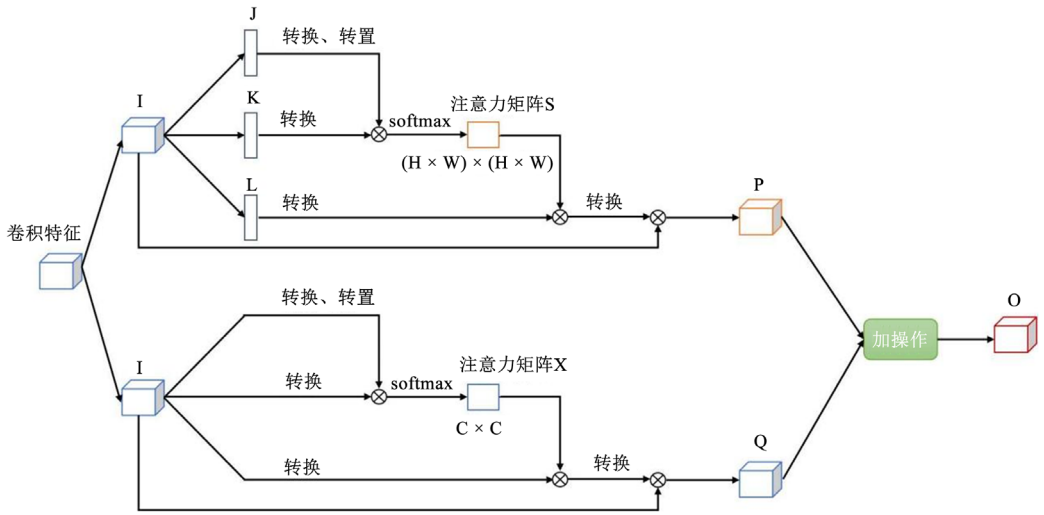

Figure 2. The overall architecture of the double attention module
图 2. 双重注意模块的整体结构

3.2.1. 空间注意力模块

遥感图像检索任务需要判别性强的特征。我们引入空间注意力模块，借助长距离上下文信息编码来增强特征的判别性。接下来，我们介绍空间注意力模块的处理过程。

我们将 χ 变换为 $I \in R^{C\times H\times W}$，然后将其输入到一个卷积层，生成两个特征 J、K 和 L，其中 $\{J,K,L\} \in R^{C\times H\times W}$。我们将它们变换成二维矩阵 $R^{C\times N}$，其中 N 为像素个数，$N = H\times W$。我们对 J 做多一个转置操作。如公式(4)，在将 J 和 K 进行乘操作之后，我们将得到的矩阵经过一个 softmax 层，得到空间注意力矩阵 $S \in R^{N\times N}$，如公式(4)所示。

$$s_{ji} = \frac{\exp\left(J_i \cdot K_j\right)}{\sum_{i=1}^{N} \exp\left(J_i \cdot K_j\right)} \tag{4}$$

其中 s_{ji} 表示第 i 个位置对第 j 个位置的影响。两个位置的特征越相似，它们之间的相关性就越高。同时，我们对变换过的 L 与 S 进行乘操作，并将结果转换为矩阵 $T \in R^{C\times H\times W}$。最后，我们将 T 乘以比例系数 α 后，与 A 进行加操作，得到最后的输出 $P \in R^{C\times H\times W}$，如公式(5)所示。

$$P_j = \alpha \sum_{i=1}^{N} \left(s_{ji} L_i\right) + I_j \tag{5}$$

其中 α 初始化为 0，并随着学习，逐渐增大权重。P 的每个位置的最终特征是所有位置的特征与原始特征的加权和。因此，它具有全局上下文信息，并根据空间注意力选择性地聚合上下文，从而提高了类内的紧凑性和语义一致性。

3.2.2. 通道注意力模块

我们引入通道注意力机制来构建通道之间的依赖关系。接下来，我们介绍空间注意力模块的处理过程。

通道注意力模块和空间注意力模块是基本一样的操作，但有两点不同。在通道注意力模块中，我们直接没有使用卷积层来处理，而是直接将 I 转换，并和通道注意力特征矩阵 $X \in R^{C\times C}$。我们将 I 转换成矩阵 $R^{C\times N}$，然后将对 I 和 I 的转置矩阵进行乘操作。最后，我们将得到的矩阵经过一个 softmax 层，得到通道注意力矩阵 $X \in R^{C\times C}$，如公式(6)所示。

$$x_{ji} = \frac{\exp\left(I_i \cdot I_j\right)}{\sum_{i=1}^{N} \exp\left(I_i \cdot I_j\right)} \tag{6}$$

其中 s_{ji} 表示第 i 个通道对第 j 个通道的影响。同时，我们对变换过的 X 与 A 进行乘操作，并将结果转换为矩阵 $T \in R^{C\times H\times W}$。最后，我们将 T 乘以比例系数 β 后，与 A 进行加操作，得到最后的输出 $Q \in R^{C\times H\times W}$，如公式(7)所示。

$$Q_j = \beta \sum_{i=1}^{C} \left(x_{ji} I_i\right) + I_j \tag{7}$$

其中 β 初始化为 0，并随着学习，逐渐增大权重。Q 是对通道间的长距离信息进行建模，从而提高了特征的可判别性。

3.3. 损失函数

Radenović 等人[37]发现使用对比损失函数的情况比使用三元组损失函数的情况，使用对比损失函数时的检索精度更高。所以，我们也使用对比损失函数，如公式(8)所示。

$$LOSS_{(i,j)}=\begin{cases}\frac{1}{2}\left\|F(i)-F(j)\right\|^2, Y(i,j)=1\\ \frac{1}{2}\left(\max\left\{0,\tau-\left\|F(i)-F(j)\right\|\right\}\right)^2, Y(i,j)=0\end{cases} \tag{8}$$

其中每个输入包括一组图片(i,j)和一组标签$Y(i,j)\in\{0,1\}$。当i和j匹配时，$Y(i,j)=1$；否则，$Y(i,j)=0$。τ是边距超参数。

4. 实验与分析

4.1. 数据集

使用3个不同的数据集来评估所提出的方法在不检索性能。数据集分别为UCMerced Land Use (UCM) [38]、Satellite Remote Sensing Image Database (SATREM) [39]和NWPU-RESISC4 (NWPU) [40]。表1列出了每个数据集的详细信息(图像大小，图像数量等)。实验使用80%的数据集图像用于训练，20%图像用于测试。

Table 1. Details of the datasets
表1. 数据集的细节

数据集	图片数量	类别数	分辨率
UCM	2100	21	256 × 256 pix
SATREM	3000	20	256 × 256 pix
NWPU	31500	45	256 × 256 pix

4.2. 评估标准

本文使用图像检索任务中的常用指标——平均检索精度(*mAP*)来评估检索性能。如果图像与查询图像属于同一类别，则认为该图像与查询图像非常匹配。平均检索精度定义见公式(9)、(10)、(11)、(12)。

$$mAP=\frac{1}{B}\sum_{i=1}^{B}(AP)_i \tag{9}$$

$$AP=\frac{1}{T}\sum_{i=1}^{Q}P_i(rel)_i \tag{10}$$

$$(rel)_i=\begin{cases}1 & \text{第}i\text{个图像与查询图像相似}\\ 0 & \text{其他}\end{cases} \tag{11}$$

$$P_i=\frac{NO_i}{i} \tag{12}$$

其中，B为查询次数，Q为检索结果中最相似的Q幅图像，T为检索N幅图像时真正与须茶图像相似的图像个数，NO_i表示检索结果中真正与待查询图像相似的排序。

4.3. 实验结果

将我们的方法，与ResNet-50、DBOW [9]、D-CNN [41]、V-DELF [29]这4个性能好的基于深度学习的方法作对比，以评估我们方法的检索性能。表2为每个方法在各个数据集上的平均检索精度。我们可以很明显观察到，除了在SATREM数据集上，我们的方法基本比其他方法的平均检索精度都要高。在NWPU数据集上，在其他方法的平均检索精度都有明显的精度下降的现象，但我们的方法的精度下降得不明显。图3是一个定性的检索结果，展示了在NWPU数据集上的一些检索示例。

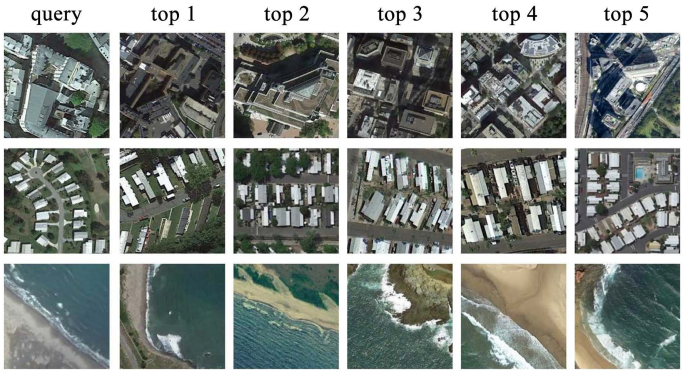

Figure 3. Example queries and retrieved images for the NWPU dataset
图 3. NWPU 数据集的查询示例查询

Table 2. Comparison of average retrieval accuracy of different methods
表 2. 不同方法的平均检索精度对比

	UCM	SATREM	NWPU
ResNet-50	0.82	0.86	0.80
DBOW	0.83	**0.93**	0.82
D-CNN	0.87	0.85	0.74
V-DELF	0.92	0.89	0.86
our method	**0.92**	0.90	**0.89**

同时，我们使用加权梯度类激活映射(Grad-CAM) [42]可视化了模型提取到的特征。图 4 中，可以观察到使用了双重注意力模块的方法提取的特征比原始 ResNet-50 提取的特征，更接近显著区域。这表明双重注意力模块可以充分利用显着区域中的信息并聚合特征。因此，实验结果表明了我们引入的双重注意力模块的有效性。

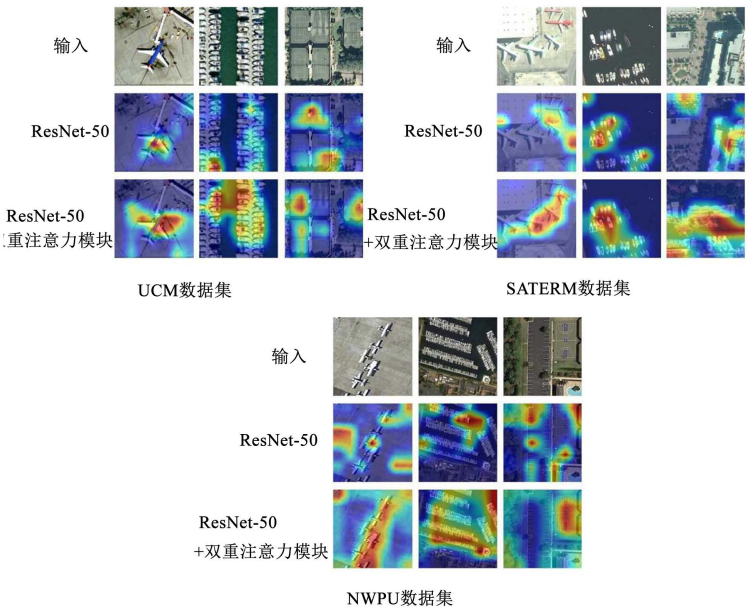

Figure 4. Features extracted from the visualisation model using Grad-CAM in the UCM, SATREM and NWPU datasets
图 4. 使用 Grad-CAM 分别可视化模型在 UCM、SATREM 和 NWPU 数据集中提取的特征

综上所述，我们的方法提高了遥感图像检索的平均检索精度。但是还有存在不足。在 NWPU 数据集上表现出的检索性能不如在其他三个数据集上的检索性能。原因可能是，NWPU 数据集的遥感图像数量多，是其他三个数据集的图片数量的 10 倍，类别也比其他三个数据集的类别多。NWPU 数据集同类别的视觉差异比其他三个数据集的大，而且一些类别之间区分度较小，例如在池塘类中的两幅图像，两者之间的视觉差异性较大，有些与农田类的图像有较大视觉相似性。而其他三个数据集中，同类别的视觉差异较小，不同类别图像区分度较好。

5. 结论

本文引入双重自注意力模块，将空间和通道上的长距离上下文信息编码为局部特征，从而增强特征的表达能力。本文在 3 个典型的数据集上做了实验。实验表明，双重自注意力模块对遥感图像检索性能的提升有显著的作用。

尽管我们提出的方法有更好的性能，但是仍然存在一些不能忽略的缺点。例如，我们的双重自注意力模块只能连接到 CNN 的卷积层。但是，全连接层和卷积层也可以用作特征表示。在某些条件下，某些 CNN 的全连接层比卷积层可以获得更好的检索性能。因此，如何克服使用双重自注意力模块的局限性是我们未来的重点之一。

基金项目

广东省信息物理融合重点实验室(2016B030301008)；国家自然科学基金(61701123)；国家高分地球观测主要项目(83-Y40G33-9001-18/20)；广东省农业科学与技术创新团队项目(2019KJ147)；广东省科技计划项目，水资源大数据项目(2016B010127005)；广东省自然科学基金项目(2018A030313195)；广州市科技 计划项目(201804010262)。

参考文献

[1] Du, P.J., Chen, Y.H., Tang, H. and Fang, T. (2005) Study on Content-Based Remote Sensing Image Retrieval. *IEEE International Geoscience & Remote Sensing Symposium*, Seoul, 29 July 2005, 4. https://doi.org/10.1109/IGARSS.2005.1525204

[2] Ning, X., Li, D. and Ye, W. (2005) Content-Based Remote Sensing Image Retrieval. *Proceedings of SPIE—The International Society for Optical Engineering*, **6044**, 60440Q. https://doi.org/10.1117/12.654549

[3] Sudha, S.K. and Aji, S. (2019) A Review on Recent Advances in Remote Sensing Image Retrieval Techniques. *Journal of the Indian Society of Remote Sensing*, **47**, 2129-2139. https://doi.org/10.1007/s12524-019-01049-8

[4] Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F. and Fraundorfer, F. (2017) Deep Learning in Remote Sensing: A Review. https://arxiv.org/abs/1710.03959v1

[5] Wan, J., Wang, D., Hoi, S.C.H., Wu, P. and Li, J. (2014) Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. *Proceedings of the 22nd ACM International Conference on Multimedia*, ACM, November 2014, 157-166. https://doi.org/10.1145/2647868.2654948

[6] Lowe, D.G. (1999) Object Recognition from Local Scale-Invariant Features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, **2**, 1150-1157. https://doi.org/10.1109/ICCV.1999.790410

[7] Bay, H., Tuytelaars, T. and Van Gool, L. (2006) SURF: Speeded up Robust Features. In: Leonardis, A., Bischof, H. and Pinz, A., Eds., *European Conference on Computer Vision*, Springer, Berlin, Heidelberg, 404-417. https://doi.org/10.1007/11744023_32

[8] Yang, J., Liu, J. and Dai, Q. (2015) An Improved Bag-of-Words Framework for Remote Sensing Image Retrieval in Large-Scale Image Databases. *International Journal of Digital Earth*, **8**, 273-292. https://doi.org/10.1080/17538947.2014.882420

[9] Tang, X., Zhang, X., Liu, F. and Jiao, L. (2018) Unsupervised Deep Feature Learning for Remote Sensing Image Retrieval. *Remote Sensing*, **10**, 1243. https://doi.org/10.3390/rs10081243

[10] Jégou, H., Douze, M., Schmid, C. and Pérez, P. (2010) Aggregating Local Descriptors into a Compact Image Repre-

sentation. 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, 13-18 June 2010, 3304-3311. https://doi.org/10.1109/CVPR.2010.5540039

[11] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, **25**, 1097-1105.

[12] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., *et al.* (2015) Going Deeper with Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[13] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[14] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017) Feature Pyramid Networks for Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[15] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y. and Berg, A.C. (2016) SSD: Single Shot Multibox Detector. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *European Conference on Computer Vision*, Springer, Cham, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[16] Redmon, J. and Farhadi, A. (2017) YOLO9000: Better, Faster, Stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 7263-7271. https://doi.org/10.1109/CVPR.2017.690

[17] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[18] Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A. (2017) Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **31**, 4278-4284.

[19] Ge, Y., Jiang, S., Xu, Q., Jiang, C. and Ye, F. (2018) Exploiting Representations from Pre-Trained Convolutional Neural Networks for High-Resolution Remote Sensing Image Retrieval. *Multimedia Tools and Applications*, **77**, 17489-17515. https://doi.org/10.1007/s11042-017-5314-5

[20] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z. and Lu, H. (2019) Dual Attention Network for Scene Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 3146-3154. https://doi.org/10.1109/CVPR.2019.00326

[21] Mnih, V., Heess, N., Graves, A. and Kavukcuoglu, K. (2014) Recurrent Models of Visual Attention. arXiv preprint arXiv:1406.6247.

[22] Gregor, K., Danihelka, I., Graves, A., Rezende, D. and Wierstra, D. (2015) DRAW: A Recurrent Neural Network for Image Generation. *Proceedings of the* 32*nd International Conference on Machine Learning*, **37**, 1462-1471.

[23] Ba, J., Mnih, V. and Kavukcuoglu, K. (2014) Multiple Object Recognition with Visual Attention. arXiv preprint arXiv:1412.7755.

[24] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. and Tang, X. (2017) Residual Attention Network for Image Classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 3156-3164. https://doi.org/10.1109/CVPR.2017.683

[25] Shao, Z., Yang, K. and Zhou, W. (2018) Performance Evaluation of Single-Label and Multi-Label Remote Sensing Image Retrieval Using a Dense Labeling Dataset. *Remote Sensing*, **10**, 964. https://doi.org/10.3390/rs10060964

[26] Roy, S., Sangineto, E., Demir, B. and Sebe, N. (2020) Metric-Learning-Based Deep Hashing Network for Content-Based Retrieval of Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, **18**, 226-230. https://doi.org/10.1109/LGRS.2020.2974629

[27] Bello, I., Zoph, B., Vaswani, A., Shlens, J. and Le, Q.V. (2019) Attention Augmented Convolutional Networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 3286-3295. https://doi.org/10.1109/ICCV.2019.00338

[28] Xiong, W., Lv, Y., Cui, Y., Zhang, X. and Gu, X. (2019) A Discriminative Feature Learning Approach for Remote Sensing Image Retrieval. *Remote Sensing*, **11**, 281. https://doi.org/10.3390/rs11030281

[29] Imbriaco, R., Sebastian, C., Bondarev, E. and de With, P.H.N. (2019) Aggregated Deep Local Features for Remote Sensing Image Retrieval. *Remote Sensing*, **11**, 493. https://doi.org/10.3390/rs11050493

[30] Yuan, Y. and Wang, J. (2018) Ocnet: Object Context Network for Scene Parsing. arXiv preprint arXiv:1809.00916.

[31] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y. and Liu, W. (2019) CCNet: Criss-Cross Attention for Semantic Segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, 27 October-2 No-

vember 2019, 603-612. https://doi.org/10.1109/ICCV.2019.00069

[32] Wang, X., Girshick, R., Gupta, A. and He, K. (2018) Non-Local Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7794-7803. https://doi.org/10.1109/CVPR.2018.00813

[33] Du, Y., Yuan, C., Li, B., Zhao, L., Li, Y. and Hu, W. (2018) Interaction-Aware Spatio-Temporal Pyramid Attention Networks for Action Classification. *Proceedings of the European Conference on Computer Vision* (*ECCV*), 373-389. https://doi.org/10.1007/978-3-030-01270-0_23

[34] Berman, M., Jégou, H., Vedaldi, A., Kokkinos, I. and Douze, M. (2019) Multigrain: A Unified Image Embedding for Classes and Instances. arXiv preprint arXiv:1902.05509.

[35] Babenko, A. and Lempitsky, V. (2015) Aggregating Deep Convolutional Features for Image Retrieval. arXiv preprint arXiv:1510.07493.

[36] Tolias, G., Sicre, R. and Jégou, H. (2015) Particular Object Retrieval with Integral Max-Pooling of CNN Activations. arXiv preprint arXiv:1511.05879.

[37] Radenović, F., Tolias, G. and Chum, O. (2018) Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**, 1655-1668. https://doi.org/10.1109/TPAMI.2018.2846566

[38] Yang, Y. and Newsam, S. (2010) Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification. *Proceedings of the* 18*th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, November 2010, 270-279. https://doi.org/10.1145/1869790.1869829

[39] Tang, X., Jiao, L., Emery, W.J., Liu, F. and Zhang, D. (2017) Two-Stage Reranking for Remote Sensing Image Retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, **55**, 5798-5817. https://doi.org/10.1109/TGRS.2017.2714676

[40] Zhao, B., Zhong, Y., Xia, G.S. and Zhang, L. (2015) Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, **54**, 2108-2123. https://doi.org/10.1109/TGRS.2015.2496185

[41] Cheng, G., Han, J. and Lu, X. (2017) Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, **105**, 1865-1883. https://doi.org/10.1109/JPROC.2017.2675998

[42] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 618-626. https://doi.org/10.1109/ICCV.2017.74.