引用格式：彭宏，王炳烨，高子惠．基于Transformer的传统纹样子图检索方法[J]．中国传媒大学学报（自然科学版），2022，29(04)：08-18.
文章编号：1673-4793(2022)04-0008-11

基于Transformer的传统纹样子图检索方法

彭宏[1*]，王炳烨[2]，高子惠[2]
（1.文化和旅游部民族民间文艺发展中心，北京 100007；2.北京邮电大学人工智能学院，北京 100876）

摘要：针对如何有效、准确地检索传统纹样子图数据原图的问题，提出了一种基于Transformer的传统纹样子图检索算法。首先构建传统纹样子图数据集，其次利用卷积神经网络提取多层次的特征图进行融合。针对数据库中的图像利用Transformer生成预测框，将预测框映射回融合特征图提取局部与全局特征图后利用特征聚合算法聚合为全局与局部特征向量；针对查询子图，仅利用特征聚合算法对融合特征图进行聚合为全局向量。之后将查询子图的特征向量与数据库图像的特征向量分别进行相似度计算，排序后得到检索结果。对比实验证明本论文子图检索模型的有效性。

关键词：子图检索；图像检索；传统纹样；注意力机制
中图分类号：TP391　　**文献标识码**：A

Traditional pattern image retrieval method based on Transformer

PENG Hong[1*], WANG Bingye[2], GAO Zihui[2]
(1.Center for Ethnic and Folk Literature and Art Development, Ministry of Culture and Tourism, Beijing 100007, China; 2. Artificial Intelligence Institute, Beijing University of Post and Telecommunication, Beijing 100876, China)

Abstract: Aiming at the problem of retrieving the original image of traditional pattern sub-image effectively and accurately, a traditional pattern sub-image retrieval algorithm based on Transformer is proposed. Firstly, the traditional sub-image dataset is constructed, and then the convolutional neural network is used to extract the multi-level feature for fusion. For the image in the database, Transformer is used to generate the prediction box, and the prediction box is mapped back to fusion feature images to extract local and global feature images, and then global and local feature vectors are aggregated by feature aggregation algorithm. For the query sub-image, the feature aggregation algorithm is used to aggregate the fusion feature graph into global vector. After that, the feature vectors of the query subgraph and the database image are respectively calculated for similarity, and the retrieval results are obtained after sorting.

Keywords: subgraph retrieval; image retrieval; traditional pattern; attention mechanism

1　引言

传统纹样是中华民族的文化符号，被称为“无字的史书和民族的图腾”，从花鸟虫鱼、飞禽走兽到如意方胜、水波云气，由简入繁，美不胜收，承载着劳动人民对于美好生活的期盼，展现着一个时代的精神面貌。随着传统文化研究工作的深入，以图片形式保存的资料越来越多，种类更加繁杂，如何在大规模的图像中找到自己所需要的图像成为一个难题。在研究传统纹样图案时，需要将有意义的纹样图案从

基金项目：揭榜挂帅重点研发课题（课题编号：2021YFF0901701）
作者简介（*为通讯作者）：彭宏（1972-），男，高级工程师，主要研究领域为文化资源数字化。Email：466985365@qq.com

特定的载体中分割提取出来，进而构成中华文化素材库，助力文化大数据体系三库（标本库、基因库、素材库）的建设，这些分割提取出的纹样图案称为子图，由于已经去除原图中的背景，子图中只有特定区域有助于构造有区别的全局特征，不同于图像检索，子图检索是一个新的挑战、新的探索。在此背景下，本文将卷积神经网络与 Transformer 结合搭建子图检索模型，构建传统纹样数据集，并提取子图数据集，最后利用所搭建的模型设计并实现了传统纹样子图检索系统。

2　国内外研究现状

图像检索是计算机视觉任务中一个长期存在的研究课题，国内外已有大量学者在图像检索领域做出了杰出贡献。

在 20 世纪 90年代初，基于内容的图像检索被提出。通过提取图像自身的轮廓、纹理等视觉信息来实现检索图像，其中最直接的方法是提取图像的全局特征表征。然而，受平移、光照、遮挡等因素的影响，全局描述符在图像检索中的应用范围受限。2000年之后诞生了基于局部描述符的图像检索技术。2004年，Lowe DG 提出了尺度不变特征变换（SIFT 描述符）[8]，局部特征描述符还包括 LBP[1]、ORB[2]、SURF[3]、BRISK[4]、FREAK[5]、空间包络特征 Gist 特征[6]等。2003年，J. Sivic 提出了 Bag-of-Word（BoW）模型[7]，与局部描述符结合作用于图像检索，BoW 利用编码的思想，以有效的局部特征描述符为基础，采用聚类等算法训练编码本获得图像的整体表达。由于 BoW 模型的高复杂度限制了这类方法的提升空间，之后出现了整合局部向量的 Fisher Vector[9]、VLAD（Vector of Aggragate Locally Descriptor）[10]等方法。由于 VLAD、FV 等算法嵌入维度过高，研究人员提出了近似近邻的 ANN 算法[11, 12]，主成分分析 PCA 也常用于降维任务。

2012年 Krizhevsky 等人使用 AlexNet[13]在 ILSRVC（ImageNet 大规模视觉识别挑战）中达到了当时最优的识别精度。随后基于卷积神经网络（Convolutional Neural Network, CNN）的图像检索方法不断地被提出，传统的提取图像描述符的方法逐渐被取代。使用卷积神经网络提取图像特征可以分为两种，一种是提取卷积神经网络最后的全连接层上的描述符，它将图像中的视觉内容概括为单个特征向量，该向量被视作全局描述符。另一种是提取中间卷积层的特征，卷积滤波器被视为局部检测器，与提取到的全连接层描述符相比局部描述符对图像的遮挡和截断更加健壮。采用编码和池化两种策略将提取到的一组局部特征描述符聚合为全局特征。编码策略常用的方法有 VLAD、FV 等算法，直接对卷积特征进行池化也可以生成具有区分度的特征。例如 R-MAC[14]在所提取的特征输出图上设计不同大小的区域窗口，通过滑动采集不同的局部特征，然后计算多个局部映射的最大值，再将多个局部向量整合成单维特征向量作为图像的全局特征。SPoC 方法[15]使用池化解决了把卷积特征图变成单维特征向量的问题。Yannis 等人提出 CroW[16]特征方法，首先提取卷积神经网络最高层的卷积图，之后对提取到的卷积图计算平面权值与每个通道的权重后加权。Philippe[17]等人针对图像检索中对局部特征的使用与聚合问题，基于所提取到称为超级特征的中级特征提出了一种新的深度图像检索架构。Hui Wu[18]等人针对大型码本量化深度局部特征占用内存过多，且特征学习和聚合的能力受限等问题，提出了一个统一的框架来共同学习局部特征表示和聚合。

3　传统纹样数据集构建

本文所使用的数据集来源主要有三种渠道：1. 图书，利用扫描仪扫描采集传统纹样的图书；2.PatternNet 资源库，对实验室已有的图像数据进行整理；3. 互联网。

对于图书的扫描，首先选取大量包含传统纹样图像的专业书籍，检查书籍页面的文字、图片是否印刷清晰。设置扫描仪设备的参数，该参数直接决定扫描所采集图像的质量，本论文调整的扫描仪参数主要有图像大小、图像色彩、图像格式、图像分辨率等 4 项。本文设置扫描仪采集图片大小大于 30MB，图像分辨率为 400dpi，存储格式为 TIF 格式，彩色模式保存。之后选择图像数据存储位置，由于书籍的 ISBN 编号具有唯一性，因此在保存图像时采用“ISBN 码-页码”的形式命名图片。为了方便实验室后续对数据集的使用和管理，需要将采集好的图像数据上传 PatterNet 资源库，根据实验室的图像采集元数据规范，对采集的图像数据进行标注[19]。

本文所用的子图检索数据集尽可能地选取图像中包含多个不同类型纹样的图像，子图检索数据集所采用的部分图像如图 1 所示。

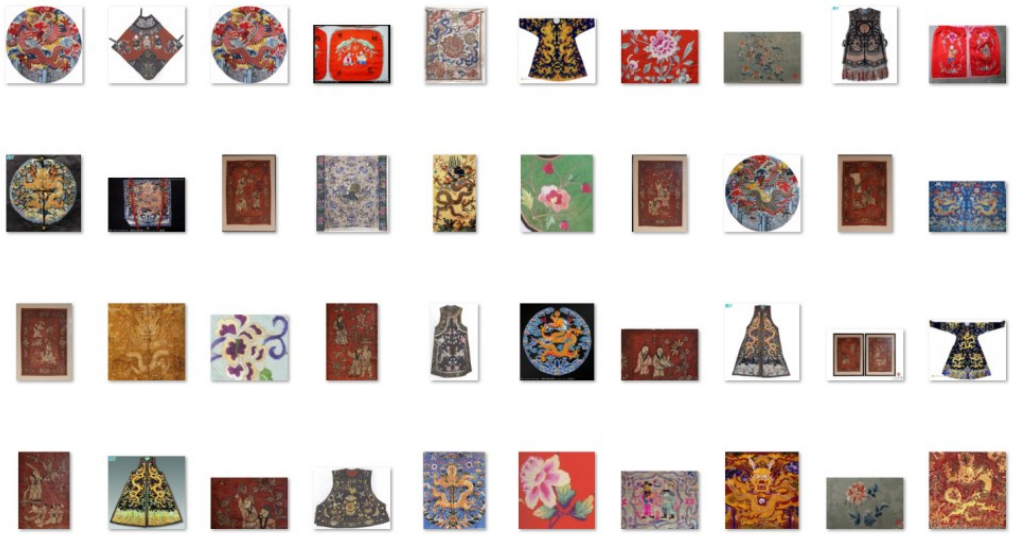


图1 子图数据集部分图像

子图检索数据集包括花、蝴蝶、龙、云、人物、鸟、蝙蝠、鱼、海水姜芽、石榴十类纹样图案,共3074张图片。将构建好的子图数据集按照7:3的比例划分训练集和测试集,训练集图像共2152张,测试集图像922张,子图数据集详细统计信息如表1所示。

对子图检索的数据集进行标注,本文采用LabelImg对子图数据集进行标注,LabelImg图像标注界面如图2所示,用矩形框标出图像中目标物体,之后进行保存。

表1 传统纹样子图数据集

目标纹样类别	花	鸟	人物	云	蝴蝶	海水姜芽	石榴	龙	鱼	蝙蝠
图像数量	1587	770	1012	120	573	47	56	152	52	101

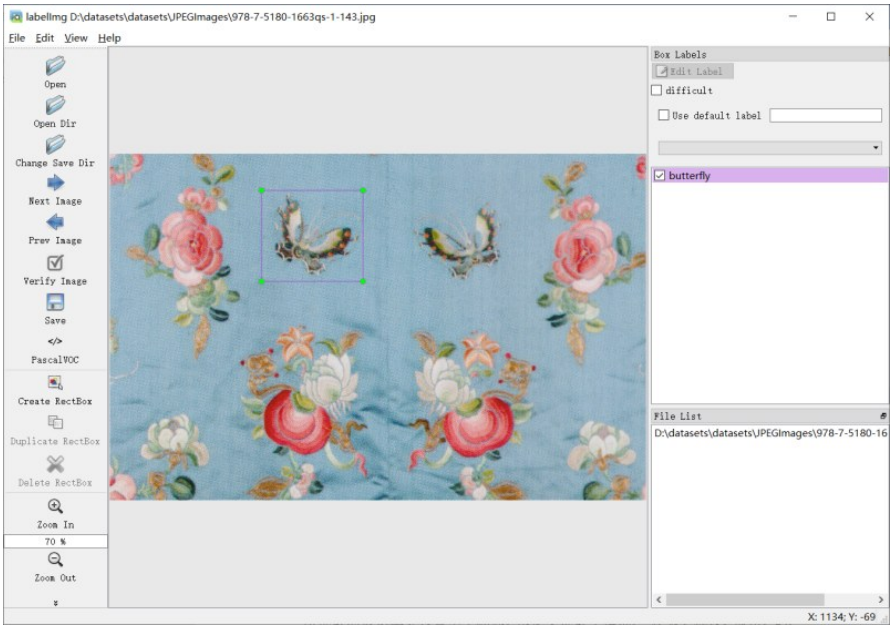


图2 LabelImg图像标注界面

对子图数据集，分割提取图像中的子图作为模型的查询子图。子图的分割提取采用 Adobe Photoshop（简称 PS），如图 3 所示，子图数据集图像删除背景，将采集的子元素保存，作为查询子图。

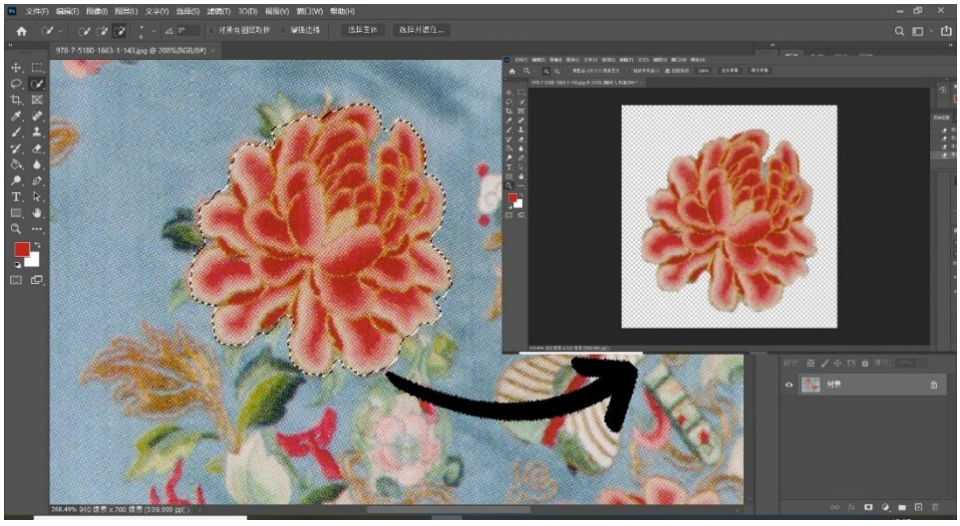


图 3　查询子图提取过程

4　基于 Transformer 的传统纹样子图检索方法

4.1　子图检索整体结构

采用 DETR[20] 模型中的 Transformer 结构，构建面向传统纹样的子图检索模型 CE-Transformer，模型整体结构如图 4 所示，分为对数据库图像的处理和查询子图的处理。

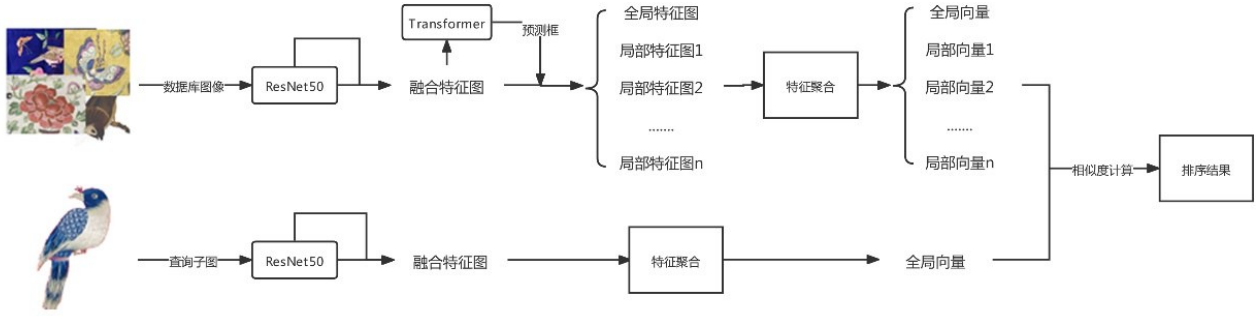


图 4　子图检索模型整体结构

本文以 RestNet50 作为特征提取的主干，提取 CNN 不同层次的卷积特征输出，在网络结构中增加特征融合模块，将提取到的多层特征输出进行融合得到融合特征图。子图输入的查询图通常只包含一个纹样图案，为了验证所提出的子图检索模型的有效性，本文所构建的子图检索数据集会尽可能地选择一张图中包含多个纹样图案的图像，因此尽可能地提取一张图像中的多个区域的局部特征。利用 DETR 模型的 Transformer 结构，将数据库图像提取到的融合后的特征图输入进 Transformer 结构中，生成预测框。

Transformer[21] 结构由编码器与解码器两部分构成，解码器的输入需要是序列，首先将提取到的融合特征图进行降维，使用多头自注意力机制和编解码注意力机制对 N 个 d 维的特征进行编解码，最终由前馈网络预测相对于原图像尺寸归一化后的框中心点坐标、高度和宽度。线性映射层使用 Softmax 层预测类别标签，并使用一个额外的特殊类别标签来表示背景，即原图像中没有对象被检测到。

模型通过一次编解码会产生一个大小为 N 的预测集，N 被设置为远大于数据集中目标物体的数量，子图检索模型的训练损失函数如下：

$$\hat{\sigma} = \min_{\sigma \epsilon GN} \sum_{i}^{N} L_{match}\left(y_i, \hat{y}_{\sigma(i)}\right) \quad (1)$$

其中y表示一系列目标对象的Ground Truth集合，$\hat{y} = \{\hat{y}_i\}_{i=1}^{N}$是模型所产生的$N$个预测，$N$的数目大于原图像中的目标数，因此把$y$也看作以空集∅填充的大小为$N$的集合，$L_{match}(y_i, \hat{y}_{\sigma(i)})$是Ground Truth y_i和相应预测的二分图损失，该优化问题通过匈牙利算法求解。

匹配损失考虑类别预测与Ground Truth Boxes之间的相似性，每个Ground Truth集合可以看作$y_i = (c_i, b_i)$，c_i表示目标对象的类别标签(也有可能是空值)，b_i是一个向量，表示目标对象的Ground Truth的中心点和宽、高，对于索引为$\sigma(i)$的预测，定义预测类别为c_i的概率为$\hat{p}_{\sigma(i)}(c_i)$，预测框为$\hat{b}_{\sigma(i)}$，匹配损失可以定义为：

$$L_{match}\left(y_i, \hat{y}_{\sigma(i)}\right) = -\mathbb{1}_{c_i \neq \varnothing} \hat{p}_{\sigma(i)}\left(c_i\right) + \mathbb{1}_{c_i \neq \varnothing} L_{box}\left(b_i, \hat{b}_{\sigma(i)}\right) \quad (2)$$

计算所有匹配对的Hungarian loss：

$$L_{Hungarian}(y, \hat{y}) = \sum_{i=1}^{N} -\log \hat{p}_{\hat{\sigma}(i)}\left(c_i\right) + \mathbb{1}_{c_i \neq \varnothing} L_{box}\left(b_i, \hat{b}_{\hat{\sigma}(i)}\right) \quad (3)$$

$\hat{\sigma}$是上一步中得到的最优匹配，计算损失函数的第二部分为对Bounding Boxes评分的$L_{box}(\cdot)$，使用L1损失与IoU损失的联合：

$$L_{box}\left(b_i, \hat{b}_{\sigma(i)}\right) = \lambda_{iou} L_{iou}(b_i, \hat{b}_{\sigma(i)}) + \lambda_{L1} \left\| b_i - \hat{b}_{\sigma(i)} \right\|_1 \quad (4)$$

$\lambda_{iou}, \lambda_{L1} \in \mathbb{R}$是超参数。

利用训练完成后模型提取子图数据集的全局与局部特征向量和查询子图的全局特征向量，之后进行相似度的计算，返回最终的计算结果。

4.2 特征融合

卷积神经网是一个层次化的结构，其不同层次的特征旨在对不同层的信息进行编码。网络结构由浅入深，语义信息越来越丰富，但卷积特征输出图的分辨率会逐渐变小，细节信息相对要少。底层特征关注的是细节信息，但所含的语义性相对低一些，模糊性强，且存在噪声和背景杂乱的问题。在深度学习的很多工作中，将不同层次的卷积特征进行融合，利用卷积神经网不同层次的互补优势成为改善模型提高性能的一个关键。如图5所示为ResNet50从输入图像中提取到的不同层次的信息。

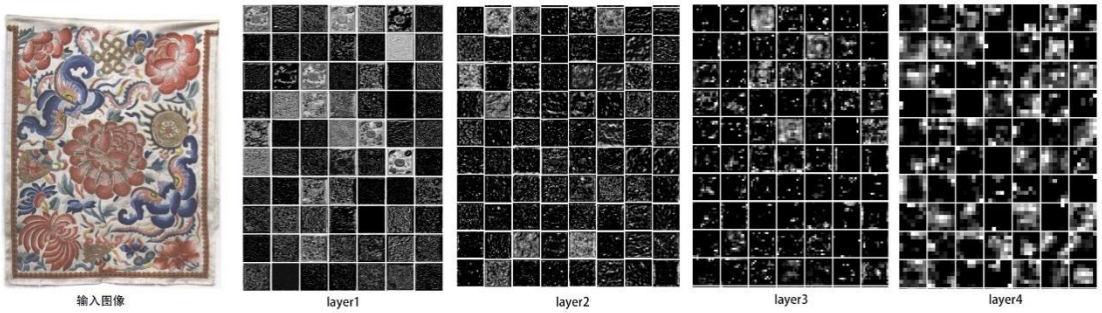


图5 ResNet50不同层次的部分特征图

在图像检索中，用底层的特征来度量图像间细粒度的相似性，用高层特征来度量语义的相似性。由于子图检索模型需要对数据集中的图像利用提取到的特征图生成预测框，底层与高层特征的融合特征图可以提高模型对小目标物体的敏感性，因此采用特征金字塔的方式对不同层次的特征图进行融合，本文所采用的图像金字塔如图6所示。

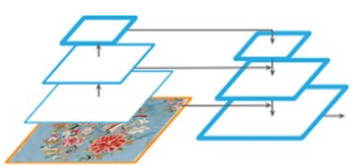


图6 图像特征提取金字塔

本文采用最近邻插值法尽可能地留住所提取的特征图中的语义信息，从而在与具有充分空间信息的较低层特征图融合时可以得到兼具空间信息和明显语义信息的融合特征图。

4.3 特征提取

子图检索需要提取数据集图像中目标区域的特征向量分别与用户输入的查询子图进行相似度的计算。将融合特征图输入Transformer中生成预测框，将模型所生成的N个预测框映射回提取到的融合特征图中，提取数据库图像的区域特征图，计算公式如下：

$$X_i = X\left(\lfloor w \times a_i \rfloor : \lfloor w \times b_i \rfloor, \lfloor h \times c_i \rfloor : \lfloor h \times d_i \rfloor\right) \quad (5)$$

式中X表示图像的融合特征图，X_i是提取到的第i

个区域，w，h 分别表示融合特征图的宽、高，a_i，b_i，c_i，d_i 表示第 i 个预测框的坐标，值在 0 到 1 之间。之后将采用改进的 R-MAC 算法对提取到的全局融合特征图和区域特征图进行聚合。

4.4　自注意力机制

子图检索与图像检索的最大区别在于用户输入的查询图像的不同，子图检索的查询子图是从图像中分离出的元素，背景已经被替换为白色或透明，当提取查询子图的视觉特征时，考虑增大目标元素所在区域的权重，而抑制白色或透明背景的权重。注意力机制类似于生物观察行为的视觉机制，能够在大量信息中有效地筛选出少量重要信息，忽略掉大量的干扰信息，聚焦到重要信息上去。自注意力机制是一种注意力机制，它在减少对外部信息依赖的同时更擅长捕获输入数据的内部关联，因此本文采用自注意力机制计算用户查询子图的注意力，增大查询子图的注意力区域的特征权重，抑制不重要区域的特征。

以自注意力为核心的 Transformer 已经被大规模地用于处理自然语言等课题任务中，Vision Transformer 是 Transformer 在计算机视觉中的成功应用。标准 Transformer 的输入是一维序列，为了处理图像 $X \in \mathbb{R}^{H \times W \times C}$，Vision Transformer[22] 将图像分割为一组固定大小为 P 的块 $X_p \in \mathbb{R}^{N \times (p^2 \cdot C)}$，H、W 是图像的宽高，C 是通道数，$N$ 表示分为 N 个块，其中 $N = HW/P^2$，Transformer 会将分割的块展平，并线性映射为 D 维的 Patch Embedding，为了保持位置信息，每个 Patch Embedding 需要附加一维的位置 Embedding 后输入进 Transformer 中，Transformer 的编码层由 L 层组成，每个层有两个主要模块：多头自注意力（MSA）层和前馈网络（FFN），多头注意力层会对输入计算自注意力，Layer Norm 会在每个结构前加入，随后是残差连接，具体结构如图 7 所示。

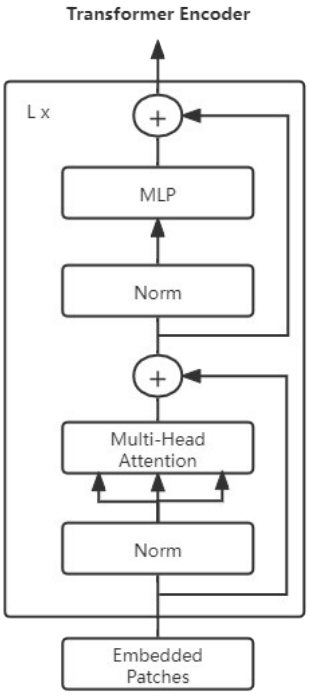
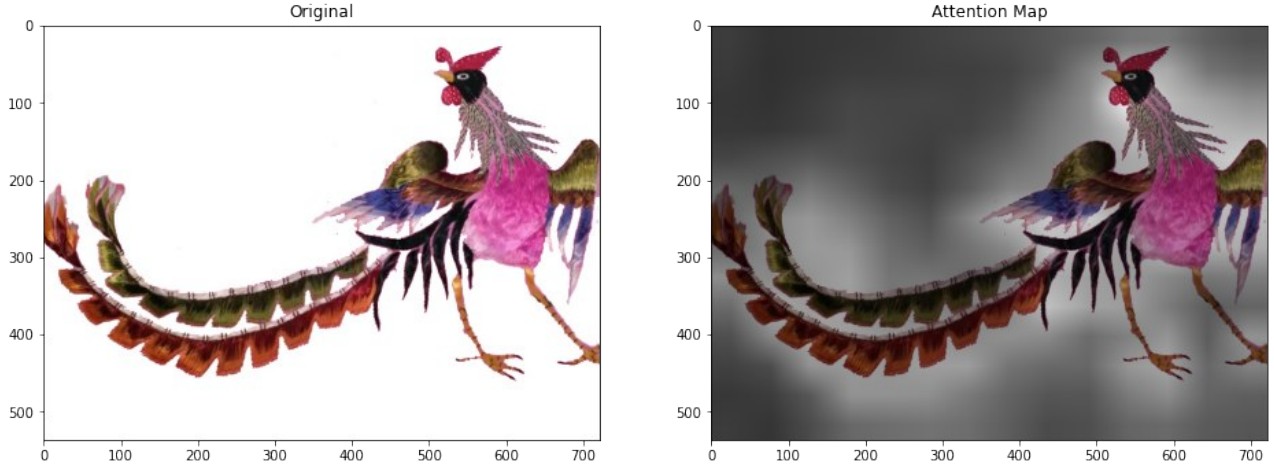


图 7　Transformer 编码器结构

用户输入的查询子图经 ResNet50 提取融合特征图后被分割为 N 个块输入进 Vision Transformer 中，设置 Vision Transformer 的 N 大小为 14*14，head 为 12，计算查询子图的注意力权重。本文将得到的 12 个头的注意力权重相加后做平均，如图 8 为查询子图生成的注意力投射到原图中对应的位置。



（a）原图　　（b）注意力映射图

图 8　查询子图与生成的注意力权重图

在图8中的注意力映射图中,利用Vision Transformer所生成的注意力映射到原图中后黑色部分为注意力较弱的区域,而亮的区域为注意力较强的区域。

4.5 特征聚合

使用卷积神经网提取图像的卷积特征可以为图像检索提供有效的描述符,针对卷积特征涌现出众多的特征聚合算法。R-MAC是基于区域的特征聚合算法,将局部特征聚合为有效的全局描述符。R-MAC采用变窗口的方式在特征平面上滑动处理。

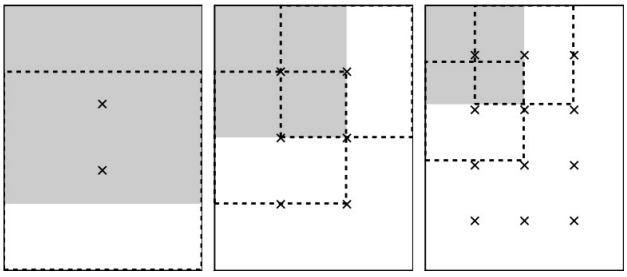


图9 R-MAC采样区域

使用CNN提取查询图像的卷积特征图,大小为$W \times H \times K$,K为通道数,W和H分别是所提取到的特征图的宽度和高度,选择区域大小为R_s的滑窗对卷积特征进行采样,其中$R_s = 2\,min\,(W,H)/(s+1)$,$s$表示在$s$个不同尺度上采样正方形滑窗的区域,在$s=1$时,采样区域要尽可能大,即宽高等于min($W$, H),图9表示在4个不同尺度下的采样区域,即s=1, 2, 3时的采样过程,两个滑动窗口间需要保持重叠区域最少为40%,采样之后,R-MAC会对所有的区域特征图进行最大池化、L2归一化和PCA,之后使用求和池化获得全局特征向量后,再进行一次L2归一化。

因为当采样区域数量越多,尺度越多时,平均池化会使得训练更加的稳定,因为平均池化会根据采样的数量调整求和的梯度,所以在本文使用平均池化代替原R-MAC算法中的求和池化。R-MAC在图像检索中有着比较高的性能,但R-MAC在聚合图像的区域特征向量时会统一的对待图像中的所有区域,而对于子图来说,图像中仅有特定的区域有助于构建有区别的全局特征,且由于上述问题,当采用更多的尺度采集更多区域时,R-MAC的性能会下降。注意力机制能够从大规模信息中有效过滤出少量重要信息,因此本文采用Transformer对R-MAC进行改进,采用Vision Transformer计算子图的注意力,作为特征图的空间权值,对于特征图的通道加权,采用类似IDF的思想,即降低文档中高频出现但是对信息的表达影响比较小的单词的权重,例如"the"。在计算特征图通道特征时,如果某个通道的特征图被强烈激活,即大部分元素是非零的,则该特征图不利于定位物体的区域,需要降低该通道的权重,而那些非零元素稀疏的通道特征图可能提供重要的信息需要增大权重,计算公式如下:

$$I_k = \log\left(\frac{K\varepsilon + \sum_h Q_h}{\varepsilon + Q_k}\right) \tag{6}$$

I_k表示通道权重,ε是常量,Q_k表示第k个通道的特征图中非零相应所占的比例。本文所采用的特征加权过程如图10所示。

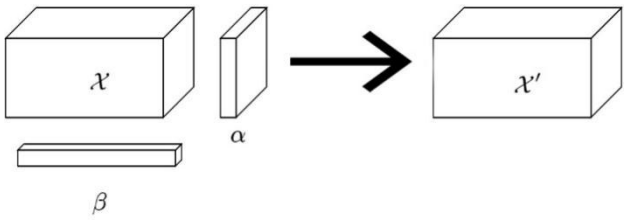


图10 卷积特征加权过程

X是提取到的融合特征图,α表示融合特征图的平面注意力权重,β表示通道权重。

4.6 相似性匹配及检索

查询子图只包括用户分割提取出的纹样图案,而数据库的图像中包含多种纹样图案,数据库的经过Transformer会生成N个预测框,Transformer生成的预测框对应到原图中的位置如图11所示。

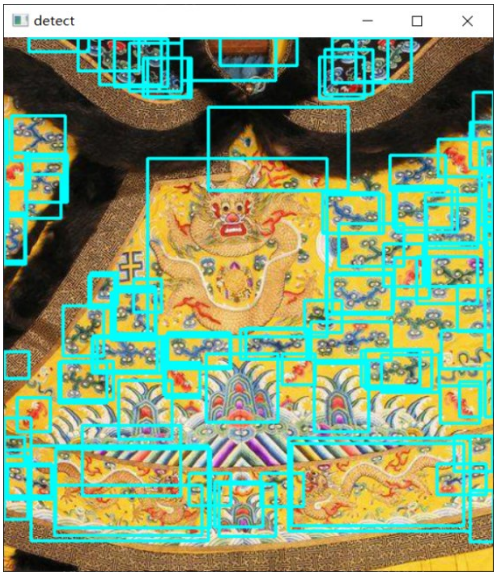


图11 Transformer生成预测框

因此，一幅图像经过模型后会提取到 N+1个特征向量（包括图像的全局特征向量）。将数据库图像所提取的全局和局部特征向量与用户输入的检索子图所提取到的特征分别计算相似性，取最大值作为查询子图与该数据库图像的相似性，即取查询子图特征向量与数据库图像特征向量距离的最小值，计算过程如下式所示：

$$sim\left(D_i,Q\right)=\min\left(dist\left(x_0,\mathrm{y}\right),dist\left(x_1,\mathrm{y}\right),\ldots,dist\left(x_n,\mathrm{y}\right)\right) \quad (7)$$

式中 y 表示提取到的查询子图的特征向量，D_i 表示数据库图像中第 i 幅图像，x_n 表示第 i 幅图像提取到的第 n 个特征向量，sim 表示查询子图与数据库图像的相似性。余弦距离是利用向量空间中两个向量夹角的余弦值来度量两者之间的距离，它对向量的绝对数值不敏感，而欧几里得距离主要从向量的数值大小分析差异。因此，本文采用余弦距离衡量查询特征于数据库向量之间的相似度，计算公式如下：

$$dist\left(x,q\right)=cos\,\theta=\frac{\vec{x}\cdot\vec{y}}{\|x\|\|y\|} \quad (8)$$

将数据库图像按照计算出的余弦距离从小到大排序，得到最终的图像检索的结果。如图 12 为本模型返回前 10 张图像的检索结果，被红框圈出的图像为查询子图所在的父图，即查询子图所在的原始图像。

图 12　前 10 张图像的检索结果

5　实验

为证明本论文提出的传统纹样子图检索模型的有效性，采用第 3 节整理的传统纹样子图检索数据集进行子图检索的相关对比实验。

5.1　实验设置

本文采用 1.8.0 版本的 Pytorch 来搭建子图检索模型，使用在 ImageNet2012 数据集上预训练的 ResNet50 作为特征提取的主干网络，采用 DETR 模型中的 Transformer 结构，并加入特征融合模块与特征聚合模块，搭建子图检索数据库图像特征提取分支模型。将模型输入图像大小调整为 800×800，使用 AdamW 来训练。本文设置 Transformer 的初始学习率为 0.001，主干网络的初始学习率为 0.0009，权重衰减系数为 0.0001，动量率设置为 0. 1，迭代训练 5000次后将学习率降低为原来的十分之一。

对于查询子图的图像特征提取网络分支，主干部分特征融合模块和与上述相同，采用 ResNet50，之后加入在 ImageNet2012 数据集上预训练的 Vision Transformer 对特征聚合模块进行改进，训练时加载数据库图像的特征提取分支模型的主干部分，并固定参数，采用 cifar 图像数据集进行微调，本文设置网络模型的学习率为 0.001，权重衰减系数为 0.0001，动量衰减系数为 0.0001，训练周期为 5000，调整模型的输入图像大小为 800×800。

5.2　实验结果分析

为了验证提出的子图检索模型的各个组件的有效性，在传统纹样子图数据集上进行消融实验，包括（1）特征融合模块的有效性；（2）改进后的 R-MAC 与 R-MAC 算法的对比；（3）Transformer 生成预选框的有效性。采用平均精度均值 mAP 来评价模型的检索

性能。

(1)特征融合模块的有效性

在使用ResNet50作为特征提取主干的情况下,为了验证增加的特征融合模块的有效性,将数据库图像与查询子图特征提取主干中的特征融合模块摘除,仅提取ResNet50中Cov4_x卷积组的输出特征图,保持Transformer、特征聚合等模块不变,利用平均检索精度进行评估,对比实验结果如表2所示。

表2 有特征融合与无特征融合效果对比

子图检索模型	平均检索精度mAP
SI-Transformer(无特征融合)	31.02%
SI-Transformer(有特征融合)	86.09%

表2中去除了特征融合模块的子图检索网络模型与添加了特征融合的网络模型相比其检索精度明显下降了55.07%,由此可以证明增加的特征融合模块可以有效地将不同卷积层的特征图融合,提高检索的精度。

(2)改进后的R-MAC与R-MAC算法

为了证明本文改进后的R-MAC算法的有效性,将子图模型中的改进后的R-MAC算法与原始R-MAC进行对比实验,除R-MAC算法之外其余模块保持不变,设置s=5,即在5中不同的尺度下进行采样,如表3所示为使用不同的R-MAC算法时,返回100张图像的平均检索精度。

表3 改进后的R-MAC算法与原始R-MAC算法对比

子图检索模型	平均检索精度mAP
SI-Transformer(R-MAC)	83.10%
SI-Transformer(改进后R-MAC)	86.09%

在表3中,使用改进后的R-MAC算法实验结果的平均检索精度比原始的平均检索精度高了2.99%,由此可以证明本文对R-MAC改进的有效性,在原始R-MAC算法基础上增加空间与通道权重可以有效改善检索结果。

(3)Transformer生成预选框的有效性

由于用户输入的查询子图通常仅包含一个查询目标或仅有查询目标的一部分,而数据库中的图像中包含多个目标区域。因此为了提高检索的准确性,本文利用DETR模型中的Transformer结构生成预测框提取局部区域特征,将数据库图像中提取到的多个局部区域与查询子图计算相似度,数据库的每幅图像中提取100个预选框,由于预选框数量较大且重合的区域较多,为减少计算量,计算每两个预选框之间的区域重叠,若重叠面积大于80%,则仅保留一个。为证明所采用的利用预测框提取区域特征的有效性,将子图检索模型中生成预选框的部分去除,表4所示为使用预选框与不使用预选框的检索结果对比。

表4 有预选框与无预选框的结果对比

子图检索模型	平均检索精度MAP
SI-Transformer(无预选框)	70.85%
SI-Transformer(有预选框)	86.09%

表4中使用预选框的模型与无预选框的模型检索精度上升了15.24%,由此可以证明Transformer所生成的预测框的有效性,利用预选框提取图像中所包含的多个目标区域的特征向量,针对所提取的多个区域特征进行相似度计算,提高子图检索的精度。

5.3 其他模型实验结果

5.2节中利用消融实验分别证明了特征融合模块、预测框、改进R-MAC算法的有效性,为了证明本文所提出的检索方法与其他论文提出的图像检索方法相比在传统纹样数据集上的有效性,选取了几个比较经典的图像检索模型进行了对比实验,采用在ILSVRC ImageNet数据集上预训练的模型,使用本文第3节整理的子图数据集作测试,如表5所示为不同模型在数据集上前200张图像的平均检索精度。

表5 其他模型实验结果

模型	特征维度	平均检索精度mAP@200
rR-MAC[23]	2048	49.22%
DELF[24]	1024	47.04%
SPoC[25]	1024	43.06%
NetVLAD[26]	2048	35.66%
Crow[27]	1024	44.52%
Ours	1024	60.74%

从表中可以看出本文所提出的子图检索模型取得了最好的结果,其平均检索精度mAP值明显优于其他指标,且在传统纹样子图数据集中进行微调后本文的算法平均精索精度可以达到86.09%,由此进一步证明了本文所提出的模型在子图检索中的显著性。如图13所示为本文算法在第3节整理的传统纹样子图数据集中微调后所对应的PR曲线。

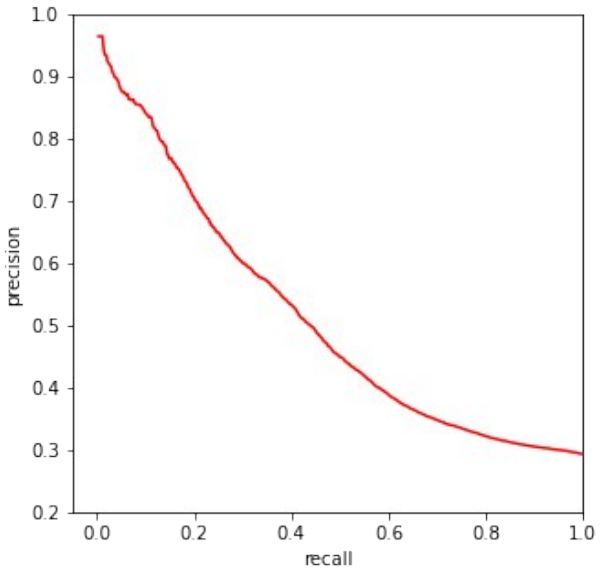


图 13　本文算法 PR 曲线

6　结束语

现有的图像检索算法与图像数据集，主要应用于单目标的图像与图像之间的查询，缺少针对传统纹样子图检索的算法与数据集。本文将传统纹样作为研究对象，构建了一套传统纹样子图数据集，并提出了一种基于 Transformer 的子图检索算法。首先利用 CNN 提取输入图不同层次的特征输出，利用特征金字塔的方式进行融合，然后利用 Transformer 生成预测框，将所生成的预测框映射回融合特征图中提取数据库图像的全局与局部特征图，并利用改进 R-MAC 算法进行聚合，对于用户输入的查询子图，由于子图是从原图中分割提取得到，已去除原图中的背景，在子图中只有特定的区域有助于构造有区别的全局特征，因此用 Vision Transformer 对 R-MAC 特征聚合算法进行改进，之后将提取到的子图特征与数据库图像的全局与局部特征计算相似度，排序后得到检索结果，最后通过对比实验证明本论文子图检索模型的有效性。但目前本文提出的子图检索还存在一些问题，首先，因为数据集制作比较耗时，因此本论文所构建的数据集的标签较粗，在未来的工作中可以考虑将数据集标签细化，实现细粒度的图像检索。其次，本文采用 DETR 模型的 Transformer 结构，会并行解码生成 N 个预测框，N 的数量会被设计为远大于图像中目标的数量，存在大量冗余，且对显存条件要求比较高，未来可以考虑改为不并行生成预测框。

参考文献（References）：

[1] Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7):

[2] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF [C]//International Conference on Computer Vision, IEEE, 2011:2564-2571.

[3] Bay H, Ess A, Tuytelaars T, et al. Speeded-Up Robust Features (SURF) [J]. Computer Vision and Image Understanding, 2008, 110(3):346-359.

[4] Yang S, Li B, Zeng K. SBRISK: speed-up binary robust invariant scalable keypoints [J]. Journal of Real-Time Image Processing, 2016, 12(3):583-591.

[5] Alahi A, Ortiz R, Vandergheynst P. FREAK: fast retina keypoint [C]//IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012:510-517.

[6] Chu J, Zhao G H. Scene classification based on SIFT combined with GIST [C]//International Conference on Information Science, Electronics and Electrical Engineering, 2014: 100-109.

[7] Sivic J, Zisserman A. Video Google: a text retrieval approach to object matching in videos [C]//Proceedings Ninth IEEE International Conference on Computer Vision, 2003: 1470-1477.

[8] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[9] Perronnin F, Sanchez J, Mensink T. Improving the Fisher Kernel for Large-Scale Image Classification [C]//Computer Vision -ECCV 2010, Springer, 2010:143-156.

[10] Jegou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation [C]//Conference on Computer Vision and Pattern Recognition, IEEE, 2010:3304-3311.

[11] Jegou H, Douze M, Schmid C. Product quantization for nearest neighbor search [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(1):117-128.

[12] Muja M, Lowe D G. Scalable nearest neighbor algorithms for high dimensional data [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(11):2227-2240.

[13] Krizhevsky A, Sutskever, I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6):84-90.

[14] Tolias G, SicreR, Jégou H. Particular object retrieval with integral max-pooling of CNN activations [DB/OL]. arXiv:

1511.05879.

[15] Babenko A, Lempitsky V. Aggregating deep convolutional features for image retrieval[DB/OL]. arXiv:1510.07493.

[16] Kalantidis Y, Mellina C, Osindero S. Cross-dimensional weighting for aggregated deep convolutional features[C]// Computer Vision -ECCV 2016 Workshops, Springer, 2016: 685-701.

[17] Wu Z, Yu J. A multi-level descriptor using ultra-deep feature for image retrieval[J]. Multimedia Tools and Applications, 2019, 78(18): 25655-25672.

[18] Wu H, Wang M, Zhou W, et al. Learning token-based representation for image retrieval[DB/OL]. arXiv:2112.06159.

[19] 周伟, 赵海英. 传统民族服饰数字化采集元数据构建[J]. 图学学报, 2018, 39(06):1183-1191.

[20] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//Computer Vision -ECCV 2020, Springer, 2020: 213-229.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017: 5998-6008.

[22] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[DB/OL]. arXiv:2010.11929.

[23] Kim J, Yoon S E. Regional Attention Based Deep Feature for Image Retrieval[C]//British Machine Vision Conference (BMVC), 2018: 209.

[24] Noh H, Araujo A, Sim J, et al. Large-scale image retrieval with attentive deep local features[C]//IEEE International Conference on Computer Vision, IEEE, 2017: 3456-3465.

[25] Babenko A, Lempitsky V. Aggregating deep convolutional features for image retrieval[DB/OL]. arXiv:1510.07493.

[26] Arandjelović R, Gronat P, Torii Akihiko, et al. NetVLAD: CNN architecture for weakly supervised place recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(06): 1437 -1451.

[27] Kalantidis Y, Mellina C, Osindero S. Cross-dimensional weighting for aggregated deep convolutional features[C]// Computer Vision -ECCV 2016 Workshops, Springer, 2016: 685-701.

编辑:王谦

(上接第7页)

mation Processing Systems, 2007.

[2] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7):1527-54.

[3] Bengio Y. Learning deep architectures for AI[J]. Foundations and Trends in Machine Learning, 2009, 2(1):1-127.

[4] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010, 11:3371-3408.

[5] Tariyal S, Majumdar A, Singh R, et al. Greedy deep dictionary learning[J]. IEEE Access, 2016, 4: 10096-10109.

[6] Zhang Q, Li B. Discriminative K-SVD for dictionary learning in face recognition[C]//2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, 2010: 2691-2698.

[7] Song J, Xie X, Shi G, et al. Multi-layer discriminative dictionary learning with locality constraint for image classification[J]. Pattern Recognition, 2019, 91: 135-146.

[8] Chun I Y, Fessler J A. Convolutional dictionary learning: Acceleration and convergence[J]. IEEE Transactions on Image Processing, 2017, 27(4): 1697-1712.

[9] Hu J, Tan Y P. Nonlinear dictionary learning with application to image classification[J]. Pattern Recognition, 2018, 75: 282-291.

[10] Xiao W, Liu H, Tang H, et al. Two-layers local coordinate coding[C]//CCF Chinese Conference on Computer Vision. Springer, Berlin, Heidelberg, 2015: 34-45.

[11] Zhang Z, Jiang W, Qin J, et al. Jointly learning structured analysis discriminative dictionary and analysis multiclass classifier[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 29(8): 3798-3814.

[12] Nguyen H V, Ho H T, Patel V M, et al. DASH-N: Joint hierarchical domain adaptation and feature learning[J]. IEEE Transactions on Image Processing, 2015, 24(12): 5479-5491.

[13] Tang H, Wei H, Xiao W, et al. Deep micro-dictionary learning and coding network[C]//Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019: 386-395.

编辑:王谦