中图法分类号:TP391　　文献标识码:A　　文章编号:1006-8961(2024)12-3543-21

论文引用格式:Zhou B W, Li Y, Ma X J, Miao Z and Zhang R. 2024. A survey of cross-view geo-localization methods based on deep learning. Journal of Image and Graphics, 29(12):3543-3563(周博文, 李阳, 马鑫骥, 苗壮, 张睿. 2024. 深度学习的跨视角地理定位方法综述. 中国图象图形学报, 29(12):3543-3563)[DOI:10.11834/jig.230858]

深度学习的跨视角地理定位方法综述

周博文,李阳*,马鑫骥,苗壮,张睿

陆军工程大学指挥控制工程学院,南京　210007

摘　要: 跨视角地理定位技术是计算机视觉领域中的重要问题之一,因其可在缺乏卫星定位环境中实现实时定位,一直受到图像配准、导航定位和图像检索等诸多领域的关注。传统的跨视角地理定位方法采用手工特征进行特征抽取,导致定位精度受限。随着深度学习技术的发展,深度学习的跨视角地理定位方法成为当前的主流技术。但由于跨视角地理定位任务涉及多个步骤、迁移知识广泛,因此本领域仍缺少相关综述。本文首次从跨视角地理定位任务框架的视角,对当前深度学习的跨视角地理定位方法进行全面综述。在问题概述的基础上,对数据预处理、深度学习网络、特征注意力模块和损失函数等技术的发展进行了归纳总结。通过对近百篇高影响力文献的梳理,本文总结出跨视角地理定位任务的特性和改进思路,有助于启发研究者设计新方法。此外,还在两个具有代表性的数据集上分别测试了10种不同深度学习的跨视角地理定位方法。从实验精度、模型的参数量和推理速度3个方面综合评估了现有方法的性能。最后,基于对上述跨视角地理定位方法的归纳分析,本文结合实际应用指出该领域存在的一些问题,并对未来发展趋势进行讨论,希望为该领域感兴趣的学者提供参考。

关键词: 跨视角;地理定位;图像检索;深度学习;注意力;无人机

A survey of cross-view geo-localization methods based on deep learning

Zhou Bowen, Li Yang*, Ma Xinji, Miao Zhuang, Zhang Rui

Command and Control Engineering College, Army Engineering University of PLA, Nanjing　210007, China

Abstract: Cross-view geo-localization aims to estimate a target geographical location by matching images from different viewpoints. This method is usually viewed as an image retrieval task that has been widely adopted in various artificial intelligence tasks, such as person re-identification, vehicle re-identification, and image registration. The main challenge of this localization task lies in the drastic changes among different viewpoints, which reduce the retrieval performance of the model. Conventional techniques for cross-view geo-localization rely on manual feature extraction, which restricts precision when determining location. With the development of deep learning techniques, deep learning-based cross-view geo-localization methods have become the current mainstream technology. However, due to the involvement of multiple steps and the extensive transfer of knowledge in cross-view geo-localization tasks, only a few studies have been conducted in this field. In this paper, we present the first review of cross-view geo-localization methods based on deep learning. We analyze the various developments in data preprocessing, deep learning networks, feature attention modules, and loss functions within the context of cross-view geo-localization tasks. To address the challenges in this field, the data preprocessing phase involves feature alignment, sampling strategies, and data augmentation. Feature alignment serves as prior knowledge for

收稿日期:2023-12-12;**修回日期:**2024-03-13;**预印本日期:**2024-03-20

*通信作者:李阳　solarleeon@outlook.com

基金项目:国家自然科学基金项目(61806220)

Supported by: National Natural Science Foundation of China(61806220)

cross-view geo-localization that contributes to improving the localization accuracy. The use of GAN networks has emerged as a prominent trend for feature alignment. Additionally, the discrepancy in sample quantities among satellite, ground, and drone images necessitates the use of effective sampling strategies and data augmentation techniques to achieve training balance. Deep learning networks play a critical role in extracting image features, and their performance directly impacts the accuracy of cross-view geo-localization tasks. In general, the methods that use Transformer as the backbone network have a higher accuracy than those that based on ResNet. Meanwhile, those methods that use the ConvNeXt network show the best performance. To further extract image features and enhance the discriminative power of the model, feature attention modules need to be designed. By learning effective attention mechanisms, these modules adaptively weight the input images or feature maps to improve their focus on task-relevant regions or features. Experimental results show that these modules can explore previously unattended feature information, further extract image features, and enhance the discriminative power of the model. Loss functions are used to improve the fit of the model to the data and to accelerate its convergence. Based on their results, these functions guide the training direction of the entire network based, thus enabling the model to learn better representations and further improve the accuracy of cross-view geo-localization tasks. Some of the most commonly used loss functions include contrastive loss and triplet loss. With the improvement in these loss functions, the number of samples extracted by the model evolves from one-to-one to one-to-many, thus allowing the model to cover all samples during training and further enhance its performance. By analyzing nearly a hundred pieces of influential literature, we summarize the characteristics and propose some ideas for improving cross-view geo-localization tasks, which can inspire researchers to design new methods. We also test 10 deep learning-based cross-view geo-localization methods on 2 representative datasets. This evaluation considers the backbone network type and input data size of these methods. In the University-1652 dataset, we evaluate the accuracy metrics R@1 and AP, the model parameters, and the inference speed. In the CVUSA dataset, we mainly evaluate four accuracy metrics, namely, R@1, R@5, R@10, and R@Top1. Experimental results show that a better backbone network and a large image data input size positively affect the performance of the model. Building upon an extensive review of the current state-of-the-art cross-view geo-localization methods, we also discuss the related challenges and provide several directions for further research on cross-view geo-localization.

Key words: cross-view; geo-localization; image retrieval; deep learning; attention; drone

0 引 言

跨视角地理定位是指将实时查询图像与带有精准地理定位的图像进行匹配，从而确定查询图像的地理位置（Shen等，2024；Zhou等，2023）。随着信息技术的飞速发展，跨视角地理定位技术已广泛应用于无人机导航（Zhu等，2021；Wei和Wang，2018）、自动驾驶（Kim和Walter，2017）和事件检测技术（Choi等，2020；Algamdi等，2020）等多个领域。

根据无人平台的拍摄视角，跨视角地理定位通常包含两种任务：地面—卫星跨视角地理定位与空中—卫星跨视角地理定位（Lin等，2022）。早期研究者主要关注地面—卫星跨视角地理定位，该任务旨在解决地面无人平台所拍摄的查询图像与具有地理标记的卫星参考图像之间的匹配问题（Zhang等，2023a）。随着无人机的广泛应用（Zhou等，2022；Rave等，2023），空中—卫星跨视角地理定位逐渐成为研究热点，该任务旨在解决空中无人平台所拍摄的查询图像与具有地理标记的卫星参考图像之间的匹配问题（Zheng等，2020）。虽然两种任务的视角有所差异，但所关注的难点都是如何提高跨视角地理定位的准确性与实时性。

早期跨视角地理定位主要采用图像检索中的关键点匹配方法进行解决（Arandjelovic等，2016；Liu和Li，2019）。不同视角图像中的关键点检测方法主要使用手工特征检测器检测关键点，如表1所示。这些方法虽然可实现跨视角地理定位，但其定位的准确性受到制约。Workman和Jacobs（2015）的实验验证了预先在Places数据集（Zhou等，2014）上训练的深度学习模型性能优于手工特征。

自2015年开始，深度学习的方法在跨视角地理定位中取得了优异表现。该类方法是跨模态检索的一个子类，主要针对不同视角的图像进行检索。由于该类方法的优异性能，基于有监督的深度学习方法成为跨视角地理定位领域的研究热点（Workman

等，2015）。研究者希望借助深度神经网络充分提取不同视角图像的特征，进而提高匹配性能。深度学习的跨视角地理定位方法的框架通常分为 4 个部分：数据预处理、深度学习网络、特征注意力模块和损失函数，如图 1 所示。

经过前期文献收集工作发现，目前跨视角地理定位领域的相关综述仍较少（黄高爽 等，2023）。为帮助读者了解跨视角地理定位领域的发展状况，我们整理归纳了现有深度学习的跨视角地理定位方法，希望针对不同方法的缺陷设计更合理的解决方案。本综述贡献主要包括以下 3 个方面：1）首次按跨视角地理定位任务框架，针对当前深度学习的跨视角地理定位方法进行全面综述。2）在 University-1652 数据集和 CVUSA（cross-view USA）数据集上测试了不同深度学习的跨视角地理定位方法。根据实验精度、参数量和推理速度等方面，综合评估了现有方法的性能，总结出跨视角地理定位方法的特性。3）讨论了该领域存在的一些问题，并对未来发展趋势进行展望，为研究者设计新方法提供了启示。

表 1　传统的手工特征检测器

Table 1　Traditional handcrafted feature detectors

方法	主要贡献
Harris（Harris 和 Stephens，1988）	通过计算图像的一阶和二阶导数来挖掘几何结构
FAST（Rosten 和 Drummond，2006）	计算点周围较亮或较暗像素的数量，以提高性能和效率，加速关键点检索
SIFT（Lowe，1999）	集成了检测器和描述符，通过在多个尺度级别上用高斯滤波器卷积图像来寻找尺度不变的角点或斑点
SURF（Bay 等，2008）	通过使用 Hessian 矩阵和积分图像的近似来加速检测过程
MSER（Matas 等，2004）	在不同的阈值下对图像进行二值化，并选择稳定区域作为关键点
KAZE（Alcantarilla 等，2012）	将 Hessian 矩阵应用于多尺度下的非线性扩散尺度空间

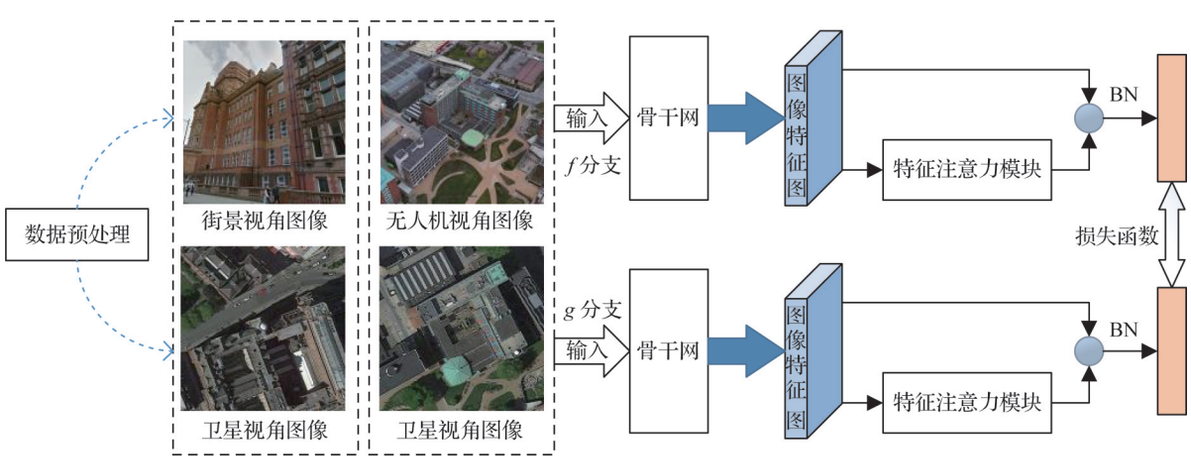


图 1　跨视角地理定位方法框架图

Fig. 1　The framework of cross-view geo-localization method

1　跨视角地理定位概述

1.1　问题定义

跨视角地理定位任务可以描述为：匹配不同视角的地理图像，通常是无 GPS（global positioning system）数据图像与有 GPS 数据图像之间的匹配。地面视角图像和无人机视角图像通常不具有 GPS 数据，而卫星视角图像具有准确的 GPS 数据（Song 等，2023）。本文将具有 GPS 数据的图像定义为 $\boldsymbol{I}_i^w$，将不具有 GPS 数据的图像定义为 $\boldsymbol{I}_i^o$，其中 $i = 1, \cdots, N$。当给定具有索引 q 的查询图像 $\boldsymbol{I}_q^o$ 时，跨视角地理定位的目的是搜索具有 GPS 数据的最佳匹配图像 $\boldsymbol{I}_a^w, a \in \{1, \cdots, N\}$，将其 GPS 数据赋予查询图像 $\boldsymbol{I}_q^o$。

深度学习的跨视角地理定位任务通常使用孪生神经网络结构作为主流框架(Chopra 等, 2005; Arandjelovic 等, 2016),将无GPS数据图像与有GPS数据图像数据{($\boldsymbol{I}_i^o$, $\boldsymbol{I}_i^w$)}传入该网络的双分支结构f(地面/无人机)与g(卫星)中。双分支结构学习输入数据的映射函数$f(\boldsymbol{I}_i^o, \boldsymbol{\Theta I}_i^o)$,将其映射到$K$维空间中,最后生成一个特征向量,从而检索到最佳匹配图像$\boldsymbol{I}_i^w$,使得$g(\boldsymbol{I}_i^w, \boldsymbol{\Theta I}_i^w)$是嵌入空间中$f(\boldsymbol{I}_i^o, \boldsymbol{\Theta I}_i^o)$的最近样本。同时,孪生神经网络中权重是否共享取决于输入图像之间是否存在显著视觉差异。在共享权重的情况下,该模型的双分支f与g学习相同的特征表示来增强模型的表达能力;在不共享权重的情况下,双分支f与g分别独立训练,以便从两个视角图像中提取不同的判别特征(见图1)。

1.2 数据集

早期跨视角地理定位研究中,由于图像集中于地面和卫星视角,因此地面—卫星跨视角地理定位任务是当时的研究热点。该研究不仅关注建筑物本身所带来的特征信息,而且泛化研究街道、乡村和城市等场景中的定位问题,因此构造出数据集CVUSA(Workman 等, 2015)、San Francisco(Workman 和 Jacobs, 2015)、CVACT(cross-view act)(Liu 和 Li, 2019)和 VIGOR(cross-view image geo-localization beyond one-to-one retrieval)(Zhu 等, 2021)。随着无人机技术的成熟,无人机在飞行任务中难免会遇到卫星拒止(global navigation satellite system-denied, GNSS-denied)环境(Choi 和 Myung, 2020; Kinnari 等, 2021),因此空中—卫星跨视角地理定位任务被提出。Zheng 等人(2020)利用地面视角下的图像作为建筑的辅助表示,以提高该任务的准确性。他们所创建的University-1652数据集中提供了无人机、卫星和地面3种视角图像。此外,还有数据集SUES-200(Shanghai University of Engineering Science-200)(Zhu 等, 2023c)可用于评估空中—卫星跨视角地理定位任务。表2总结了常见的6种数据集的基本信息。

表2 常见数据集的基本信息

Table 2 Basic information about common datasets

数据集	发布时间	图像尺寸/像素	图像类型	数据量	地点
CVUSA(Workman 等, 2015)	2015	1 232×224(地面)/750×750(卫星)	地面图像和卫星图像	44 416对地面和卫星图像	美国
San Francisco (Workman 和 Jacobs, 2015)	2015	256×256	地面图像和卫星图像	278 561幅地面图像和174 217幅卫星图像	旧金山弗朗西斯科
CVACT(Liu 和 Li, 2019)	2019	多尺寸	地面图像和卫星图像	44 416对地面和卫星图像以及92 802幅额外测试图像	澳大利亚堪培拉
VIGOR(Zhu 等, 2021)	2021	2 048×1 024(地面)/640×640(卫星)	地面图像和卫星图像	90 618幅卫星图像和238 696幅地面图像	纽约市(曼哈顿)、旧金山弗朗西斯科、芝加哥和西雅图
University-1652 (Zheng 等, 2020)	2020	512×512	地面图像、无人机图像和卫星图像	1 652组地面、无人机和卫星图像	世界各地72所大学
SUES-200(Zhu 等, 2023c)	2023	1 080×1 080(无人机)/512×512(卫星)	无人机图像和卫星图像	24 120对无人机和卫星图像	上海工程科技大学周围

1)CVUSA数据集(Workman 等, 2015)是一个包含44 416对地面和卫星图像的数据集。该数据集包含美国各地的建筑以及部分街道、乡村和城市场景,其中,地面图像来自谷歌街景全景图,匹配的卫星图像来自微软必应地图,如图2(a)所示。数据集包括35 532对训练样本和8 884对测试样本,每个卫星图像都有一个对应的地面图像,主要用于跨视角地理定位任务。

2) San Francisco 数据集(Workman 和 Jacobs, 2015)是一个包含278 561幅地面图像和174 217幅卫星图像的数据集。该数据集的图像集中于旧金山弗朗西斯科周围一个200平方公里的区域范围内。

数据集的卫星图像是从Bing地图上收集而来，对应的地面图像来源于Flickr和谷歌街景图像。该数据集是较早用于评估地面—卫星跨视角地理定位任务的数据集之一。

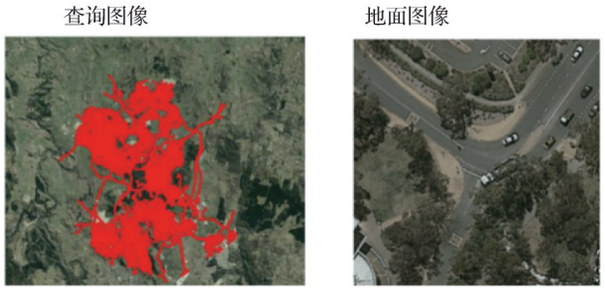


CVUSA数据集中匹配的地面和卫星图像示例

(a) CVUSA数据集(Workman 等，2015)



查询图像　　地面图像



卫星数据(带有GPS信息)　　卫星图像

(b) CVACT数据集(Liu和Li，2019)



地面图像　　卫星图像



真实无人机图像　　合成的无人机图像

(c) University-1652数据集(Zheng 等，2020)

图2　常用数据集展示

Fig. 2　Examples of common datasets

((a) CVUSA(Workman et al. ,2015); (b) CVACT(Liu and Li,2019); (c) University-1652(Zheng et al. ,2020))

3)CVACT数据集(Liu和Li,2019)是一个训练集和测试集的规模与CVUSA相同的大规模跨视角数据集。该数据集使用的是澳大利亚堪培拉地区的建筑物，包含大部分都市建筑，其中地面图像来自谷歌街景全景图，相匹配的卫星图像来源于谷歌地图。为了支持细粒度的城市地理定位测试，该数据集还提供了一个包含92 802幅图像的额外测试集，如图2(b)所示。与CVUSA相同，卫星图像与地面图像之间的映射关系也是一对一的。

4)VIGOR数据集(Zhu 等，2021)包含90 618幅卫星图像和238 696幅地面图像。该数据集包含来自美国4个主要城市(旧金山、纽约、西雅图和芝加哥)的建筑场景。数据集中卫星图像和地面图像均是从谷歌地图中采集，其中一幅查询图像对应多幅参考图像。该数据集用于评估两种不同的跨视角地理匹配形式：相同区域和交叉区域。在相同区域的跨视角地理定位任务中，所有城市的图像都用于训练和验证；而在交叉区域的跨视角地理定位任务中，设置纽约和西雅图为训练集，并设置旧金山弗朗西斯科和芝加哥为测试集，进一步验证跨视角地理定位方法的性能。

5)University-1652数据集(Zheng 等，2020)是一个包含1 652组地面、无人机和卫星图像的多视角数据集。该数据集没有选择地标建筑作为目标位置，而是选择了世界各地72所大学的1 652个普通建筑作为目标，其中地面图像和卫星图像来源于谷歌地图，无人机图像是利用Google Earth提供的3D模型来模拟真实的无人机相机拍摄而成，如图2(c)所示。该数据集中平均每栋楼有1幅卫星图像和54幅无人机图像。训练集中有33所大学的701个建筑目标，而另外39所大学的951个建筑目标构成测试集。训练集和测试集的大学建筑物之间没有重叠，具体分布如表3所示。

表3　图像数据在University-1652数据集上的分布

Table 3　Distribution of image data on the University-1652 dataset

数据集	类别	图像数目	建筑数量	大学数量
训练集	无人机	37 854	701	33
	卫星	701	701	
	地面	11 640	701	
测试集	无人机(图库)	51 355	951	39
	卫星(图库)	951	951	
	无人机(查询集)	37 854	701	
	卫星(查询集)	701	701	

University-1652数据集可用于评估两个新的跨视角地理定位任务。任务(1)无人机视角目标定位。该任务给定一幅无人机视角图像,搜寻与之最相似的卫星图像。由于卫星图像具有GPS数据,所以可对无人机视角中的目标进行定位。该任务的查询集有37 854幅无人机图像,图库中有951幅匹配的卫星图像。任务(2)无人机导航。该任务给定一幅卫星图像,让无人机找寻它所拍摄过的无人机视角图像。如果找到就让无人机回到它所拍摄的位置,完成一次导航操作。该任务中,查询集有701幅卫星图像,图库中有51 355幅匹配的无人机图像。

6)SUES-200数据集(Zhu等,2023c)是一个包含24 120对无人机和卫星图像的数据集。该数据集采集于上海工程科技大学周围200个地点,图像信息不仅限于校园建筑,还包括公园、学校、湖泊和公共建筑。数据集中共收集了200个地点,其中无人机图像被划分为4个高度(150 m、200 m、250 m和300 m),每个无人机高度有50个无人机视角图像,每个地点有1个相应的卫星视角图像。SUES-200中的图像分为训练集和测试集,其中,60%用于训练,40%用于测试,以评测不同高度的空中—卫星跨视角地理定位任务。

由于跨视角地理定位领域常用的3个数据集分别为CVUSA、CVACT和University-1652。因此,本文在常用数据集展示中针对上述3个数据集进行了展示,如图2所示。

1.3 评价指标

在深度学习的跨视角地理定位任务中,使用召回精度(Recall@K,R@K)和平均精度(average precision, AP)来评估模型的性能。Recall@K也称为K点召回率,从查询集(query)中选择一幅查询图像,到参考图库集(gallery)中进行检索,通过计算在前K个结果中返回的正确图像的百分比来评估检索算法的有效性。例如,如果K的值是5,那么Recall@5是前5个条目中检索到的正确图像的百分比,K常取值为1,5,10;如果K的值是1%,则指占参考图库集(Gallery)数量的1%。召回率计算为

$$\text{Recall@}K=\sum_{i=1}^{n}\frac{S_{i,k}}{n} \tag{1}$$

式中,n代表查询图像的数量,i代表标签为i的查询图像。如果返回的前K个结果中有正确图像,则$S_{i,k}$为1,否则为0。Recall@K值越高,检索算法的性能越好。

精确率(Precision)是指检索出的相关结果数与结果总数之间的比率,衡量的是检索系统的查准性能。

由于单独使用召回率和精确率的评估不够全面,因此提出将精确率—召回率(precision-recall,P-R)曲线下的面积作为更全面的性能衡量指标,即平均精度AP。该指标同时反映了检索性能的查准率和召回率。

在数据集CVUSA和CVACT中采用K = 1, 5, 10和Top1(R@1,R@5,R@10,R@Top1)作为评价指标。在数据集University-1652中使用了K = 1,即R@1情况下的精度和平均精度作为评价指标。

2 跨视角地理定位方法总结

根据深度学习的跨视角地理定位框架(如图1所示),本文将跨视角地理定位方法分为数据预处理、深度学习网络、特征注意力模块以及损失函数4个部分。在数据预处理部分(2.1节),根据跨视角地理定位任务的特性分析了特征对齐问题、数据集的抽样策略和数据增强方法;在深度学习网络部分(2.2节),根据特征提取网络的不同属性,将其分为基于卷积神经网络(convolutional neural network, CNN)的方法、基于Transformer的方法和基于混合结构的方法;在特征注意力模块部分(2.3节),按时间顺序归纳了现有的特征注意力模块,用于进一步处理经过骨干网络提取的特征图;在损失函数部分(2.4节),分析了目前常用的损失函数,包括对比损失、三元组损失和其他损失。

2.1 数据预处理

由于跨视角地理定位数据的采集视角、采集时间和采集方式不同。因此,其采集的数据样本存在特征偏差、样本不均衡和样本不充足等问题。为解决以上问题,本节主要从特征对齐、抽样策略和数据增强3个方面介绍该领域的数据预处理方法。

2.1.1 特征对齐

在特征对齐工作中,Liu和Li(2019)以及Shi等人(2020a)指出现有方法将跨视角地理定位任务视为一个纯粹的位置估计问题,但忽略了方向对齐。Liu和Li(2019)进一步指出,如果能够完成方向对齐操作,则查询图像可以放置在相对于卫星图像的几

何参考坐标系中，进而减少寻找匹配图像的搜索空间，实现更快的搜索和更准确的匹配。因此，方向对齐步骤可以作为跨视角地理定位任务中的补充信息。

Shi等人（2020b）指出现有的大多数深度学习方法忽略了特征空间的布局信息。因此，受人类在定位时会记住建筑物相对空间位置的启发，设计出CVFT（cross-view feature transport）模块。该模块插入在分支f（地面/无人机）中用于学习特征转换，使得匹配的不同视角的图像特征可以在空间上对准，进而缩小两个视角之间的域差异。Wang等人（2022d）指出现有方法忽略了邻近区域的上下文信息，因此采用方形回环特征划分策略，划分对象为数据集事先对齐的邻域区域。实验结果表明，使用特征对齐作为先验知识可以有效提升定位精度。

此外，Regmi和Shah（2019）在研究极坐标变换工作的过程中，发现经过极坐标变换后的卫星图像虽然视角域发生转换，但其与现实的地面图像相差较大（见图3）。因此，他们利用生成对抗网络（generative adversarial network，GAN）从一个视角合成指定视角图像（Goodfellow等，2014），可以获得更优的生成图像信息用于跨视角地理定位任务。Toker等人（2021）在Regmi和Shah（2019）的工作基础上，研究卫星图像到地面图像的合成工作，提出将极坐标变换后的图像作为GAN的输入数据，可以生成更逼真的地面图像，如图3所示。因此，由GAN生成的图像与目标域图像进行训练，可以有效地提升跨视角地理定位性能（Hu等，2022a）。利用GAN实现特征对齐的方法逐渐成为该领域的发展趋势。特征对齐方法总结如表4所示。



图3　不同方法生成图像与真实图像差异（Toker等，2021）

Fig. 3　Examples of the difference between the image generated by different methods and the ground truth（Toker et al. ，2021）

表4　特征对齐方法总结

Table 4　Summary of feature alignment methods

特征对齐类型	时间	论文来源	效果	特点
方向对齐（Liu和Li，2019）	2019	CVPR	效果较差	完成方向对齐操作，可实现更快搜索和更准确匹配
极坐标变换（Shi等，2019）	2019	NeurIPS	效果一般	对视角差异问题的初步探索，但与真实的指定视角图像相差较大
GAN（Regmi和Shah，2019）	2019	ICCV	效果较好	使用GAN生成指定视角域的图像，有效地缓解了语义鸿沟问题
极坐标变换 + GAN（Toker等，2021）	2021	CVPR	效果最佳	将极坐标变换后的图像作为GAN网络的输入，可生成较为逼真的指定视角图像

2.1.2　抽样策略

跨视角地理定位中抽样操作是指从同类目标的不同视角中抽取图像，将其送入网络模型的双分支f与g中。该操作的目的是获取来自不同平台具有相

同地理标签的正样本对，用于构建训练批次（Wang等，2022a）。然而，由于图像收集源的不同，致使样本数量之间存在失衡。其中，卫星图像较难获取，该类图像数量偏少；而地面图像和无人机图像采集具有便利性，因此该类图像数量较多。

为了解决样本不均衡的问题，Zhu等人（2021）在制作数据集VIGOR的过程中，使用了更为密集的采样操作。该方法设计了两种均衡策略，如表5所示。策略2表示确保每幅卫星图像对应2个正样本地面图像；策略3表示确保每幅卫星图像对应3个正样本地面图像。

从表5的R@1评价指标可以看出，设置均衡策略3可以有效地提升模型的性能。Ding等人（2020）使用随机旋转、裁剪和擦除等方式，生成新的卫星视角图像，将目标位置的卫星图像数量从1扩展到54，最终使得训练数据集的目标位置均有54幅无人机和卫星图像。Wang等人（2022a）设计出对称抽样策略，该策略包括卫星视角抽样和无人机视角抽样。卫星视角抽样利用随机卫星图像作为锚点，找到具有相同地理标签的无人机图像形成正样本对；同理，无人机视角抽样利用随机无人机图像作为锚点，找到具有相同地理标签的卫星图像形成正样本对。该策略确保模型可以使用所有训练数据，同时保持类别抽样相对平衡。Dai等人（2022）借鉴Ding等人（2020）的做法，提出了一种多重抽样策略。该方法在University-1652数据集中获取卫星视角下的图像，生成k幅增强卫星图像。其中，增强方法包括随机移位、随机填充、随机剪切和随机颜色增强等。Deuser等人（2023）借鉴Dai等人（2022）的方法，提出了动态相似性抽样（dynamic similarity sampling，DSS）。DSS抽取每个查询图像的前k个最近邻生成一个集合，并将k个邻居根据相似性进行排序，选择k/2个最近的样本作为批次，剩余的k/2个样本从k中的其他样本随机选择（其中，k设置为小于或等于批次大小）。这种抽样方式进一步解决了样本不均衡问题，而且经过DSS操作后的样本提升了模型的泛化能力。

表5　VIGOR数据集中的均衡策略

Table 5　Balancing strategies in the VIGOR dataset

策略	每幅卫星图像中的正样本数	地面图像数量	卫星图像数量	R@1/%
原始	/	238 696	90 618	/
策略2	2	105 214	90 618	41.1
策略3	3	149 869	90 618	48.5

注："/"表示未提供该信息。

2.1.3　数据增强

由于跨视角地理定位任务中存在样本不均衡问题，因此数据增强方法也可用于解决该问题。数据增强首先可以避免模型训练过拟合（Chen等，2022a）。当数据集具有某种明显的特征时，使用裁剪等方法可避免模型学到与目标无关的信息（Chen等，2022b）。其次，通过使用遮挡、亮度和模糊等方法，给训练数据加上噪声，可使训练后的模型具有更强的鲁棒性（Gong等，2021；Wang等，2022b）。此外，还可以通过对数据增强后的图像进行测试，验证模型的鲁棒性。

在数据增强领域，人工设计数据增强方法逐渐被自动选择增强策略替代。Cubuk等人（2019）指出手工配置数据增强方法效率低且与图像最适合的数据增强方法难以选择，因此提出Auto Augment自动搜索算法改进数据增强方法。最终，通过搜索算法找到最佳策略，使数据集能够达到更好的数据增强效果。Chen等人（2021）提出了一种名为生成对抗U-Net的数据增强方法，该方法利用U-Net结构作为生成器生成多样化的图像，解决了标注图像不足的问题。

在跨视角地理定位任务中，数据增强方法也有相关运用。Zhang等人（2023d）指出图像经过翻转后虽然保持了细节信息，但修改了几何布局。因此，该方法在数据增强的基础上提出布局模拟的增强方法。原始的数据增强采用颜色抖动来修改图像中的亮度、对比度和饱和度；布局模拟的增强是在图像上应用随机翻转和随机旋转（90°、180°或270°）来修改图像的空间信息，进而产生具有不同几何布局的匹配图像丰富训练样本。Dai等人（2022）指出跨视角地理定位应增强对位置偏移的鲁棒性，因此提出FSRA（feature segmentation and region alignment）方法。该方法中使用BP（BlackPad）和FP（FlipPad）两种偏移操作进行数据增强。BP操作是在图像左侧填充宽度为P的黑色块、右侧剪切宽度为P的图像；FP操作是翻转图像左侧宽度为P的部分、剪切右侧宽度为P的图像。上述两个数据增强操作的组合，

在保持细节信息的同时，进一步挖掘了图像的空间信息。Zhu 等人（2023a）在 Dai 等人（2022）的基础上，仍然使用翻转填充变换作为数据增强的方法。他们在翻转填充时，翻转图像左侧的 n 个像素且裁剪图像右侧的 n 个像素。实验中，对增强后的图像进行模型测试，结果表明该方法提升了模型对位置偏移的鲁棒性，验证了数据增强可以丰富数据集并提升视觉抗干扰能力。

2.2 深度学习网络

由于跨视角地理定位需要利用不同视角的图像特征，进行匹配以实现定位任务，因此，图像的特征提取方法成为影响其性能的主要因素。当前深度学习特征提取网络的方法主要包括：基于 CNN 的方法、基于 Transformer 的方法和基于混合结构的方法。表 6 是对现有不同网络结构的跨视角地理定位方法的分类和归纳。

表 6 不同跨视角地理定位方法分析

Table 6 Analysis of different cross-view geo-localization methods

方法	来源	年份	类型	数据集	特点
Workman 等人（2015）	ICCV	2015	卷积神经网络	CVUSA	优点：特征表达能力强，能更好适应视角差异 缺点：解释性不强，需要大量图像进行训练
Workman 和 Jacobs（2015）	CVPR	2015	卷积神经网络	San Francisco	
Lin 等人（2015）	CVPR	2015	卷积神经网络	自建数据集	
Kan 等人（2016）	CVPR	2016	卷积神经网络	自建数据集	
Zheng 等人（2017）	TOMM	2017	卷积神经网络	Oxford5k（Philbin 等，2007）	
Liu 和 Li（2019）	CVPR	2019	卷积神经网络	CVUSA 和 CVACT	
Zheng 等人（2020）	ACM MM	2020	卷积神经网络	University-1652	
Zhuang 等人（2021）	Remote Sensing	2021	卷积神经网络	University-1652	
Wang 等人（2022d）	TCSVT	2021	卷积神经网络	University-1652、CVUSA、CVACT、Oxford5k（Philbin 等，2007）	
Dai 等人（2022）	TCSVT	2021	Transformer	University-1652	优点：特征表达能力和全局上下文推理能力更强，有效地减少视觉歧义 缺点：模型参数量偏大
Zhu 等人（2022）	CVPR	2022	Transformer	CVUSA 和 VIGOR	
Zhuang 等人（2022）	ACCESS	2022	Transformer	University-1652	
Yang 等人（2021）	NeurIPS	2021	混合结构	CVUSA 和 CVACT	
Zhao 等人（2022）	arXiv	2022	混合结构	CVUSA 和 CVACT	
Zhang 等人（2023d）	AAAI	2022	混合结构	CVUSA 和 CVACT	
Zhu 等人（2023b）	arXiv	2023	混合结构	University-1652、CVUSA、CVACT、VIGOR 和 Oxford5k	
Zhu 等人（2023a）	Sensors	2023	混合结构	University-1652 和 SUES-200	
Shen 等人（2024）	TCSCV	2023	混合结构	University-1652 和 SUES-200	
Deuser 等人（2023）	ICCV	2023	混合结构	CVUSA、CVACT、University-1652 和 VIGOR	

2.2.1 基于 CNN 的方法

基于卷积神经网络的方法首先通过 CNN 网络提取图像特征信息，生成输入图像的特征图。然后比对目标图像与数据库中图像的特征关系，进而推断出目标图像的位置信息。

Workman 等人（2015）首次将深度学习网络用于跨视角地理定位任务，所提出的网络架构用于学习地面与卫星图像的联合语义特征表示，并建立了第一个跨视角地理定位数据集 CVUSA。同年该团队进一步验证了深度学习模型可以提取具有区分性的图像特征，并且该特征具有区分两个地理区域的能力（Workman 和 Jacobs，2015），为该领域打下坚实基

础。Lin等人(2015)结合卷积神经网络在人脸验证中的成功,将其引入跨视角地理定位任务中,提出了Where-CNN的方法,并展示了该方法在跨视角地理定位中的有效性。Kan等人(2016)指出跨视角地理定位领域中不同视角的图像之间存在较大差异的问题,因此提出了多视图深度网络MvDN(multi-view deep network)用于寻找并消除复杂(高度非线性)的视图差异。但是,使用独特的网络分支寻找并消除多视图差异的方法会在训练和测试中产生较大的开销(Hu等,2018)。Zheng等人(2017)指出现有的跨视角地理定位方法区分度不够,因此提出一种能够同时计算识别损失和验证损失的孪生神经网络结构。该网络结合验证模型和识别模型两种卷积神经网络在行人再识别(Hermans等,2017)中的应用,将其引入跨视角地理定位任务中,进一步增加了模型区分度。为解决模型训练收敛较慢的问题,Hu等人(2018)提出CVM-Net(cross-view matching network)方法。该方法使用全卷积层提取图像局部特征,并利用NetVLAD将其编码为全局图像描述符进行图像检索,加快了模型收敛速度并提高了检索精度。Liu和Li(2019)指出方位信息可作为空间定位任务的辅助信息,提出了一种从图像外表和方向几何信息中联合学习特征嵌入的孪生神经网络结构。该方法建立了一种新的任务方式,有效提升了跨视角地理定位任务的精度。Zheng等人(2020)提出空中—卫星的跨视角地理定位任务。为了验证不同视角的图像是否具有不同的原始细节,Zheng等人(2020)在孪生神经网络的基础上设计出3分支网络,其输入分别为卫星、地面和无人机图像。但该方法目前主要集中于探索图像全局信息,而忽略了上下文信息(Wang等,2022d)。Zhuang等人(2021)指出现有跨视角地理定位模型缺乏对偏移量和尺度变化的鲁棒性,因此设计了一种具有全局和局部分支的多尺度块注意力网络(multiscale block attention network,MSBA)。网络的两个分支将图像分割成不同尺度的多个部分同时训练,将全局信息与局部信息结合,使该模型的精度超过了现有的最佳模型。Wang等人(2022d)指出现有方法通常集中于挖掘图像中心地理目标的细粒度特征,而忽略了邻近区域的上下文信息。因此,他们提出局部模式网络(local pattern network,LPN),该网络利用分区与图像中心的距离设计注意力分配,进而简化部分匹配并实现了逐部分的表示学习。实验结果表明,该方法对旋转变化具有良好的可扩展性,丰富了地理定位的判别特征。

2.2.2 基于Transformer的方法

近年来,ViT(vision Transformer)模型在计算机视觉领域受到广泛关注。ViT任务(Dosovitskiy等,2010)开创性地将自然语言处理(natural language processing,NLP)中的Transformer架构(Vaswani等,2017)应用于图像分类任务中。Touvron等人(2021)提出的DeiT(data-efficient image Transformer)方法证明了ViT模型在ImageNet数据集(Deng等,2009)上的图像分类问题优于CNN对应的方法。因此,使用Transformer模型的跨视角地理定位方法迅速取代了基于卷积神经网络的方法,并成为该领域的主流。

Dai等人(2022)指出现有方法低估了提取鲁棒特征表示的重要性和特征对齐的作用,因此提出了一个特征分割和区域对齐的网络FSRA。该方法利用Transformer提取特征图的热力分布并根据分布划分区域,然后在不同视图中一一对齐多个特定区域,从而实现更精确的匹配。但该方法忽略了场景中部分区域的位置信息并且缺乏对整个图像中上下文信息的探索(Shen等,2024)。Zhu等人(2022)提出了一个纯粹的基于Transformer的跨视角地理定位方法TransGeo(Transformer is all you need for cross-view image geo-localization),该方法不依赖于坐标对齐,并使用注意力引导的非均匀裁剪策略,消除了输入图像中大量无信息的斑块,减少了跨视角地理定位任务的计算量。Zhuang等人(2022)提出一个包含语义引导模块(semantic guidance module,SGM)的Transformer网络结构,该网络利用自注意力机制提取上下文信息,进一步解决了卷积神经网络中感受野受限的问题。

2.2.3 基于混合结构的方法

虽然直接将Transformer引入跨视角地理定位任务中可以提升该任务的检索精度,但是纯粹的ViT模型缺乏灵活性(Liu等,2021a)。基于混合结构的方法是指利用卷积神经网络或Transformer网络进行特征提取(Zhang等,2023b),然后使用基于卷积神经网络或Transformer设计的特征注意力模块(Dong等,2022)进一步丰富判别信息。

Yang等人(2021)将基于混合结构的方法运用于跨视角地理定位。该网络包含一个层到层Transformer网络模型(layer-to-layer Transformer,L2LTR),

并利用自注意力特性建模全局依赖关系，从而显著降低了跨视角地理定位中视觉模糊性问题。虽然该方法是基于混合结构的方法，但由于该方法采用CNN作为特征提取器，而自注意力机制和位置嵌入信息仅用于高层CNN特征，因此该方法没有充分利用第1层的全局建模能力和位置信息(Zhu等，2022)。Zhao等人(2022)指出现有方法中的孪生神经网络没有考虑两个分支之间的互利关系，因此提出了一个互补生成式学习网络(mutual generative Transformer learning，MGTL)。该网络的双分支可以生成互利的空中和地面知识，最终产生丰富的判别特征。Zhang等人(2023d)指出现有方法缺乏提取视觉特征的空间几何能力。因此，他们提出了GeoDTR(geometric descriptor Transformer)网络，从原始特征中分离几何信息，并且通过几何布局提取器模块学习来自不同视角图像的空间相关性，有效提升了模型性能。Zhu等人(2023b)指出现有特征提取网络是通用的，所以利用其解决跨视角地理定位这一特殊任务时性能不佳。因此，他们设计了基于简单注意力的跨视角地理定位网络(simple attention-based image geo-localization network，SAIG)。Zhu等人(2023a)指出现有方法忽视了外部因素特征和各种表征间的相互作用，因此提出将孪生神经网络的双分支分成两部分。第1部分采用ResNetV2学习细粒度特征；第2部分采用ViT网络从图像中提取全局特征，提升了模型的识别性能。

在网络模型的进一步发展中，Liu等人(2022)结合ResNet和Inception模型的优势重新设计卷积神经网络，得到ConvNeXt网络模型。该网络在处理速度和精确度方面与ViT网络具有相似的性能(Dosovitskiy等，2010；Liu等，2021a)，但在网络结构上更简单。Shen等人(2024)将ConvNeXt网络引入跨视角地理定位任务中，其中ConvNeXt模块如图4所示。该网络还包含一个跨维交互模块，用于提取不同维度的判别信息，使得该网络获取了更全面的特征表示。

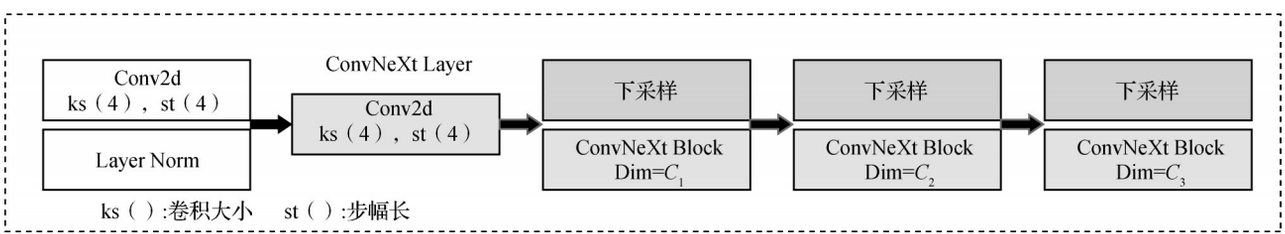


图4　ConvNeXt模块结构(Shen等，2024)

Fig. 4　Structure of ConvNeXt(Shen et al.，2024)

2.3　特征注意力模块

单纯地使用卷积神经网络或Transformer网络进行特征提取，只能得到原始图像的部分简单特征信息。为了进一步提取图像特征并增强模型的判别能力，需要设计特征注意力模块(Arar等，2022)。该类型模块通过学习有效的注意力机制，对输入图像或特征图进行自适应地加权，以便更加关注与任务相关的区域或特征(Naranjo-Alcazar等，2021)。

Cai等人(2019)提出了一个轻量级的双重注意力模块(feature context-based attention module，FCAM)，该模块包括通道注意力和空间注意力两部分。其中，通道注意力采用了卷积块注意力模块(convolutional block attention module，CBAM)，而空间注意力模块集成了一种上下文感知的特征重新加权策略，进一步提取了具有区分能力的特征。Lin等人(2022)提出了一种迫使模型聚焦于显著区域的单元减法注意力模块(unit subtraction attention module，USAM)，该模块使用单元减法卷积(unit subtraction convolution，USC)从特征图中提取关键点，进一步提升了挖掘细节特征的能力。Zhang等人(2023d)提出了几何布局提取器模块，用于挖掘视觉特征之间的全局上下文信息。该方法可产生有效的几何布局描述符用于调制原始特征，进一步提高了跨视角地理定位的辨别能力。Zhu等人(2023b)设计了一种空间混合特征聚合模块(spatial-mixed feature aggregation module，SMD)，在任务中可以自由地对长距离信息建模。该模块将空间信息混合并投影到低维空间中生成特征描述符，减少训练时的过度特征损失。

Shen等人(2024)受到三元组注意模块(Misra等,2021)的启发,设计了一种跨维度交互模块。该模块利用通道信息和空间信息处理特征图,建立了维度间依赖关系,实现了更佳的跨维度信息交互。实验结果表明,经过特征注意力模块处理的特征图可以提供更丰富更有效的特征信息。

2.4 损失函数

为了帮助模型更好地拟合数据,加快模型的收敛速度,需要不断改进损失函数。对比损失(Vo和Hays,2016;Varior等,2016)和三元组损失(Schroff等,2015)是跨视角地理定位领域中常见的距离损失函数。损失函数通常使用于模型训练阶段,用来度量模型的预测值与真实值之间的差异(Wood等,2022)。在本节中,我们将分析对比损失、三元组损失和其他损失在跨视角地理定位任务中的特性。

2.4.1 对比损失

对比损失相对于其他损失函数具有较少的计算量,并且可以有效地处理孪生神经网络中成对数据的关系。该函数最早是由Hadsell等人(2006)提出,主要用于度量两个样本之间的相似性或差异性,增加分类器的类间差异,计算为

$$L(\boldsymbol{x},\boldsymbol{y},l)=(1-l)\frac{1}{2}lD^2+\frac{1}{2}l\{\max(0,m-D)\}^2 \tag{2}$$

式中,$l\in\{0,1\}$表示输入图像$\boldsymbol{x},\boldsymbol{y}$是否与标签匹配,如果匹配,那么$l=1$;否则$l=0$。$m>0$是不匹配图像的函数间隔值,$D=\left\|f(\boldsymbol{x})-f(\boldsymbol{y})\right\|_2$表示图像在特征空间中的欧氏距离。

Lin等人(2015)为了优化现有的跨视角地理定位网络,将对比损失引入该任务中,让匹配的图像具有接近0的较小欧氏距离,而让不匹配的图像具有大于间隔值m的较大欧氏距离。为了测试不同损失函数对模型训练的影响,Zheng等人(2017)在任务中分别使用交叉熵损失以及对比损失作为损失函数进行训练。结果表明,使用对比损失训练的模型在R@1和AP指标上效果分别下降了4.39%和6.55%;Hu等人(2018)使用对比损失训练CVM-Net模型并在CVUSA数据集上进行测试,最终结果均不如三重损失和四重损失。因此,在后续的跨视角地理定位任务研究中,对比损失逐渐被其他损失所取代。

2.4.2 三元组损失

由于对比损失每次仅考虑两个样本的类间差异,因此在训练任务中存在性能不足的问题。在跨视角地理定位任务中,三元组损失随机选择一幅图像作为锚点,然后拉近同类标签图像并推远不同标签图像。该损失有利于减小来自不同视角但具有相同标签值的特征向量之间的距离(Ding等,2022)。

Schroff等人(2015)首次提出三元组损失,该损失需利用3个样本组成的样本对(包括一个锚点样本、一个正样本和一个负样本)。对于每组样本中的锚点样本,需从同一标签中随机选取一个正样本和另一个不同标签中的负样本,使得它们之间的距离差异比较大。然后,通过对比这3个样本之间的距离关系,更新模型中的参数,使得锚点样本与正样本之间的距离尽可能小,而与负样本之间的距离尽可能大(Xu等,2020),计算为

$$L=\max(d(a,n)-d(a,p)+m,0) \tag{3}$$

式中,a表示锚点样本;p表示与a是同一标签的样本,即正样本;n表示与a是不同标签的样本,即负样本;m是一个大于0的超参数。

Vo和Hays(2016)指出基于欧氏距离的常见损失函数不是最佳的损失函数,因此提出一种类似于标准softmax的损失函数,用于加快三元组损失的收敛速度,计算为

$$L=\log(1+\mathrm{e}^{d_{\mathrm{pos}}-d_{\mathrm{neg}}}) \tag{4}$$

式中,d_{pos}和d_{neg}是所有正样本和负样本到所选锚点样本的距离。

在提高损失函数收敛性的研究中,Hu和Lee(2020)指出原始三元组损失中的阈值m需要经过仔细选择以获得模型优化的最佳结果(见式(4))。因此,他们在Vo和Hays(2016)的基础上,提出了一种软边界三元组损失。该方法为避免选取确定阈值m造成模型训练过慢的问题,通过系数α缩放距离提高模型收敛速度,计算为

$$L=\ln(1+\mathrm{e}^{\alpha(d_{\mathrm{pos}}-d_{\mathrm{neg}})}) \tag{5}$$

2.4.3 其他损失

三元组损失在跨视角地理定位任务中有着广泛的运用,为了进一步增强模型对样本的区分能力以及提高模型收敛速度,Chen等人(2017)在三元组损失(Schroff等,2015)的基础上提出了四元组损失(Quadruplet)。四元组损失在三元组损失的基础上引入了负样本的选择,训练时需使用4个样本,分别为锚点样本a、正样本p、第1个负样本n_1和第2个负

样本n_2。计算为

$$L=(d_{a,p}-d_{a,n_1}+\alpha)_+ +(d_{a,p}-d_{n_1,n_2}+\beta)_+ \quad (6)$$

式中，α和β是手动设置的超参数，通常设置α小于β，前一项称为强推动项，后一项称为弱推动项；$(\bullet)_+$是指$\max(\bullet,0)$操作。相比于三元组损失只考虑正负样本间的相对距离，四元组通过新添加一个负样本，使模型进一步学习到更好的表征。

现有的损失函数大多使用距离作为相似性度量，而距离度量在尺度变化时比较敏感。为进一步改善模型性能并提高模型收敛速度，Wang 等人(2017)提出一种角度损失。其主要思想是根据负样本处的角度对三元组内的三阶关系进行编码，通过限制角度的上界，推动负样本远离正簇的中心，并拉动正样本之间的距离，计算为

$$L_{\text{ang}}=\left\| \boldsymbol{X}_{\text{a}}-\boldsymbol{X}_{\text{p}} \right\|^2-4\tan^2\alpha\left\| \boldsymbol{X}_{\text{n}}-\boldsymbol{X}_c \right\|^2 \quad (7)$$

式中，$\boldsymbol{X}_{\text{a}}$表示锚点样本；$\boldsymbol{X}_{\text{p}}$表示正样本；$\boldsymbol{X}_{\text{n}}$表示负样本，$\boldsymbol{X}_c=1/2(\boldsymbol{X}_{\text{a}}+\boldsymbol{X}_{\text{p}})$，$\alpha$是一个大于0的超参数。该损失使用角度关系作为相似性度量，通过引入尺度不变性提高目标对于特征差异的鲁棒性并加快了收敛速度。

Van Den Oord 等人(2018)和 Radford 等人(2021)提出 InfoNCE(noise contrastive estimation)损失，它是对比损失的一种改进方案。Deuser 等人(2023)将其引入跨视角地理定位任务中，该损失利用批次中所有可用的负样本，丰富模型训练过程，计算为

$$L(\boldsymbol{q},R)_{\text{InfoNCE}}=-\log\frac{\exp\left(\frac{\boldsymbol{q}\cdot\boldsymbol{r}_+}{\tau}\right)}{\sum_{i=0}^{R}\exp\left(\frac{\boldsymbol{q}\cdot\boldsymbol{r}_i}{\tau}\right)} \quad (8)$$

式中，$\boldsymbol{q}$表示查询样本(地面或无人机图像)的编码向量；R表示参考图像(卫星样本)数量。$\boldsymbol{r}_+$表示与查询样本q具有相同标签值的正样本编码向量，当$\boldsymbol{r}_+$与$\boldsymbol{q}$匹配时，只有一个正的样本$\boldsymbol{r}_i$。InfoNCE 损失使用点积来计算查询图像和参考图像之间的相似度，当查询和正匹配相似时，该损失较低；而当负样本$\boldsymbol{r}_i$与$\boldsymbol{q}$不相似时，该损失较高；其中温度参数τ是一个超参数，可以设置为动态学习(Radford 等，2021)或设置为静态值。该方法在训练过程中使用更多负样本，通过一(正样本)对多(负样本)的训练关系进一步提升模型性能。

3 跨视角地理定位方法效果分析

本节根据跨视角地理定位任务中深度模型骨干网类别，将现有主流方法在两个主流数据集 University-1652 和 CVUSA 上进行测评。在 University-1652 数据集中测评内容包括：骨干网类型、输入尺寸、无人机—卫星之间两类检索任务的两种精度指标(R@1 和 AP)、参数量和推理速度，如表 7 所示。在 CVUSA 数据集中测评内容包括：骨干网类型、输入尺寸和地面—卫星的 4 个精度指标，如表 8 所示。实验结果从总体上可反映当前跨视角地理定位方法的水平。

3.1 University-1652 数据集结果分析

数据集 University-1652 主要执行空中—卫星跨视角地理定位任务。从表 7 中两类任务的两种精度指标(R@1 和 AP)可以看出，Sample4Geo(Deuser 等，2023)方法在两种匹配模式下召回精度 R@1 以及平均精度 AP 均最高。MCCG(multiple classifier for cross-view geo-localization)方法(Shen 等，2024)取得了次优的精度。值得注意的是，两类方法均使用了 ConvNeXt 网络结构，同时图像输入尺寸均为较大的 384 × 384 像素。由此可得到两个结论：1)新型网络结构能为跨视角地理定位任务提供更好的特征提取能力。具体而言，Sample4Geo 方法使用了能力更强的 ConvNeXt-B 网络模型，在无人机至卫星的任务模式下，召回精度 R@1 比 MCCG 方法(ConvNeXt 模型)提升了 3.01%、平均精度 AP 提升了 2.49%；在卫星至无人机的匹配模式下，召回精度 R@1 比 MCCG 方法提升了 0.74%、平均精度 AP 提升了 2%。2)采用较大的图像输入尺寸可使数据中包含更多的图像细节信息，进而改善跨视角地理定位能力。具体而言，MCCG 方法(Shen 等，2024)在输入尺寸为 384 × 384 像素的条件下比 256 × 256 像素的条件下，无人机至卫星的匹配任务中，召回精度 R@1 提升了 0.36%，平均精度 AP 提升了 0.31%；而卫星至无人机的匹配任务中，召回精度 R@1 提升了 0.01%，平均精度 AP 提升了 0.1%。

此外，从表 7 的骨干网类型可以看出，现有骨干网包括 ResNet 网络、Transformer 网络、ConvNeXt 网络以及它们的一些变形。总体来说，使用 Transformer 作为骨干网的方法精度大多高于使用 ResNet

表7　跨视角地理定位方法在University-1652数据集上的结果比较

Table 7　The results of cross-view geo-localization methods on the University-1652 dataset

方法	骨干网类型	输入尺寸/像素	无人机→卫星		卫星→无人机		参数量/M	推理速度
			R@1/%	AP/%	R@1/%	AP/%		
Baseline(Zheng等,2020)	ResNet-50	256×256	58.23	62.91	74.47	59.45	26	1.00×
LPN*(Wang等,2022d)	ResNet-101	256×256	76.13	79.29	85.45	75.45	45	1.51×
FSRA(Dai等,2022)	ViT-S	256×256	84.51	86.71	88.45	83.37	51	1.05×
MSBA(Zhuang等,2021)	ResNet-50	256×256	82.33	84.78	90.58	82.06	36	1.05×
SGM(Zhuang等,2022)	Swin Tiny	256×256	82.14	84.72	88.16	81.81	28	1.04×
PAAN(Bui等,2022)	SE-ResNet50	256×256	84.51	86.78	91.01	82.28	43.8	1.07×
Swin-B+DWDR(Wang等,2022a)	Swin Transformer	256×256	86.41	88.41	91.30	86.02	/	/
MCCG(Shen等,2024)	ConvNeXt	256×256	89.28	91.01	94.29	89.29	/	/
MCCG(Shen等,2024)	ConvNeXt	384×384	<u>89.64</u>	<u>91.32</u>	<u>94.30</u>	<u>89.39</u>	/	/
Sample4Geo(Deuser等,2023)	ConvNeXt-B	384×384	**92.65**	**93.81**	**95.14**	**91.39**	88	/
PCL(Tian等,2022)	ResNet-50	512×512	83.27	87.32	91.78	82.18	/	/

注:加粗、下划线字体分别表示各列最优、次优结果,"/"表示未提供该信息,"*"表示方法使用极坐标变换作为数据预处理中特征对齐方式,推理速度的"×"表示比例关系,基准方法为1.00×,其他为相对基准方法的倍数。

表8　跨视角地理定位方法在CVUSA数据集上的结果比较

Table 8　The results of cross-view geo-localization methods on the CVUSA dataset

方法	骨干网类型	输入尺寸/像素	R@1/%	R@5/%	R@10/%	R@Top1/%
Toker*(Toker等,2021)	AlexNet	/	92.56	97.55	98.33	99.67
SAFA+LPN(Wang等,2022d)	VGG-16	256×256	92.83	98.00	98.85	99.78
Polar-L2LTR*(Yang等,2021)	ResNet	256×256	94.05	98.27	98.99	99.67
SIRNet*(Lu等,2022)	VGG-16	/	93.74	98.02	98.85	99.76
TransGeo(Zhu等,2022)	Transformer	256×256	94.08	98.36	99.04	99.77
GeoDTR(Zhang等,2023d)	ResNet-34	256×256	93.76	98.47	99.22	99.85
GeoDTR*(Zhang等,2023d)	ResNet-34	256×256	95.43	98.86	99.34	<u>99.86</u>
LPN*+DWDR(Wang等,2022a)	ResNet-50	256×256	94.33	98.54	99.09	99.80
SAIG-D*(Zhu等,2023b)	ResNet-50	256×256	<u>96.34</u>	<u>99.10</u>	<u>99.50</u>	<u>99.86</u>
MGTL*(Zhao等,2022)	VGG-16	384×384	94.11	98.30	99.03	99.74
Sample4Geo(Deuser等,2023)	ConvNeXt-B	384×384	**98.68**	**99.68**	**99.78**	**99.87**

注:加粗、下划线字体分别表示各列最优、次优结果,"/"表示未提供该信息,"*"表示方法使用极坐标变换作为数据预处理中特征对齐方式,推理速度的"×"表示比例关系,基准方法为1.00×,其他为相对基准方法的倍数。

作为骨干网的方法,而使用ConvNeXt网络的方法是所有方法中性能更优的。从图像输入尺寸可以看出,大多数跨视角地理定位方法输入尺寸为256×256像素,还包括384×384像素和512×512像素两种输入尺寸。较大尺寸的输入可以带来较好的结果。从参数量可以看出,Baseline(Zheng等,2020)方法的参数量最少;而使用ConvNeXt网络虽然带来了更好的性能,但由于该网络是结合了ResNet和Inception模型的优势重新设计的新型网络,因此参数量相对较大。Sample4Geo方法比先前参数量最大的FSRA方法(51 M)还多了接近50%。从推理速度可以看出,Baseline(Zheng等,2020)方法的推理速

度最快，而 Wang 等人（2022d）提出的 LPN 方法推理速度慢。

3.2 CVUSA 数据集结果分析

数据集 CVUSA 主要执行地面—卫星跨视角地理定位任务，表 8 给出了 4 个指标（R@1、R@5、R@10 和 R@Top1）的评测结果。从表 8 可以看出，Sample4Geo（Deuser 等，2023）方法在 4 个指标中均取得最佳性能，SAIG-D*（Zhu 等，2023b）方法取得次优结果。值得注意的是，Sample4Geo 方法使用的图像输入尺寸是 384 × 384 像素，而 SAIG-D*方法使用的图像输入尺寸是 256 × 256 像素。同时，两类方法使用的骨干网类型也不同。

除此之外，在数据集 CVUSA 上大多数方法均使用极坐标变换作为数据预处理步骤，用“*”标出。例如，GeoDTR 方法在骨干网类型、图像输入尺寸相同条件下，使用极坐标变换后，召回精度 R@1 提升了 1.67%，召回精度 R@5 提升了 0.19%，召回精度 R@10 提升了 0.12%，召回精度 R@Top1 提升了 0.01%。由此可见，使用极坐标变换进行数据预处理有利于提升跨视角地理定位精度。

4 挑战与未来展望

当前，深度学习带动了跨视角地理定位技术的发展，但该领域仍存在一些挑战。本节将进一步讨论本领域的未来发展方向，相信这将有助于进一步推进跨视角地理定位技术的发展。

1）受人类视觉系统启发的跨视角地理定位方法。人类通过人类视觉系统（human visual system，HVS）来获取外界信息（Ge 等，2022；Dong 等，2023），由于人眼在观察图像时，对于原始图像失真部分的敏感程度不同，因此，对于不敏感的信息允许有较大的失真；而对于敏感的信息，则需要减少失真。目前 HVS 中包含对色彩亮度的异常感知、对物体运动的研究以及从粗到细判别过程等机制（Van de Ven 等，2020）。例如从粗到细的相似物体判别过程，启发 Lin 等人（2022）在表征学习中考虑全局和局部信息，所设计的模型融合了全局特征（粗）和局部特征（细）。因此通过对该领域的研究，可以设计更符合人类视觉系统与深度学习结合的判别方法，从而能够高效准确地解决跨视角地理定位工作。

2）基于视觉大模型的跨视角地理定位。当前的数据增强方法虽然可以智能地获取数据增强策略（Cubuk 等，2020；Ni 等，2021；Shu 等，2023），但如何为给定任务自动定制一组最佳数据增强方案仍有待研究。随着虚拟现实和大语言模型技术的快速发展（Yue 等，2023；Zhang 等，2023c），使用大模型技术生成的图像作为新的数据来源已经被提出（Hao 等，2023；Yang 等，2023）。Mandlekar 等人（2023）在模拟仿真环境中，使用数字孪生技术复制真实世界中人类的操作数据，生成了真实世界中没有的操作。因此，通过该技术模拟多视角的真实场景图像，可为跨视角地理定位任务提供具有不受现实采样约束的数据增强方案。除此之外，鉴于大模型在自然语言处理、计算机视觉和语音识别等领域的成功（Pal 等，2023），可将其引入跨视角地理定位任务中。大模型具有较好的泛化能力，可以应对实际场景中的鲁棒性以及可迁移性问题（Zhang 等，2023e）。在实际部署中，由于跨视角地理定位需要应对各种变化的自然场景，而大模型的模型泛化能力可以较好地解决变换场景所带来的鲁棒性问题。同时，现有的跨视角地理定位方法是有监督的方法，而现实场景中的跨视角定位任务往往是半监督或无监督条件。这导致现有的方法在理想的实验条件下效果较好，在实际场景中效果略差。大模型所具备的迁移学习能力可以有效地帮助本领域由有监督学习向域适应、半监督和无监督的方向发展（Yin 等，2023）。

3）轻量化模型的跨视角地理定位方法。随着无人机技术的广泛运用，在无人机上部署跨视角地理定位技术成为可能。但该任务要求模型小且推理性能高（Deng 等，2020；Qin 等，2020；赵洁 等，2023），因此利用轻量化方法获得高精度轻量级模型逐渐成为发展趋势（Luo 等，2023）。当前模型轻量化方法主要采用知识蒸馏（Chen 等，2022b；司兆峰和齐洪钢，2023）、模型量化（Cai 等，2020；Liu 等，2021b）、重参数化（Hu 等，2022b；Huang 等，2022）等技术。KeepEdge 方法已经将模型轻量化技术运用于跨视角地理定位，该方法首次使用知识蒸馏技术在无人机上部署了一个高精度模型，实现了边缘化计算（Luo 等，2023）。由于该领域目前研究处于起步阶段，如何综合利用模型轻量化方法在不损失精度的情况下，提升模型在边缘设备上的部署能力仍值得研究。

4）可持续学习的跨视角地理定位方法。跨视角

地理定位任务中最大的挑战是不同视角图像的视觉差异(Zhu等,2023b)。在无人机跨视角地理定位场景中,图像采集装置获得的图像往往比数据集更恶劣,并且会受自然因素的干扰。这就导致跨视角地理定位在实际运用中需要克服鲁棒性、泛化性、可迁移性以及针对各种in the wild变化场景的问题。持续学习是指在训练过程中连续不断地学习一系列任务(Ke等,2020),并利用从旧任务中积累的知识来更好地学习新任务(Wang等,2022c,2023)。因此,将持续学习的理念引入跨视角地理定位任务中,使其在新场景中具有较高的鲁棒性、泛化性和可迁移性,将成为未来发展趋势。

5 结 语

作为一项重要的视觉定位技术,跨视角地理定位在图像检索、遥感监测、视频监控、导航定位和军事侦察中发挥着至关重要的作用。本文主要阐述了跨视角地理定位的研究背景和意义,首次按照深度学习的跨视角地理定位框架对国内外研究现状进行了整理和总结。首先,简述了跨视角地理定位的任务来源,总结了早期任务中手工特征提取的解决方案。其次,通过阐述该任务的问题定义,介绍了现有数据集和评价指标。随后,根据该任务的框架归纳出现有方法在数据预处理、深度学习网络、特征注意力模块和损失函数方面的发展。最后,结合当下人工智能发展的趋势,讨论了该领域的挑战与未来发展趋势。

经过综合的分析归纳,发现跨视角地理定位任务作为图像检索问题的延伸,各项研究仍处于不断发展的阶段且具有较大科研前景。本文希望通过对深度学习的跨视角地理定位方法综述,让更多学者关注这一领域内容,进而推动该领域的发展与进步。

参考文献(References)

Alcantarilla P F, Bartoli A and Davison A J. 2012. KAZE features//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer: 214-227 [DOI: 10.1007/978-3-642-33783-3_16]

Algamdi A M, Sanchez V and Li C T. 2020. Dronecaps: recognition of human actions in drone videos using capsule networks with binary volume comparisons//Proceedings of 2020 IEEE International Conference on Image Processing (ICIP). Abu Dhabi, United Arab Emirates: IEEE: 3174-3178 [DOI: 10.1109/ICIP40778.2020.9190864]

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]

Arar M, Shamir A and Bermano A H. 2022. Learned queries for efficient local attention//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10841-10852 [DOI: 10.1109/CVPR52688.2022.01057]

Bay H, Ess A, Tuytelaars T and Van Gool L. 2008. Speeded-up robust features (SURF). Computer Vision and Image Understanding, 110(3): 346-359 [DOI: 10.1016/j.cviu.2007.09.014]

Bui D V, Kubo M and Sato H. 2022. A part-aware attention neural network for cross-view geo-localization between UAV and Satellite. Journal of Robotics, Networking and Artificial Life, 9(3): 275-284 [DOI: 10.57417/jrnal.9.3_275]

Cai S D, Guo Y L, Khan S, Hu J W and Wen G J. 2019. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8391-8400 [DOI: 10.1109/ICCV.2019.00848]

Cai Y H, Yao Z W, Dong Z, Gholami A, Mahoney M W and Keutzer K. 2020. Zeroq: a novel zero shot quantization framework//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13169-13178 [DOI: 10.1109/CVPR42600.2020. 01318]

Chen J N, Sun S Y, He J, Torr P, Yuille A and Bai S. 2022a. Transmix: attend to mix for vision Transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12135-12144 [DOI: 10.1109/CVPR52688. 2022.01182]

Chen W H, Chen X T, Zhang J G and Huang K Q. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 403-412 [DOI: 10.1109/CVPR.2017.145]

Chen X C, Li Y, Yao L N, Adeli E and Zhang Y. 2021. Generative adversarial u-net for domain-free medical image augmentation [EB/OL]. [2023-12-12]. https://arxiv.org/pdf/2101.04793.pdf

Chen Y D, Wang S, Liu J J, Xu X W, de Hoog F and Huang Z. 2022b. Improved feature distillation via projector ensemble//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 12084-12095

Choi J and Myung H. 2020. BRM localization: UAV localization in GNSS-denied environments based on matching of numerical map

and UAV images//Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, USA: IEEE: 4537-4544 [DOI: 10.1109/IROS45743.2020.9341682]

Choi J, Sharma G, Chandraker M and Huang J B. 2020. Unsupervised and semi-supervised domain adaptation for action recognition from drones//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass, USA: IEEE: 1717-1726 [DOI: 10.1109/WACV45572.2020.9093511]

Chopra S, Hadsell R and LeCun Y. 2005. Learning a similarity metric discriminatively, with application to face verification//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE: 539-546 [DOI: 10.1109/CVPR.2005.202]

Cubuk E D, Zoph B, Mane D, Vasudevan V and Le Q V. 2019. AutoAugment: learning augmentation strategies from data//Proceedings of 2019 IEEE/CVF Computer Society Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 113-123 [DOI: 10.1109/CVPR.2019.00020]

Cubuk E D, Zoph B, Shlens J and Le Q V. 2020. Randaugment: practical automated data augmentation with a reduced search space//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE: 702-703 [DOI: 10.1109/CVPRW50498.2020.00359]

Dai M, Hu J H, Zhuang J D and Zheng E H. 2022. A Transformer-based feature segmentation and region alignment method for UAV-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology, 32(7): 4376-4389 [DOI: 10.1109/TCSVT.2021.3135013]

Deng B L, Li G Q, Han S, Shi L P and Xie Y. 2020. Model compression and hardware acceleration for neural networks: a comprehensive survey. Proceedings of the IEEE, 108(4): 485-532 [DOI: 10.1109/JPROC.2020.2976475]

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Deuser F, Habel K and Oswald N. 2023. Sample4geo: hard negative sampling for cross-view geo-localization [EB/OL]. [2023-12-12]. https://arxiv.org/pdf/2303.11851.pdf

Ding L R, Zhou J, Meng L X and Long Z Y. 2020. A practical cross-view image matching method between UAV and satellite for UAV-based geo-localization. Remote Sensing, 13(1): #47 [DOI: 10.3390/rs13010047]

Ding Y, Yu J, Liu B, Hu Y, Cui M X and Wu Q. 2022. MuKEA: multimodal knowledge extraction and accumulation for knowledge-based visual question answering//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5089-5098 [DOI: 10.1109/CVPR52688.2022.00503]

Dong X Y, Bao J M, Chen D D, Zhang W M, Yu N H, Yuan L, Chen D and Guo B N. 2022. CSWin Transformer: a general vision Transformer backbone with cross-shaped windows//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12124-12134 [DOI: 10.1109/CVPR52688.2022.01181]

Dong Z, Lin B J and Xie F. 2023. Optimizing remote sensing image scene classification through brain-inspired feature bias estimation and semantic representation analysis. IEEE Access, 11: 34764-34771 [DOI: 10.1109/ACCESS.2023.3264502]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2010. An image is worth 16x16 words: Transformers for image recognition at scale [EB/OL]. [2023-12-12]. https://arxiv.org/pdf/2010.11929.pdf

Ge Y H, Xiao Y, Xu Z, Wang X R and Itti L. 2022. Contributions of shape, texture, and color in visual recognition//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 369-386 [DOI: 10.1007/978-3-031-19775-8_22]

Gong C Y, Ren T Z, Ye M and Liu Q. 2021. MaxUP: lightweight adversarial training with data augmentation improves neural network training//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2474-2483 [DOI: 10.1109/CVPR46437.2021.00250]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680

Hadsell R, Chopra S and LeCun Y. 2006. Dimensionality reduction by learning an invariant mapping//Proceedings of 2006 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE: 1735-1742 [DOI: 10.1109/CVPR.2006.100]

Hao X J, Liu L, Yang R J, Yin L Z Y, Zhang L and Li X H. 2023. A review of data augmentation methods of remote sensing image target recognition. Remote Sensing, 15(3): #827 [DOI: 10.3390/RS15030827]

Harris C and Stephens M. 1988. A combined corner and edge detector//Taylor C J, ed. Proceedings of the Alvey Vision Conference. Manchester, UK: Alvety Vision Club. 1-6 [DOI: 10.5244/C.2.23]

Hermans A, Beyer L and Leibe B. 2017. In defense of the triplet loss for person re-identification [EB/OL]. [2023-12-12]. https://arxiv.org/pdf/1703.07737.pdf

Hu M, Feng J Y, Hua J S, Lai B S, Huang J Q, Gong X J and Hua X S. 2022b. Online convolutional re-parameterization//Proceedings of 2022 IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 568-577 [DOI: 10.1109/

CVPR52688. 2022.00065]

Hu S X, Feng M D, Nguyen R M H and Lee G H. 2018. CVM-Net: cross-view matching network for image-based ground-to-aerial geo-localization//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7258-7267 [DOI: 10.1109/CVPR.2018.00758]

Hu S X and Lee G H. 2020. Image-based geo-localization using satellite imagery. International Journal of Computer Vision, 128(5): 1205-1219 [DOI: 10.1007/s11263-019-01186-0]

Hu W M, Zhang Y C, Liang Y X, Yin Y F, Georgescu A, Tran A, Kruppa H, Ng S K and Zimmermann R. 2022a. Beyond geo-localization: fine-grained orientation of street-view images by cross-view matching with satellite imagery//Proceedings of the 30th ACM International Conference on Multimedia. New York, USA: Association for Computing Machinery: 6155-6164 [DOI: 10.1145/3503161.3548102]

Huang G S, Zhou Y, Hu X F, Zhao L Y and Zhang C L. 2023. A survey of the research progress in image geo-localization. Journal of Geo-Information Science, 25(7): 1336-1362 (黄高爽, 周杨, 胡校飞, 赵璐颖, 张呈龙. 2023. 图像地理定位研究进展. 地球信息科学学报, 25(7): 1336-1362) [DOI: 10.12082/dqxxkx.2023.230073]

Huang T, You S, Zhang B H, Du Y X, Wang F, Qian C and Xu C. 2022. DyRep: bootstrapping training with dynamic re-parameterization//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 588-597 [DOI: 10.1109/CVPR52688.2022.00067]

Kan M N, Shan S G and Chen X L. 2016. Multi-view deep network for cross-view classification//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4847-4855 [DOI: 10.1109/CVPR.2016.524]

Ke Z X, Liu B and Huang X C. 2020. Continual learning of a mixed sequence of similar and dissimilar tasks//Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 18493-18504

Kim D K and Walter M R. 2017. Satellite image-based localization via learned embeddings//Proceedings of 2017 IEEE International Conference on Robotics and Automation. Singapore: IEEE: 2073-2080 [DOI: 10.1109/ICRA.2017.7989239]

Kinnari J, Verdoja F and Kyrki V. 2021. GNSS-denied geolocalization of UAVs by visual matching of onboard camera images with orthophotos//Proceedings of the 20th International Conference on Advanced Robotics (ICAR). Ljubljana, Slovenia: IEEE: 555-562 [DOI: 10.1109/ICAR53236.2021.9659333]

Lin J L, Zheng Z D, Zhong Z, Luo Z M, Li S Z, Yang Y and Sebe N. 2022. Joint representation learning and keypoint detection for cross-view geo-localization. IEEE Transactions on Image Processing, 31: 3780-3792 [DOI: 10.1109/TIP.2022.3175601]

Lin T Y, Cui Y, Belongie S and Hays J. 2015. Learning deep representations for ground-to-aerial geolocalization//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 5007-5015 [DOI: 10.1109/CVPR.2015.7299135]

Liu L and Li H D. 2019. Lending orientation to neural networks for cross-view geo-localization//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5624-5633 [DOI: 10.1109/CVPR.2019.00577]

Liu Y A, Zhang W and Wang J. 2021b. Zero-shot adversarial quantization//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 1512-1521 [DOI: 10.1109/CVPR46437.2021.00156]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021a. Swin Transformer: hierarchical vision Transformer using shifted windows//Proceedings of 2021 IEEE International Conference on Computer Vision. Montreal, Canada: IEEE: 10012-10022 [DOI: 10.1109/ICCV48922.2021.00986]

Liu Z, Mao H Z, Wu C Y, Feichtenhofer C, Darrell T and Xie S N. 2022. A ConvNet for the 2020s//Proceedings of 2022 IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11976-11986 [DOI: 10.1109/CVPR52688.2022.01167]

Lowe D G. 1999. Object recognition from local scale-invariant features//Proceedings of 1999 IEEE International Conference on Computer Vision. Kerkyra, Greece: IEEE: 1150-1157 [DOI: 10.1109/ICCV.1999.790410]

Lu X F, Luo S Q and Zhu Y Y. 2022. It's okay to be wrong: cross-view geo-localization with step-adaptive iterative refinement. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-13 [DOI: 10.1109/TGRS.2022.3210195]

Luo H Y, Chen T X, Li X J, Li S Y, Zhang C, Zhao G S and Liu X. 2023. KeepEdge: a knowledge distillation empowered edge intelligence framework for visual assisted positioning in UAV delivery. IEEE Transactions on Mobile Computing, 22(8): 4729-4741 [DOI: 10.1109/TMC.2022.3157957]

Mandlekar A, Nasiriany S, Wen B W, Akinola I, Narang Y, Fan L X, Zhu Y K and Fox D. 2023. MimicGen: a data generation system for scalable robot learning using human demonstrations [EB/OL]. [2023-12-12]. https://arxiv.org/pdf/2310.17596.pdf

Matas J, Chum O, Urban M and Pajdla T. 2004. Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing, 22(10): 761-767 [DOI: 10.1016/j. imavis. 2004.02.006]

Misra D, Nalamada T, Arasanipalai A U and Hou Q B. 2021. Rotate to attend: convolutional triplet attention module//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, US: IEEE: 3139-3148 [DOI: 10.1109/WACV48630.2021.00318]

Naranjo-Alcazar J, Perez-Castanos S, Lopez-Garcia A, Zuccarello P, Cobos M and Ferri F J. 2021. Squeeze-excitation convolutional

recurrent neural networks for audio-visual scene classification［EB/OL］.［2023-12-12］. https://arxiv.org/pdf/2107.13180.pdf

Ni R K, Goldblum M, Sharaf A, Kong K Z and Goldstein T. 2021. Data augmentation for meta-learning//Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: PMLR: 8152-8161

Pal A, Karkhanis D, Roberts M, Dooley S, Sundararajan A and Naidu S. 2023. Giraffe: adventures in expanding context lengths in LLMs［EB/OL］.［2023-12-12］. https://arxiv.org/pdf/2308.10882.pdf

Philbin J, Chum O, Isard M, Sivic J and Zisserman A. 2007. Object retrieval with large vocabularies and fast spatial matching//Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE: 1-8［DOI: 10.1109/CVPR.2007.383172］

Qin H T, Gong R H, Liu X L, Shen M Z, Wei Z R, Yu F W and Song J K. 2020. Forward and backward information retention for accurate binary neural networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2250-2259［DOI: 10.1109/CVPR42600.2020.00232］

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Kruege G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: PMLR: 8748-8763

Rave A, Fontaine P and Kuhn H. 2023. Drone location and vehicle fleet planning with trucks and aerial drones. European Journal of Operational Research, 308(1): 113-130［DOI: 10.1016/j.ejor.2022.10.015］

Regmi K and Shah M. 2019. Bridging the domain gap for ground-to-aerial image matching//Proceedings of 2019 IEEE International Conference on Computer Vision. Seoul, Korea (South): IEEE: 470-479［DOI: 10.1109/ICCV.2019.00056］

Rosten E and Drummond T. 2006. Machine learning for high-speed corner detection//Proceedings of the 9th European Conference on Computer Vision. Graz, Austria: Springer: 430-443［DOI: 10.1007/11744023_34］

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］

Shen T R, Wei Y M, Kang L, Wan S S and Yang Y H. 2024. MCCG: a ConvNeXt-based multiple-classifier method for cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology, 34(3): 1456-1468［DOI: 10.1109/TCSVT.2023.3296074］

Shi Y J, Liu L, Yu X and Li H D. 2019. Spatial-aware feature aggregation for cross-view image based geo-localization//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.

Shi Y J, Yu X, Campbell D and Li H D. 2020a. Where am I looking at? Joint location and orientation estimation by cross-view matching//Proceedings of 2020 IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4064-4072［DOI: 10.1109/CVPR42600.2020.00412］

Shi Y J, Yu X, Liu L, Zhang T and Li H D. 2020b. Optimal feature transport for cross-view image geo-localization//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 11990-11997［DOI: 10.1609/aaai.v34i07.6875］

Shu J, Yuan X, Meng D Y and Xu Z B. 2023. DAC-MR: data augmentation consistency based meta-regularization for meta-learning［EB/OL］.［2023-12-12］. https://arxiv.org/pdf/2305.07892.pdf

Si Z F and Qi H G. 2023. Survey on knowledge distillation and its application. Journal of Image and Graphics, 28(9): 2817-2832 (司兆峰, 齐洪钢. 2023. 知识蒸馏方法研究与应用综述. 中国图象图形学报, 28(9): 2817-2832)［DOI: 10.11834/jig.220273］

Song H S, Wang Z, Lei Y, Shi D X, Tong X C, Lei Y X and Qiu C P. 2023. Learning visual representation clusters for cross-view geo-Location. IEEE Geoscience and Remote Sensing Letters, 20: #6011805［DOI: 10.1109/LGRS.2023.3326005］

Tian X Y, Shao J, Ouyang D Q and Shen H T. 2022. UAV-satellite view synthesis for cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology, 32(7): 4804-4815［DOI: 10.1109/TCSVT.2021.3121987］

Toker A, Zhou Q J, Maximov M and Leal-Taixé L. 2021. Coming down to earth: satellite-to-street view synthesis for geo-localization//Proceedings of 2021 IEEE Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 6488-6497［DOI: 10.1109/CVPR46437.2021.00642］

Touvron H, Cord M, Douze M, Massa F, Sablayrolles A and Jégou H. 2021. Training data-efficient image Transformers & distillation through attention//Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: PMLR: 10347-10357

Van de Ven G M, Siegelmann H T and Tolias A S. 2020. Brain-inspired replay for continual learning with artificial neural networks. Nature Communications, 11(1): #4069［DOI: 10.1038/s41467-020-17866-2］

Van Den Oord A, Li Y Z and Vinyals O. 2018. Representation learning with contrastive predictive coding［EB/OL］.［2023-12-12］. https://arxiv.org/pdf/1807.03748.pdf

Varior R R, Haloi M and Wang G. 2016. Gated Siamese convolutional neural network architecture for human re-identification//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 791-808［DOI: 10.1007/978-3-319-46484-8_48］

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5998-

6008

Vo N N and Hays J. 2016. Localizing and orienting street views using overhead imagery//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 494-509 [DOI: 10.1007/978-3-319-46448-0_30]

Wang H H, Huang Z Y, Wu X D and Xing E. 2022b. Toward learning robust and invariant representations with alignment regularization and data augmentation//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, USA: Association for Computing Machinery: 1846-1856 [DOI: 10.1145/3534678.3539438]

Wang J, Zhou F, Wen S L, Liu X and Lin Y Q. 2017. Deep metric learning with angular loss//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2593-2601 [DOI: 10.1109/ICCV.2017.283]

Wang L Y, Zhang X X, Su H and Zhu J. 2023. A comprehensive survey of continual learning: theory, method and application [EB/OL]. [2023-11-28]. https://arxiv.org/pdf/2302.00487.pdf

Wang T Y, Zheng Z D, Yan C G, Zhang J Y, Sun Y Q, Zheng B L and Yang Y. 2022d. Each part matters: local patterns facilitate cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology, 32(2): 867-879 [DOI: 10.1109/TCSVT.2021.3061265]

Wang T Y, Zheng Z D, Zhu Z J, Gao Y H, Yang Y and Yan C G. 2022a. Learning cross-view geo-localization embeddings via dynamic weighted decorrelation regularization [EB/OL]. [2023-12-12]. https://arxiv.org/pdf/2211.05296.pdf

Wang Z F, Zhang Z Z, Lee C Y, Zhang H, Sun R X, Ren X Q, Su G L, Perot V, Dy J and Pfister T. 2022c. Learning to prompt for continual learning//Proceedings of 2022 IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 139-149 [DOI: 10.1109/CVPR52688.2022.00024]

Wei H and Wang L P. 2018. Visual navigation using projection of spatial right-angle in indoor environment. IEEE Transactions on Image Processing, 27(7): 3164-3177 [DOI: 10.1109/TIP.2018.2818931]

Wood D, Mu T T and Brown G. 2022. Bias-variance decompositions for margin losses//Proceedings of the 25th International Conference on Artificial Intelligence and Statistics. Valencia, Spain: PMLR: 1975-2001 [DOI: 10.48550/arXiv.2204.12155]

Workman S, Souvenir R and Jacobs N. 2015. Wide-area image geolocalization with aerial reference imagery//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 3961-3969 [DOI: 10.1109/ICCV.2015.451]

Workman S and Jacobs N. 2015. On the location dependence of convolutional neural network features//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Boston, USA: IEEE: 70-78 [DOI: 10.1109/CVPRW.2015.7301385]

Xu F R, Zhang W, Cheng Y and Chu W. 2020. Metric learning with equidistant and equidistributed triplet-based loss for product image search//Proceedings of 2020 Web Conference. New York, USA: Association for Computing Machinery: 57-65 [DOI: 10.1145/3366423.3380094]

Yang H J, Lu X F and Zhu Y Y. 2021. Cross-view geo-localization with evolving Transformer//Proceedings of the 35th Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc.: 29009-29020 [DOI: 10.48550/arXiv.2107.00842]

Yang Z H, Zhan F N, Liu K H, Xu M Y and Lu S J. 2023. AI-generated images as data source: the dawn of synthetic era [EB/OL]. [2023-11-28]. https://arxiv.org/pdf/2310.01830.pdf

Yin X J, Huang B Z and Wan X J. 2023. Alcuna: large language models meet new knowledge [EB/OL]. [2023-12-12]. https://arxiv.org/pdf/2310.14820.pdf

Yue M R, Zhao J, Zhang M, Du L and Yao Z Y. 2023. Large language model cascades with mixture of thoughts representations for cost-efficient reasoning [EB/OL]. [2023-12-12]. https://arxiv.org/pdf/2310.03094.pdf

Zhang J Y, Krishna R, Awadallah A H and Wang C. 2023c. EcoAssistant: using LLM assistant more affordably and accurately [EB/OL]. [2023-11-28]. https://arxiv.org/pdf/2310.03046.pdf

Zhang S, Wang Q Z, Bian J and Xiong H Y. 2023b. TiC: exploring vision Transformer in convolution [EB/OL]. [2023-12-12]. https://arxiv.org/pdf/2310.04134.pdf

Zhang T Y, Wang Z, Huang J, Tasnim M M and Shi W. 2023e. A survey of diffusion based image generation models: issues and their solutions [EB/OL]. [2023-12-12]. https://arxiv.org/pdf/2308.13142.pdf

Zhang X H, Li X Y, Sultani W, Zhou Y and Wshah S. 2023d. Cross-view geo-localization via learning disentangled geometric layout correspondence. Proceedings of 2023 AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 3480-3488 [DOI: 10.1609/aaai.v37i3.25457]

Zhang X H, Sultani W and Wshah S. 2023a. Cross-view image sequence geo-localization//Proceedings of 2023 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2914-2923 [DOI: 10.1109/WACV56688.2023.00293]

Zhao J, Yuan Y S, Zhang P Y and Wang D. 2023. An efficient Transformer-based object-capturing video annotation method. Journal of Image and Graphics, 28(10): 3176-3190 (赵洁, 袁永胜, 张鹏宇, 王栋. 2023. 轻量化Transformer目标跟踪数据标注算法. 中国图象图形学报, 28(10): 3176-3190) [DOI: 10.11834/jig.220823]

Zhao J W, Zhai Q, Zhao P B, Huang R and Cheng H. 2022. Co-visual pattern augmented generative Transformer learning for automobile geo-localization [EB/OL]. [2023-12-12]. https://arxiv.org/pdf/2203.09135.pdf

Zheng Z D, Wei Y C and Yang Y. 2020. University-1652: a multi-view multi-source benchmark for drone-based geo-localization//Proceedings of the 28th ACM international conference on Multimedia. New

York，USA：Association for Computing Machinery：1395-1403［DOI：10.1145/3394171.3413896］

Zheng Z D，Zheng L and Yang Y. 2017. A discriminatively learned CNN embedding for person reidentification. ACM Transactions on Multimedia Computing，Communications，and Applications（TOMM），14(1)：#13［DOI：10.1145/3159171］

Zhou B L，Lapedriza A，Xiao J X，Torralba A and Oliva A. 2014. Learning deep features for scene recognition using places database//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge，USA：MIT Press：487-495

Zhou J Z，Yan Y M，Gu G H and Su N. 2023. Multi-scale geo-localization based on local similarity area distance measurement method//IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. Pasadena，USA：IEEE：5069-5072［DOI：10.1109/IGARSS52108.2023.10282536］

Zhou L P，Huang G Q，Mao Y N，Wang S Z and Kaess M. 2022. EDPLVO：efficient direct point-line visual odometry//Proceedings of 2022 IEEE International Conference on Robotics and Automation. Philadelphia，USA：IEEE：7559-7565［DOI：10.1109/ICRA46639.2022.9812133］

Zhu P F，Zheng J Y，Du D W，Wen L Y，Sun Y M and Hu Q H. 2021. Multi-drone-based single object tracking with agent sharing network. IEEE Transactions on Circuits and Systems for Video Technology，31(10)：4058-4070［DOI：10.1109/TCSVT.2020.3045747］

Zhu R Z，Yang M Z，Yin L，Wu F and Yang Y C. 2023a. UAV's status is worth considering：a fusion representations matching method for geo-localization. Sensors，23(2)：#720［DOI：10.3390/s23020720］

Zhu R Z，Yin L，Yang M Z，Wu F，Yang Y C and Hu W B. 2023c. SUES-200：a multi-height multi-scene cross-view image benchmark across drone and satellite. IEEE Transactions on Circuits and Systems for Video Technology，33(9)：4825-4839［DOI：10.1109/TCSVT.2023.3249204］

Zhu S J，Shah M and Chen C. 2022. TransGeo：Transformer is all you need for cross-view image geo-localization//Proceedings of 2022 IEEE Conference on Computer Vision and Pattern Recognition. New Orleans，USA：IEEE：1162-1171［DOI：10.1109/CVPR52688.2022.00123］

Zhu S J，Yang T J N and Chen C. 2021. Revisiting street-to-aerial view image geo-localization and orientation estimation//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa，US：IEEE：756-765［DOI：10.1109/WACV48630.2021.00080］

Zhu Y Y，Yang H J，Lu Y X and Huang Q. 2023b. Simple，effective and general：a new backbone for cross-view image geo-localization［EB/OL］.［2023-11-28］. https://arxiv.org/pdf/2302.01572.pdf

Zhuang J D，Chen X R Y，Dai M，Lan W B，Cai Y H and Zheng E H. 2022. A semantic guidance and Transformer-based matching method for UAVs and satellite images for UAV geo-localization. IEEE Access，10：34277-34287［DOI：10.1109/ACCESS.2022.3162693］

Zhuang J D，Dai M，Chen X R Y and Zheng E H. 2021. A faster and more effective cross-view matching method of UAV and satellite images for UAV geolocalization. Remote Sensing，13(19)：#3979［DOI：10.3390/rs13193979］

作者简介

周博文，男，硕士研究生，主要研究方向为深度学习的跨视角地理定位。E-mail：zhoubw510@163.com

李阳，通信作者，男，副教授，主要研究方向为机器视觉与图像检索。E-mail：solarleeon@outlook.com

马鑫骥，男，硕士研究生，主要研究方向为机器学习。E-mail：3294745961@qq.com

苗壮，男，教授，主要研究方向为图像视频处理。E-mail：emiao_beyond@163.com

张睿，男，教授，主要研究方向为数据工程和信息融合。E-mail：3959966@qq.com