Opto-Electronic Engineering
光 电 工 程

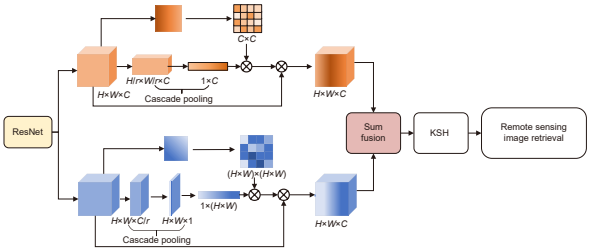
Article
2022 年, 第 49 卷, 第 12 期

DOI: 10.12086/oee.2022.220029

# 面向遥感图像检索的级联池化自注意力研究

吴 刚[1,2]，葛 芸[1,2*]，储 珺[1,2]，叶发茂[3]

[1]南昌航空大学软件学院，江西 南昌 330063；
[2]南昌航空大学江西省图像处理与模式识别重点实验室，江西 南昌 330063；
[3]东华理工大学测绘工程学院，江西 南昌 330013




**摘要：**高分辨率遥感图像检索中，由于图像内容复杂，细节信息丰富，以致通过卷积神经网络提取的特征难以有效表达图像的显著信息。针对该问题，提出一种基于级联池化的自注意力模块，用来提高卷积神经网络的特征表达。首先，设计了级联池化自注意力模块，自注意力在建立语义依赖关系的基础上，可以学习图像关键的显著特征，级联池化是在小区域最大池化的基础上再进行均值池化，将其用于自注意力模块，能够在关注图像显著信息的同时保留图像重要的细节信息，进而增强特征的判别能力。然后，将级联池化自注意力模块嵌入到卷积神经网络中，进行特征的优化和提取。最后，为了进一步提高检索效率，采用监督核哈希对提取的特征进行降维，并将得到的低维哈希码用于遥感图像检索。在 UC Merced、AID 和 NWPU-RESISC45 数据集上的实验结果表明，本文方法能够有效提高检索性能。

**关键词：**遥感图像检索；级联池化；自注意力模块；监督核哈希；卷积神经网络

**中图分类号：**TP391　　**文献标志码：**A

吴刚，葛芸，储珺，等. 面向遥感图像检索的级联池化自注意力研究 [J]. 光电工程，2022，**49**(12): 220029
Wu G, Ge Y, Chu J, et al. Cascade pooling self-attention research for remote sensing image retrieval[J]. *Opto-Electron Eng*, 2022, **49**(12): 220029

# Cascade pooling self-attention research for remote sensing image retrieval

Wu Gang[1,2], Ge Yun[1,2*], Chu Jun[1,2], Ye Famao[3]

[1]School of Software, Nanchang Hangkong University, Nanchang, Jiangxi 330063, China;
[2]Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition, Nanchang Hangkong University, Nanchang, Jiangxi 330063, China;
[3]School of Surveying and Mapping Engineering, East China University of Technology, Nanchang, Jiangxi 330013, China

**Abstract:** In high-resolution remote sensing image retrieval, due to the complex image content and rich detailed information, it is difficult for the features extracted by a convolutional neural network to effectively express the salient information of the image. In response to this issue, a self-attention module based on cascade pooling is proposed to improve the feature representation of convolutional neural networks. Firstly, a cascade pooling self-attention module is designed, and the self-attention module can learn key salient features of images on the basis of

收稿日期：2022-03-30；收到修改稿日期：2022-07-22
基金项目：国家自然科学基金资助项目 (42261070, 41801288, 41261091, 62162045)；江西省自然科学基金资助项目 (20202BAB212011)
*通信作者：葛芸，geyun@nchu.edu.cn。

establishing semantic dependencies. Cascade pooling uses max pooling based on a small region, and then adopts average pooling based on the max pooled feature map. The cascade pooling is exploited in the self-attention module, which can keep important details of the image while paying attention to the salient information of the image, thereby enhancing feature discrimination. After that, the cascade pooled self-attention module is embedded into the convolutional neural network for feature optimization and extraction. Finally, in order to further improve the retrieval efficiency, supervised hashing with kernels is applied to reduce the dimensionality of features, and then the obtained low-dimensional hash code is utilized for remote sensing image retrieval. The experimental results on the UC Merced, AID and NWPU-RESISC45 data sets show that the proposed method can improve the retrieval performance effectively.

**Keywords:** remote sensing image retrieval; cascade pooling; self-attention module; supervised hashing with kernels; convolutional neural network

## 1 引　言

随着遥感卫星技术的发展以及遥感图像应用市场的扩大，基于内容的遥感图像检索在经济社会发展、资源环境监测、城市生活管理等众多领域起到不可替代的作用。遥感图像的背景信息比较复杂，导致同一类图像具有较大的差异，不同类别的图像之间存在一定的相似性，所以提取判别能力强的特征是提高检索性能的关键。

卷积神经网络 (Convolutional neural network，CNN) 采用多层网络架构来学习图像特征，CNN 提取的高层特征能较好地表达图像的语义信息并有效缩小图像检索中的语义鸿沟，提高图像的检索性能[1]。Ge 等人[2] 将 ImageNet 上预训练的 CNN 应用到遥感图像数据集上，表明 CNN 提取的特征明显优于传统的手工特征。

为了进一步提高 CNN 中的特征表达，优化特征和改进网络结构是两种有效的手段。以优化特征为主来提高图像检索性能的方法取得了较好的进展。葛芸等人[3] 通过结合判别相关分析来增强同类特征的联系，并突出不同类别特征之间的差异，再选择串联与相加两种方法来对不同特征进行融合，从而提高特征的判别能力。Hou 等人[4] 提取 Inception V4 网络不同层的特征，将不同层的特征进行融合来代表图像的整体特征，提高图像特征表达的准确性。江曼等人[5] 从多个尺度将图像表面的几何曲率信息融合到改进方向梯度特征中，在此基础上进一步融合图像的颜色信息，将融合特征进行图像检索，提高了检索性能。

改进网络结构来提高遥感图像检索性能也有较多的研究成果，Liu 等人[6] 提出一种分类相似性网络模型，在分类的同时进行相似性预测，结合深度特征和相似性分数来衡量两个图像之间的最终相似度。Zhang 等人[7] 构建了一个结合深度度量学习和非局部操作的三元组非局部神经网络模型，提出了双锚三元组损失函数，以充分利用输入样本的信息。

在改进网络结构的基础上进行特征优化可以进一步改进检索效果。Cheng 等人[8] 提出了一种基于残差注意力的深度度量学习的集成架构，在 CNN 的基础上引入并改进了残差注意力，然后对特征使用池化方法融合，进一步提高检索性能。Zhou 等人[9] 提出了一种以 ResNet 为骨干网络的 Gabor-CA-ResNet 网络，利用 Gabor 表示图像的空间频率结构，结合通道注意力机制来获得判别性更强的深层特征，之后利用基于 Split 的深度特征变换网络对特征进行降维。

虽然 CNN 能够较好地用于图像检索，但是 CNN 是对整张图像进行特征提取，不能有效突出图像中的显著特征，并且背景噪声也会对特征产生一定的干扰。在 CNN 中引入注意力机制可以进一步增强特征的判别能力，注意力机制对图像中不同区域的特征进行权重分配，能够有效地区分前景信息和背景信息，有助于提取图像的显著特征。Hu 等人[10] 提出的 SE 网络模块对特征图进行“挤压”和“激励”操作，通过特征重标定的方式来自适应地调整通道之间的特征响应，从而提高了特征的性能。Woo 等人[11] 设计了卷积注意力模块 (convolutional block attention module，CBAM)，CBAM 在通道注意力的基础上，连接了空间注意力，然后将注意力权重与输入的特征图相乘来进行特征的自适应学习。Wang 等人[12] 提出了一种有效的通道注意力，通过一维卷积来完成跨通道之间的信息交互。Hou 等人[13] 提出了一种高效的注意力机

制，将位置信息嵌入到通道信息中，避免引入较大的开销。

自注意力是一种特殊的注意力机制，为了减少对外部信息的依赖，自注意力根据图像上下文的相关信息来学习显著特征。Wang 等人[14]将自然语言处理领域的自注意力应用到计算机视觉领域，并提出了 Non-Local 模块。Fu 等人[15]提出了双重注意力网络 (dual attention network，DANet)，将 Non-Local 的思想应用在空间域和通道域，分别将特征空间以及特征通道作为查询对象进行上下文建模。Huang 等人[16]利用两个十字交叉注意力来替代基于全局像素点的建模，降低了运算复杂度。

在遥感图像检索领域，注意力机制同样备受关注。Wang 等人[17]在 CNN 中引入双线性池化，并且通过通道注意力和空间注意力来细化特征，将注意力机制输出特征作为双线性池化的输入，最后使用主成分分析 (principal component analysis，PCA)[18]对特征进行降维，实验结果表明该方法的检索结果较好。Yang 等人[19]在深度哈希的基础上引入通道注意力和位置注意力，以提高特征的表达能力，取得良好的检索效果。

受自注意力机制的启发，本文提出一种基于级联池化的自注意力模块 (cascade pooling self-attention module，CPSM) 与哈希相结合的遥感图像检索方法，从改进网络结构和优化 CNN 特征两个方面来提高检索性能。基于级联池化的自注意力模块从通道和空间两个方面减少图像背景噪声的干扰，监督核哈希 (supervised Hashing with kernels，KSH)[20]将特征映射成紧凑的哈希码，从而有效地降低特征维数。本文的主要贡献如下：

1) 改进了通道自注意力和空间自注意力。通道自注意力通过关联所有通道图之间的特征信息，学习具有内容相关性的显著通道特征，空间自注意力通过所有空间位置的特征加权提取具有位置相关性的显著空间特征。将这两者结合，从通道域和空间域两个方面进行全局建模，减少图像背景噪声的干扰。

2) 提出在自注意力模块中使用级联池化[21]来代替全局池化。级联池化结合最大池化和均值池化的优点，在提取图像显著特征的同时保留了图像重要的细节信息，适用于内容复杂的高分辨率遥感图像。

3) 采用 KSH 来进一步优化特征。在 CNN 中加上 CPSM 模块有助于学习到判别能力更强的特征，但 CNN 中提取的特征一般维数较高，因此使用 KSH 将特征映射成紧凑的哈希码，在降低特征冗余性的同时提高特征的检索效率。

## 2 级联池化自注意力的遥感图像检索

### 2.1 检索流程

传统的 CNN 容易存在梯度弥散和梯度消失等问题，ResNet50[22]主要由深度残差结构组成，残差结构使得网络层次更深、收敛速度更快，较好地解决了梯度弥散和梯度消失等问题。本文在 ResNet50 网络中引入级联池化自注意力模块，提出 ResNet50-CPSM 网络，并将网络中的特征用于遥感图像检索。检索流程如图 1 所示，具体检索步骤如下：首先，在大规模数据集 ImageNet 预训练网络的基础上，用遥感数据集训练 ResNet50-CPSM 网络，分别提取训练集和测试集的高层特征。然后，对训练集的特征用 KSH 方法进行监督学习，并根据学习的参数将测试集的特征映射为紧凑的哈希码。最后，用该哈希码用于检索遥感图像，采用汉明距离计算查询图像与数据集中图像的相似度，返回最相似的若干幅图像作为检索结果。

### 2.2 级联池化自注意力网络结构

遥感图像内容复杂，背景信息丰富，空间语义信息也丰富。为了突出图像的显著信息，提出了级联池化自注意力模块。为了充分利用 ResNet50 预训练的参数，将级联池化自注意力模块加载到 ResNet50 网络最后一层卷积层的后面。级联池化自注意力中采用了级联池化来代替传统的全局池化，级联池化首先对特征图进行小区域的最大池化，得到最大池化后的特征图，再对该特征图进行平均池化。与传统的全局池化相比，级联池化结合了最大池化和均值池化的优点，既关注了遥感图像的显著信息，又保留了重要的细节信息。级联池化自注意力包含通道自注意力和空间自注意力，自注意力可以自适应地关联远程上下文信息，更关注特征之间的相关性，通过对特征进行权重分配进一步学习遥感图像的显著特征，从而提高特征的判别性。

#### 2.2.1 通道自注意力

通道自注意力通过联系上下文信息对不同的通道进行权重分配，每一个通道可以看作是一类特征的响应，对贡献大的特征分配更大的权重，从而增强对显著特征的判别能力。通道自注意力模块如图 2 所示，

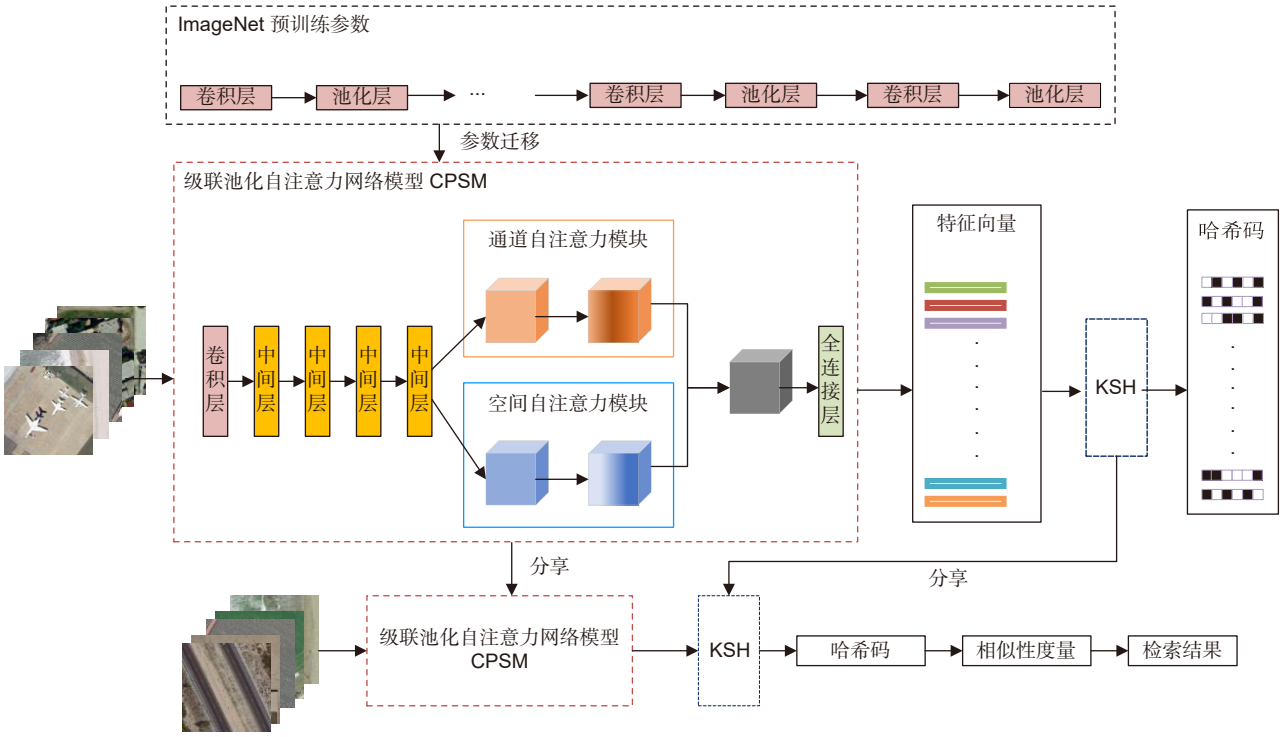



图 1 级联池化自注意力的检索流程图

Fig. 1 Retrieval flowchart for cascade pooling self-attention

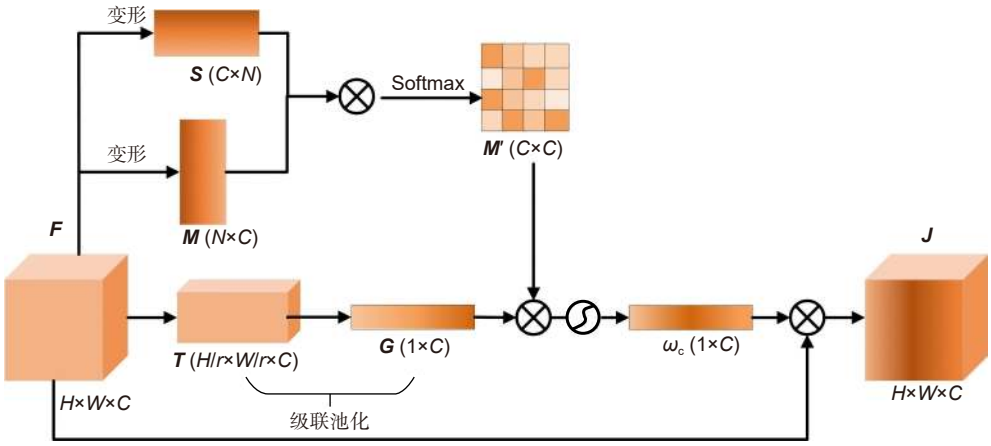



图 2 通道自注意力模块

Fig. 2 Channel self-attention module

令卷积特征 $\boldsymbol{F}$ 的尺寸为 $H\times W\times C$，$H$ 和 $W$ 是每个通道对应特征图的高度和宽度，$C$ 为是通道的数目。传统的通道注意力模块忽略了同一个通道不同位置的相互关系，为了得到不同通道间的相互关系，将 $\boldsymbol{F}$ 通过变形转换变为两个二维矩阵 $\boldsymbol{S}$ 和 $\boldsymbol{M}$，$\boldsymbol{S}$ 的维数是 $C\times N$，$\boldsymbol{M}$ 的维度是 $N\times C$，其中 $N=H\times W$。然后，矩阵 $\boldsymbol{S}$ 与 $\boldsymbol{M}$ 进行矩阵乘法，得到通道自注意力矩阵 $\boldsymbol{M}'$，通道自注意力矩阵 $\boldsymbol{M}'$中的元素 $m'_{ij}$ 的计算如式 (1) 所示：

$$m'_{ij}=\frac{\exp(M_i\cdot M_j)}{\sum_{i=1}^{C}\exp(M_i\cdot M_j)}. \tag{1}$$

$m'_{ij}$ 表达了通道 $i$ 和通道 $j$ 之间细化的通道关系，该值越大，说明两个通道之间的联系越紧密。通过计算同一个通道不同位置的相互关系，能够有效获得通道的上下文语义信息。卷积特征 $\boldsymbol{F}$ 采用级联池化能够进一步提高特征的判别能力。级联池化的思想是首先对特征图采用重叠的小区域最大池化，获得多个小区

域的显著特征图，将其构成的显著特征标记为 $\boldsymbol{T}$，$\boldsymbol{T}$ 的维数是$H/r\times W/r\times C$，$r$ 为特征图缩小的倍数。对该显著特征再均值池化，得到输出通道信息，将其构成的矩阵标记为 $\boldsymbol{G}$，维数是 $1\times C$。相关矩阵 $\boldsymbol{M}'$的维数为 $C\times C$，将 $\boldsymbol{M}'$和 $\boldsymbol{G}$ 相乘，得到融入了上下文依赖关系的显著通道信息，经过激活函数映射输出通道权重 $\omega_{\mathrm{c}}$，$\omega_{\mathrm{c}}$的计算如式 (2) 所示，$g_i$为矩阵 $\boldsymbol{G}$ 的元素：

$$\omega_{\mathrm{c}}=\sigma\left(\sum\nolimits_{i=1}^{C} g_i\cdot m'_{ij}\right). \tag{2}$$

$\omega_{\mathrm{c}}$再与特征图 $\boldsymbol{F}$ 进行矩阵乘法运算，得到融合了通道信息和相关信息的特征图 $\boldsymbol{J}$，$\boldsymbol{J}$ 的计算如式 (3) 所示：

$$\boldsymbol{J}=\omega_{\mathrm{c}}\otimes\boldsymbol{F}. \tag{3}$$

### 2.2.2 空间自注意力

高层特征的位置可以看作是图像上相同区域对不同卷积核的响应，因此提出了一个空间自注意力模块，加强不同区域之间的关系。空间自注意力模块通过空间权重来增强感兴趣的特定目标区域，并弱化不相关的背景区域，从而改进特征描述能力。空间自注意力模块如图 3 所示。

根据 CNN 局部感受野的特点，特征图上的值反映了 $H\times W$ 个局部小块位置的信息，通过计算位置之间的相关性来反映不同位置之间的关系。为了更直观地获得位置信息，将三维张量 $\boldsymbol{F}$ 通过变形和转置转换为两个二维矩阵 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$，$\boldsymbol{Q}$ 的维数是 $N\times C$，$\boldsymbol{K}$ 的维数是 $C\times N$，其中 $N=H\times W$。然后，求 $\boldsymbol{S}$ 的空间自注意力矩阵 $\boldsymbol{S}'$，相关矩阵中的元素 $s'_{ij}$ 的计算如式 (4) 所示：

$$s'_{ij}=\frac{\exp(S_i\cdot S_j)}{\sum_{i=1}^{N}\exp(S_i\cdot S_j)}. \tag{4}$$

$s'_{ij}$ 反映了位置 $i$ 和位置 $j$ 之间细化的位置关系，该值越大，说明两个位置之间越相关。通过计算同一个位置不同通道的相互关系，能够有效获得距离较远位置的依赖关系。此外，为了得到 $\boldsymbol{F}$ 的显著空间信息，对 $\boldsymbol{F}$ 对应的空间向量进行级联池化，先对 $\boldsymbol{F}$ 的通道维度进行重叠区域的最大池化，获得多个小区域的显著特征图，将其构成的显著特征标记为 $\boldsymbol{B}$，$\boldsymbol{B}$ 的维数是 $H\times W\times C/r$，$r$ 为特征图缩小的倍数。对这些显著特征再均值池化，得到输出空间信息，求得显著空间信息构成的矩阵为 $\boldsymbol{P}$，其维度为 $H\times W$，将其变形为维度为 $1\times N$ 的矩阵 $\boldsymbol{E}$。将 $\boldsymbol{S}'$和 $\boldsymbol{E}$ 这两个矩阵相乘，得到融入了位置相关性的显著空间信息，经过激活函数 $\sigma$ 映射得到空间权重$\omega_{\mathrm{s}}$，$\omega_{\mathrm{s}}$的计算如式 (5) 所示，$e_i$ 为矩阵 $\boldsymbol{E}$ 的元素：

$$\omega_{\mathrm{s}}=\sigma\left(\sum\nolimits_{i=1}^{N} e_i\cdot s'_{ij}\right). \tag{5}$$

空间权重$\omega_{\mathrm{s}}$再与特征图 $\boldsymbol{F}$ 进行矩阵乘法运算，得到融合了空间信息和相关信息的特征图 $\boldsymbol{Z}$，$\boldsymbol{Z}$ 的计算如式 (6) 所示：

$$\boldsymbol{Z}=\omega_{\mathrm{s}}\otimes\boldsymbol{F}. \tag{6}$$

## 2.3 特征优化

在 ResNet50 中添加基于级联池化的自注意力模块提取判别能力强的特征，但是特征维数过高，依然存在冗余，因此需要对特征进行降维。KSH 是一种监督核哈希方法，其目标是将数据映射为紧凑的二进制哈希码。该方法在避免特征高维度的同时，提高类

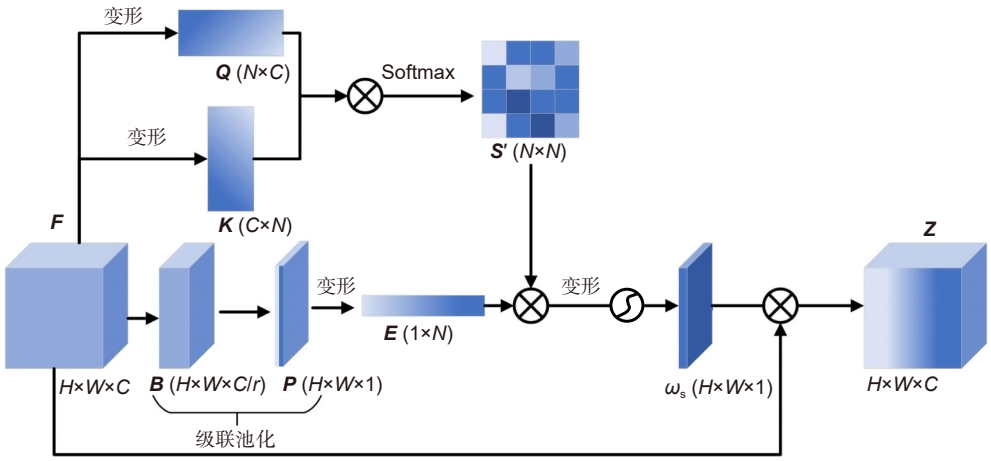



图 3 空间自注意力模块

Fig. 3 Spatial self-attention module

内的相似性，并扩大类间的差异性，在计算相似度学习的任务中，可以有效地提高特征的判别能力。

### 2.4 时间效率分析

本文提出的方法主要由 ResNet50、基于级联池化的自注意力、KSH 算法组成，总体复杂度为三者复杂度之和。ResNet50 的时间复杂度为 $O\left(n\sum_{l=1}^{D}P^2K^2C_{l-1}C_l\right)$[23]，其中，$n$ 为训练样本数，$D$ 为网络的卷积层数，$l$ 为第 $l$ 个卷积层，$P$ 为卷积核输出特征图的边长，$K$ 为卷积核的边长，$C_l$ 为网络第 $l$ 层的输出通道数，$C_{l-1}$ 为网络第 ($l$-1) 层的输出通道数。通道自注意力的时间复杂度为 $O(nC^2HW)$，空间自注意力矩阵的时间复杂度为 $O(nCH^2W^2)$。KSH 训练的时间复杂度为 $O((nm+p^2m+m^2p+m^3)b)$[20]，其中 $m$ 为随机均匀选取样本数，用于得到基于核的哈希函数，$p$ 为监督学习样本数，$b$ 为哈希码位数，其中 $n>>p>m$，注意力模块只加在最后一个池化层之前，因此整体的模块参数增加量较少。

## 3 实验及结果分析

为了验证本文方法的有效性，在不同分辨率、不同尺寸的高分辨遥感数据集上进行实验，与其他注意力机制方法和遥感图像检索方法进行对比与分析。实验框架为 Pytorch1.6.0，GPU 为 RTX2060s，训练时每个类别随机抽取 80% 的图像作为训练集，剩下 20% 的图像作为测试集，采用的优化器为随机梯度下降，学习率为 0.0035，动量为 0.9，学习率衰减系数为 0.1。

### 3.1 数据集以及评价指标介绍

本文选用 UC Merced 数据集[24]、AID 数据集[25]和 NWPU-RESISC45 数据集[26]进行实验。图 4 中显示了 UC Merced 数据集、AID 数据集与 NWPU-RESISC45 数据集的部分示例图像。UC Merced 数据集包含 21 个类别，每个类别 100 张图像，每张图像的大小是 256 pixels×256 pixels，图像的分辨率为 0.3 m。AID 数据集包含 30 个类别，每个类别图像数量从 220 ~ 400 张不等，一共有 10000 张，每张图像的大小是 600 pixels×600 pixels，图像的分辨率介于 0.5 m~8.0 m 之间。NWPU-RESISC45 数据集包含 45 个类别，每个类别 700 张图像，每张图像的大小是 256 pixels×256 pixels，图像的分辨率介于 0.2 m ~ 30 m 之间。采用平均精度均值 (mean average precision, mAP，在式用 $A$ 表示) 和 P@$k$ 对检索结果进行评价。mAP


图 4 示例图像。(a) UC Merced 数据集示例图像；(b) AID 数据集示例图像；(c) NWPU-RESISC45 数据集示例图像

Fig. 4 Sample images. (a) Sample images of the UC Merced data set; (b) Sample images of the AID data set; (c) Sample images of the NWPU-RESISC45 data set

是图像检索算法的主要评估指标，是所有查询图像平均精度的平均值，图像与查询图像的相关性越高，它的排名就越高。P@$k$ 则关注前 $k$ 幅检索结果中相关图像的数量 ($N_R$)。具体计算结果如式 (7)，式 (8) 和式 (9) 所示。

$$\mathrm{P@}k = \frac{N_R}{k}, \tag{7}$$

$$\overline{P(k)} = \frac{\sum_{k=1}^{N} P(k)\times rel(k)}{N}, \tag{8}$$

$$A = \frac{\sum_{q=1}^{Q} \overline{P(q)}}{Q}, \tag{9}$$

其中：$Q$ 为查询图像的总数，$P(k)$为当前检索出来 $k$ 个结果的准确率，$rel(k)$表示第 $k$ 个检索结果是否与查询图像有关，1 表示有关，0 表示无关。$N$ 表示检索出来的图像数量。

### 3.2 不同注意力方法的比较

为了验证自注意力模块的有效性，在不同数据集上与 CBAM 和 DANet 注意力进行比较。表 1 和图 5 分别显示了不同注意力模块的 mAP 值和 P@$k$ 值，ResNet50 为基础网络模型，ResNet50_CBAM 为基于 CBAM 的 ResNet50 网络，ResNet50_DANet 为基于 DANet 的 ResNet50 网络，ResNet50_CSM (ResNet50_channel self-attention module) 为基于通道自注意力模块的 ResNet50 网络，ResNet50_SSM (ResNet50_spatial self-attention module) 为基于空间自注意力模块的 ResNet50 网络，ResNet50_DSM (ResNet50_dual self-attention module) 为基于双重自注意力模块的 ResNet50 网络，自注意力模块中均使用的是全局平均池化。提取的特征为网络中最后一个卷积层的特征，为了更直观地对比不同注意力模块之间的性能，对该特征没有进行降维。

由表 1 可以发现，在 UC Merced 数据集中，ResNet50_DSM 使 mAP 值从 ResNet50 的 91.17% 提升到 92.67%，提升了 1.5%；在 AID 数据集中，ResNet50_DSM 使 mAP 值从 ResNet50 的 87.35% 提升到 93.48%，提升了 6.13%；在 NWPU-RESISC45 数据集中，ResNet50_DSM 使 mAP 值从 ResNet50 的 60.07% 提升到 78.28%，提升了 18.21%。由图 5 可知，在不同的数据集中，ResNet50_DSM 的 P@$k$ 值均高于 ResNet50_CSM 和 ResNet50_SSM，提升效果比单个自注意力模块要好。

从不同的实验结果可以看出，注意力模块对特征的表达有明显的提升，并且数据集越复杂提升效果越明显。多个注意力模块结合效果比单一的注意力模块效果更好，这是因为通道域和空间域的注意力模块通过不同的域对特征进行加权，有效降低背景信息的干扰。ResNet50_DSM 与其他注意力模块相比较，同样

表 1　不同注意力模块的 mAP 值

Table 1　mAP value of different attention modules

| 网络结构 | UC Merced/% | AID/% | NWPU-RESISC45/% |
|---|---|---|---|
| ResNet50 | 91.17 | 87.35 | 60.07 |
| ResNet50_DANet | 92.13 | 88.13 | 73.02 |
| ResNet50_CSM | 91.50 | 92.38 | 77.12 |
| ResNet50_SSM | 92.60 | 92.91 | 77.13 |
| ResNet50_CBAM | **93.72** | 93.33 | 78.19 |
| ResNet50_DSM | 92.67 | **93.48** | **78.28** |

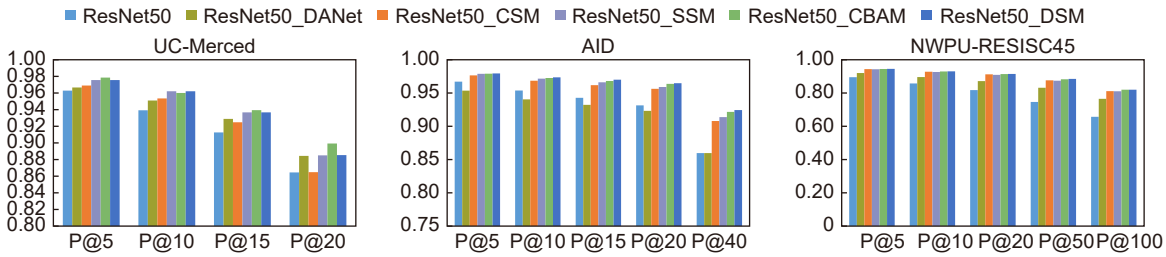



图 5　不同注意力模块的 P@$k$ 值

Fig. 5　P@$k$ value of different attention moduless

存在优势。由表 1 可以看到，在 AID 和 NWPU-RESISC45 数据集中，ResNet50_DSM 的 mAP 值均大于 ResNet50_DANet 和 ResNet50_CBAM 的 mAP 值；在 UC Merced 数据集中，ResNet50_DSM 的 mAP 值大于 ResNet50_DANet，但是小于 ResNet50_CBAM 的 mAP 值。图 5 中的 P@$k$ 值比较出现了与 mAP 值类似的趋势，这是因为自注意力模块不仅可以在通道和空间上进行特征响应，而且还考虑同一个通道之间的位置关系，以及同一个位置的通道关系，而且自注意力模块更擅长捕捉特征的内部相关性，当特征量偏少的时候，捕捉特征内部相关性的能力会有所减弱；当特征量足够的时候，捕捉特征内部相关性的能力就可以充分体现。相比于 CBAM，DSM 关注特征上下文依赖关系，细节信息更丰富；相比于 DANet，DSM 在细化特征关注度的基础上，减少了参数的冗余，从而提高了模型的泛化能力。

### 3.3 不同池化方法的比较

为了验证级联池化和全局均值池化两种方法对自注意力模块的影响，在不同的数据集上使用不同的池化方法，mAP 值结果如表 2 所示。ResNet50_CPCSM (ResNet50_cascade pooling channel self-attention module) 为基于级联池化通道自注意力模块的 ResNet50 网络，ResNet50_CPSSM (ResNet50_cascade pooling spatial self-attention module) 为基于级联池化空间自注意力模块的 ResNet50 网络，ResNet50_CPSM (ResNet50_cascade pooling self-attention module) 为基于级联池化自注意力模块的 ResNet50 网络。

由表 2 可知，在不同的数据集中，ResNet50_CPCSM 的 mAP 值高于 ResNet50_CSM；ResNet50_CPSSM 的 mAP 值高于 ResNet50_SSM；ResNet50_CPSM 的值高于 ResNet50_DSM。因此，在不同的注意力中，级联池化的结果优于全局均值池化。图 6 显示了不同池化方法的 P@$k$ 值。从图 6 可以看到，在不同数据集中，$k$ 取不同值的时候，ResNet50_CPSM 的 P@$k$ 值都明显高于 ResNet50_DSM 的 P@$k$。由不同的实验结果可以得到，在其他条件相同的情况下，采用级联池化方式的自注意力模块检索效果优于全局均值池化方式的自注意力模块的检索效果。与普通的光学图像不同，遥感图像主要是自然地理场景信息，内容丰富，信息复杂，而且遥感图像因为拍摄角度和位置的原因，图像上的很多关键信息尺寸不会太大，普通全局池化均不能较好地对特征进行采样。全局最大池化可以提取特征图中关键特征，但是忽略了一些重要的细节信息，而且容易受噪声干扰，全局平均池化可以综合所有特征，但是不能有效提取显著特征。级联池化是对特征图进行重叠区域的最大池化，提取

表 2　不同池化方法的 mAP 值

Table 2　mAP value of different pooling methods

| 网络结构 | UC Merced/% | AID/% | NWPU-RESISC45/% |
|---|---|---|---|
| ResNet50 | 91.17 | 87.35 | 60.07 |
| ResNet50_CSM | 91.50 | 92.38 | 77.12 |
| ResNet50_SSM | 92.60 | 92.91 | 77.13 |
| ResNet50_DSM | 92.67 | 93.48 | 78.28 |
| ResNet50_CPCSM | 93.35 | 93.23 | 78.17 |
| ResNet50_CPSSM | 93.62 | 93.15 | 77.19 |
| ResNet50_CPSM | **94.12** | **93.76** | **79.00** |

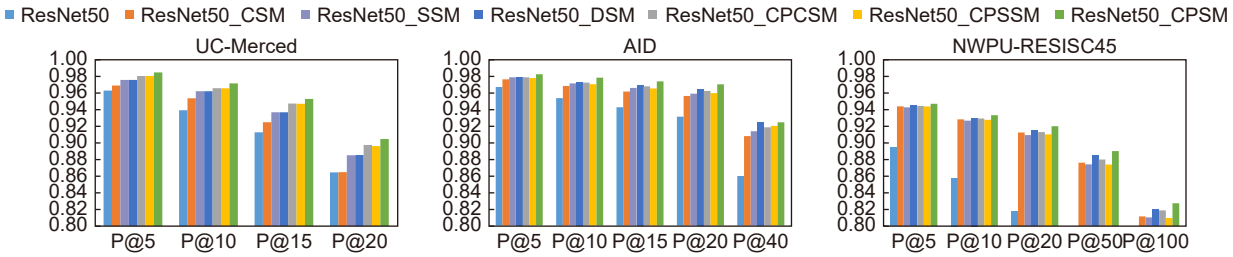



图 6　不同池化方法的 P@$k$ 值

Fig. 6　P@$k$ value of different pooling methods

重叠区域显著特征，然后对显著特征进行平均池化。因此级联池化综合了最大池化和平均池化的优点，有效地减少遥感图像特征的冗余信息，同时也能保留一些区分度明显的特征信息。综合表 1 和表 2 可知，ResNet50_DSM 的性能明显优于 ResNet50_DANet，尤其是在较大的数据集 AID 和 NWPU-RESISC45 中，ResNet50_DSM 的 mAP 值也略高于 ResNet50_CBAM。引入级联池化的 ResNet50_CPSM，其检索性能相比 ResNet50_DSM 得到进一步提升，ResNet50_CPSM 的 mAP 值在三个数据集上均超过了 ResNet50_DANet 和 ResNet50_CBAM，因此结合级联池化的自注意力模块检索结果最好。

### 3.4 不同降维方法的比较

CNN 特征维数较高，存在一定的冗余信息，特征降维能够进一步优化 CNN 特征。

为了验证不同的降维方法的效果和性能，使用 PCA、线性判别分析 (linear discriminant analysis, LDA)[27] 和 KSH 进行对比，在 UC Merced 数据集上使用级联池化自注意力网络提取特征，并对该特征采用不同方法进行降维，得出不同方法的 mAP 值和每幅图片平均检索时间，具体结果如表 3 所示。LDA 方法降维的最大维度数为类别数减一，因此维度只能降到 20。由表 3 可知，降维到 20 维度的时候，LDA 方法比 KSH 的 mAP 值要高 0.64%，比 PCA 的 mAP 值要高 2.22%，但是平均检索时间明显慢于 PCA 和 KSH；降维到 64 维度的时候，KSH 方法比 PCA 的 mAP 值高 4.67%，而且平均检索时间要快于 PCA。综合 mAP 值和平均检索时间，KSH 方法的检索性能较好。

### 3.5 与其他方法的比较

为了评估本文方法的泛化能力，在不同的数据集上进行实验，并与其他图像检索方法进行比较。表 4、表 5、表 6 分别是 UC Merced 数据集、AID 数据集、NWPU-RESISC45 数据集的实验结果，其中 ResNet50_CPSM_KSH 是指对 ResNet50_CPSM 提取的特征使用 KSH 降维。

在 UC Merced 数据集中，本文方法的 mAP 值相比于大多数其他方法的 mAP 值高，比 FAH (feature

表 3 不同降维方法的比较

Table 3 Comparison of different dimensionality reduction methods

| 方法 | mAP/% | 时间/ms | 维度 |
|---|---|---|---|
| - | 94.12 | 24.06 | 100352 |
| PCA | 95.85 | 23.11 | 20 |
| PCA | 93.56 | 23.46 | 64 |
| LDA | 98.07 | 29.67 | 20 |
| KSH | 97.43 | 23.32 | 20 |
| KSH | **98.23** | 23.44 | 64 |

表 4 UC Merced 数据集中不同方法的比较

Table 4 Comparison of different methods on UC Merced data set

| 参考文献 | 方法 | mAP/% |
|---|---|---|
| Ye等[28] | Pool5_W | 94.86 |
| Ye等[29] | Pool5 | 95.62 |
| 叶发茂等[30] | Pool5_ACO | 96.71 |
| Roy等[31] | MiLaN | 91.60 |
| Wang等[17] | W-CAN | 95.10 |
| Zhang等[7] | - | 97.89 |
| Song等[32] | DHCNN | 98.02 |
| Liu等[33] | FAH | **98.30** |
| Cheng等[8] | - | 97.77 |
| Zhou等[9] | Gabor-CA-ResNet | 97.50 |
| 本文 | ResNet50_CPSM_KSH | 98.23 |

表 5 AID 数据集中不同方法的比较

Table 5 Comparison of different methods on AID data set

| 参考文献 | 方法 | mAP/% |
|---|---|---|
| Liu等[33] | FAH | 91.75 |
| Roy等[31] | MiLaN | 92.60 |
| Song等[32] | DHCNN | 94.27 |
| Cheng等[8] | - | 93.84 |
| Zhou等[9] | Gabor-CA-ResNet | 94.34 |
| 本文 | ResNet50_CPSM_KSH | **94.96** |

表 6 NWPU-RESISC45 数据集不同方法比较

Table 6 Comparison of different methods on NWPU-RESISC45 data set

| 参考文献 | 方法 | mAP/% |
|---|---|---|
| Tang等[34] | RFM | 25.63 |
| Yang等[24] | SBOW | 37.02 |
| Demir等[35] | Hash | 34.49 |
| Marmanis等[36] | DN7 | 60.54 |
| Marmanis等[36] | DN8 | 59.47 |
| Demir等[34] | DBOW | 82.15 |
| Imbriaco等[37] | QE-S | 85.70 |
| Imbriaco等[37] | V-DELF | 84.00 |
| Liu等[33] | FAH | 70.41 |
| Hou等[38] | - | 83.07 |
| Wang等[39] | JSST | 80.39 |
| Fan等[40] | N-pair-mc Loss | 93.06 |
| Fan等[40] | Triplet Loss | 93.82 |
| 本文 | ResNet50_CPSM_KSH | **94.53** |

and hash)[33]方法仅低 0.07%，因此在该小规模数据集上的检索性能较好。在 NWPU-RESISC45 数据集和 AID 数据集中，本文方法都有效提高了遥感图像检索的准确率，特别是，本文方法的 mAP 值比 FAH 方法分别提高了 3.21% 和 24.12%。

因此，基于级联池化自注意力的方法在多种不同规模的数据集中均能取得较优的检索结果，尤其是在图像类别数目较多的大规模数据集中，检索准确率的提升效果更为明显。

## 4 结 论

本文针对遥感图像同一类图像具有较大的差异，不同类别的图像之间存在一定相似性的问题，提出一种基于级联池化自注意力的遥感图像检索方法。该方法通过级联池化和自注意力来优化网络结构，级联池化综合了最大池化和平均池化的优点，自注意力能自适应地联系上下文信息对显著特征图加权，使得学习到的特征具有更强的判别性。CNN 特征的维数较高，因此还采用 KSH 对特征进行降维，进一步提高特征的性能。实验结果表明，通过级联池化自注意力网络提取的遥感图像特征判别能力强，在不同的数据集上都具有较高的检索精度。与其他的遥感图像检索方法相比，本文方法表现出较优的检索性能。

## 参考文献

[1] Husain S S, Bober M. REMAP: Multi-layer entropy-guided pooling of dense CNN features for image retrieval[J]. *IEEE Trans Image Process*, 2019, **28**(10): 5201−5213.

[2] Ge Y, Jiang S L, Xu Q Y, et al. Exploiting representations from pre-trained convolutional neural networks for high-resolution remote sensing image retrieval[J]. *Multimed Tools Appl*, 2018, **77**(13): 17489−17515.

[3] Ge Y, Ma L, Chu J. Remote sensing image retrieval combining discriminant correlation analysis and feature fusion[J]. *J Image Graphics*, 2020, **25**(12): 2665−2676.

葛芸, 马琳, 储珺. 结合判别相关分析与特征融合的遥感图像检索[J]. 中国图象图形学报, 2020, **25**(12): 2665−2676.

[4] Hou F, Liu B, Zhuo L, et al. Remote sensing image retrieval with deep features encoding of Inception V4 and largevis dimensionality reduction[J]. *Sens Imaging*, 2021, **22**(1): 20.

[5] Jiang M, Zhang H X, Cheng D Q, et al. Multi-scale image retrieval based on HSV and directional gradient features[J]. *Opto-Electron Eng*, 2021, **48**(11): 210310.
江曼, 张皓翔, 程德强, 等. 融合HSV与方向梯度特征的多尺度图像检索[J]. 光电工程, 2021, **48**(11): 210310.

[6] Liu Y S, Chen C H, Han Z Z, et al. High-resolution remote sensing image retrieval based on classification-similarity networks and double fusion[J]. *IEEE J Sel Top Appl Earth Obs Remote Sens*, 2020, **13**: 1119−1133.

[7] Zhang M D, Cheng Q M, Luo F, et al. A triplet nonlocal neural network with dual-anchor triplet loss for high-resolution remote sensing image retrieval[J]. *IEEE J Sel Top Appl Earth Obs Remote Sens*, 2021, **14**: 2711−2723.

[8] Cheng Q M, Gan D Q, Fu P, et al. A novel ensemble architecture of residual attention-based deep metric learning for remote sensing image retrieval[J]. *Remote Sens*, 2021, **13**(17): 3445.

[9] Zhuo Z, Zhou Z. Remote sensing image retrieval with Gabor-CA-ResNet and split-based deep feature transform network[J]. *Remote Sens*, 2021, **13**(5): 869.

[10] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7132−7141. doi: 10.1109/CVPR.2018.00745.

[11] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 3–19. doi: 10.1007/978-3-030-01234-2_1.

[12] Wang Q L, Wu B G, Zhu P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 11531–11539. doi: 10.1109/CVPR42600.2020.01155.

[13] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 13708–13717. doi: 10.1109/CVPR46437.2021.01350.

[14] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7794–7803. doi: 10.1109/CVPR.2018.00813.

[15] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 3141–3149. doi: 10.1109/CVPR.2019.00326.

[16] Huang Z L, Wang X G, Huang L C, et al. CCNet: Criss-cross attention for semantic segmentation[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 603–612. doi: 10.1109/ICCV.2019.00069.

[17] Wang Y M, Ji S P, Lu M, et al. Attention boosted bilinear pooling for remote sensing image retrieval[J]. *Int J Remote Sens*, 2020, **41**(7): 2704−2724.

[18] Wold S, Esbensen K, Geladi P. Principal component analysis[J]. *Chemom Intell Lab Syst*, 1987, **2**(1–3): 37−52.

[19] Yang W J, Wang L J, Cheng S L, et al. Deep hash with improved dual attention for image retrieval[J]. *Information*, 2021, **12**(7): 285.

[20] Liu W, Wang J, Ji R R, et al. Supervised hashing with kernels[C]//*IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 2074 –2081. doi: 10.1109/CVPR.2012.6247912.

[21] Ge Y, Tang Y L, Jiang S L, et al. Region-based cascade pooling of convolutional features for HRRS image retrieval[J]. *Remote Sens Lett*, 2018, **9**(10): 1002−1010.

[22] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[23] Sun Y C, Li G. Abnormal behavior detection of crowds based on nested model of convolutional neural network[J]. *Comput Appl Software*, 2019, **36**(3): 196−201, 276.
孙月驰, 李冠. 基于卷积神经网络嵌套模型的人群异常行为检测[J]. 计算机应用与软件, 2019, **36**(3): 196−201, 276.

[24] Yang Y, Newsam S. Geographic image retrieval using local invariant features[J]. *IEEE Trans Geosci Remote Sens*, 2013, **51**(2): 818−832.

[25] Xia G S, Hu J W, Hu F, et al. AID: A benchmark data set for performance evaluation of aerial scene classification[J]. *IEEE Trans Geosci Remote Sens*, 2017, **55**(7): 3965−3981.

[26] Cheng G, Han J W, Lu X Q. Remote sensing image scene classification: benchmark and state of the art[J]. *Proc IEEE*, 2017, **105**(10): 1865−1883.

[27] Izenman A J. Linear discriminant analysis[M]//Izenman A J. *Modern Multivariate Statistical Techniques*. New York: Springer, 2013: 237–280. doi: 10.1007/978-0-387-78189-1_8.

[28] Ye F M, Xiao H, Zhao X Q, et al. Remote sensing image retrieval using convolutional neural network features and weighted distance[J]. *IEEE Geosci Remote Sens Lett*, 2018, **15**(10): 1535−1539.

[29] Ye F M, Dong M, Luo W, et al. A new re-ranking method based on convolutional neural network and two image-to-class distances for remote sensing image retrieval[J]. *IEEE Access*, 2019, **7**: 141498−141507.

[30] Ye F M, Meng X L, Dong M, et al. Remote sensing image retrieval with ant colony optimization and a weighted image-to-class distance[J]. *Acta Geod Cartogr Sin*, 2021, **50**(5): 612−620.
叶发茂, 孟祥龙, 董萌, 等. 遥感图像蚁群算法和加权图像到类距离检索法[J]. 测绘学报, 2021, **50**(5): 612−620.

[31] Roy S, Sangineto E, Demir B, et al. Metric-learning-based deep hashing network for content-based retrieval of remote sensing images[J]. *IEEE Geosci Remote Sens Lett*, 2021, **18**(2): 226−230.

[32] Song W W, Li S T, Benediktsson J A. Deep hashing learning for visual and semantic retrieval of remote sensing images[J]. *IEEE Trans Geosci Remote Sens*, 2021, **59**(11): 9661−9672.

[33] Liu C, Ma J J, Tang X, et al. Deep hash learning for remote sensing image retrieval[J]. *IEEE Trans Geosci Remote Sens*, 2021, **59**(4): 3420−3443.

[34] Tang X, Jiao L C, Emery W J. SAR image content retrieval based on fuzzy similarity and relevance feedback[J]. *IEEE J Sel Top Appl Earth Obs Remote Sens*, 2017, **10**(5): 1824−1842.

[35] Demir B, Bruzzone L. Hashing-based scalable remote sensing image search and retrieval in large archives[J]. *IEEE Trans Geosci Remote Sens*, 2016, **54**(2): 892−904.

[36] Marmanis D, Datcu M, Esch T, et al. Deep learning earth observation classification using ImageNet pretrained

networks[J]. *IEEE Geosci Remote Sens Lett*, 2016, **13**(1): 105−109.

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

[38] Hou D Y, Miao Z L, Xing H Q, et al. Exploiting low dimensional features from the MobileNets for remote sensing image retrieval[J]. *Earth Sci Inform*, 2020, **13**(4): 1437−1443.

[39] Wang Y M, Ji S P, Zhang Y J. A learnable joint spatial and spectral transformation for high resolution remote sensing image retrieval[J]. *IEEE J Sel Top Appl Earth Obs Remote Sens*, 2021, **14**: 8100−8112.

[40] Fan L L, Zhao H W, Zhao H Y. Distribution consistency loss for large-scale remote sensing image retrieval[J]. *Remote Sens*, 2020, **12**(1): 175.

## 作者简介

吴刚 (1997-)，男，硕士研究生，研究方向为遥感图像处理与机器学习。

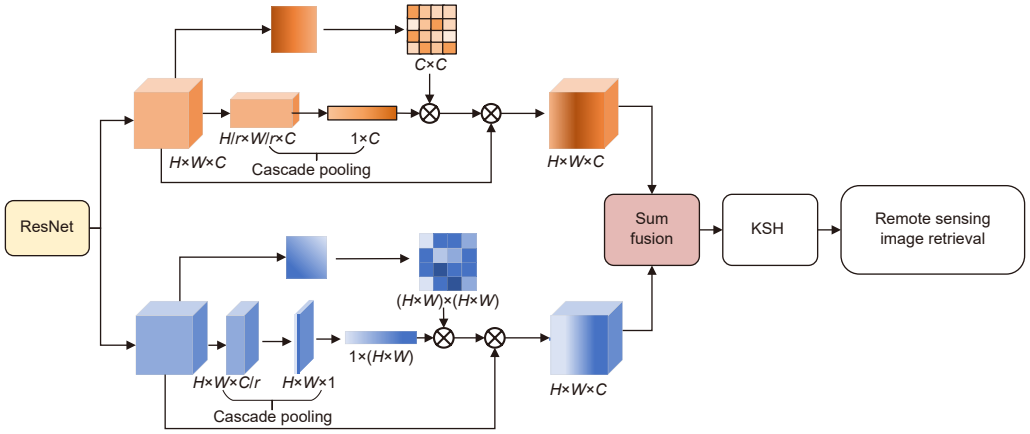
E-mail: wugang_peter@foxmail.com

【通信作者】葛芸 (1983-)，女，博士，副教授，研究方向为遥感图像处理与机器学习。

E-mail: geyun@nchu.edu.cn

# Cascade pooling self-attention research for remote sensing image retrieval

Wu Gang[1,2], Ge Yun[1,2*], Chu Jun[1,2], Ye Famao[3]




Cascade pooling self-attention for remote sensing image retrieval

**Overview:** With the development of remote sensing satellite technology and the expansion of the market in remote sensing images (RSIs), content-based remote sensing image retrieval (RSIR) plays an irreplaceable role in many fields, such as economic and social development, resource and environmental monitoring, and urban life management. However, there are complex content and rich background information in the high-resolution remote sensing images, whose features extracted by convolutional neural networks are difficult to effectively express the salient information of the RSIs. For this problem in high-resolution RSIR, a self-attention mechanism based on cascading pooling is proposed to enhance the feature expression of convolutional neural networks. Firstly, a cascade pooling self-attention module is designed. Cascade pooling uses max pooling based on a small region, and then adopts average pooling based on the max pooled feature map. Compared with traditional global pooling, cascade pooling combines the advantages of max pooling and average pooling, which not only pays attention to the salient information of the RSIs, but also retains crucial detailed information. The cascade pooling is employed in the self-attention module, which includes spatial self-attention and channel self-attention. The spatial self-attention combines self-attention and spatial attention based on location correlation, which enhances specific object regions of interest through spatial weights and weakens irrelevant background regions, to strengthen the ability of spatial feature description. The channel self-attention combines self-attention and content correlation-based channel attention, which assigns weights to different channels by linking contextual information. Each channel can be regarded as the response of one class of features, and more weights are assigned to the features with large contributions, thereby the ability to discriminate the salient features of the channel is enhanced. The cascade pooling self-attention module can learn crucial salient features of the RSIs based on the establishment of semantic dependencies. After that, the cascade pooled self-attention module is embedded into the convolutional neural networks to extract features and optimize features. Finally, in order to further increase the retrieval efficiency, supervised Hashing with kernels is applied to reduce the dimensionality of features, and then the obtained low-dimensional hash code is utilized in the RSIR. Experiments are conducted on the UC Merced, AID and NWPU-RESISC45 datasets, the mean average precisions reach 98.23%, 94.96% and 94.53% respectively. The results show that compared with the existing retrieval methods, the proposed method improves the retrieval accuracy effectively. Therefore, cascade pooling self-attention and supervised hashing with kernels optimize features from two aspects of network structure and feature compression respectively, which enhances the feature representation and improves retrieval performance.

Wu G, Ge Y, Chu J, et al. Cascade pooling self-attention research for remote sensing image retrieval[J]. *Opto-Electron Eng*, 2022, **49**(12): 220029; DOI: 10.12086/oee.2022.220029

Foundation item: National Natural Science Foundation of China (42261070, 41801288, 41261091, 62162045), and Natural Science Foundation of Jiangxi Province (20202BAB212011)

[1]School of Software, Nanchang Hangkong University, Nanchang, Jiangxi 330063, China; [2]Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition, Nanchang Hangkong University, Nanchang, Jiangxi 330063, China; [3]School of Surveying and Mapping Engineering, East China University of Technology, Nanchang, Jiangxi 330013, China

* E-mail: geyun@nchu.edu.cn