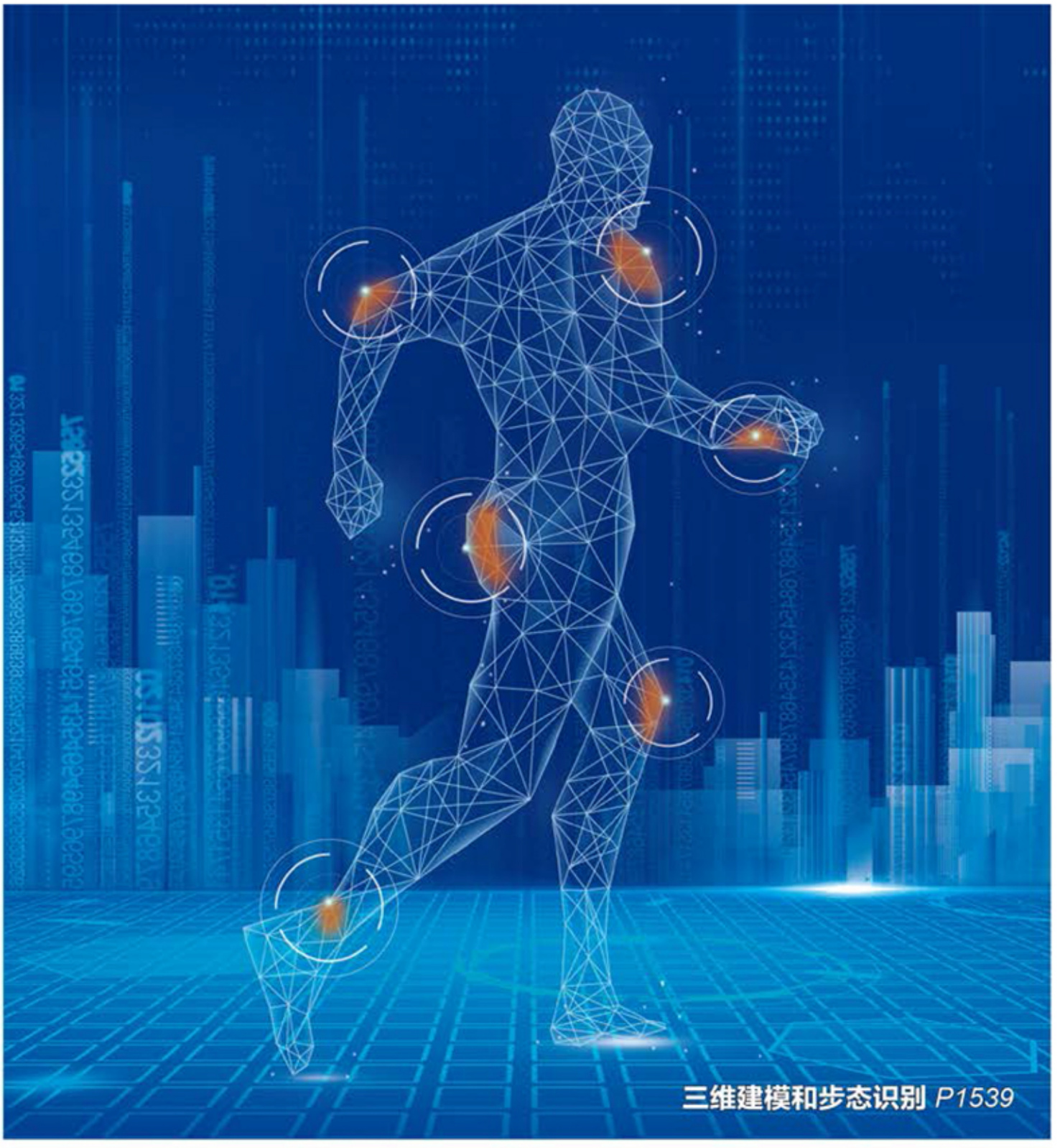
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2020
08
VOL.25

ISSN1006-8961
CN11-3758/TB



三维建模和步态识别 *P1539*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第25卷第8期（总第292期）
2020年8月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 顾行发
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Gu Xingfa
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

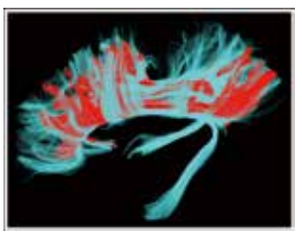
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

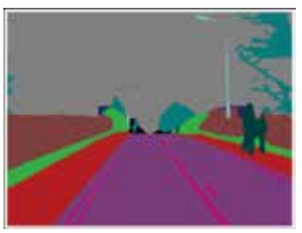
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2020年8月 第25卷 第8期（总第292期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

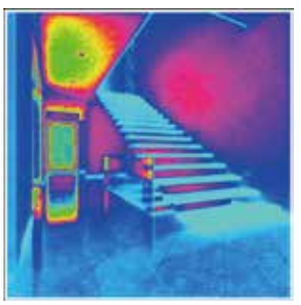
目次



神经纤维跟踪算法研究进展(第1513页)



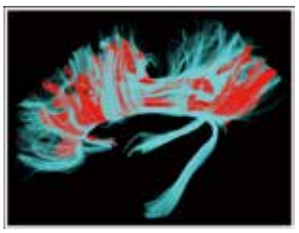
结合注意力机制的双路径语义分割(第1627页)



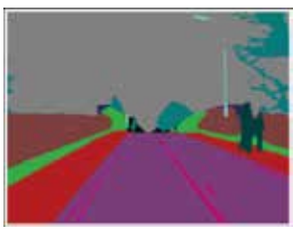
突变策略下多通路Metropolis光照传播(第1658页)

Volume 25, Number 8
Published August 16, 2020

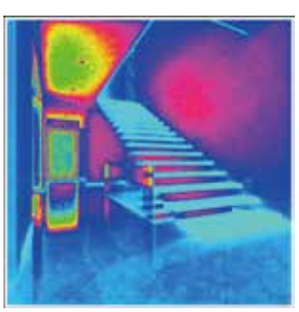
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Research progress of neural fiber tracking(P1513)



Two-path semantic segmentation algorithm combining attention mechanism(P1627)



Improved multiplexed Metropolis light transport based on fusion mutation strategy(P1658)

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2020)08-1567-11

论文引用格式:Mao X Y and Peng Y B. 2020. Landmark recognition based on ArcFace loss and multiple feature fusion. Journal of Image and Graphics, 25(08):1567-1577(毛雪宇,彭艳兵. 2020. 增量角度域损失和多特征融合的地标识别. 中国图象图形学报,25(08):1567-1577)[DOI:10.11834/jig.190418]

增量角度域损失和多特征融合的地标识别

毛雪宇[1,2],彭艳兵[2]

1. 武汉邮电科学研究院,武汉　430074;　2. 南京烽火天地通信科技有限公司,南京　210019

摘　要:**目的**　地标识别是图像和视觉领域一个应用问题,针对地标识别中全局特征对视角变化敏感和局部特征对光线变化敏感等单一特征所存在的问题,提出一种基于增量角度域损失(additive angular margin loss, ArcFace 损失)并对多种特征进行融合的弱监督地标识别模型。**方法**　使用图像检索取 Top-1 的方法来完成识别任务。首先证明了 ArcFace 损失参数选取的范围,并于模型训练时使用该范围作为参数选取的依据,接着使用一种有效融合局部特征与全局特征的方法来获取图像特征以用于检索。其中,模型训练过程分为两步,第 1 步是在谷歌地标数据集上使用 ArcFace 损失函数微调 ImageNet 预训练模型权重,第 2 步是增加注意力机制并训练注意力网络。推理过程分为 3 个部分:抽取全局特征、获取局部特征和特征融合。具体而言,对输入的查询图像,首先从微调卷积神经网络的特征嵌入层提取全局特征;然后在网络中间层使用注意力机制提取局部特征;最后将两种特征向量横向拼接并用图像检索的方法给出数据库中与当前查询图像最相似的结果。**结果**　实验结果表明,在巴黎、牛津建筑数据集上,特征融合方法可以使浅层网络达到深层预训练网络的效果,融合特征相比于全局特征(mean average precision,mAP)值提升约 1%。实验还表明在神经网络嵌入特征上无需再加入特征白化过程。最后在城市级街景图像中本文模型也取得了较为满意的效果。**结论**　本模型使用 ArcFace 损失进行训练且使多种特征相似性结果进行有效互补,提升了模型在实际应用场景中的抗干扰能力。

关键词:地标识别;增量角度域损失函数;注意力机制;多特征融合;卷积神经网络(CNN)

Landmark recognition based on ArcFace loss and multiple feature fusion

Mao Xueyu[1,2], Peng Yanbing[2]

1. Wuhan Research Institute of Posts and Telecommunications, Wuhan　430074, China;

2. Nanjing FiberHome World Communication Technology Co., Ltd., Nanjing　210019, China

Abstract: **Objective** Landmark recognition, which is a new application in computer vision, has been increasing investigated in the past several years and has been widely used to implement landmark image recognition function in image retrieval. However, this application has many problems unsolved, such as the global features are sensitive to view change, and the local features are sensitive to light change. Most existing methods based on convolutional neural network (CNN) are used to extract image features for replacing traditional feature extraction methods, such as scale-invariant feature transform (SIFT) or speeded up robust feature (SURF). At present, the best model is deep local feature (DeLF), but its retrieval needs the combination of product quantization (PQ) and K-dimensional (KD) trees. The process is complex and consumes

收稿日期:2019-08-11;**修回日期**:2019-12-13;**预印本日期**:2019-12-20

基金项目:国家重点研发计划项目(2017YFB1400704)

Supported by: National Key Research and Development Plan for China(2017YFB1400704)

approximately 6 GB of display memory, which is unsuitable for rapid deployment and use, and the most time-consuming process is random sample consensus. **Method** A multiple feature fusion method is needed when focusing on the problems of a single feature, and multiple features can be horizontally connected to create a single vector for improving the performance of CNN global features. For large-scale landmark data, manual labeling of images is time consuming and laborious, and artificial cognitive bias exists in labeling. To minimize human work in labeling images, weakly supervised loss, such as the additive angular margin loss function (ArcFace loss function), which is improved from standard cross-entry loss and changes the Euclidean distances to angular domain, is used to train the model in image-level annotations. The ArcFace loss function performs well in facial recognition and image classification and is easy to use in other deep learning applications. This paper provides the values of the parameters in ArcFace loss function and the proof process. Thus, a weakly supervised recognition model based on ArcFace loss and multiple feature fusion is proposed for landmark recognition. The proposed model uses ResNet50 as its trunk and has two steps in model training, including the trunk's finetuning and attention layer's training. Finetuning uses the Google landmark image dataset, and the trunk is finetuned on the weights pretrained on the ImageNet dataset. The average pooling layer is replaced by a generalized mean (GeM) pooling layer because it is proven useful in image retrieval. The attention mechanism is built using two convolutional layers that use 1×1 kernel to train the features focusing on the local features needed. Image preprocessing is required before training. The preprocessing consists of three stages, including center crop/resize and random crop. People usually prefer to place buildings and themselves in the center of images. Thus, a center crop method is suitable to ignore the problems occurring in padding or resizing. The proposed model uses classification training to complete the image retrieval task. The final input image size is set to 448^2. This value is a compromise value because the input image size in image retrieval is usually $800 \times 800 \sim 1\,500 \times 1\,500$ pixels and the classification size is $224 \times 224 \sim 300 \times 300$ pixels. The image is center cropped first, and then its size is resized to 500×500 pixels because it is a useful method to enhance the data through random cropping. For inference, the image is center cropped and directly resized to 448×448 pixels because it only needs to be processed twice. The inference of this model is divided into three parts, namely, extracting global features, obtaining local features, and feature fusion. For the inputted query image, the global feature is first extracted from the embedding layer of CNN fine-tuned by ArcFace loss function; Second, the attention mechanism is used to obtain local features in the middle layer of the network, and the useful local features must be larger than the threshold; finally, two features are fused, and the results that are the most similar with the current query image in the database are obtained through image retrieval. **Result** We compared the proposed model with several state-of-the-art models, including the traditional approaches and deep learning methods on two public reviewed datasets, namely, Oxford and Paris building datasets. The two datasets are reconstructed in 2018 and are classified into three levels, namely, easy, medium, and hard. Three groups of comparisons are used in the experiment, and they are all compared on the reviewed Oxford and Paris datasets. The first group is to compare the proposed model's performance with other models, such as HesAff-rSIFT-VLAD and VggNet-NetVLAD. The second group is designed to compare the performance of single global feature with the performance of fused features. The last group compares the results obtained from the whiting of the extracted features at different layers of the proposed model. Results show that the feature fusion method can make the shallow network achieve the effect of deep pretrained network, and the mean average precision (mAP) increases by approximately 1% compared with the global features on the two previously mentioned datasets. The proposed model achieves satisfactory results in urban street view images. **Conclusion** In this study, we proposed a composite model that contains a CNN, an attention model, and a fusion algorithm to fuse two types of features. Experimental results show that the proposed model performs well, the fusion algorithm improves its performance, and the performance in urban street datasets ensures the practical application value of the proposed model.

Key words: landmark recognition; additive angular margin loss function (ArcFace loss function); attention mechanism; multiple features fusion; convolutional neural network (CNN)

0 引　言

地标通常依据其形成方式分为自然景观和人造建筑，著名并且容易辨认的地标一直是人们旅游参观及拍照的重点。随着摄像终端和移动互联网的普及，包含著名地标的图像在互联网上已经到达不可忽视的数量并且大部分图像不包含 GPS 信息，在此情况下，地标识别受到国内外学者的普遍关注，因为使用地标识别技术能够有效解决互联网上此类地标图像资源不能合理整合的难题。其次，地标识别是一个可靠且顾及人类空间认知规律形式的位置定位方法，可以弥补 GPS 定位方法易受到遮挡效应影响而造成精度不足的缺陷（周沙，2017）。

谷歌于 2016 年公开一种使用卷积神经网络（convolutional neural network，CNN）的方法 PlaNet（Weyand 等，2016）对图像进行分类来完成地标识别。将世界地图划分为 26 263 个区域，并以此划分作为类别标签进行训练和预测。其效果并不理想，街道级的准确率仅有 3.6%，城市级 10.1%。

实现图像地理位置定位功能常用的另一种方法是图像检索（Lowry 等，2016）。图像检索依据查询目标的不同可以分为同类检索、实例检索和副本检索 3 种（Jain 等，2018），从粗到细进行划分，同类检索是指寻找某种类型的对象而不用在意类内的变化，实例检索是指对类别中某一具体对象进行查询，副本检索是查找特定图像经过编辑或处理后的图像。地标检索对图像中出现的具体地标物体进行检索，无需扩展至建筑和自然风景等大类类别检索，因此对地标图像的检索属于实例检索。在当前使用图像检索进行图像位置定位的方法中都有一个前提假设，即待查询位置的图像需要在查询图像库中存在同实例图像和标签（Lowry 等，2016），即查询图像中的地标实体需要有其他含有该地标的图像存在于检索数据库中。特征提取是图像检索中的关键步骤，直接影响检索效果的优劣，图像特征通常可以分为局部特征和全局特征两种。

在传统方法中，图像局部特征通常比全局特征更适合在大规模数据库中进行匹配，例如 2003 年 SIFT（scale-invariant feature transform）特征开创性地与词袋模型（bag of word，BoW）（Sivic 和 Zisserman，2003）相结合，此种局部特征提取与匹配的方法至今仍在广泛应用。随着卷积神经网络与 ImageNet 大规模图像分类数据集的出现，在分类数据集上训练的 CNN 对于各种计算机视觉任务，包括图像分类和图像检测等，已经成为非常有效的工具。2012 年 AlexNet 首次在大规模社会图像数据集（ImageNet large scale visual recognition challenge，ILSVRC）（Russakovsky 等，2015）上使用，数据集中包含 500 类 200 万幅图像，其结果与使用 SIFT 描述符的词袋模型进行了比较，CNN 的识别率为 23.88%，而 BoW 的识别率仅为 9.5%。但是 CNN 分类识别效果优于局部特征识别效果并不代表 CNN 抽取出的特征在检索中的效果优于 SIFT 特征检索方法，直到多种 CNN 特征提取方法的提出转变了这一状况。

Radenović等人（2018a）对多种特征提取模型进行测试，总结出当前效果较优的方法是利用 CNN 提取局部特征再结合几何验证（random sample consensus，RANSAC）（Fischler 和 Bolles，1981）进行重排序的方法。此类方法以深度局部特征（deep local features，DeLF）（Noh 等，2017）为典型代表，且 DeLF 首次尝试在微调 CNN 中间层特征图上使用注意力机制，从图像特征的高层语义中提取局部特征，但是其特征检索过程使用乘积量化（product quantization，PQ）（Jégou 等，2011）检索与 KD 树（K-dimensional tree）进行多次组合，过于复杂，在资源有限和快速请求响应的条件下无法部署使用。局部特征可以在视角变化、背景混乱和部分遮挡的情况下于图像中进行局部匹配和检索，相比之下，全局特征比局部特征更依赖于姿态，但当照明条件发生变化时，全局特征的效果优于局部特征（Lowry 等，2016）。

对于 CNN 全局特征的使用也有相关研究。全局特征通常直接使用预训练模型进行提取或者对预训练模型在目标数据集上进行微调后从网络固定节点层提取。许多实验（Zheng 等，2018）结果表明在与预测数据近似分布的数据集上微调的效果将明显优于直接使用预训练模型的效果。模型微调时通常使用标准交叉熵损失、对比损失等损失函数进行训练，如 Radenović等人（2018b）使用对比损失在预训练 ResNet101 网络（He 等，2016）上利用相关地标数据集进行微调。Appalaraju 和 Chaoji 等人（2019）使用多尺寸特征图融合与三元组损失训练模型以获取相似性判决模型。Ng 等人（2015）提取 CNN 全局特征来代替传统特征聚合技术（如 vector of locally

aggregated descriptors(VLAD))中的手工特征。Berman 等人(2019)在训练时使用标准交叉熵损失和三元组损失的加权融合损失函数,在推理时保留分类层权重,使模型可以同时完成分类和检索两种任务。Gordo 等人(2016)提出一种新的针对地标数据集进行检索的网络模型,在此模型中增加目标检测网络来提高模型对图像全局特征的描述能力。且 Gordo 在文中强调使用无噪声且含有标注框的数据集进行训练是该方法的关键,但是对于现实场景中的地标数据,获取大规模标注框费时费力,且此方法的效果和传统局部特征相比依旧存在一定的差距。

这些方法依然存在以下问题:1)模型使用单一特征,在 CNN 方法中多数学者只关注局部特征或者全局特征,即仅使用全局特征或者局部特征,全局特征对图像中存在的视角变化问题敏感,而局部特征则对光线变化敏感。如何整合两种特征是一个难题。2)对于大规模地标数据,对图像进行人工标注费时费力,且标注存在人工认知偏差。3)损失函数在神经网络训练中可以有效提高模型对数据的拟合能力,但是分类训练中标准交叉熵损失函数在类别数量远大于嵌入特征维度时表现欠佳,度量学习中三元组损失(Schroff 等,2015)则因为三元组挖掘的效率问题使模型训练难以快速收敛。

针对以上 3 个问题,本文提出一种基于增量角度域损失(additive angular margin loss)——ArcFace 损失(Deng 等,2019),并对全局特征与局部特征进行融合且使用弱监督训练的地标识别模型。该模型在牛津和巴黎建筑数据集上进行测试,并使用平均检索精度(mean average precision, mAP)指标与其他公开模型进行对比。在东京城市数据集(Arandjelovic 等,2016)上进行地点识别测试,用以验证该模型在城市级别建筑群图像中的定位识别精度。

1 ArcFace 损失及参数选取

1.1 ArcFace 损失函数

ArcFace 损失函数在人脸识别中取得了优异的效果,该损失函数训练难度低于三元组损失且在大规模数据集上的效果优于标准交叉熵损失,因此本文采用 ArcFace 损失函数对模型进行训练。在训练神经网络模型时,标准交叉熵损失是常用的损失,标准交叉熵损失公式为

$$L_{\text{softmax}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i + b_{y_i})}{\sum_{j=1}^{n}\exp(\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{x}_i + b_j)} \tag{1}$$

式中,$\boldsymbol{x}_i \in \mathbf{R}^d$ 表示第 y 类别中第 i 个样本的 d 维特征向量,$\boldsymbol{W}_j \in \mathbf{R}^d$ 表示权重矩阵 $\boldsymbol{W} \in \mathbf{R}^{d\times n}$ 的第 j 行,$b_j \in \mathbf{R}^n$ 则表示偏置,N 为批处理训练样本数,n 为训练类别数。ArcFace 损失在角度域对标准交叉熵损失进行了改进。

当偏置 $b_j = 0$ 时,有 $\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{x}_i = \|\boldsymbol{W}_j\|\|\boldsymbol{x}_i\|\cos\theta_j$,其中 $\cos\theta_j$ 表示权重向量 $\boldsymbol{W}_j$ 和特征向量 $\boldsymbol{x}_i$ 的夹角余弦值。在训练时对 $\boldsymbol{W}_j$ 和 $\boldsymbol{x}_i$ 分别进行 L2 归一化处理,有 $\|\boldsymbol{W}_j\| = 1$ 和 $\|\boldsymbol{x}_i\| = 1$,此时 $\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{x}_i = \cos\theta_j$。至此,特征向量和权重向量可以认为是分布在半径为 1 的超球面上。如图 1 所示,ArcFace 损失的改进在于两点,首先将超球面半径扩大 s 倍,使其适用于大规模类别数量的特征嵌入;其次,在角度域增加边界值 m,使特征向量的分布更加集中于权重中心,以此来提高对特征向量的聚合能力,使其达到在角度域中相似的效果。因此,ArcFace 损失函数为

$$L_{\text{ArcFace}} = -\frac{1}{N}\sum_{i=1}^{N}\ln\frac{\varphi_{y_i}}{\varphi_{y_i} + \sum_{j=1,j\neq y_i}^{n}\exp(s\cos\theta_j)} \tag{2}$$

$$\varphi_{y_i} = \exp(s(\cos(\theta_{y_i} + m)))$$

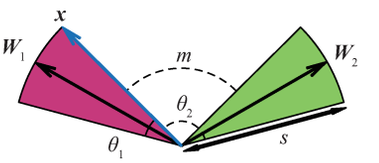


图 1 ArcFace 特征空间几何解释

Fig. 1 Geometric interpretation of ArcFace feature space

1.2 ArcFace 参数选取

在 ArcFace 损失中,超球面放大倍数 s 和角度域增值 m 的选取是需要设计的参数。在 Deng 等人(2019)的方法中没有给出两个参数选取的条件,本文将对此部分进行详细的描述,由此提出定理 1 和定理 2。

定理 1 ArcFace 损失中,假设类别数为 n,类别中心的最小后验概率为 P_W,则超球面放大倍数 s 有最小值,即

$$s \geqslant \frac{n-1}{n}\ln\frac{(n-1)P_W}{1-P_W} \tag{3}$$

证明 假设 $\boldsymbol{W}_i$ 是第 i 个类别单位权重向量,$\forall i$,有不等式

$$\frac{\exp(s)}{\exp(s)+\sum_{j=1,j\neq i}^{n}\exp(s(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j))}\geqslant P_W \tag{4}$$

$$\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j=\Big(\sum_{i=1}^{n}\boldsymbol{W}_i\Big)^2-\Big(\sum_{i=1}^{n}\boldsymbol{W}_i^2\Big)\geqslant -n \tag{5}$$

对不等式(4)两边同时取倒数得

$$1+\exp(-s)\sum_{j=1,j\neq i}^{n}\exp(s(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j))\leqslant\frac{1}{P_W} \tag{6}$$

再对类别 n 进行累加得不等式

$$\sum_{i=1}^{n}\Big(1+\exp(-s)\sum_{j=1,j\neq i}^{n}\exp(s(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j))\Big)\leqslant\frac{n}{P_W} \tag{7}$$

再同除类别数 n 得不等式

$$1+\frac{\exp(-s)}{n}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\exp(s(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j))\leqslant\frac{1}{P_W} \tag{8}$$

函数 $\exp(sx)$ 是一个凹函数，由琴生不等式得到

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\exp(s(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j))\geqslant \exp\Big(\frac{s}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j\Big) \tag{9}$$

与不等式(5)结合可以得不等式

$$1+(n-1)\exp\Big(-\frac{sn}{n-1}\Big)\leqslant\frac{1}{P_W} \tag{10}$$

最终可以得到超球面放大倍数 s 的取值范围式(3)。当且仅当每一个 $\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j(i\neq j)=0$ 且 $\sum_{i=1}^{n}\boldsymbol{W}_i=0$ 时等号成立。在 d 维超球面上，有 $d+1$ 个向量可以满足上述条件，因此等式仅在 $n\leqslant d+1$ 的时候成立。

证毕。

定理 2　ArcFace 损失中，假设特征空间维度为 d，类别数为 n，则角度域增值参数 m 的取值范围为

$$\begin{cases}0\leqslant m\leqslant\dfrac{2\pi}{n} & d=2\\ 0\leqslant m\leqslant\arccos\Big(\dfrac{1}{1-n}\Big) & n\leqslant d+1\\ 0\leqslant m\ll\arccos\Big(\dfrac{1}{1-n}\Big) & n>d+1\end{cases} \tag{11}$$

证明　特征空间维度为 2 时，在单位圆上对类别数 n 进行特征空间划分(Wang 等，2018)，对于最佳分类状态有 $\max(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j)=\cos(2\pi/n)$，所以边界值范围有 $0\leqslant m\leqslant 2\pi/n$。

对于高维特征空间 $d>2$ 时，对于权重向量中心夹角 $\psi_{i,j}$ 有 $\cos(\psi_{i,j})=\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j$，类别之间的角度域增量 m 必须小于类别权重向量中心夹角 $\psi_{i,j}$，且有不等式

$$n(n-1)\max(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j)\geqslant\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j=\Big(\sum_{i=1}^{n}\boldsymbol{W}_i\Big)^2-\Big(\sum_{i=1}^{n}\boldsymbol{W}_i^2\Big)\geqslant -n \tag{12}$$

因此有 $\max(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j)\geqslant 1/(1-n)$，由于 $y=\arccos(x)$ 是单调递减函数，于是有 $0\leqslant m\leqslant\min(\psi_{i,j})$，$\min(\psi_{i,j})=\arccos(\max(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j))$ 和 $\arccos(\max(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j))\leqslant\arccos(1/(1-n))$。当且仅当每一个 $\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{W}_j(i\neq j)$ 为 0，$\sum_{i=1}^{n}\boldsymbol{W}_i=0$ 且 $n\leqslant d+1$ 时等号成立，当等号成立时，每两个权重向量之间夹角的度数相等。从单纯形角度理解，如在 2 维单位圆中，构成内接等边三角形，重心是原点，权重向量为从重心出发指向 3 个顶点的向量，夹角 120°；在3 维空间中，构成单位球面内接正四面体，权重中心向量的分布可以等同于甲烷分子模型，其碳氢键夹角为 109.47°。在 $n>d+1$ 时，很难准确求解 m 的上限值，于是在 $n\gg d$ 时即类别数远大于特征维度时，有不等式 $0\leqslant m\ll\arccos(1/(1-n))$。

证毕。

2　模型结构设计与多特征融合

ResNet50 网络(He 等，2016)是当前卷积神经网络模型中常用基础网络之一，本文地标识别模型也采用此网络结构作为基础网络并在其基础上修改。整体的流程分为训练和推理两个步骤，本文使用两种不同的模型结构来分别完成训练过程和推理过程。

2.1　训练过程模型结构

训练时，模型结构分为两个主要部分，如图 2 所示，第 1 部分是 ResNet50 网络与 ArcFace 损失相结合的微调网络，第 2 部分是 ResNet50 浅层共用网络与注意力机制(attention mechanism)进行融合的注意力训练网络。模型结构决定训练过程分为两个步骤，分别是 ResNet50 基础网络微调和注意力机制训练。

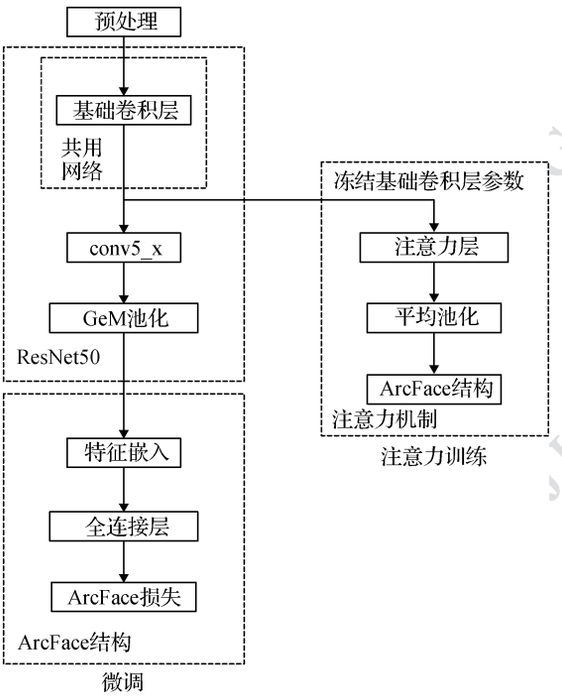


图 2　训练过程模型结构

Fig. 2　Model structure in training process

在训练过程的两个步骤中，图像预处理方法相同，都由中心裁剪、尺寸缩放和随机裁剪组成。对于检索，通常情况下其输入图像大小在 800 × 800 ～ 1 500 × 1 500 像素，而在分类应用中，输入图像大小通常在 224 × 224 ～ 300 × 300 像素，本文模型对输入图像大小进行折中处理，模型的输入图像大小为 448 × 448 像素。经过中心裁剪后，尺寸缩放过程统一将图像缩放至 500 × 500 像素，再用随机裁剪将图像裁剪至 448 × 448 像素。因为模型采用弱监督训练，所以训练时只需要将图像数据与类别标签数据输入网络即可。

模型微调过程使用预训练模型进行微调，即 ResNet50 网络的初始化参数为 ImageNet 预训练参数。本文模型丢弃平均池化层和 ImageNet 分类层权重，将平均池化层替换为 GeM（generalized mean）池化层，GeM 池化层由 Radenović 等人（2018a）首先提出，该池化层能够有效提高检索效果，因此本文模型也使用该池化层，其中池化层参数 p 为 2.3。ArcFace 结构中网络嵌入特征设计为 512 维，即特征嵌入层有 512 个神经网络单元。

注意力机制在网络中对当前模型的输出特征图进行加权计算，使模型更加关注于想要获取的特征，因此在注意力机制的训练中将仅对注意力网络层与注意力 ArcFace 结构进行训练。卷积神经网络中层抽取的特征对外观变化鲁棒，高层则对视角变化鲁棒且富含语义（Babenko 等，2014），即中层网络具有更加通用的特征提取能力，而深层网络将更加关注于具体类别的特征提取，因此本模型从中间层抽取通用特征来构建注意力层。在注意力结构中将使用 CNN 卷积层来完成对模型中间层网络输出特征的加权操作，注意力网络层结构如表 1 所示。在训练注意力机制时，首先冻结 ResNet50 共用网络层权重参数，其次随机初始化注意力网络层权重参数，然后从 ResNet50 网络的 conv4_x 层取出特征图送入注意力层（attention layers），最后使用 ArcFace 损失函数来对注意力机制进行训练。

表 1　注意力层参数

Table 1　Parameters of attention layer

层名	卷积核/宽 × 高，通道	步长	输入尺寸/宽 × 高，通道	输出尺寸/宽 × 高，通道
卷积层 1	1 × 1，1 024	1	28 × 28，1 024	28 × 28，1 024
卷积层 2	1 × 1，10	1	28 × 28，1 024	28 × 28，10

2.2　推理过程模型结构

在推理时模型结构如图 3 所示，模型将分别抽取全局特征和局部特征，再进行融合。

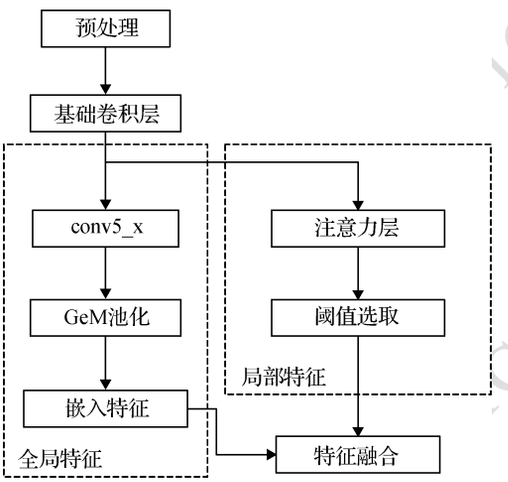


图 3　推理过程模型结构

Fig. 3　Model structure in inference process

图像预处理过程与训练过程中的处理方式不同，只包含中心裁剪和缩放两步，在缩放时将图像直接缩放至 448 × 448 像素。在提取特征时，全局特征去除全连接层，注意力机制去除训练过程中的平均池化层与 ArcFace 结构，增加阈值选取层。全局特

征从 ArcFace 结构中的特征嵌入层提取，局部特征由注意力层提取，再依据测试数据选取合理阈值对局部特征进行筛选。

本文模型在推理过程中使用改进的 ReLU（rectified linear unit）激活函数代替常用传统阈值筛选方法，维持在 GPU 中模型的特征抽取速度，并确保所有图像局部特征的维度固定，以方便检索，改进 ReLU 函数参考图 4，图中 th 为阈值。

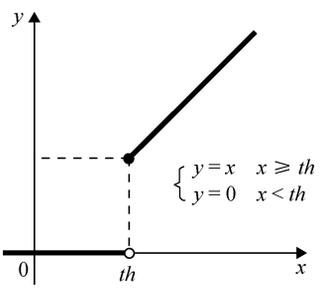


图 4　改进 ReLU 函数

Fig. 4　Modified ReLU function

2.3　多特征融合

多特征融合针对全局特征和局部特征进行有效融合，为解决两种特征分别存在的问题来提高检索的效果。对于单幅图像，全局特征由512 维向量表示，局部特征是改进 ReLU 函数的输出结果，由一个 $28 \times 28 \times 10$ 维度的张量表示，将此张量展平后为 7 840 维向量。全局特征和局部特征需要分别经过 L2 归一化，使各自映射在超球面上，从而将相似性量化到两种角度距离的叠加。全局特征在 ArcFace 结构中经过 L2 归一化处理。局部特征通过阈值选取步骤，已去除无用特征，此部分无用特征用 0 表示，且需要归一化后依旧对相似性评分结果没有贡献。对局部特征进行 L2 归一化能够仅对非 0 的特征数据进行归一化，特征为 0 的部分没有影响，满足上述归一化的需求。分别归一化之后将两种特征向量进行横向拼接，如图 5 所示，并将得到的特征向量用余弦相似性进行特征检索，相似性结果范围为 $-2 \sim 2$。

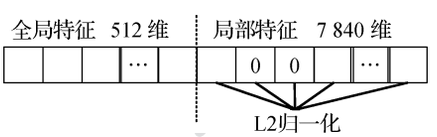


图 5　特征向量横向拼接示意图

Fig. 5　Features splice horizontally

3　实验与结果

3.1　实验相关设置

实验在 GTX 1060 8 GB 显存显卡、24 GB 内存和 i7-7600 CPU 构建的硬件平台上进行，软件系统为 ubuntu 16.04，神经网络框架为 PyTorch 1.0.0。

训练模型时，微调过程学习率为 4×10^{-3}，使用随机梯度下降方法，超球面半径 s 设置为 12，角度域增值 m 设置为 0.3，GeM 池化层参数 p 为 2.3。注意力训练时学习率为 4×10^{-5}，超球面半径 s 设置为 10，角度域增值 m 设置为 0.5。

模型训练时，微调过程与注意力机制，都是训练集和验证集损失稳定时即可停止训练。

3.2　数据集准备

实验中共使用了 4 种数据集，选用谷歌地标数据集（Russakovsky 等，2015）作为训练数据集，巴黎和牛津建筑数据集是标准数据集，可作为模型效果测试集，测试结果用于与其他方法进行比较，最后使用东京城市街景数据模拟实际应用场景来验证模型的有效性。

谷歌地标数据集作为当前最大的公开地标数据集，其训练数据有 14 951 个类别的地标，共 1 225 029 幅图像，包含诸多著名建筑和自然景观的图像。数据于 2018 年 Kaggle 谷歌地标识别及地标检索大赛上公开，大赛中数据集分为 3 个部分，分别是训练、测试和查询，其中仅有训练部分是包含有标签的数据，因此将训练数据再次拆分出含有标签的验证集和测试集。数据集中图像数量在类别上分布非常不平衡，图像数量最多类别有 2 万多幅图像，而类别最少的仅有 1 幅图像，所以仅使用图像数量最多的前 2 000 个类别的图像数据进行数据筛选和扩增，构建训练集和验证集。构建后的数据集共有 445 617 幅图像用于训练，22 281 幅图像用于验证，各个类别中图像的数量都将维持在 200 ～ 250 之间，以达到平衡数据集的效果。

牛津建筑数据集有 5 062 幅图像，11 个类别，巴黎数据集包含 6 412 幅图像，11 个类别，将这两个数据集按检索难度进行重构，分为简单、中等和困难 3 种级别。重构后的数据集中，牛津数据集有 4 993 幅图像用于构建检索图像数据库，巴黎数据集有 6 322 幅图像用于数据库的构建，两个数据集分别

有 70 幅图像用于检索查询。本文模型将使用重构后的数据集进行测试,为模拟真实应用场景,测试时输入图像为整幅图像,而不是经过标注框裁剪后的关键区域图。

东京城市街景数据由 Arandjelovic 等人(2016)公开,该数据从谷歌地图街景模式中进行采集,共有 1 441 个地点与 GPS 对应,数据库共有 49 056 幅图像,查询图像 7 186 幅,其中同一个地点包含多种时间节点的图像,每一个地点每一个时间节点有 5 幅图像。同类数据集还有匹兹堡城市数据集等其他数据集,由于其他数据获取难度较大,本文选用当前已经公开的东京城市街景数据集来验证本文提出的模型效果。

3.3 巴黎和牛津数据集实验结果

该实验中共设置 3 组对比,第 1 组是本文模型与现有检索方法进行效果对比,结果如表 2 所示,第 2 组是对比全局特征与特征融合的效果,结果如表 3 所示,最后一组是在全局特征中引入白化前后的结果对比,如表 4 所示。在此 3 组对比中,模型都在重构的巴黎、牛津数据集上进行评价且使用检索常用的平均检索精度(mAP)指标与 mAP@10 指标。

表 2 各模型 mAP 与 mAP@10 效果对比

Table 2 Comparison of mAP and mAP@10 in different models

/%

方法	M				H			
	roxford		rparis		roxford		rparis	
	mAP	mAP@10	mAP	mAP@10	mAP	mAP@10	mAP	mAP@10
HesAff-rSIFT-VLAD (Radenović等,2018b)	33.9	54.9	43.6	90.9	13.2	18.1	17.5	50.7
VggNet-NetVLAD (Radenović等,2018b)	37.1	56.5	59.8	94.0	13.8	23.3	35.0	73.7
Res101-O-GeM (Radenović等,2018b)	45.0	66.2	**70.7**	**97.0**	17.7	32.6	**48.7**	**88.0**
Res101-O-MAC (Radenović等,2018b)	41.7	65.0	66.2	96.4	18.0	**32.9**	44.1	86.3
本文(fused features, golbal512 + attention)	**50.9**	**70.0**	68.1	94.6	**19.2**	32.5	43.4	77.7

注:加粗字体为每列最优值,M 和 H 分别表示中等部分(Medium)和困难部分(Hard)。

表 3 全局特征与融合特征的效果对比

Table 3 Comparison of global features and fusion features

/%

数据集	特征	E	E10	M	M10	H	H10
rparis	全局特征	78.7	89.4	68	**94.7**	**43.5**	76.7
	融合特征	**79.5**	**90.2**	**68.1**	94.6	43.4	**77.7**
roxford	全局特征	70.1	76.6	48.8	67.1	16.2	27.6
	融合特征	**71.5**	**78.2**	**50.9**	**70.0**	**19.2**	**32.5**

注:加粗字体为每列最优值,E、M 和 H 分别表示简单(Easy)、中等和困难部分,E10、M10 和 H10 分别表示各部分的 mAP@10。

3.3.1 结果对比

表 2 中,roxford 和 rparis 表示重构的巴黎、牛津数据集。VggNet-NetVLAD 使用的是 Vgg 预训练网络后接入 NetVLAD 层进行特征检索。对于 Res101-O-GeM,Res101 表示 ResNet101 基础网络结构;O 表示 off-the-shelf,即使用 ImageNet 预训练网络权重;GeM 表示将平均池化层更改为 GeM 池化层,再将从池化层抽取特征送入白化处理。对于 Res101-O-MAC,MAC 则表示将平均池化层改为最大值池化层。

表 3 中融合特征为全局特征和 attention 机制抽取特征融合后的特征。在牛津建筑数据集上,3 种查询难度中融合特征的指标均优于基于 ArcFace 训练的 512 维全局特征,在巴黎数据集中,融合特征效果接近或略高于全局特征。总体上融合特征相比于单一全局特征,在 mAP 和 mAP@10 指标上平均高

表 4　不同特征白化效果对比

Table 4　Comparison of whitening on different features

/%

方法	M				H			
	roxford		rparis		roxford		rparis	
	mAP	mAP@10	mAP	mAP@10	mAP	mAP@10	mAP	mAP@10
global512	48.8	67.1	**68.0**	**94.7**	16.2	27.6	**43.5**	**76.7**
global512 + whitening	19.4	45.7	25.7	79.0	4.8	11.4	12.4	46.0
GeM + whitening	**52.9**	**71.2**	65.7	92.1	**25.0**	**38.1**	42.3	75.4

注：加粗字体为每列最优值，M 和 H 表示中等部分和困难部分。

出约 1% 。

表 4 中，global512 是指全局特征，即从微调网络特征嵌入层提取出的 512 维 L2 归一化后的特征。对比项 global512 + whitening 是指在抽取出全局特征基础上加入特征白化过程，特征白化数据集依然是谷歌地标数据集。GeM + whitening 是指对本文模型的 GeM 池化层特征加入特征白化过程后的结果。

3.3.2　结论与分析

由表 2 可以得出当前方法效果与 ResNet101 预训练模型效果相近的结论。当前模型基础网络仅使用 50 层微调卷积层可以达到 101 层预训练卷积网的效果，证明融合特征可以提升模型检索的性能，但同时对于困难检索，模型效果仍有待提升空间。后期改进则可以引入图像多尺度与特征多尺度，以对抗图像中目标物体过小或过大后难以检索的情况。

从表 3 中可以看出，特征融合的效果优于全局特征，表明了融合特征可以有效提高检索效果。使用特征融合方法使得最终得分是全局相似性得分与局部特征相似性得分的求和。该模型可以认为是一种多特征共同预测的方法，局部特征的引入在全局特征相似性得分的基础上贡献了局部信息相似的得分，从而提高了检索效果。

从表 4 可以得出，特征嵌入层之后并不适合再加入特征白化过程。嵌入后的 512 维特征加入白化后与 GeM 池化层 2 048 维特征白化后在效果上有着明显的差异。嵌入特征由全连接层输出，全连接层已经对特征维度进行了一次降维处理，GeM 有 2 048 维特征，通过网络自动迭代学习，映射到了 512 维；特征白化过程是一个分解并提取特征有效值的过程，该过程与嵌入层作用重复，因此对于全连接层的嵌入特征，无需再进行特征白化过程，加入白化过程后甚至会有反向作用效果。而对于 GeM 池化层提取出的特征数据，相比于全连接层，神经网络 2 048 个通道之间依然存在大量耦合现象，可以通过白化处理来对特征值进行解耦合并降低特征维度。

3.4　东京城市街景数据集效果

本文方法在东京城市数据集上进行测试，统计模型召回率 Recall@ N。现实场景中街景车大部分是在道路中部采集图像，这样会产生 7 m 的误差，所以 GPS 信息转换为欧氏距离后，距离在 25 m 范围内都认为是同一个地点。在计算召回率时，top-N 幅图像内出现同一个地点的图像即被认为召回，其结果如图 6 所示，本模型在召回率效果上优于最大值池化模型 Fmax 所提取的特征。

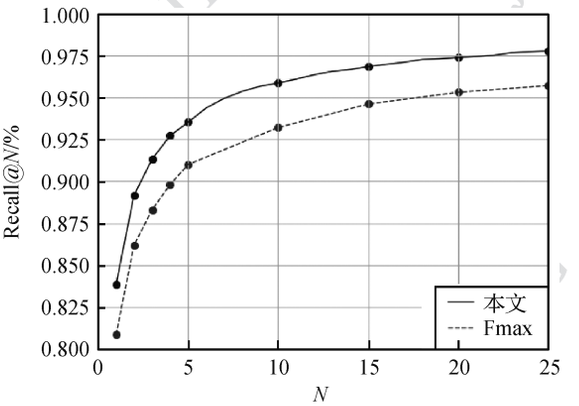


图 6　Recall@ N 效果图

Fig. 6　Results of Recall@ N

本文方法在实际应用中的检索效果如图 7 所示，图中绿色标记框图像表示识别命中图像，红色标记框图像表示识别未命中图像，本文方法能够对视角变化有一定的抵抗能力，同时对于细节相似的图

(a) 输入原图　　(b) 检索结果

图 7　东京城市数据集检索效果图

Fig. 7　Retrieval results on Tokyo dataset((a) input images; (b) retrieval results)

像也能够进行召回。因此本文方法能够在城市街景数据中有良好的应用。

4　结　论

针对地标识别的实际应用需求并结合全局特征和局部特征的优势，本文提出了一种基于 ArcFace 损失并对全局特征与局部特征进行融合且使用弱监督训练的地标识别模型。该模型仅使用单一模型抽取两种特征以完成推理过程，单幅图像推理过程占用显存低于 2 GB，对低显存平台也同样具有可用性。实验结果对比表明，在巴黎、牛津数据集上，改进的融合特征效果优于单一全局特征，融合特征效果逼近于深层 ResNet101-O-GeM 模型。本文方法使用著名建筑和自然风景的谷歌地标数据集训练的模型在城市级街景数据集上展现了较好的鲁棒性和泛化能力。

在后期的改进中可以引入图像金字塔方法，即输入图像多尺度或特征图多尺度，用来对抗图像中目标物体因为在图像中占比变化而难以检索的情况。对于谷歌地标数据集，采用人工或机器筛选(Ozaki 和 Yokoo，2019)过于复杂和耗时，还可以考虑将 NoiseFace 方法(Hu 等，2019)引入 ArcFace 中，进一步降低角度域损失对干净数据的依赖性。

参考文献(References)

Appalaraju S and Chaoji V. 2019. Image similarity using deep CNN and curriculum learning[EB/OL]. [2019-05-17]. https://arxiv.org/pdf/1709.08761.pdf

Arandjelovic R, Gronat P, Torii A, Pajdla T and Sivic J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 5297-5307 [DOI: 10.1109/CVPR.2016.572]

Babenko A, Slesarev A, Chigorin A and Lempitsky V. 2014. Neural codes for image retrieval//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 584-599 [DOI: 10.1007/978-3-319-10590-1_38]

Berman M, Jégou H, Vedaldi A, Kokkinos I and Douze M. 2019. MultiGrain: a unified image embedding for classes and instances[EB/OL]. [2019-05-17]. https://arxiv.org/pdf/1902.05509.pdf

Deng J K, Guo J, Xue N N and Zafeiriou S. 2019. ArcFace: additive angular margin loss for deep face recognition[EB/OL]. [2019-05-17]. https://arxiv.org/pdf/1801.07698.pdf

Fischler M A and Bolles R C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6): 381-395 [DOI: 10.1145/358669.358692]

Gordo A, Almazán J, Revaud J and Larlus D. 2016. Deep image retrieval: learning global representations for image search//Proceedings of the 14th European Conference on Computer Vision. Amsterdam,

The Netherlands: Springer: 241-257 [DOI: 10.1007/978-3-319-46466-4_15]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hu W, Huang Y Y, Zhang F and Li R R. 2019. Noise-tolerant paradigm for training face recognition CNNs [EB/OL]. [2019-09-26]. https://arxiv.org/pdf/1903.10357.pdf

Jain H, Zepeda J, Pérez P and Gribonval R. 2018. Learning a complete image indexing pipeline//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4933-4941 [DOI: 10.1109/CVPR.2018.00518]

Jégou H, Douze M and Schmid C. 2011. Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(1): 117-128 [DOI: 10.1109/TPAMI.2010.57]

Lowry S, Sünderhauf N, Newman P, Leonard J J, Cox D, Corke P and Milford M J. 2016. Visual place recognition: a survey. IEEE Transactions on Robotics, 32(1): 1-19 [DOI: 10.1109/TRO.2015.2496823]

Ng J Y H, Yang F and Davis L S. 2015. Exploiting local features from deep networks for image retrieval//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Boston, USA: IEEE: 53-61 [DOI: 10.1109/CVPRW.2015.7301272]

Noh H, Araujo A, Sim J, Weyand T and Han B. 2017. Large-scale image retrieval with attentive deep local features//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3476-3485 [DOI: 10.1109/ICCV.2017.374]

Ozaki K and Yokoo S. 2019. Large-scale landmark retrieval/recognition under a noisy and diverse dataset [EB/OL]. [2019-09-26]. https://arxiv.org/pdf/1906.04087.pdf

Radenović F, Iscen A, Tolias G, Avrithis Y and Chum O. 2018b. Revisiting oxford and Paris: large-scale image retrieval benchmarking//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5706-5715 [DOI: 10.1109/CVPR.2018.00598]

Radenović F, Tolias G and Chum O. 2018a. Fine-tuning CNN image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(7): 1655-1668 [DOI: 10.1109/TPAMI.2018.2846566]

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, Huang Z H, Karpathy A, Khosla A, Bernstein M, Berg A C and Li F F. 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252 [DOI: 10.1007/s11263-015-0816-y]

Schroff F, Kalenichenko D and Philbin J. 2015. Facenet: a unified embedding for face recognition and clustering//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 815-823 [DOI: 10.1109/CVPR.2015.7298682]

Sivic J and Zisserman A. 2003. Video Google: a text retrieval approach to object matching in videos//Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France: IEEE: 1470-1477 [DOI: 10.1109/ICCV.2003.1238663]

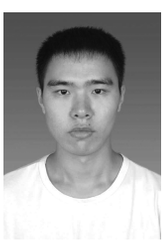
Wang H, Wang Y T, Zhou Z, Ji X, Gong D H, Zhou J C, Li Z F and Liu W. 2018. CosFace: large margin cosine loss for deep face recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5265-5274 [DOI: 10.1109/CVPR.2018.00552]

Weyand T, Kostrikov I and Philbin J. 2016. PlaNet-photo geolocation with convolutional neural networks//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 37-55 [DOI: 10.1007/978-3-319-46484-8_3]

Zheng L, Yang Y and Tian Q. 2018. SIFT meets CNN: a decade survey of instance retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(5): 1224-1244 [DOI: 10.1109/TPAMI.2017.2709749]

Zhou S. 2017. Research on the Key Technologies of Landmark-based Pedestrian Navigation. Wuhan: China University of Geosciences (周沙. 2017. 基于地标的行人导航关键技术研究. 武汉: 中国地质大学)

作者简介



毛雪宇，1995 年生，男，硕士研究生，主要研究方向为图像检索、目标检测。
E-mail: maoxueyu2010@126.com

彭艳兵，男，博士，高级工程师，主要研究方向为模式识别、海量数据挖掘。E-mail: pengyanbing@nj.fiberhome.com.cn