

一种基于 ViT 的局部图像拷贝检测算法

朱 辰¹, 陈于勋², 陈禹坤¹, 王总辉²

¹浙江大学国家卓越工程师学院 杭州 中国 310015

²浙江大学计算机科学与技术学院 杭州 中国 310027

摘要 近年来,随着图像编辑技术的不断发展,数字图像的知识产权保护形势日益严峻。为有效遏制图像内容的非法窃取、复制与篡改行为,切实保障用户的合法权益,图像拷贝检测作为内容审核体系的关键组成部分,其重要性不言而喻。其中,一种常见的规避检测手段是攻击者将源图像的部分内容巧妙地植入目标图像中,形成“画中画”式的局部拷贝。针对这一挑战,本文提出了一种基于 ViT 的局部图像拷贝检测算法。该算法旨在利用 ViT 强大的局部特征提取能力,在原始全局特征的基础上提取出图像的局部特征,揭示图像间潜在的依赖关系,并通过重新排列相关图像的顺序,提高具有更高相似度的图像在检测结果中的排序,从而精准实现对图像局部信息的拷贝检测。本文与基于自监督描述符的图像拷贝检测算法(A Self-Supervised Descriptor for Image Copy Detection)相比,在 DISC2021 数据集上取得显著提升,局部拷贝图像识别准确率提高 10%, μ AP(Micro Average Precision)指标提高 10%~15%。此外,本文还利用热力图进行了直观的定性分析,证实了 ViT 能够敏锐地捕捉到图像中存在局部拷贝的区域,进一步验证了算法的有效性。本文提出的基于 ViT 的局部图像拷贝检测重排序算法能对使用画中画图像增强方式的拷贝图像进行有效检测,并在 DISC2021 数据集上取得了优异成绩,拓展了数字图像取证领域研究的新思路。

关键词 图像拷贝检测; ViT; 图像增强; 重排序; 热力图

中图分类号 TP309.2 DOI 号 10.19363/J.cnki.cn10-1380/tn.2025.11.07

A ViT-based Algorithm for Localized Image Copy Detection

ZHU Chen¹, CHEN Yuxun², CHEN Yukun¹, WANG Zonghui²

¹Polytechnic Institute, Zhejiang University, Hangzhou 310015, China

²College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

Abstract In recent years, with the continuous development of image editing technology, the situation of intellectual property protection of digital images has become increasingly severe. In order to effectively curb the illegal stealing, copying and tampering of image content, and effectively protect the legitimate rights and interests of users, image copy detection, as a key component of the content audit system, is of great importance. One of the common means to avoid detection is that the attacker skillfully implants part of the source image into the target image, forming a "picture-in-picture" type of partial copy. To address this challenge, this paper proposes A ViT-based Algorithm for Localized Image Copy Detection. The algorithm aims to utilize the powerful local feature extraction capability of ViT to extract the local features of an image on the basis of the original global features, to reveal the potential dependencies between the images, and to improve the ordering of images with higher similarity in the detection results by rearranging the order of related images, so as to accurately realize the copy detection of local information of an image. Compared with A Self-Supervised Descriptor for Image Copy Detection, this paper achieves a significant improvement on the DISC2021 dataset, with 10% increase in localized copy image recognition accuracy, and 10% - 15% increase in μ AP (Micro Average Precision) metrics. In addition, this paper also carries out an intuitive qualitative analysis using heat maps, which confirms that ViT is able to acutely capture the region of the image where local copies exist, further validating the effectiveness of the algorithm. The ViT-based localized image copy detection reordering algorithm proposed in this paper is able to effectively detect copy images using picture-in-picture image enhancement, and achieves excellent results on the DISC2021 dataset, which expands new ideas for research in the field of digital image forensics.

Key words image copy detection; ViT; image enhancement; reordering; heat map

通信作者: 王总辉, 博士, 高级工程师, Email: zhwang@zju.edu.cn。

本课题得到国家重点研发计划项目(No. 2020AAA0140000)和浙江省教育厅一般科研项目(No. Y202351231)资助。

收稿日期: 2024-01-15; 修改日期: 2024-04-19; 定稿日期: 2025-10-14

1 引言

近年来, 社交媒体上的图像和视频内容广泛传播, 在丰富了人们日常生活的同时, 也为数据信息的传播和存储提供了保障。然而, 这导致了图像数据内容的盗取、复制和篡改行为, 给知识产权保护带来了挑战。如何有效维护版权所有者的隐私和利益已成为亟待解决的问题。

在数字知识产权保护领域, 图像拷贝检测^[1-4]是一项重要方法。这种基于内容的检测方法无须嵌入任何标记或水印, 通过提取图像的颜色、纹理等特征信息, 并将其转化为唯一的描述符来验证。类似于实例级图像检索^[5-8], 图像拷贝检测技术通过比较查询图像描述符和预先提取的数据库中的参考图像描述符来判断相似度, 从而识别图像中存在的拷贝。

目前的图像拷贝检测方法主要分为传统方法和基于深度学习的方法。传统的图像拷贝检测方法通常基于局部手工制作的图像特征, 如尺度不变特征变换^[9]、基于 SIFT 的主成分分析^[10]、加速鲁棒特征^[11]和视觉词袋^[12]等。传统方法通常需要从每张图像中提取大量的高维手工特征, 再对这些特征进行逐一匹配, 这一过程非常耗时。

随着深度学习的迅速发展, 从神经网络, 特别是卷积神经网络(CNN)中提取的特征已经被证明在图像拷贝检测算法中优于手工制作的特征。Liu 等人^[13]提出了一种基于空间域和频域几何不变矢量距离的有效图像哈希方法, 将图像随机划分成多个重叠块并提取特征矩阵, 利用特征矩阵中向量之间的不变性距离, 形成一个紧凑的哈希码。Liang 等人^[14]设计了一种基于旋转不变性的低维特征提取模型, 将图像提取为一个低维的特征向量, 并利用距离向量构造具有较好鲁棒性和紧凑型的哈希短序列进行图像之间的比较。Zhou 等人^[15]提出了一种基于上下文哈希嵌入的局部 CNN 特征图像复制检测方法, 在由 CNN 提取的图像局部特征周围相对较大的区域捕获上下文哈希序列。Aria 等人^[16]提出了一种基于深度学习方法的图像质量独立方法, 能够同时检测多个伪造区域, 以及确定伪造的来源和目标。Wang 等人^[17]提出了一个检测和定位图像的模型, 提取图像的高频特征, 用于捕捉在 RGB 域中不再可见的细微操作痕迹。Pizzi 等人^[18]设计了一种建立在自监督对比训练基础上的模型, 通过引入微分熵正则化等优化方式, 在 DISC2021 数据集上效果显著。

现有的深度学习方法主要专注于优化图像的全局特征, 因此在图像局部信息的拷贝检测方面存在

一定的限制。Zhou 等人^[15]在提取图像局部特征时, 对每张图像的局部特征进行分别提取, 但未考虑到不同图像比较时局部特征之间的差异, 导致在局部图像检测任务中提取的局部特征不够合理。

本文针对现有工作在局部图像拷贝数据集上准确率不足的问题, 提出了一种基于 ViT 的局部图像拷贝检测重排序算法。传统的卷积神经网络提取的全局特征包含了整张图像的信息, 但在局部图像拷贝检测任务中, 拷贝区域只占图像的一小部分。这导致全局特征包含了大量不必要的信息, 在计算时浪费资源, 且影响检测算法的准确性。因此, 本文利用 ViT 中的注意力机制对全局特征进行筛选和提取, 得到更能代表拷贝区域的局部特征, 并利用这些局部特征指导模型进行图像重排序, 从而提高局部图像拷贝检测任务的准确性。

本文主要贡献有以下几点: 1) 在基于自监督描述符的图像拷贝检测模型(A Self-Supervised Descriptor for Image Copy Detection, SSCD)^[18]的基础上, 引入了 ViT 模块。该模块根据不同图像对之间的联系, 对从 ResNet50 中提取的全局特征进行进一步提取, 得到更具有代表性的局部特征。2) 利用得到的局部特征对 SSCD 算法初步排序结果中排名靠前的图像进行重排序, 从而提高局部图像拷贝检测的准确率。3) 通过可视化分析结果, 证实了添加的 ViT 模块能够关注图像局部拷贝的位置, 具备一定的可解释性。实验结果表明, 本文算法在局部图像拷贝检测方面优于现有的 SSCD 算法。

2 相关工作

2.1 基于 SSCD 模型的图像拷贝检测算法

SSCD 是一种建立在最新的自监督对比训练目标基础上的模型。它通过改变架构和训练目标, 将实例匹配领域中的池化算子纳入其中, 并将对比学习调整为结合图像的增强, 从而适用于图像拷贝检测任务。图 1 是 SSCD 模型的基本框架图。

在训练过程中, 首先对原始图像进行随机增强, 模拟实际环境中的图像拷贝手段。随后, 对增强后的图像进行随机裁剪和正则规范化, 为输入数据做准备。接着, 使用 ResNet50 卷积神经网络提取图像特征, 然后通过广义池化层、全连接层和 L2 规范化层对这些特征进行处理。最后, 将处理后的特征输入损失函数进行计算, 并进行反向传播, 以完成整个模型的训练。

在验证过程中, 需要对查询图像和参考图像进行特征提取。首先, 经过简单的裁剪和规范化处理

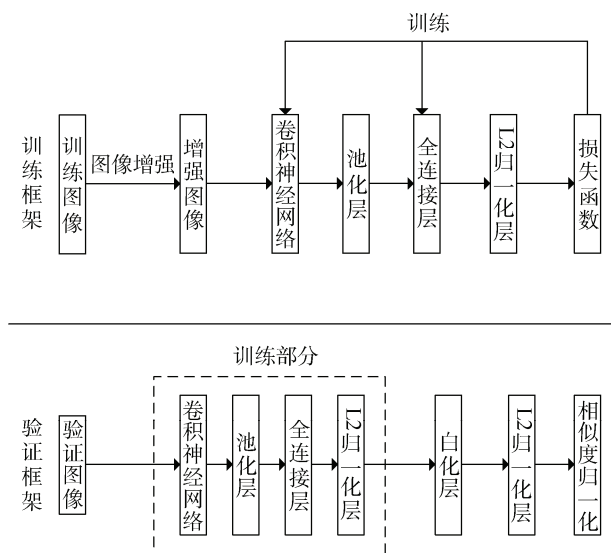


图 1 SSCD 模型的基本框架图

Figure 1 Basic framework diagram of the SSCD model

后, 这些图像被输入到 ResNet50 卷积神经网络中, 以提取主干特征。然后, 对这些特征张量进行池化、正则化和白化等数据处理操作, 得到经过处理的特征张

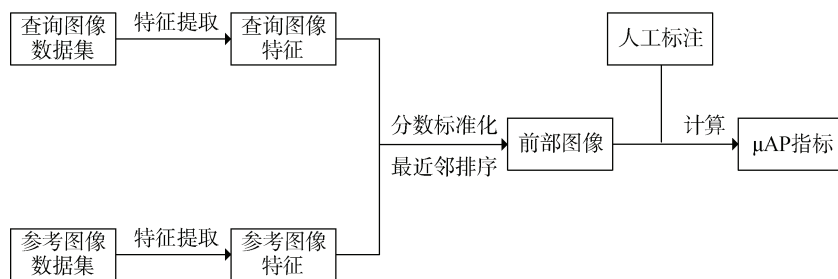


图 2 SSCD 模型的验证过程图

Figure 2 Diagram of the validation process of the SSCD model

2.3 注意力机制

注意力机制是当前研究的热门方向, 模拟人类观察事物的视觉特点, 尝试为图像的各个部分分配不同的关注度, 以更好地捕捉图像的各部分信息。本文采用的是 Vision Transformer(ViT)模型来对 SSCD 模型进行改进。在过往, Transformer 在自然语言处理(NLP)取得了巨大的成功, 但是在计算机视觉领域仍受到限制。随后, Dosovitskiy 等人^[32]对 Transformer 进行改进, 并取得了巨大的成就。ViT 受到 NLP 领域的启发, 将图像序列化分成小块, 并为每个块加上位置信息, 然后将它们连接起来形成类似于语言的序列。ViT 还借鉴了 BERT(Bidirectional Encoder Representations from Transformers)^[33]的方式, 定义了一个额外的输出层, 用于对图像对进行相似性分析^[34], 以便在比较不同图像对时确定相互关注的区域。为

量。随后, 通过合并的方式, 将 score normalization 操作与特征张量进行结合处理, 利用最近邻排序的方法, 在引用数据集中检索出与之相似度最高的前部图像。最后, 将这些检索结果与人工标注结果进行比较, 计算得出 μAP 分数。SSCD 模型的验证过程如图 2 所示。

2.2 图像重排序

在一些传统的项目^[19-22]中, 几何验证被视为主要的图像重排序方法, 并被广泛应用。后来, 受到文本检索方法的启发, 查询技术^[23-26]也被引入图像检索中。与几何验证不同, 查询技术在验证过程中分析每张查询图像的局部最近邻图, 依赖于这些局部关系来完成重排序。此外, 还有基于扩散^[27-31]的方法, 通过在全局亲和图上进行相似性传播来学习数据流形的结构, 从而更好地捕捉图像之间的内在关系和结构。总体而言, 图像重排序旨在通过利用相关技术细化检索结果, 优化图像的重排序结果。本文的重点是学习图像对之间的相似性, 并通过这种相似性进一步细化图像特征; 再利用这些细化后的特征, 对粗排阶段的检索结果进行重新排序, 使得与查询图像更相关的图像能够在最终排名中获得更高的位置。

为了方便统计和分割输出结果, 引入了<CLS>和<SEP>两个特殊的 token, 在图像数据输入时与图像的 patches 相连接。在本文实验中, <CLS>用来总结两幅图像之间的相关信号, <SEP>则用于标明两幅图像之间的分隔。

3 本文算法

3.1 总体框架

本文以 SSCD 模型预训练的 ResNet50 作为全局特征提取器, 通过增加卷积层 Conv2d 和 ViT 来提取局部特征, 并对全局特征进行降维和注意力权重的重新分配。这样的设计能够优化全局特征的初步排序结果。注意力权重指的是每个特征维度的重要程度。在全局特征中, 每个特征维度被视为同等重要, 但在局部拷贝图像中, 局部拷贝区域的特征应该更受重视。因此, 重新分配注意力权重的目的是增加表

示局部拷贝区域的特征权重,同时减少其余非拷贝区域的特征权重。总体框架如图 3 所示。

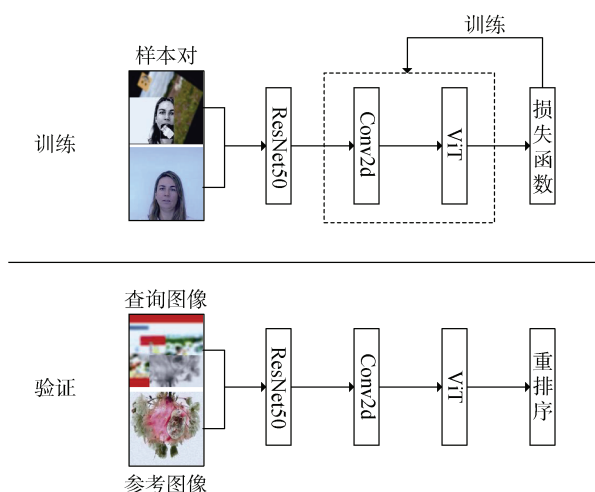


图 3 总体框架图

Figure 3 General framework diagram

首先,本文选用 SSCD 中预训练完成的 ResNet50 作为主干网络来提取图像的全局特征。这一选择不仅是因为 ResNet50 在特征提取方面表现良好,而且方便与基准模型 SSCD 进行比较。

其次,为了减少 ViT 模型的算力成本开销,需要对 ResNet50 提取的全局特征进行降维处理。本文采

用二维卷积层 Conv2d 对全局特征进行降维处理。Conv2d 作为卷积神经网络的一部分,不仅能减少耗时,同时与生物视觉传导相似。尽管特征降维可能会对准确率产生影响,但 Conv2d 已被证明能够考虑到图像各个部分特征之间的关系,构建高维结构化特征。

然后,针对局部拷贝图像,在进行图像比较时应着重比较其拷贝的部分。因此,本文采用 ViT 将提取的特征聚焦在拷贝区域。与 CNN 相比,ViT 依靠出色的建模能力,在 ImageNet、COCO 和 ADE20k 等多个基准数据集上达到了非常优异的效果,并且在迁移学习任务中显示出比 CNN 更优的性能。ViT 的可扩展性也满足了本文对图像对进行相同处理的需求。

3.2 图像增强方法

图 4 展示了本文使用的图像增强操作。除了最常见的图像缩放、图像水平翻转以及灰度图操作(后续将称为原始方法),本文还根据 DISC2021 数据集的处理方式,添加了图像模糊、图像随机旋转、随机添加文字、随机添加表情、图像属性调整以及图像压缩(后续将称为新方法)。此外,本文最重要的增强方式是 Cutmix 画中画增强操作,它将图像的一部分剪切并添加到随机选择的背景图像中,从而构造出一幅具有画中画效果的局部拷贝图像。Cutmix 增强方式将在本文实验中得到广泛使用。

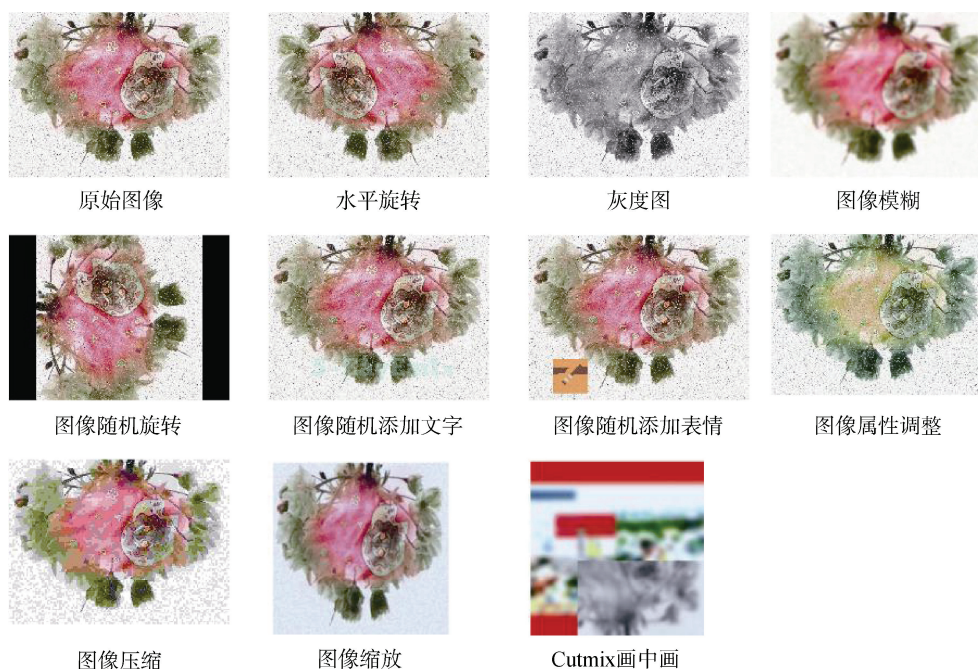


图 4 图像增强操作示意图

Figure 4 Schematic diagram of the image enhancement operation

值得一提的是,以上所述的增强操作是按照线性顺序逐步应用的,如图 5 所示,每个步骤都具有一

定的增强概率。这些不同的增强操作在图像上融合在一起,以增强模型拷贝检测的效果。

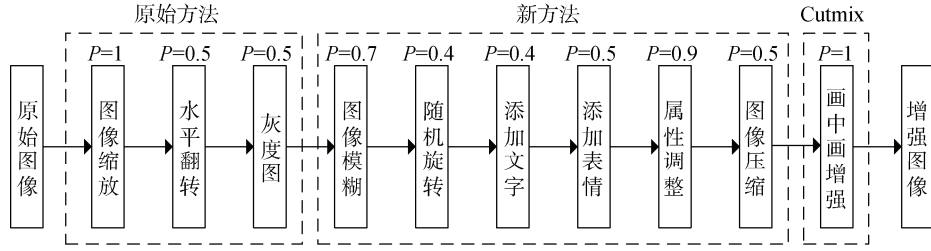


图 5 图像增强过程示意图

Figure 5 Schematic diagram of the image enhancement process

3.3 训练框架

由于 DISC2021 数据集缺乏相关标注且图像拷贝情况复杂, 本文采用自监督学习技术, 并通过数据增强来模拟实际的图像拷贝攻击。对于每张原始训练图像, 利用上述图像增强方式生成一张增强图像, 形成一组图像正样本对。由于训练集中不存在相似图像, 因此每张原始训练图像将与其他任意图像的增强图像形成多组图像负样本对。在实验中, 按照 1:1 比例构建图像正负样本对, 并输入到训练框架中进行训练。

如图 6 所示, 训练框架首先将原始训练图像经过上述的图像样本对构造后送入 SSCD 模型的主干卷积神经网络, 即 ResNet50, 以进行全局特征的提取。随后, 通过卷积层 Conv2d 对特征进行降维。降维后的图像对特征被输入到 ViT 模型中, 通过计算注意力权重捕捉图像对之间的关系。最后, 通过结合注意力权重和图像对的特征, 计算得到两幅图像之间的相似度分数 S_n 。其中, n 表示第 n 组图像对, S_n 表示对第 n 组图像对进行重排序之后的相似度。

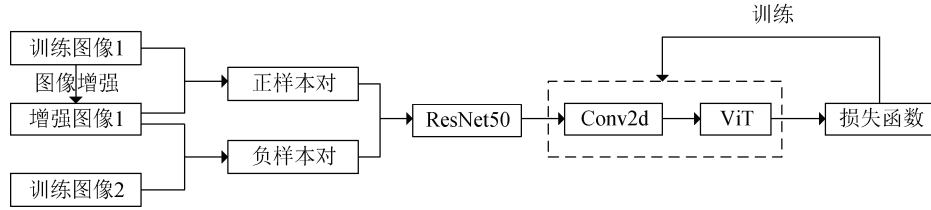


图 6 训练框架图

Figure 6 Training framework diagram

得到相似度分数 S_n 后, 将对损失函数进行计算。在构造正负样本对时, 会对图像对的正负标签进行标记, 即将正样本对标记为 1, 负样本对标记为 0, 用符号 M_n 表示, 这里 n 表示第 n 对图像。训练框架会利用相似度分数 S_n 和标签标记结果 T_n 进行损失函数的计算, 并进行梯度反向传播, 调整模型参数。这里采用的是二值交叉熵损失函数, 其主要用于多标签分类, 适用于此处的正负样本对分类任务。该函数利用二值交叉熵, 预测每一个样本的损失值并利用权重进行赋值, 具体形式如下:

$$l(S, M) = L = l_1, \dots, l_N^T \quad (1)$$

其中

$$l_n = -w_n [M_n \cdot \log(S_n) + (1 - M_n) \cdot \log(1 - S_n)] \quad (2)$$

其中, S_n 是模型输入的分值, M_n 是正负样本对的标记, w_n 是样本在损失函数中所占的比重, $n \in \dots N$ 。

在训练过程中, 本文将直接利用 SSCD 训练完成的主干卷积神经网络 ResNet50 进行全局特征提取, 而不对其进行训练。本文只针对卷积层 Conv2d 和

ViT 两个模块进行训练。

3.4 验证框架

图 7 所示是验证框架, 各步骤解析如下。

(1) 根据查询图像和参考图像集获取原始 SSCD 模型得到的初步排序结果, 并得出前部参考图像集。考虑到验证部分不需要进行数据增强操作, 框架直接对查询图像 Q 和参考图像 $R_1, \dots, R_n$ 分别调整大小, 并进行规范化。随后将两者输入到原始 SSCD 模型预训练过的 ResNet50 网络进行特征提取, 得到全局特征描述符张量, F_Q 以及 $F_{R_1}, \dots, F_{R_n}$ 。将特征 F_Q 以及 $F_{R_1}, \dots, F_{R_n}$ 经过 GeM Pooling 层、全连接层以及 L2 规范化层等处理得到最终特征张量。利用查询图像特征和参考图像特征进行图像距离的计算, 并进行 KNN 最近邻排序, 得到查询图像 Q 对引用数据集中 $R_1, \dots, R_n$ 图像的排序情况。根据排序情况得到排名靠前的前部参考图像集。

(2) 将 SSCD 模型排序得到的前部参考图像集

与查询图像利用预训练过的 ResNet50 网络进行特征提取, 分别得到全局特征描述符张量 F_Q 以及

$F_{R_1}, \dots, F_{R_k}$, k 值是保存的前部重排序范围, 也是前部参考图像集的图像数量。

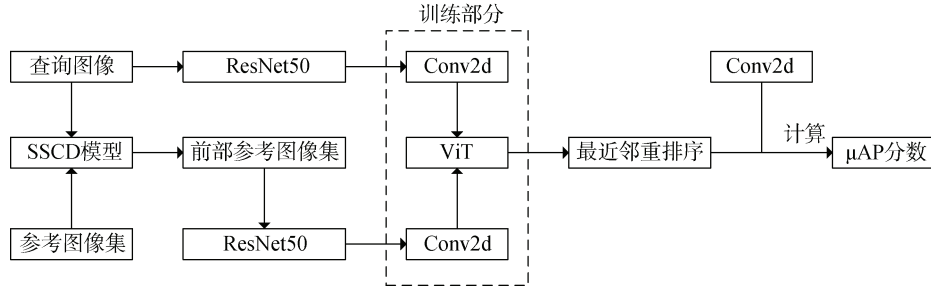


图 7 验证框架图

Figure 7 Validation framework diagram

(3) 利用卷积层 Conv2d 对(2)中得到的全局特征描述符张量进行降维。将降维得到的查询图像特征和参考图像特征, 以图像特征对的形式共同输入到 ViT 模型中进行注意力权重的分配。图 8 是 ViT 的工作原理, 本文引入一个<CLS>的特殊标记, 它用于表征两幅图之间的注意力权重信号; 还引入一个<SEQ>的标记用于分隔两幅图像。将特征进行扁平序列化后, 再将该特征和<CLS>、<SEQ>进行线性链接共同输入到 ViT 中。在 ViT 中, 特征首先经过多头注意力层, 由<CLS>标记来表征两幅图像比较过程中注意力机制的权重。具体计算公式如下:

$$T^1 w_1^1 w_2^1 \dots w_n^1 T^2 w_1^2 w_2^2 \dots w_n^2 = \text{ViT}(\langle \text{CLS} \rangle x_1^1 x_2^1 \dots x_n^1 \langle \text{SEQ} \rangle x_1^2 x_2^2 \dots x_n^2) \quad (3)$$

其中, x_i^1 是序列化的查询图像特征, x_i^2 是序列化的参考图像特征, w_i^1 是指对应 x_i^1 特征所占的权重, w_i^2 是指对应 x_i^2 特征所占的权重, T^1 和 T^2 作为总结以及分隔的符号并不参与注意力权重的重新分配。

(4) 在 ViT 模型后续的残差网络计算中, 借助注意力机制权重, 计算出两幅图像之间的相似度分数 x_i 。然后, 利用相似度分数 x_i 和引用数据集中参考图像 R_i 之间的对应关系, 将参考图像集进行重排序。由于 ViT 模型的复杂性较高, 直接通过 ViT 模型提取特征将严重影响验证框架的效率。因此, 在验证框架中, 重排序操作旨在保证模型准确率的前提下有效提高框架的效率。

(5) 在得到查询图像对应的前部参考图像集新排序后, 根据重新排序的结果和人工标注的结果, 进行 μAP 指标的计算。

4 实验结果与分析

4.1 实验数据集和评价指标

本文所采用的数据集 DISC2021 是一种大规模

图像相似度检测的基准测试数据集。原数据集包含 50,000 张查询图像、1,000,000 张参考图像以及 1,000,000 张训练图像。其中, 参考图像包含了从查询图像中得到的 10,000 张拷贝图像。训练数据集和引用数据集属于同一数据分布, 具有良好的关联性。然而, 两者不包含完全相同的图像, 且训练数据集上也不含任何标注和副本, 以确保模型的真实性和真实性。

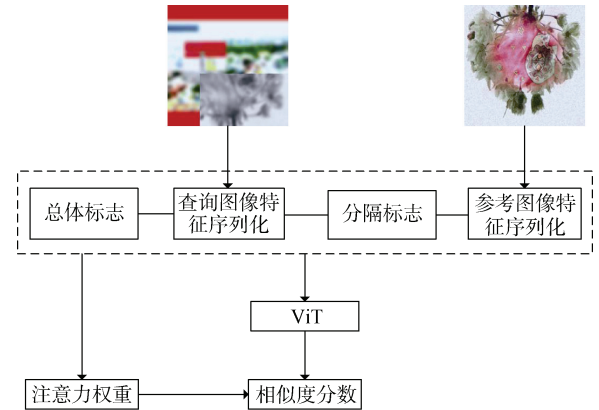


图 8 ViT 工作原理

Figure 8 ViT working principle

考虑到 DISC 数据集中局部拷贝情况较少, 本文在重新构建了实验数据集。训练图像数据集从 DISC2021 的训练图像数据集中随机选取了 100,000 张图像, 参考图像数据集则由 DISC2021 原有的 50,000 张查询图像构成。查询图像数据集则通过对 DISC2021 原有的 50,000 张查询图像进行图像增强处理, 按照 1:1 的比例, 将一部分加入到无关的背景图中, 生成了 50,000 张参考图像。如图 9 所示, (a)是原始图像, (b)是 Advanced 增强图像, (c)是将图像(b)的一部分添加到随机背景图中构成的 Advanced + Cutmix 增强的图像。最后, 对新构建的查询图像数据集和参考图像数据集进行了人工标注, 确保了查

询图像和参考图像之间的一一对应关系。

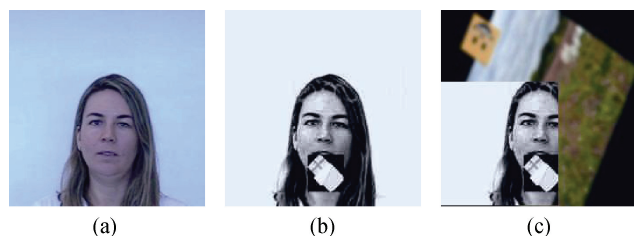


图 9 查询图像构造示意图

Figure 9 Schematic diagram of query image construction

在图像拷贝检测领域, μAP 指标是最重要的评价指标, 全称为 Micro Average Precision。该指标能够展现模型在图像拷贝检测任务中的整体效果。然而, 考虑到计算所有结果的计算量过大, 且不符合重排优化的情况, 因此本论文引入 $\mu\text{AP}@k$ 指标来替代。 $\mu\text{AP}@k$ 和 μAP 的差别在于, 每次只取单张查询图片的 Top k 个结果图像来进行排名和计算, 其他计算过程均相同。同时, 下文所有提及的 μAP 之处都视为 $\mu\text{AP}@k$ 。

μAP 的计算公式如下:

$$\mu\text{AP} = \sum P_i \times R_i \quad (4)$$

其中, P_i 为准确度, R_i 为召回率, i 是图像排名编号。

为展示优化结果, 本文将对每张查询图像得到的前 k 个可能存在拷贝的图像进行人工标注结果比对。通过统计比对成功的图像(即真正的拷贝图像)在不同排名名次上的数量, 并根据排名第一的图像数量来计算准确度, 以实现直观的对比分析。

4.2 对比实验

对比实验采用 3 层 ViT 进行实验。在训练过程中, 本实验采用 AdamW 的优化器以及 LinearWarmup CosineAnnealingLR 的调度器, 学习率设置为 0.001, Weight decay 设置为 0.0001, 同时采用 10 个 Epoch 的训练周期, 以及 200 的批处理大小来进行训练。根据上述查询图像数据集的构造过程, 训练中对每一张训练图像用 Advanced+Cutmix 的线性增强方式进行增强, 并使用原图像和增强图像等比例地构造训练图像对作为模型训练输入。

在验证过程中, 每张查询图像和与之对应的待检测参考图像数据集在经过 SSCD 模型的特征提取后, 计算出一个相似度分数, 然后根据相似度分数从大到小排序, 得到与之对应的一个参考图像数据集的内部排名。实验中, 将重排序的范围限定为每张查询图像对应内部排名前 10 的参考图像。接着, 将

该查询图像和这些排名前 10 的参考图像组成图像对, 输入到卷积层进行降维, 再输入到 ViT 中进行注意力机制权重的计算。得到的特征再进行相似度计算, 按照分数从大到小, 重新排序排名前 10 的参考图像。随后, 分别取 Top 1 至 Top 10 的图像来进行 μAP 指标的计算。此外, 还将对每张查询图像检测后得到的前 10 个可能存在拷贝情况图像进行人工标注结果比对, 试图找到查询图像在参考图像数据集中真正对应的拷贝图像, 并记录和统计对应图像所在的排名, 再与原始 SSCD 模型的结果进行对比。结果表明, 在 3-layer ViT 的帮助下, SSCD 模型检测出正确的对应拷贝图像在前 10 范围内排位有所提前, 尤其是对应拷贝图像位于排名第一的总数在原有基础上增加了近 4000 张。表中的“other”一栏表示未能参与重排序的图像数目, 即查询图像对应的参考图像数据集在经过 SSCD 模型初步排序后, 其对应拷贝图像位于第 10 名以后的数目。

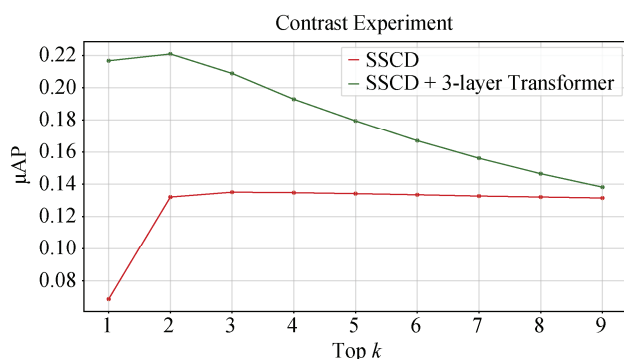
表 1 查询图像在模型改进前后检测出对应拷贝图像排名的分布表

Table 1 Distribution of the ranking of the query image before and after the model improvement to detect the corresponding copy images

排名	SSCD	SSCD+3-layer ViT
1	12776	16714
2	9088	4441
3	1601	2171
4	819	1239
5	616	841
6	461	507
7	371	321
8	337	202
9	314	125
10	246	68
other	23371	23371

图 10 反映了模型改进前后的 μAP 指标对比。从曲线可以看出, 改进的 SSCD+3-layer ViT 新模型在 μAP 指标上始终高于原有的 SSCD 模型。说明了在进行 Cutmix 图像增强操作的情况下, 新模型拥有比原有模型更好的效果, 进而证明 ViT 框架在经过 Cutmix 增强操作的图像排序中起到重要作用。

在本次对比实验中, 采用的是 3 层 ViT 进行训练, 并且取得了一定效果。下面, 将就 ViT 层数问题展开消融实验, 探究在相同训练条件下, 不同层数的 ViT 对重排序结果的影响。

图 10 模型改进前后的 μAP 指标对比图Figure 10 Comparison of μAP metrics before and after model improvement

4.3 消融实验

这里的训练同样采用对比实验中相同的参数设定, ViT 的层数设定为 1~6。首先, 观察训练过程中不同层数的 ViT 的 Loss 变化。从图 11 中可以发现, 当层数较少的时候, Loss 更容易收敛; 而当层数增大时, Loss 的下降需要在学习率变小时才能进行。同时可以观察到, 对于 6 层 ViT, 层数过多发生了过拟合现象, 使它的 Loss 值无法降低到一个很好的水平, 模型无法完成一个很好的训练, 因此, 后续将不使用 6 层 ViT 的模型进行数据对比。对于层数较少的 ViT, 在相同的训练条件下, 虽然 Loss 衰减更加迅速, 但它的收敛值相较于其他比它层数多的 ViT 收敛值会略高一些。

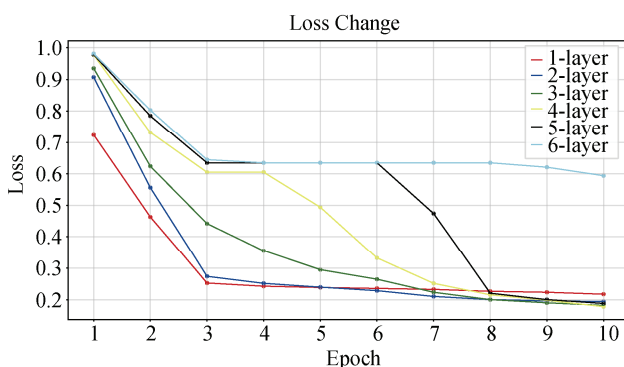
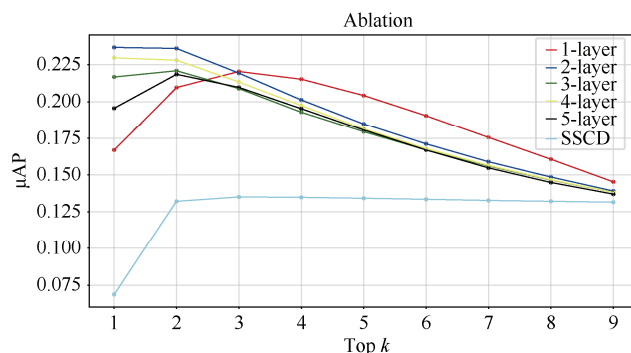


图 11 不同层数的 ViT 训练时, Loss 随 Epoch 的变化图

Figure 11 Variation of Loss with Epoch for ViT training with different number of layers

图 12 所示是 μAP 值随所取 Top k 范围的变化规律。由图可知, 1~5 层的 ViT 都能够起到优化的作用, 其中 1 层由于参数较少, 可能训练模型时无法进行正常优化, 发生欠拟合现象。至于 2~5 层, 在对比实验构造的数据集上都能够有一个比较稳定的提升,

尤其是 2 层表现最为突出, 其原因应该是和训练的过程合适程度较高。至于 6 层的 ViT 模型, 在目前实验设定的训练条件下, 难以实现有效的训练, 因此无法和 1~5 层的 ViT 模型以及原有 SSSD 模型(图中为 0-layer)的结果进行比较。

图 12 μAP 值随所取 Top k 范围的变化规律图Figure 12 Plot of the pattern of variation of μAP values with the range of Top k taken

同时, 实验还观察了不同层数的 ViT 相对于 SSSD 模型(0 层 ViT)之间的排名分布情况。表 2 中的数据可以发现, 除 1 层外, 各层检测出对应拷贝图像排名第 1 的数量都相较于 SSSD 模型有明显提升, 由此可以证明, ViT 在 Cutmix 增强操作过的数据集上确实可以提高检测出对应拷贝图像的准确率。从表 3 的数据也不难看出, 除 1 层外, 各层相对于 SSSD 模型检测出对应拷贝图像排名第 1 的准确率都提高了 5%~10%, 其中 2~5 层的提高效果最为明显, 排名第 1 的图像相比于 SSSD 模型提高了 3000~5000 张。同时, 从总体的趋势上看, 使用了 ViT 的模型检测出对应拷贝图像位于排名 7~9 的数量都相较于 SSSD

表 2 使用不同层数 ViT 检测出的排名结果表

Table 2 Table of ranking results detected using different layers of ViT

排名	0 层	1 层	2 层	3 层	4 层	5 层
1	12776	11866	17666	16714	17489	14996
2	9088	5454	4263	4441	4247	5243
3	1601	3304	1875	2171	2029	2604
4	819	2107	1079	1239	1203	1526
5	616	1388	647	841	688	967
6	461	1037	450	507	402	565
7	371	696	309	321	259	331
8	337	459	190	202	159	216
9	314	216	82	125	102	138
10	246	102	41	68	51	43
other	23371	23371	23371	23371	23371	23371

表 3 使用不同层数 ViT 检测出对应拷贝图像排名第 1 的准确率结果表

Table 3 Table of accuracy results for detecting the corresponding copy image ranked 1st using different layers of ViT

模型	准确率
0 层	25.50%
1 层	23.70%
2 层	35.30%
3 层	33.40%
4 层	35.00%
5 层	30.00%

模型都有所减少。就 1 层而言, 在排名方面与 SSCD 相比并没有明显优势, 然而在 μ AP 指标方面却呈现出相对较大的提升。这表明在各个排名区间内, 检测出的拷贝图像置信度分值较高, 使得在阈值排名时能够获得更好的排位, 进而整体效果相对于 SSCD 有所提升。

在消融实验中, 2~5 层表现出较好的效果。相比之下, 1 层网络规模较小, 容易过早收敛, 而过多的训练可能导致过拟合。另外, 对于 6 层而言, 其网络规模相对于本实验的训练规模较大, 因此很难获得良好的训练效果。结果表明, 在既定实验条件下, 不同层数的 ViT 对实验结果并没有产生很大的影响, 它们都能对现有模型实现很好的优化效果。

4.4 可视化实验

为了进一步验证注意力机制是否起到相应的作用, 实验随机选取了一张图片, 根据 ViT 中多头注意力机制层提供的注意力权重矩阵数值来进行注意力机制热力图的绘制。

图 13 是采用训练完成的 1 层 ViT 得到的注意力机制可视化结果。图中可以观察到, 图(b)存在图(a)的拷贝部分, 图(b)右下角的部分图像是图(a)左上角的图像经过增强处理后得到的结果。在图(d)中可以看到, 注意力机制很好地注意到了右下角的部分, 同时图(c)的左上角部分也被注意力机制所察觉。

5 结论

本文对基于注意力机制的局部图像拷贝检测方法进行了研究。以在 DISC2021 数据集上表现良好的 SSCD 模型为基础, 借助基于注意力机制的 ViT 框架, 对 SSCD 模型进行改进。通过注意力机制对 SSCD 模型产生的全局特征进行进一步的特征提取得到局部特征, 根据局部特征的计算结果实现对 SSCD 模型排序结果的重排序操作。本文主要针对注意力机

制的特性设计了对比实验, 用 3 层的 ViT 来检验优化情况, 总体上可以使排名结果产生一个较为良好的提升。同时, 本文还设计了关于 ViT 不同层数的消融实验, 对 ViT 的机制进行进一步的研究, 都得到了不错的效果。

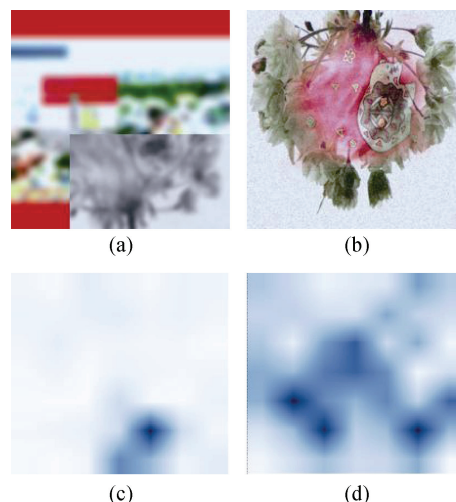


图 13 图像与其对应的热力图
Figure 13 Image with its corresponding heat map

本文认为将来可以在以下个方面进行更加深入的研究。其一, 可以将注意力机制引入 SSCD 模型的预训练过程中, 实现端到端的预训练, 从而去除耗时的重排序过程。其二, 可以探索更轻量级的图像拷贝检测模型, 并考虑如何将其扩展到视频拷贝检测问题。

参考文献

- [1] Kim C. Content-Based Image Copy Detection[J]. *Signal Processing: Image Communication*, 2003, 18(3): 169-184.
- [2] Zhou Z L, Wang Y L, Jonathan Wu Q M, et al. Effective and Efficient Global Context Verification for Image Copy Detection[J]. *IEEE Transactions on Information Forensics and Security*, 2017, 12(1): 48-63.
- [3] Zhou Z L, Jonathan Wu Q M, Sun X M. Multiple Distance-Based Coding: Toward Scalable Feature Matching for Large-Scale Web Image Search[J]. *IEEE Transactions on Big Data*, 2021, 7(3): 559-573.
- [4] Zhou W G, Li H Q, Lu Y J, et al. SIFT Match Verification by Geometric Coding for Large-Scale Partial-Duplicate Web Image Search[J]. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2013, 9(1): 1-18.
- [5] Amato G, Carrara F, Falchi F, et al. Large-Scale Instance-Level Image Retrieval[J]. *Information Processing & Management*, 2020, 57(6): 102100.
- [6] Bai C, Li H K, Zhang J L, et al. Unsupervised Adversarial Instance-Level Image Retrieval[J]. *IEEE Transactions on Multimedia*, 2021, 23: 2199-2207.

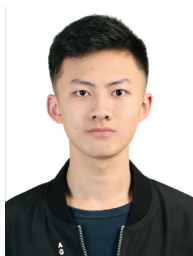
- [7] Li K, Pang K Y, Song Y Z, et al. Fine-Grained Sketch-Based Image Retrieval: The Role of Part-Aware Attributes[C]. *2016 IEEE Winter Conference on Applications of Computer Vision*, 2016: 1-9.
- [8] Hwan S, Joo H, Avrithis Y. All the Attention You Need: Global-Local, Spatial-Channel Attention for Image Retrieval[C]. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022: 439-448.
- [9] Lowe D G. Distinctive Image Features from Scale-Invariant Key-points[J]. *International Journal of Computer Vision*, 2004, 60(2): 91-110.
- [10] Ke Y, Sukthankar R. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors[C]. *The 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, 2004: II.
- [11] Bay H, Ess A, Tuytelaars T, et al. Speeded-up Robust Features (SURF)[J]. *Computer Vision and Image Understanding*, 2008, 110(3): 346-359.
- [12] Sivic Z. Video Google: A Text Retrieval Approach to Object Matching in Videos[C]. *Proceedings Ninth IEEE International Conference on Computer Vision*, 2008: 1470-1477.
- [13] Liu S G, Huang Z Q. Efficient Image Hashing with Geometric Invariant Vector Distance for Copy Detection[J]. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2019, 15(4): 1-22.
- [14] Liang X P, Tang Z J, Huang Z Q, et al. Efficient Hashing Method Using 2D-2D PCA for Image Copy Detection[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2023, 35(4): 3765-3778.
- [15] Zhou Z L, Wang M M, Cao Y, et al. CNN Feature-Based Image Copy Detection with Contextual Hash Embedding[J]. *Mathematics*, 2020, 8(7): 1172.
- [16] Aria M, Hashemzadeh M, Farajzadeh N. QDL-CMFD: A Quality-Independent and Deep Learning-Based Copy-Move Image Forgery Detection Method[J]. *Neurocomputing*, 2022, 511: 213-236.
- [17] Wang J K, Wu Z X, Chen J J, et al. ObjectFormer for Image Manipulation Detection and Localization[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 2354-2363.
- [18] Pizzi E, Roy S D, Ravindra S N, et al. A Self-Supervised Descriptor for Image Copy Detection[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 14512-14522.
- [19] Philbin J, Chum O, Isard M, et al. Object Retrieval with Large Vocabularies and Fast Spatial Matching[C]. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007: 1-8.
- [20] Siméoni O, Avrithis Y, Chum O. Local Features and Visual Words Emerge in Activations[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 11643-11652.
- [21] 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*, 2017: 3476-3485.
- [22] Cao B Y, Araujo A, Sim J. Unifying Deep Local and Global Features for Image Search[C]. *Computer Vision – ECCV 2020*, 2020: 726-743.
- [23] Chum O, Philbin J, Sivic J, et al. Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval[C]. *2007 IEEE 11th International Conference on Computer Vision*, 2007: 1-8.
- [24] Chum O, Mikulik A, Perdoch M, et al. Total Recall II: Query Expansion Revisited[C]. *CVPR 2011*, 2011: 889-896.
- [25] Tolias G, Jégou H. Visual Query Expansion with or without Geometry: Refining Local Descriptors by Feature Aggregation[J]. *Pattern Recognition*, 2014, 47(10): 3466-3476.
- [26] Gordo A, Radenovic F, Berg T. Attention-Based Query Expansion Learning[C]. *Computer Vision – ECCV 2020*, 2020: 172-188.
- [27] Donoser M, Bischof H. Diffusion Processes for Retrieval Revisited[C]. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 1320-1327.
- [28] Zhang S T, Yang M, Cour T, et al. Query Specific Rank Fusion for Image Retrieval[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(4): 803-815.
- [29] Iscen A, Tolias G, Avrithis Y, et al. Efficient Diffusion on Region Manifolds: Recovering Small Objects with Compact CNN Representations[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 926-935.
- [30] Bai S, Tang P, Torr P H S, et al. Re-Ranking via Metric Fusion for Object Retrieval and Person Re-Identification[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 740-749.
- [31] Bai S, Bai X, Tian Q, et al. Regularized Diffusion Process on Bidirectional Context for Object Retrieval[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(5): 1213-1226.
- [32] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale[EB/OL]. 2020: arXiv: 2010.11929. <https://arxiv.org/abs/2010.11929>.
- [33] Devlin J, Chang M W, Lee K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding[EB/OL]. 2018: arXiv: 1810.04805. <https://arxiv.org/abs/1810.04805>.
- [34] Tan F W, Yuan J B, Ordonez V. Instance-Level Image Retrieval Using Reranking Transformers[C]. *2021 IEEE/CVF International Conference on Computer Vision*, 2022: 12085-12095.



朱辰 于 2013 年在浙江工业大学计算机应用技术专业获得硕士学位。现任浙江大学工程师学院工程师。研究领域为网络安全、物联网安全实践教学和科研。Email: zhuc@zju.edu.cn



陈于勋 于 2023 年在浙江大学计算机科学与技术专业获得工学学士学位。现在浙江大学计算机科学与技术学院电子信息专业攻读硕士学位。研究领域为人工智能。Email: 22321306@zju.edu.cn



陈禹坤 于 2022 年在浙江大学计算机科学与技术专业获得工学学士学位。现在浙江大学工程师学院电子信息专业攻读硕士学位。研究领域为网络安全、后门攻防。
Email: yukunchen@zju.edu.cn



王总辉 工学博士, 浙江大学计算机科学与技术学院高级工程师, 博士生导师。研究领域为内容安全、多媒体溯源、系统安全、计算机体系结构。Email: zhwang@zju.edu.cn