计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100066

跨视角地理视觉定位

刘旭东[1]　余　平[2]
1 国家能源投资集团新疆能源有限责任公司乌东煤矿　乌鲁木齐 830000
2 国能网信科技(北京)有限公司　北京 100011
(415576201@qq.com)

摘　要　伴随着智能终端设备的爆炸性增长和移动互联网的迅速崛起,在许多场景下,例如地广人稀的偏远山区,基于位置的服务需求越来越凸显。但由于这些区域GPS信号遮挡或信号基站难以覆盖,GPS定位无法正常发挥作用。图像地理定位指仅根据视觉信息确定图像的拍摄位置。在没有任何先验知识的情况下,预测照片的地理位置是一项非常艰巨的任务,因为不同条件下(例如,不同的天气,物体或相机设置)拍摄的图像会呈现出巨大的变化。文中尝试探索图像的跨视角地理视觉定位方法,首先利用逆极坐标转换将街景视角转换为空域视角图像,以此减少两者间的域差异,再利用深度学习的方法来对不同视角的图像进行编码以获得更加鲁棒的图像全局向量描述子,然后在此基础之上进行图像匹配和街景视角查询图像的定位。在图像特征提取方面,采用了VGG16模型,利用层数更深的小卷积核的方式去增大网络模型的感受视野并节省参数。在特征编码方面,将多尺度注意力机制融入NetVLAD模型,将骨架模型提取到的特征编码成更加鲁棒的全局特征描述子向量。实验结果显示,上述方法能够实现较高精度的街景视角的匹配与定位,同目前已有的方法相比,匹配精度更高。而且无须专业设备采集的高清街景视图,普通智能手机拍摄的街景视图即可获得较好的匹配定位精度。

关键词:跨视角定位;逆极坐标系转换;NetVLAD;多尺度注意力
中图法分类号　TP391

Cross-view Geo-visual Localization

LIU Xudong[1] and YU Ping[2]
1 Wudong Colliery,CHN ENERGY,Urumqi 830000,China
2 Chn Energy Network Infomation Technology,Co.,Ltd.,Beijing 100011,China

Abstract　With the explosive growth of smart terminal equipment and the rapid rise of mobile Internet,in many scenarios,such as indoor environments and remote mountainous areas with sparse population,the demand for location-based services has become more and more prominent. However,because GPS signals in these areas are blocked or the signal base stations are difficult to cover,GPS location can not working properly. Image based geo-location refers to determine the location of an image based only on visual information. Without any prior knowledge,predicting the geographic location of a photo is a very difficult task,because the images taken from the earth will show huge changes with different weather,objects or camera settings. This paper attempts to explore the cross-view geo-localization method. First,the inverse polar coordinate transformation is used to convert the street view perspective to the spatial perspective image,so as to reduce the domain gap between the two. Then deep learning is used to encode images from different perspectives to obtain more robust global vector descriptors. Finally,performing image matching on this basis. In the aspect of image feature extraction,the VGG16 model is adopted,and a smaller convolution kernel with deeper layers is used to increase the perception field of the network model and save parameters. In terms of feature encoding,the multi-scale attention mechanism is integrated into the NetVLAD model,and the features extracted from the backbone model are encoded into a more robust global feature descriptor vector. Experimental results show that the above-mentioned method can achieve higher accuracy,compared with the existing methods. And without the high-definition street view captured by professional equipment,the street view captured by ordinary smart phones can obtain good matching accuracy.

Keywords　Cross-view geo-localization,Inverse polar transform,NetVLAD,Multi-scale attention

1　引言

随着信息社会的不断发展,智能手机等设备逐渐普及,人们的日常出行基本可以依靠定位导航来实现,因此位置信息越来越受到人们的重视,现代社会对于位置信息的需求也越来越大[1]。近年来,基于图像的地理定位技术因其在自动驾驶和增强现实领域的潜在应用而受到计算机视觉领域的广泛关注[2]。同时,在人烟稀少、信号基站难以覆盖到的偏远山区却难以直接依靠GPS进行定位导航。

随着遥感卫星的不断发展,大量带有地理数据标签的卫星图像被采集到。因此,估计拍摄图像地理位置的问题被转换为了街景视角图像与空域视角卫星图像的匹配问题,即通过匹配到的带有地理坐标的卫星图像去确定拍摄图像的地理位置。2014年8月19日,中国成功发射可自主获取全色

通信作者:余平(20049948@ceic.com)

1m、多光谱 4m 的高分辨率卫星影像——高分二号卫星。随后相继发射了高分三号、四号、五号、六号卫星在轨与一号、二号相互配合，共同推动高分辨率数据应用，标志着中国遥感卫星进入高分辨率图像应用的快速发展阶段[3]。随着遥感卫星的不断发展，大量带有地理数据标签的图像被采集到。因此，估计拍摄图像地理位置的问题被转换为了地面视角图像与空域视角卫星图像的匹配问题，即通过匹配到的带有地理坐标的卫星图像去确定拍摄图像的地理位置。

鉴于上述的跨视角地理视觉定位为自动驾驶、增强现实领域以及无网络环境下的定位问题提供了一种辅助解决方案，因此其具有理论价值和实践意义。本文以跨视角地理视觉定位问题为研究对象，综合 RS、GIS 技术和深度学习方法，分析跨视角地理视觉定位的技术难点，从实际应用的角度构建解决此类问题的模型，为后续该领域的研究提供一种新颖的解决方案。

2 跨视角地理定位算法

在深度学习技术被引入这一领域之前，人工设计的特征被广泛用于交叉视图图像匹配领域[4-5]。Bansal 等从倾斜的航拍图像中提取建筑物外墙，然后通过匹配建筑地面视角拍摄的建筑外墙斑块进行地理定位[5]。Li 用 SIFT 算法提取图像的候选位置特征点，再利用 KNN 算法和 RANSAC 算法对匹配过程中产生的错配点进行剔除以提高图像室内定位的准确性[6]。随着深度学习技术的不断发展，Workman 等首先将深度特征引入到跨视角匹配任务中[7]。Vo 等评估了一系列用于匹配跨视角图像匹配的 CNN 体系结构[8]。为了让网络结构学习到图像的朝向信息，Hu 等在孪生 CNN 网络上嵌入 NetVLAD，用于交叉视角图像匹配[9]。Cai 提出了一个难样本的加权三元组损失函数以提高网络训练的质量[10]。Sun 等使用胶囊网络对图像的空间特征分层进行编码以获得更加鲁棒的图像向量描述子[11]。为了弥补地面视角和空域视角的域差距，Regmi 等利用 GAN 模型从地面图像合成航空图像，然后融合地面图像和合成航空图像的特征作为检索描述子向量[12]。总体来说，地理视觉定位的方向逐渐由 ground-to-ground 的单一视角匹配过渡到 ground-to-aerial 这一类更加困难但应用面更加广泛的跨视角匹配。

3 跨视角地理定位方法设计

3.1 算法模型结构

本文采用端到端的方式实现基于深度学习的跨视角图像匹配定位系统。如图 1 所示，整个跨视角匹配的过程分为图像描述子向量提取的训练过程和跨视角图像匹配的推理过程。在图像描述子向量提取过程中，首先对 CVUSA 数据集进行预处理操作，主要包括归一化和张量转换。其次，利用逆极坐标转换将地面视角近似转换为空域视角，再将视角转换后的地面视角图像、与其匹配的空域图像、不匹配的地面视角或空域视角构建成三元组，输入到孪生神经网络之中，得到对应图像的全局向量描述子。最后计算损失函数，利用反向传播算法去更新孪生神经网络参数，让匹配的两视角图像全局向量描述子更加接近，而不匹配的两视角图像全局描述子更加远离。在跨视角图像匹配过程中，首先利用训练好的孪生神经网络构建空域图像全局向量描述子数据库，再将需要查询的地面视角图像输入训练好的孪生神经网络之中得到该图像的全局向量描述子，最后将该向量描述子与数据库中的描述子进行匹配，将最接近的空域图像检索出来即可获得查询街景图像的地理坐标。

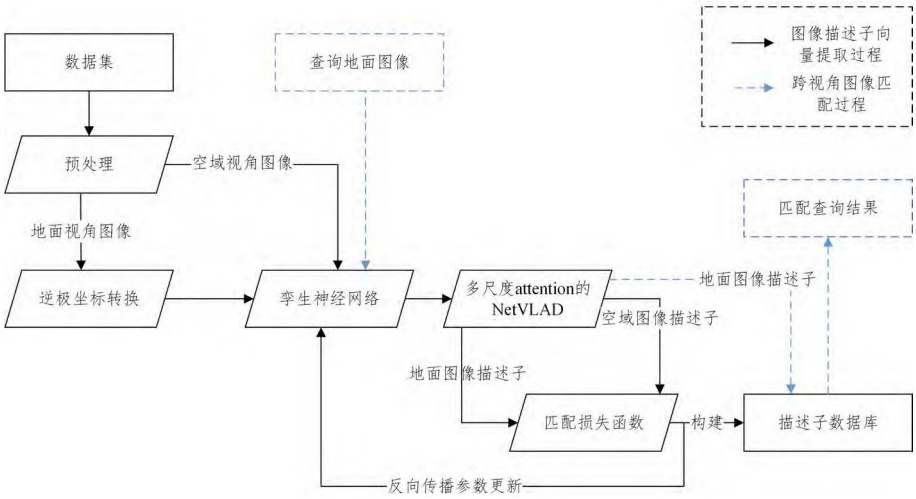

图 1 模型总体架构图

Fig. 1 Overall architecture of the proposed model

3.2 图像描述子向量提取

3.2.1 特征提取骨架网络

Krizhevsky 等提出的 AlexNet[13]在 2012 年 ImageNet 的大规模视觉识别挑战赛(ILSVRC2012)中以 Top-5(test)84.7%的准确率远超当时基于 SIFT 算子的手动特征提取算法，开启了卷积神经网络的时代。越来越多的研究人员开始尝试改进 AlexNet 的网络架构，例如 Sermanet 等的 OverFeat 网络尝试在卷积层的第一层使用较小感受窗口和步长去提取浅层网络的细粒度特征并获得了 ILSVRC2013 的冠军。本文采用的特征提取骨架网络是牛津大学的 Visual Geometry Group 提出的 VGG 网络。该模型从网络深度入手，固定使用 3×3 的小尺度卷积核代替大卷积核，在保证网络模型参数量减少的同时，逐步增加卷积网络层数。

本文在训练时输入的骨架网络 VGG 是固定大小为 288×288 的 RGB 图像。在数据预处理方面，仅仅使用数据集的均值与方差将图像归一化。然后图像会被输入到一组使用 3×3 小卷积核的卷积层中提取特征，由于 3 个 3×3 的卷积核与 1 个 7×7 的感受视野一致，当特征的通道数目为 C

时，前者的参数量为 $27C$，而后者为 $47C$，因此 VGG 使用多组小卷积核在保证精度的同时还能大大地减少网络参数。在经过卷积层提取特征后，特征图又会被输入窗口大小为 2×2、步长为 2 的最大池化层，该池化层会将窗口内最显著的特征保留下来。与此同时，每一个卷积层的最后都会连接一个 ReLU 线性整流函数，在增强网络非线性能力的同时缓解梯度消失现象。

VGG 网络配置参数如表 1 所列。

表 1　VGG 网络配置表格

Table 1　VGG network configuration

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
Input(288×288 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-256 conv3-256 conv3-256 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-256 conv3-256 conv3-256 **conv3-512**
maxpool					

3.2.2　特征编码网络 NetVLAD

NetVLAD 是一种广义的 VLAD（局部聚合描述子向量）[14] 算法，它将 VLAD 算法成功迁移到深度学习领域，利用 soft assignment 的方式将 VLAD 算法变成一个可微分的算法，再利用反向传播算法去更新参数。

NetVLAD 的目的就是将图像经过卷积网络提取到的 $W\times H\times D$ 的特征图池化为长度为 $K\times D$ 可代表整幅图像特征的一维向量描述符。其公式如下：

$$V(j,k)=\sum_{i=1}^{N}a_k(x_i)(x_i(j)-c_k(j)) \tag{1}$$

其中，$N=W\times H$ 即长度为 D 的特征向量 $\boldsymbol{x}_i$ 的个数；k 为 K 个簇中心的第 k 个中心；$a_k(\boldsymbol{x}_i)$ 为 $\boldsymbol{x}_i$ 向量分派给第 k 个簇中心的权重；$\boldsymbol{c}_k$ 为特征向量 $\boldsymbol{x}_i$ 聚类得到的第 k 个簇中心向量。

NetVLAD 将 $a_k(x_i)$ 的计算从原来的 hard assignment 改进为 soft assignment，从而让 VLAD 算法变成了一个可微分的算法，$a_k(\boldsymbol{x}_i)$ 的计算式如下：

$$a_k(\boldsymbol{x}_i)=\frac{\mathrm{e}^{-\alpha\|\boldsymbol{x}_i-\boldsymbol{c}_k\|^2}}{\sum_{k'}\mathrm{e}^{-\alpha\|\boldsymbol{x}_i-\boldsymbol{c}_{k'}\|^2}} \tag{2}$$

NetVLAD 整个网络算法流程如图 2 所示。

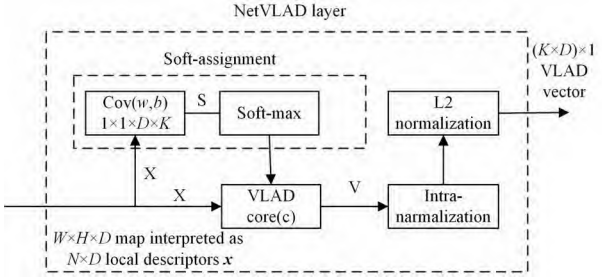

图 2　NetVLAD 算法流程图

Fig. 2　NetVLAD algorithm flowchart

首先输入 $W\times H\times D$ 的特征图即 N 个 D 维的特征向量，再利用卷积核大小为 1×1 的卷积层计算出 K 个聚类中心，再利用式(2)计算出各个向量到聚类中心的分派权重 $a_k(\boldsymbol{x}_i)$，最后使用式(1)根据分派权重将各个特征向量到聚类中心的残差加和即可得到一个长度为 $K\times D$ 的一维图像全局向量描述符。

3.3　跨视角图像匹配

3.3.1　训练过程

在训练过程中，图像匹配的目的就是让三元组中本来是一对的地空视角图像的全局描述向量越来越接近，让不是一对的地空视角图像的全局描述向量越来越远离。对于街景图像来说，有 1 个与之匹配的遥感图像正样本和 $B-1$ 个与之不匹配的遥感图像负样本。反之，对于遥感图像来说也是一样。这就意味着一个批次就可以构建出 $2B(B-1)$ 组三元组。为了充分挖掘这 $2B(B-1)$ 组三元组之间的距离关系，本文采用了 weighted soft-margin ranking triplet loss[9]，其计算式如式(3)所示：

$$L_{\text{weighted}}=\ln(1+\mathrm{e}^{\alpha(d_{\text{pos}}-d_{\text{neg}})}) \tag{3}$$

其中，α 为权重超参数，可以加快算法收敛，本文与文献[9]中的设置保持一致，将其设置为 10；d_{pos} 为正地空视角图像样本对的描述向量距离；d_{neg} 为负地空视角图像样本对的描述向量距离。

3.3.2　检索过程

在检索过程中，需要利用训练完成的模型构建向量描述子数据库，再将查询图像输入模型得到查询图像向量描述子。将该描述子与数据库中的描述子进行对比，从而将距离最近的数据库图像检索出来，以其中心坐标作为查询图像的地理位置，检索过程如图 3 所示。

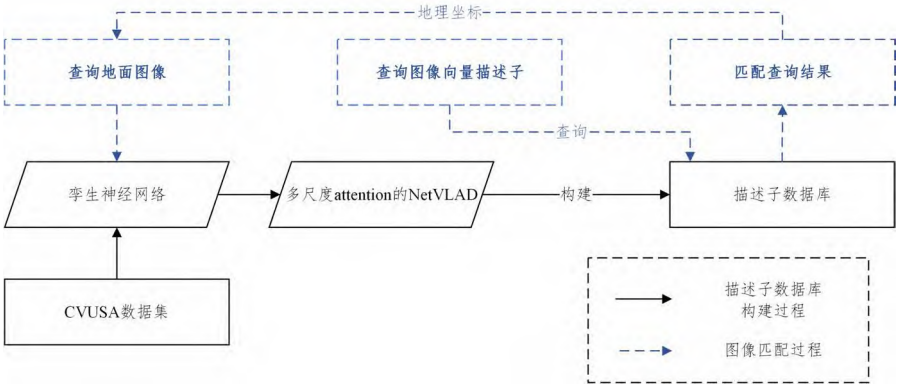

图 3 跨视角图像匹配流程图

Fig. 3 Cross view image matching flowchart

3.4 跨视角地理定位创新算法

3.4.1 逆极坐标转换

本文的研究发现，街景图像和卫星空视图像有两个十分重要的几何关联：1）街景图像中的水平线具有相同的深度，即街景图像的水平线对应着卫星空视图像的同心圆线。2）街景图像的垂直线上，深度随着 y 坐标的增大而增加，这对应于卫星空视图像的径向射线。两者几何关联示意图如图 4 所示，街景图像的黄色线段对应卫星空视图像的黄色圆环，街景图像的红色线段对应卫星空视图像的红色圆环。

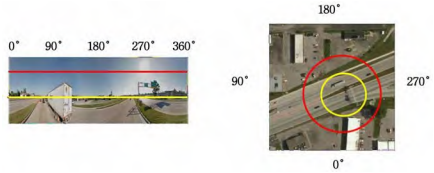

图 4 地空图像几何关联图（电子版为彩图）

Fig. 4 Geometric correlation map of ground air image

极坐标系转换的目的就是将笛卡尔坐标系的图像转换为极坐标系图像，其表达式如式（4）所示：

$$\begin{cases}\rho=\sqrt{(x-x_0)^2+(y-y_0)^2}\\ \theta=\text{atan2}\left(\dfrac{y-y_0}{x-x_0}\right)\end{cases} \tag{4}$$

通过显式地挖掘地面视角与卫星空域视角的几何关系，可以显著地降低跨视角图像之间的域差异，逆极坐标转换示意图如图 5 所示。

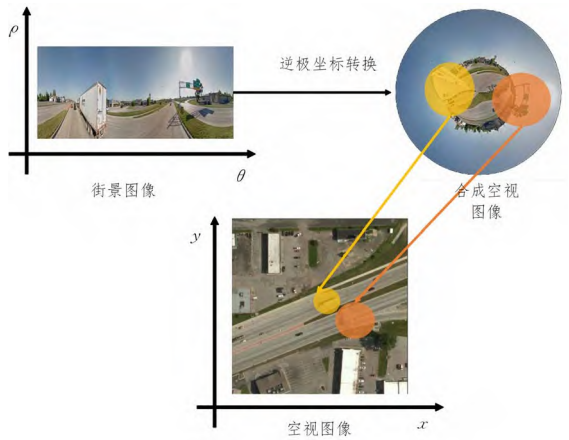

图 5 地空图像几何关联图

Fig. 5 Geometric correlation map of ground air image

不同于传统的跨视角匹配方法，本文方法通过显式地建立起地面视角图像和卫星空域视角图像的几何关系，先让这两个域信息粗略地对齐，再让神经网络学习判别特征。尽管神经网络在理论上可以学习任何几何变换关系[15]，但是通过本文方法先粗略对齐跨视角图像的域信息可以让神经网络模型更加关注于其他判别信息的学习，从而加快网络训练的收敛速度。

3.4.2 多尺度注意力机制

经过逆极坐标转换可以在几何上减少地面视角和空域视角图像之间的域差异，但是由于逆极坐标系转换仍然无法考虑到街景图像的场景深度信息，因此将街景图像转换为合成空视图像时仍然存在图像形变的问题，并且这种形变也无法通过函数变换显示地消除。因此，如图 6 所示，本文将多尺度机制引入 NetVLAD，让 NetVLAD 可以更加关注于编码骨架神经网络提取的重要特征，而抑制那些由于图像形变而产生的无效特征。

本文的多尺度注意力模块使用多组固定大小的卷积核 g_p 去显式地挖掘特征空间的上下文信息，而为了挖掘多尺度信息，使用了一组 3×3，5×5 和 7×7 的卷积核去捕捉不同尺度之下的特征空间上下文信息，多尺度信息提取表达式如式（5）所示：

$$s=\bigcup_p(g_p(d)+c) \tag{5}$$

其中，$\bigcup_p()$ 代表通道连接操作，d 代表输入特征图，$g_p(d)$ 代表第 p 组卷积核的输出，c 代表偏置常数。

得到多尺度上下文信息 s 之后，再使用一组 1×1 的卷积核逐通道地将每个空间位置的信息进行加权求和，得到一个通道数目为 1 的注意力掩模，最后使用上采样将该注意力掩模大小恢复至特征图的大小，并利用恢复后的注意力掩模对 VLAD 的 soft assignment 进行赋权操作，便可以得到一个考虑了多尺度空间信息且关注重要特征的 soft assignment。

3.4.3 全局难例挖掘策略

本文的研究发现，随着训练过程的不断推进，模型训练精度不断提高，由于批次的大小限制，在一个 batch 之内的负样本对于 loss 的贡献逐步趋向于 0，模型整体收敛速度渐渐变慢。Hu 等挖掘一个批次中的难负样本进行训练，发现这种难例挖掘策略对于模型训练精度有着明显的提高[9]。但是这种方法对硬件设备要求很高，在批次大小较小时并不适用。为了让模型在训练过程中尽可能地考虑到全局的难负样本，

使用 FIFO(First in First Out)队列将卫星空视图像经过模型前向传播得到的全局描述子向量保存起来。同样，为了节省计算开销，该队列被设置为一个固定长度且只保存当前批次的最难负样本，用于后续批次的损失计算。

本文的全局难例挖掘策略在每一个批次计算完成之后保存最难的负样本的全局描述向量子，所产生的额外计算几乎可以忽略不计。当训练集较小时，可以将全部难负样本保存下来，这样就可以考虑到全局的难负样本。

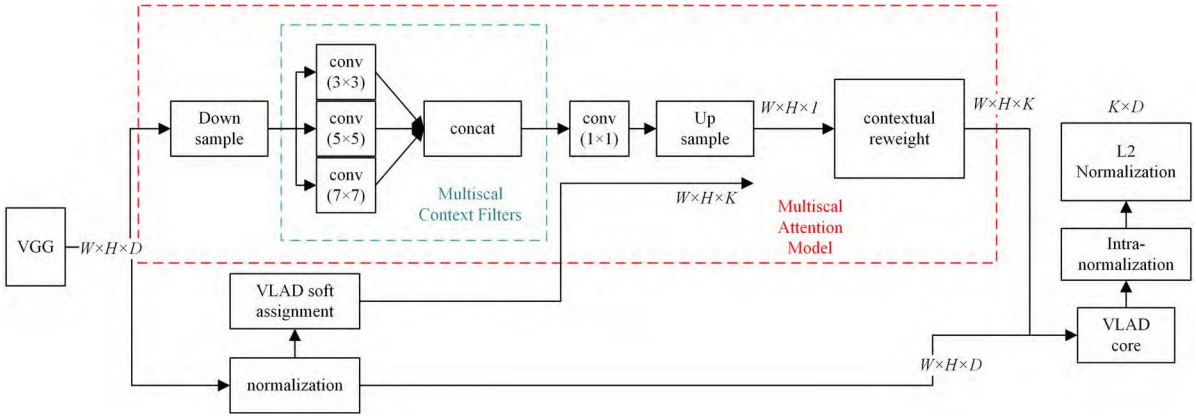

图 6　多尺度注意力 NetVLAD

Fig. 6　Multiscale attention NetVLAD

4　实验结果及分析

4.1　实验条件

本文实验数据为 Crossview USA(CVUSA)数据集[16](见图 7)，它是一个包含美国各地数万对地面和航空/卫星图像的大型数据集。该数据集带有地理坐标的遥感空域图像和对应的地面街景视角图像是从谷歌街景和必应地图网站上搜集的，其训练集包含了 35532 对街景图像与遥感图像，测试集包含了 8884 对用于验证的街景图像与遥感图像。除此之外，CVUSA 中的街景图像还提供了语义分割标签，由于本文方法不依赖于其他任何附加信息，因此不使用该语义分割标签。

图 7　CVUSA 数据集

Fig. 7　Dataset CVUSA

本文使用的编程框架为 Pytorch。Pytorch 是一个开源的 Python 机器学习框架，最初由 Facebook 人工智能研发小组开发，由于其动态计算图的设计便于开发者调试与编写程序，成为了学术界研究深度学习的首选框架。

对于跨视角地理定位问题，本文将 top-n 召回率作为 CVUSA 数据集的评估指标，对于每一个查询的街景视角图像，如果与其正确匹配的空域遥感视角图像在前 n 个检索结果之中，则认为本次检索正确。top1%是一个较弱约束的性能指标，目前已有的工作[9-10]已经将 top1%指标提升到了 95%以上，因此当 top1%再提高时不再具有很好的分辨性。而 top1 精度则是跨视角地理定位中最终需要解决的问题，即给出一张地面视角的查询图像，在数据库中找到唯一与之匹配并带有地理坐标的遥感图像。因此，top1 相比 top1%更具有实际应用意义。所以，本文在此不仅使用 top1%的准确性作为指标，还测试了本文模型的 top1 的准确性。

本文实验在 Linux 系统环境上安装数据集并运行代码，其他实验系统环境条件如表 2 所列。

表 2　实验系统环境

Table 2　Experimental system environment

环境条件	参数信息
操作系统	Ubuntu 16.04.6
深度学习框架	Pytorch
GPU	2×RTX2080Ti
cuDNN	7.6.5
CUDA	10.2
第三方库及软件	Anaconda，VScode

4.2　跨视角图像匹配测试

本文采用 Pytorch 框架，利用 Pytorch 官方提供的在 ImageNet 中预训练好的 VGG 模型权重作为模型初始权重，以加快模型训练速度。整个跨视角图像匹配测试实验主要分为 3 个部分：骨架网络模型配置、编码网络配置、训练参数配置。

骨架网络模型配置如表 1 所列，本实验选择使用拥有 16 层卷积层的 VGG16 网络，该网络整体划分为 5 个 Block，第一个 Block 由 2 个卷积层和 1 个最大池化层组成，其输入为 288×288 的 3 通道 RGB 图像，输出为 144×144 的 64 通道特征图像；第二个 Block 由 2 个卷积层和 1 个最大池化层组成，其输出为 72×72 的 128 通道特征图；第三个 Block 由 3 个卷积层和 1 个最大池化层组成，其输出为 36×36 的 256 通道特征图；第四个 Block 由 3 个卷积层和 1 个最大池化层组成，输出为 18×18 的 512 通道特征图；第五个 Block 同样由 3 个卷积层和 1 个最大池化层组成，其输出特征图通道数与第四个 Block 保持一致，均为 512，仅仅将特征图宽高缩小为原来的一半，以提取高层次的语义信息。

编码网络模型 NetVLAD 的聚类中心个数参考文献[17]选择为 64，由于 NetVLAD 需要在大尺度视觉识别数据集上预训练之后才能正常工作，因此本文使用 NetVLAD 在大尺度视觉识别数据集上预训练的权重作为实验的初始权重。

NetVLAD 中心个数为 64、输入特征图通道数目为 512 时，其编码向量长度为 32768，为了节省显存并加快推理速度，本文在 NetVLAD 之后连接了一个 BottleNeck，将 NetVLAD 输出的编码向量维度缩减为 4096。

网络训练优化器选择自适应学习率的 Adam 优化器，其一二阶矩阵估计的指数衰减动量参数 β_1 和 β_2 分别为 0.9 和 0.999，初始学习率为 5×10^{-5}。为了让网络在训练过程中不会因学习率过大而出现跳出全局最优解的问题，本实验学习率衰减策略如式(6)所示：

$$lr'=\left(1-\frac{epoch}{epochs}\right)*lr \tag{6}$$

其中，lr' 为下一个 epoch 的学习率，$epoch$ 为当前迭代次数，$epochs$ 为总迭代次数。本次实验中 $epochs$ 总迭代次数为 100。

本次实验中批次数据大小 B 为 54，全局难例挖掘策略的队列大小为 540，即本文可以考虑的难例范围被扩大到了 540 个样本，拥有更宽的样本视野才能更准确地估计难负样本，从而提升网络模型匹配精度。

本次实验输入网络的目的是增强网络模型的泛化能力，本文引入了 ImageNet 中的自动数据增强策略，即在图像归一化之后加入了随机旋转、随机水平翻转、随机对比度变换、随机剪切变换等数据增强操作。

4.3 匹配结果和对比

本节展示和分析实验结果，如表 3 所列。本文所提出的跨视角图像匹配定位方法在不使用更强大的骨架网络[18](ResNet)的前提之下，匹配精度超越了现有方法。

表 3 跨视角图像定位方法精度对比

Table 3 Precision comparison of cross view image location methods

Method	Publication	CVUSA	
		Top-1%	Top-1
Liu et al.[19]	CVPR2019	96.1	40.8
Regmi et al.[12]	ICCV2019	95.9	48.8
Zheng et al.[20]	ACM MM2020	91.8	43.9
CVFT[21]	AAAI2020	99.0	61.4
Ours	—	99.3	63.9

图 8 给出了本文提出的跨视角地理定位方法的性能曲线，从 Top-K 召回率曲线可以看出，随着 Top-K 中候选 K 数目的增加，模型召回率也越来越高。尤其是 Top1% 召回率已经高达 99.3%，即模型选择数据库中 1% 的遥感图像作为候选地点，只要这 1% 中的图像存在与查询街景图像相匹配的遥感图像，就认为这次匹配任务是成功的。显然，Top1% 作为一个约束较弱的指标，在以前跨视角图像定位问题难以解决时，可以给模型性能的度量提供一个参考。但随着目前跨视角图像定位方法的不断发展，Top1% 已经无法满足高精度、高准确度的定位要求了，因此之后的研究将逐渐减少 Top-K 中的 K 数目，以期最终能做到街景图像与空视遥感图像一对一匹配的检索。

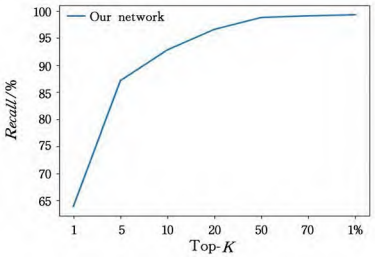

图 8 Top-K 召回率曲线

Fig. 8 Top-K recall rate curve

为了能够更加详细地展示模型跨视角图像检索匹配的能力，随机选择 3 张查询街景图像输入模型，并选择出模型编码与该街景图像最接近的 5 张空视遥感图像，如图 9 所示。其中，红色边框的遥感图像为错误匹配图像，绿色边框的遥感图像为正确匹配的图像。可以看出，每一张查询街景图像的 Top-5 匹配空视遥感图像都十分接近，特别是其道路走向基本一致，因此只能从空视遥感图像上的建筑分布加以区分，而当遥感图像中的建筑分布也十分相似时，模型检索匹配出错的概率就会大大增加。例如，第二张查询街景图像这种数据库中，前两张空视遥感图像不仅道路走向一致，连建筑分布都十分相似，并且其地物种类丰富环境信息较为复杂，因此模型对这种查询图像的匹配效果会大打折扣。

图 9 Top-5 匹配结果

Fig. 9 Top-5 matching results

结束语 图像检索任务长期以来都是计算机视觉领域研究的热点，随着深度学习技术的不断发展，图像检索精度和准确度在这新一波人工智能浪潮中迎来了进一步的提升。相比传统以 SIFT 特征为代表的人工特征检索技术，基于深度学习技术的图像特征检索技术在特征提取的准确度和特征本身的鲁棒性上都有着较为明显的优势。

由于跨视角地理图像检索不同于其他普通的检索问题，街景视角与卫星遥感图像存在的巨大视角差异给这类问题带来了巨大挑战。本文在孪生神经网络的基础上提出了 3 种关键的算法，分别是逆极坐标转换、多尺度注意力机制和全局难例挖掘策略，从而有效地缩小地空视角图像之间的域差异，并且提取到两视角图像更加鲁棒的全局图像描述子向量用于

跨视角地理定位任务。尽管本文研究了如何初步减少地空视角图像之间的域差异和提取更加鲁棒的全局图像描述子向量的问题，但是，跨视角图像定位仍然存在新的问题和挑战。首先，定位不止匹配，目前主流的跨视角图像定位方式便是构建在图像匹配的基础之上，显然，查询街景图像只能对应到遥感图像中的一小部分，这也就意味着直接将遥感图像中心点的坐标赋值给该街景图像难以做到准确的定位和应用。其次，大部分街景图像的主要描述内容是街景道路，但是也有着非静态物体如树木、车辆和天空等，这些物体给跨视角图像检索带来了干扰信息。因此，如何利用一些数据预处理手段，如图像分割的方式剔除这些干扰内容，从而达到提升精度并且减少数据量的目的也是本领域需要解决的问题。

参 考 文 献

[1] ZHANG N. Research on image matching algorithm in indoor visual location [D]. Shenyang: Shenyang University of Technology, 2020.

[2] MCMANUS C, CHURCHILL W, MADDERN W, et al. Shady dealings: Robust, long-term visual localisation using illumination invariance[C]// 2014 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2014: 901-906.

[3] LIAO C J, HOU Y K, XIN L. Research on the Operation and Service Mechanism of China's High Resolution Remote Sensing Application Satellite [J]. Satellite Applications, 2014(2): 57-61.

[4] MIDDELBERG S, SATTLER T, UNTZELMANN O, et al. Scalable 6-dof localization on mobile devices [C] // European Conference on Computer Vision. Cham: Springer, 2014: 268-283.

[5] BANSAL M, SAWHNEY H S, CHENG H, et al. Geo-localization of street views with aerial image databases[C]// Proceedings of the 19th ACM International Conference on Multimedia. 2011: 1125-1128.

[6] LI S. Indoor positioning system based on position feature image detection [D]. Wuhan: Huazhong University of Science and Technology, 2019.

[7] WORKMAN S, JACOBS N. On the location dependence of convolutional neural network features [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015: 70-78.

[8] VO N N, HAYS J. Localizing and orienting street views using overhead imagery[C]// European Conference on Computer Vision. Cham: Springer, 2016: 494-509.

[9] HU S, FENG M, NGUYEN R M H, et al. CVM-net: Cross-view matching network for image-based ground-to-aerial geo-localization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7258-7267.

[10] CAI S, GUO Y, KHAN S, et al. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8391-8400.

[11] SUN B, CHEN C, ZHU Y, et al. GEOCAPSNET: Ground to aerial view image geo-localization using capsule network[C]// 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019: 742-747.

[12] REGMI K, SHAH M. Bridging the domain gap for ground-to-aerial image matching[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 470-479.

[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] ARANDJELOVIC R, GRONAT P, TORII A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5297-5307.

[15] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(5): 359-366.

[16] WORKMAN S, SOUVENIR R, JACOBS N. Wide-area image geolocalization with aerial reference imagery[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 3961-3969.

[17] GE Y, WANG H, ZHU F, et al. Self-supervising fine-grained region similarities for large-scale image localization[C]// European Conference on Computer Vision. Cham: Springer, 2020: 369-386.

[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[19] LIU L, LI H. Lending orientation to neural networks for cross-view geo-localization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5624-5633.

[20] ZHENG Z, WEI Y, YANG Y. University-1652: A multi-view multi-source benchmark for drone-based geo-localization[C]// Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1395-1403.

[21] SHI Y, YU X, LIU L, et al. Optimal feature transport for cross-view image geo-localization[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(7): 11990-11997.

LIU Xudong, born in 1980, senior engineer. His main research interests include mining technology and intelligent technology.

YU PING, born in 1978, bachelor. His main research interests include deep learning and computer vision.