计算机科学
COMPUTER SCIENCE

跨视角地理定位中的三维交互机制

周博文, 李阳, 王家宝, 苗壮, 张睿

引用本文

周博文, 李阳, 王家宝, 苗壮, 张睿. 跨视角地理定位中的三维交互机制[J]. 计算机科学, 2025, 52(3): 86-94.

ZHOU Bowen, LI Yang, WANG Jiabao, MIAO Zhuang, ZHANG Rui. Triplet Interaction Mechanism in Cross-view Geo-localization [J]. Computer Science, 2025, 52(3): 86-94.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

一种引入核心实体关注度评估的KBQA算法

KBQA Algorithm Introducing Core Entity Attention Evaluation

计算机科学, 2024, 51(11): 239-247. https://doi.org/10.11896/jsjkx.231000015

基于弱监督语义分割的道路裂缝检测研究

Study on Road Crack Detection Based on Weakly Supervised Semantic Segmentation

计算机科学, 2024, 51(11): 148-156. https://doi.org/10.11896/jsjkx.231000148

重参数化增强的双模态实时目标检测模型

Re-parameterization Enhanced Dual-modal Realtime Object Detection Model

计算机科学, 2024, 51(9): 162-172. https://doi.org/10.11896/jsjkx.230700106

任务感知的多尺度小样本SAR图像分类方法

Task-aware Few-shot SAR Image Classification Method Based on Multi-scale Attention Mechanism

计算机科学, 2024, 51(8): 160-167. https://doi.org/10.11896/jsjkx.230500171

基于特征手性的数据无关模型评估方法

Data-free Model Evaluation Method Based on Feature Chirality

计算机科学, 2024, 51(7): 337-344. https://doi.org/10.11896/jsjkx.230500179

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240500020

跨视角地理定位中的三维交互机制

周博文　李　阳　王家宝　苗　壮　张　睿
陆军工程大学指挥控制工程学院　南京 210007
(solarleeon@outlook.com)

摘　要　跨视角地理定位是一种图像检索任务,其目的是在不同视角下使用无地理坐标的图像与数据库中有地理坐标的图像进行检索匹配,从而获取目标图像的地理位置信息。然而,现有方法大多忽略了全局位置信息和特征完整性,导致模型无法捕获深层语义信息;另外,现有的二维交互方式未充分利用维度间关系,导致跨维交互不充分。为解决上述问题,设计了一种跨视角地理定位三维交互机制。该方法利用 ConvNeXt 作为特征提取网络,随后使用所提出的三维交互机制(Triplet Interaction Mechanism,TIM)进行特征丰富操作,最后利用联合损失函数指导模型训练。所提方法在模型内进行了多次三维交互,缓解了二维特征投影部分信息丢失的问题。同时,所提出的三维交互机制在 3 个通道中使用不同的注意力,使模型对跨视角图像的平移、缩放、旋转具有鲁棒性。实验结果表明,所提方法在 University-1652 数据集上针对无人机视角定位和无人机导航两个任务均取得了最优性能。

关键词:跨视角;地理定位;交互机制;特征注意力
中图分类号　TP391

Triplet Interaction Mechanism in Cross-view Geo-localization

ZHOU Bowen,LI Yang,WANG Jiabao,MIAO Zhuang and ZHANG Rui
College of Command and Control Engineering ,Army Engineering University of PLA,Nanjing 210007,China

Abstract　Cross-view geo-localization refers to inferring the geographical location from images of different viewpoints,which is usually viewed as an image retrieval task. However,most existing methods neglect the global position information and feature completeness,which makes the model can not conducive to capturing deep semantic information. Additionally,the current two-dimensional interaction methods do not fully utilize the relationships between dimensions,leading to insufficient cross-dimensional interaction. To address these issues,this paper designs a triplet interaction mechanism for cross-view geo-localization. This method uses ConvNeXt as the feature extraction network,followed by a proposed triplet interaction mechanism,for feature enrichment operations. Finally,a joint loss function is utilized to guide model training. It performs multiple dimensional interactions within the model,reducing the problem of information loss in the two-dimensional feature projection. The proposed method includes a triplet interaction mechanism that uses different attention mechanisms in three channels,making the model robust to translations,scaling,and rotations for different cross-view images. Experimental results demonstrate that the proposed method can significantly outperforms other methods for both drone view localization and drone navigation tasks on University-1652 dataset.

Keywords　Cross-view,Geo-localization,Interaction mechanism,Feature attention

1　引言

跨视角地理定位是一种图像检索任务,其目标是在不同视角下(如地面、无人机和卫星视角)将无地理坐标的图像与数据库中有地理坐标的图像进行检索匹配,从而获取目标图像的地理位置信息[1-3]。该技术可广泛应用于无人机导航[4]、自动驾驶[5]和无人机精准配送[6-7]等。早期的跨视角地理定位任务主要研究地面和卫星视角匹配任务,但地面图像与卫星图像视角垂直导致该任务存在明显的视觉差异问题[8]。随着无人机技术的不断发展,无人机-卫星跨视角地理定位任务逐渐成为当下主流[4]。该技术目前存在以下难点:1)无人机设备与目标之间是斜角状态,导致无人机视角中的目标外观是扭曲的,特征信息不易被提取;2)同一区域的大多数建筑具有相似的建筑风格和同质外观,这导致类间差异较小,特征区分度较低。

近年来,随着深度学习的发展,跨视角地理定位技术有了显著的突破。Workman 等[9]首次引入深度学习技术来学习跨视角图像地理定位图像的特征,自此,基于深度学习的跨视角

到稿日期:2024-05-07　返修日期:2024-10-14
基金项目:江苏省自然科学基金(BK20200581)
This work was supported by the Natural Science Foundation of Jiangsu Province,China(BK20200581).
通信作者:李阳(solarleeon@outlook.com)

地理定位方法成为该领域的主要研究趋势。现有的大部分研究使用卷积神经网络提取图像特征，进一步解析输入图像的特征分布并学习特征表示[10-12]。其中，部分方法采用特征划分策略对特征图进行分割解析，从而更加细致地关注输入图像的每一部分，但是该类方法破坏了原始图像的全局位置信息和特征完整性[11]。此外，还有部分方法通过添加额外的先验知识作为特征融合信息的一部分。该类方法虽然提升了跨视角地理定位的精度，但是由于使用了额外的先验知识，因此模型计算量增加并且存在不公平比较现象[12]。近期，Shen等[2]设计了一个二维交互模块，充分利用图像的全局特征信息提升模型能力，如图1(a)所示。该二维交互模块虽然有效地完成了特征丰富，但未充分利用不同维度间的信息，导致特征交互不充分；同时，压缩操作虽然增强了特征关系，但损害了语义完整性，导致部分特征丢失。

为解决以上问题，本文设计了一种跨视角地理定位三维交互机制，在不破坏图像完整性以及不增加额外先验知识的基础上，利用完整的通道和空间交互实现了更精确的特征提取。该方法包含一个即插即用的三维交互机制(TIM)，如图1(b)所示。该机制由3个平行分支组成，在通道中分别利用 C，H 和 W 这3个维度之间任意两个特征之间的关系，使其具备交互的完整性并在各分支的输出阶段进行三维交互。与先前方法不同，本文提出的三维交互模块促使特征进行了多次交互，缓解了二维特征投影部分信息丢失的问题；同时，三维特征具有视角不变性，图像经过平移、缩放、旋转后其三维特征仍然保持稳定，可以进一步提高模型的泛化能力和鲁棒性。

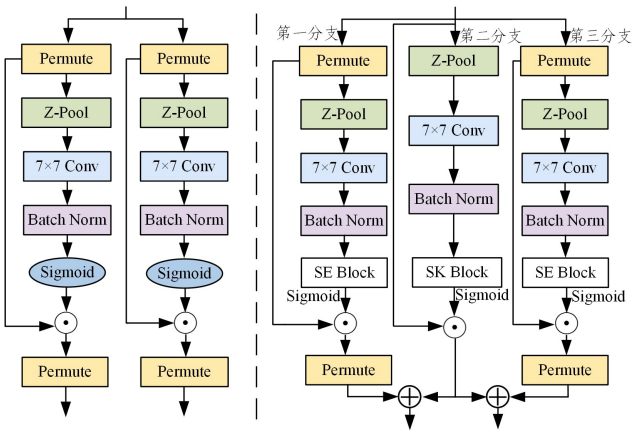


图1 交互机制比较

Fig. 1 Comparison of different interaction mechanisms

综上所述，本文方法的主要贡献主要如下：

1)设计了一种跨视角地理定位三维交互机制的方法。与现有方法不同，本文方法更加关注图像的完整结构和特征的三维语义信息，利用完整的空间交互实现了更精确的特征提取。

2)提出了一个即插即用的三维交互机制(TIM)，该机制在3个通道中使用不同的注意力机制，使模型对跨视角图像的平移、缩放、旋转具有鲁棒性。

3)在University-1652数据集上针对无人机视角定位和无人机导航两个任务进行了实验，实验结果表明，与目前主流的12种方法相比，本文方法在所有指标上均取得了最优性能。

2 相关工作

2.1 跨视角地理定位方法

早期的跨视角地理定位方法主要利用手工特征[13]进行跨视角图像匹配，但是跨视角地理图像之间存在着巨大的视角差异，手工特征提取方法提取的特征差异过大，因此检索精度不高[1]。深度学习网络的出现促进了跨视角地理定位研究的进步。现有的跨视角地理定位方法可以大致分为基于卷积神经网结构和基于混合网络结构的方法。

1)基于卷积神经网结构的方法

该类方法采用孪生网络对不同视角的图像进行训练，通过卷积神经网络提取图像特征进行图像检索，从而得到所需视角下的图像。Tian等[14]采用了速度更快的区域卷积神经网络(R-CNN)[15]检测查询图像和参考图像中的建筑物，并使用了基于优势集的多近邻匹配策略，提高了定位精度。Lin等[16]采用了在人脸验证领域表现优异的Where-CNN方法对跨视角图像进行检索。与现有方法所学习的深度特征相比，该方法通过添加一个额外的均值方差归一化操作，减少了模型受负例样本的影响，显著提升了跨视角视图匹配效果。Hu等[17]提出了一种名为CVM-Net(Cross-View Matching Network)的网络结构，该网络结构将NetVLAD层与孪生神经网络相结合，使用卷积网络提取跨视角图像的局部特征，并利用NetVLAD层构建全局描述符。鉴于孪生神经网络可以共享参数的结构，该方法能够学习到不同视角图像之间的共同特征，实现对视角变化的鲁棒性，并且显著提高了检索效率。Shi等[18]提出了一种新的跨视角特征传输(Cross-View Feature Transport，CVFT)方法，该方法利用不同视角间的特征传递，实现了特征对齐。

2)基于混合网络结构的方法

Transformer强大的全局建模能力有效地解决了CNN中感受野有限的问题。Yang等[19]开发了一种基于Transformer的层到层结构，利用Transformer的自注意力机制来建立全局依赖关系，显著减轻了视觉模糊性。Zhu等[20]将卷积神经网络与Transformer相结合，第一部分采用ResnetV2[21]学习细粒度特征，第二部分采用Transformer[22]从图像中提取全局特征。该方法首次利用多模态和双线性的融合机制来解决无人机和卫星视图之间的差异问题，提高了鲁棒性[23]。

2.2 跨视角地理定位中的注意力机制

注意力机制是一种有效重新分配网络模型可用资源的技术，可以使神经网络自动地关注输入数据的重要信息，提高模型的性能和泛化能力[24]。在现有的研究中，注意力机制已被广泛用于人工智能任务中，如行人重识别[25-26]、细粒度图像识别[27]和地理定位[2]。

在跨视角地理定位领域，Cai等[28]提出了一个轻量级注意力模块(Feature Context-based Attention Module，FCAM)，通过通道和空间注意力子模块重新加权特征，实现了更加丰富的特征表示。Zhuang等[29]提出了多尺度块注意力(Multi-scale Block Attention Network，MSBA)，将图像分割成多个

不同尺度的部分，并引入自注意力机制有效减少了跨视角地理定位任务对尺度和位移的影响。Zhuang 等[30]提出了一个语义引导模块(Semantic Guidance Module，SGM)来促进特征对齐并将相同的语义部分进行对齐，提升跨视角地理定位任务的精度。Zhu 等[31]提出了空间混合特征聚合模块(Spatial-mixed Feature Aggregation Module，SMD)，该模块将空间信息混合投影到低维空间中，进而生成特征描述符，有效地聚合了网络中的局部特征。

值得注意的是，现有的通道注意力和空间注意力是分离计算、彼此独立的，缺乏对联合特征的探索。受现有注意力机制中跨维度特征交互的启发，本文方法基于跨维交互的概念，提出三维交互机制，实现了对完整语义信息的三维特征交互，如图 1(b)所示。

3 本文方法

为了探索完整语义条件下不同维度特征的有效性，本章将对所提出的方法进行介绍，完整的网络结构如图 2(a)所示。该网络结构包含特征提取、三维交互机制和损失函数 3 个部分。第一部分使用 ConvNeXt 网络[32]对一组无人机和卫星图像 $\boldsymbol{x}_i$，$i=\{1,2\}$进行特征提取，获得输入数据的特征图 $\boldsymbol{T}_i$，$i=\{1,2\}$，经过全局平均池化操作输出包含原始特征信息的特征向量 $\boldsymbol{L}_{1i}$，$i=\{1,2\}$。第二部分利用三维交互机制将特征图 $\boldsymbol{T}_i$ 复制为 3 份，分别送入 3 个不同分支进行处理，从而得到包含 3 个不同维度信息的特征向量 $\boldsymbol{y}_c$，$c=\{1,2,3\}$。该机制最后输出两个三维交互特征 $\boldsymbol{g}_{ik}$，$i=\{1,2\}$，$k=\{1,2\}$。第三部分使用三元组损失和交叉熵损失作为最终的损失函数。

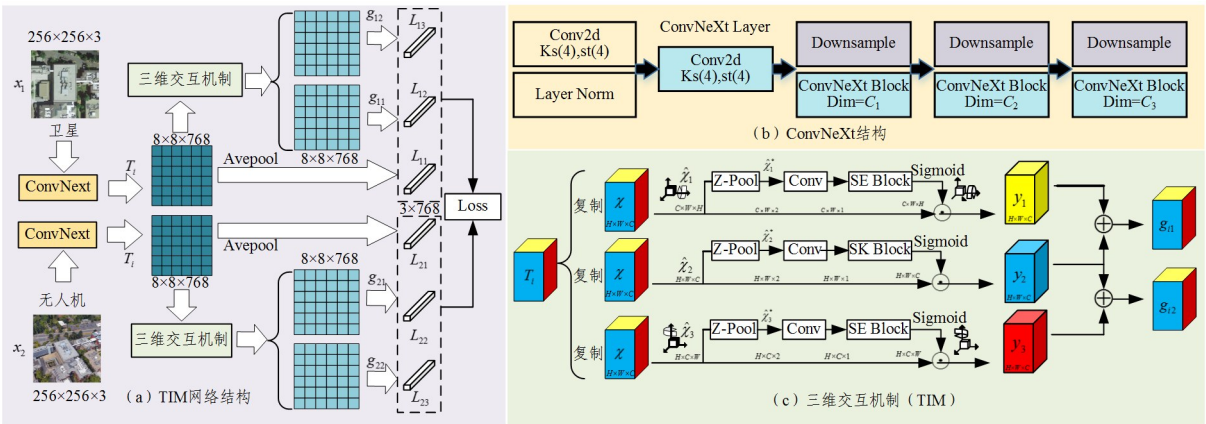


图 2 三维交互机制网络结构

Fig. 2 Framework of triplet interaction mechanism

3.1 网络结构

跨视角地理定位任务中，给定一个地理定位数据集，将输入图像表示为 $\boldsymbol{x}_i$，$i=\{1,2\}$。为了提取丰富的上下文信息，本文方法在特征提取过程中使用 ConvNeXt 网络作为骨干网，如图 2(b)所示。ConvNeXt 结合了 ResNet[33]和 Inception[34]模型的优势，减少了参数量并提高了训练稳定性。$\boldsymbol{x}_i$，$i=\{1,2\}$首先经过 ConvNeXt 网络提取，得到特征图 $\boldsymbol{T}_i$，$i=\{1,2\}$，如式(1)所示：

$$F_{\text{ConvNeXt}}(x_i)=\boldsymbol{T}_i \tag{1}$$

其中，i 表示输入图像 $\boldsymbol{x}_i$ 是从哪种无人平台收集的，$i=\{1,2\}$。具体地，$\boldsymbol{x}_1$ 表示卫星视角图像，$\boldsymbol{x}_2$ 表示无人机视角图像。

随后，将特征图 $\boldsymbol{T}_i$ 分为两部分处理，第一部分直接经过全局平均池化得到特征向量 $\boldsymbol{L}_{i\,1}$，$i=\{1,2\}$；第二部分送入本文所提出的三维交互机制进行特征交互(见 3.2 节)，将其复制成 3 份相同的张量 $\boldsymbol{\chi}$ 并送入各分支进行处理，输出为两个三维交互注意力引导的特征张量 $\boldsymbol{g}_{i\,k}$，$i=\{1,2\}$，$k=\{1,2\}$。三维交互机制的计算过程如式(2)所示：

$$[\boldsymbol{g}_{i\,1},\boldsymbol{g}_{i\,2}]=\zeta_{\text{三元维度交互机制}}(\boldsymbol{T}_i) \tag{2}$$

最后，得到的特征图 $\boldsymbol{T}_i$，$\boldsymbol{g}_{ik}$，$i=\{1,2\}$，$k=\{1,2\}$通过全局平均池化操作得到最终特征向量 $\boldsymbol{L}_{ij}$，$i=\{1,2\}$，$j=\{1,2,3\}$。具体特征向量计算式如式(3)所示：

$$\begin{cases}\boldsymbol{L}_{i1}=Avepool(\boldsymbol{T}_i)\\ \boldsymbol{L}_{i2}=Avepool(\boldsymbol{g}_{i1})\\ \boldsymbol{L}_{i3}=Avepool(\boldsymbol{g}_{i2})\end{cases} \tag{3}$$

其中，$\boldsymbol{T}_i$ 表示输入图像 $\boldsymbol{x}_i$ 经过 ConvNeXt 网络提取的特征图，$i=\{1,2\}$；$\boldsymbol{g}_{ik}$，$i=\{1,2\}$，$k=\{1,2\}$表示三维交互机制输出的特征图。

训练过程中将最终特征向量 $\boldsymbol{L}_{ij}$，$i=\{1,2\}$，$j=\{1,2,3\}$送入损失函数(见 3.3 节)，即可完成模型训练。本文使用的损失函数由三元组损失 TL 和交叉熵损失 CL 构成。

3.2 三维交互机制

如图 1(a)所示，现有的二维交互模块[2]中两个分支分别压缩了 H 和 W 维特征，捕获通道维度 C 与空间维度 H 或 W 之间的维度相互作用，实现了(C，W)和(C，H)之间的交互。但是，二维交互模块没有利用(H，W)维度之间的交互作用，对二维交互的利用缺乏完整性；此外，二维交互模块的两个分支通道中使用的压缩操作虽然增强了 C 维与 H 和 W 维之间的特征关系，但是损失了被压缩维度的特征信息。因此，二维交互模块还破坏了语义信息。

综合考虑现有方法的缺陷，本文方法设计了三维交互机制，如图 2(c)所示。该机制在通道中综合探索 C，H 和 W 3 个维度之间任意两个特征之间的关系，使其具备交互的完整性。具体地，该机制由 3 个平行分支组成，第一分支

用于捕获通道维度 C 与空间维度 H 之间的交互作用,第三分支用于捕获通道维度 C 与空间维度 W 之间的交互作用,而第二分支用于建立空间注意力。整个三维交互机制用于提取空间注意力与通道注意力之间的特征关系。具体构建过程中,首先特征图 $\boldsymbol{T}_i$ 被复制成 3 份一样的张量 $\boldsymbol{\chi},\boldsymbol{\chi}\in\mathbb{R}^{H\times W\times C}$,并传递给三维交互机制中的每一个分支中。3 个分支经过处理之后,最终输出为两个特征张量 $\boldsymbol{g}_{ik}$,$i=\{1,2\}$,$k=\{1,2\}$,经过全局平均池化操作得到包含维度交互信息的两个特征向量 $\boldsymbol{L}_{i2}$,$\boldsymbol{L}_{i3}$,$i=\{1,2\}$。下面将具体介绍三维交互机制的具体过程。

在第一分支中,其输入是由特征图 $\boldsymbol{T}_i$ 复制的张量 $\boldsymbol{\chi},\boldsymbol{\chi}\in\mathbb{R}^{H\times W\times C}$,将其沿 H 轴逆时针旋转 90°,记为 $\hat{\boldsymbol{\chi}}_1$,其形状为 $(C\times W\times H)$。为了减少资源算力,使用 Z-pool[35] 操作压缩张量的维度。Z-pool 操作在第 0 维上拼接平均池化特征和最大池化特征,将张量 $\hat{\boldsymbol{\chi}}_1$ 第零维的深度从 H 减少到 2,使得该层不仅能够保留实际张量的丰富表示而且还能缩小其深度,从而降低计算量。Z-pool 的计算过程如式(4)所示:

$$Z\text{-pool}(\boldsymbol{\chi})=[MaxPool_{0d}(\boldsymbol{\chi}),AvgPool_{0d}(\boldsymbol{\chi})] \tag{4}$$

其中,$0d$ 是第 0 维,拼接最大池化和平均池化的操作发生在张量的第 0 维上。具体地,在第一分支中形状为 $(C\times W\times H)$ 的张量 $\hat{\boldsymbol{\chi}}_1$ 经过 Z-pool 操作后形状变为 $(C\times W\times 2)$,记为 $\hat{\boldsymbol{\chi}}_1^*$。

随后,$\hat{\boldsymbol{\chi}}_1^*$ 通过内核大小为 $k\times k$ 的标准卷积层和批归一化层,输出维度为 $(C\times W\times 1)$ 的张量。由于 SE(Squeeze-and-Excitation)注意力[36] 可以在通道中对特征分布进行重新校准,让模型在该通道中强化判别信息较强的特征并抑制较弱的特征,因此,本方法使用 SE 注意力对通道中翻转后的新特征进行特征校准工作,使得新特征在通道中快速收敛并增强。维度 $(C\times W\times 1)$ 的张量特征经过 SE 注意力重新校准后,利用 Sigmoid 激活层生成该分支的注意力权重并通过 Broadcast 操作恢复为 $(C\times W\times H)$ 大小。所生成的注意力权重与张量 $\boldsymbol{\chi}_1$ 进行特征加权,加权后沿 H 轴顺时针旋转 90°,得到特征张量 $\boldsymbol{y}_1$。具体计算过程如式(5)所示:

$$\boldsymbol{y}_1=\hat{\boldsymbol{\chi}}_1\sigma(att_1(\psi_1(\hat{\boldsymbol{\chi}}_1^*))) \tag{5}$$

其中,$\hat{\boldsymbol{\chi}}_1$ 是第一分支中翻转后的张量,σ 为 Sigmoid 激活层,$\hat{\boldsymbol{\chi}}_1^*$ 表示经过压缩后的张量,att_1 表示 SE 注意力模块,ψ_1 表示第一个通道中的卷积操作。

类似地,在第二分支中,输入张量 $\boldsymbol{\chi}$ 不进行翻转生成 $\hat{\boldsymbol{\chi}}_2$,随后利用式(4)将张量 $\hat{\boldsymbol{\chi}}_2$ 的维度压缩为 $(H\times W\times 2)$,记为 $\hat{\boldsymbol{\chi}}_2^*$。张量 $\hat{\boldsymbol{\chi}}_2^*$ 通过内核大小为 $k\times k$ 的标准卷积层和批归一化层,最终的输出维度为 $(H\times W\times 1)$。由于 SK(Selective Kernel)注意力[37] 在通道中可以自适应地使用不同大小的卷积核对特征图进行特征提取,捕获不同尺度的目标物体,因此,在没有翻转操作的特征通道中,使用 SK 注意力进一步捕捉原始复杂图像空间的多尺度特征信息,丰富具有判别性的特征。维度 $(H\times W\times 1)$ 的张量经过 SK 注意力[37] 提取关键特征后,利用 Sigmoid 激活层生成该分支的注意力权重并通过 Broadcast 操作恢复为 $(H\times W\times C)$ 大小。所生成的注意力权重与张量 $\hat{\boldsymbol{\chi}}_2$ 进行特征加权,得到特征张量 $\boldsymbol{y}_2$。具体计算过程如式(6)所示:

$$\boldsymbol{y}_2=\hat{\boldsymbol{\chi}}_2\sigma(att_2(\psi_2(\hat{\boldsymbol{\chi}}_2^*))) \tag{6}$$

其中,$\hat{\boldsymbol{\chi}}_2$ 是第二分支中不翻转的张量,σ 为 Sigmoid 激活层,$\hat{\boldsymbol{\chi}}_2^*$ 表示经过压缩后的张量,att_2 表示 SK 注意力模块,ψ_2 表示第二个通道中的卷积操作。

在第三分支中,输入张量 $\boldsymbol{\chi}$ 沿着 W 轴逆时针旋转 90°,记为 $\hat{\boldsymbol{\chi}}_3$,其维度为 $(H\times C\times W)$。然后对其进行与第一分支相同的操作,将所得特征注意力权重通过 Broadcast 操作恢复为 $(H\times C\times W)$ 大小并与输入特征进行加权操作。最后沿 W 轴顺时针旋转 90°,得到特征张量 $\boldsymbol{y}_3$。具体计算过程如式(7)所示:

$$\boldsymbol{y}_3=\hat{\boldsymbol{\chi}}_3\sigma(att_1(\psi_3(\hat{\boldsymbol{\chi}}_3^*))) \tag{7}$$

其中,$\hat{\boldsymbol{\chi}}_3$ 是第一分支中翻转后的张量,σ 为 Sigmoid 激活层,$\hat{\boldsymbol{\chi}}_3^*$ 表示经过压缩后的张量,att_1 表示 SE 注意力模块,ψ_3 表示第三个通道中的卷积操作。

三维交互机制分别探索了不翻转、沿 H 轴翻转和沿 W 轴翻转的特征信息,完整提取了 C,H 和 W 这 3 个维度中任意两个特征之间的关系,并生成了对应的特征张量 $\boldsymbol{y}_1$,$\boldsymbol{y}_2$,$\boldsymbol{y}_3$。本文方法不仅探索了二维的特征交互,而且关注三维交互。$\boldsymbol{y}_1$,$\boldsymbol{y}_2$,$\boldsymbol{y}_3$ 的维度均为 $(H\times W\times C)$,使用第二分支张量分别与第一、三分支张量进行融合。具体地,三维交互机制最终输出维度特征张量 $\boldsymbol{g}_{i1}$,$\boldsymbol{g}_{i2}$ 的计算过程如式(8)所示:

$$\begin{cases}\boldsymbol{g}_{i1}=\dfrac{1}{2}(\boldsymbol{y}_1+\boldsymbol{y}_2)\\ \boldsymbol{g}_{i2}=\dfrac{1}{2}(\boldsymbol{y}_2+\boldsymbol{y}_3)\end{cases} \tag{8}$$

其中,$i=\{1,2\}$,$\boldsymbol{y}_1$,$\boldsymbol{y}_2$,$\boldsymbol{y}_3$ 表示三维交互机制中 3 个分支中输出的特征张量。

3.3 损失函数

在跨视角地理定位任务中,每个目标可以被视为一个单独的类别,因此图像检索问题可以转化为图像分类问题,即该训练过程的目标是最大程度地提高分类结果的准确性。交叉熵损失是衡量真实和预测概率分布之间差异和优化网络参数的损失函数,其定义如式(9)所示:

$$CL_j=-\sum_{r=1}^{R}p(x_{ir})\log q(x_{ir}) \tag{9}$$

其中,R 是类别的数量,$p(x_{ir})$ 和 $q(x_{ir})$ 分别是真实概率和预测概率。如果 x_{ir} 是对应目标的特征(label 的值为 1),则 $p(x_{ir})=1$,否则 $p(x_{ir})=0$。

三元组损失用于计算来自不同视图相同类别特征向量之间的距离。该过程使用传统的欧几里得距离[38] 作为度量指标,其定义为如式(10)所示:

$$TL=\max(\|\boldsymbol{F}_a-\boldsymbol{F}_p\|_2-\|\boldsymbol{F}_a-\boldsymbol{F}_n\|_2+M,0) \tag{10}$$

其中,$\|\cdot\|_2$ 表示 L_2 范数,$\boldsymbol{F}_a$ 是锚点图像 a 的特征向量,$\boldsymbol{F}_p$ 是与 a 类别相同的图像特征向量,$\boldsymbol{F}_n$ 是与 a 的类别不同的图像特征向量。在实验设置中,计算三元组损失的

间隔值 $M=0.3$。

综上所述，本文的总损失由交叉熵损失和三元组损失组成。其中，交叉熵损失用于为每个输入图像分配正确的地理位置标签[20,31]，三元组损失用于帮助网络学习到更具区分度的特征表示[2,23]。其计算过程如式(11)所示：

$$LS=\frac{1}{3}\sum_{q=1}^{3}\sum_{i=1}^{3}(CL_{iq}+TL_{iq}) \tag{11}$$

其中，CL_{iq} 和 TL_{iq} 分别是 $\boldsymbol{x}_i$ 的第 q 个特征的交叉熵损失和三元组损失。

4 实验结果与分析

4.1 节介绍本文方法的基本实验设置；4.2 节介绍本文使用的跨视角地理定位数据集；4.3 节介绍针对该数据集的评估指标；4.4 节进行实验结果分析。

4.1 实验设置

在实验中，参数初始化方法使用 Kaiming 初始化[39]，训练和测试阶段的输入图像大小调整为 256×256，并执行图像增强，如随机填充、随机裁剪和随机翻转。模型采用随机梯度下降(SGD)进行训练，动量参数设置为 0.9，权重衰减系数设置为 0.000 5，批量大小设置为 8。骨干网络的学习率为 0.003，其余层的学习率为 0.01。共训练 200 个 epoch，学习率在 80 和 120 个 epoch 后分别降低 10%。在测试阶段，本文方法使用特征的余弦距离作为查询图像和图库集中的候选图像之间的相似性度量。实验使用 PyTorch 框架在配备 48 GB 内存的 Nvidia GTX A6000 GPU 上完成，配置实验环境如表 1 所列。

表 1 实验环境

Table 1 Experimental environment

硬件环境	软件环境
CPU:Intel Xeon Gold 5218	操作系统:Ubuntu 20.04.1
GPU:NVIDIA RTX A6000	框架:PyTorch
RAM:48 GB	语言:python

4.2 University-1652 数据集

University-1652 数据集[4]是一个涵盖多视角、多数据源的综合数据集，它整合了来自 3 个不同平台的图像数据，包括卫星图像、无人机图像以及地面图像。与选取具有显著地标特征的建筑作为目标位置的传统方法不同，该数据集创新性地选取了全球范围内 72 所大学的 1 652 栋普通建筑作为目标对象。数据集中，同一地点和相同视角下的不同图像是通过平移、旋转和缩放等操作而得来的，每栋建筑对应 1 张卫星图像和 54 张无人机图像。数据集被划分为训练集和测试集，其中训练集包含 33 所大学的 701 栋建筑目标，而测试集则由另外 39 所大学的 951 栋建筑目标构成。值得注意的是，训练集和测试集中的大学建筑目标不存在重叠，从而确保了评估结果的客观性和有效性。此外，由于采集图像时使用了多角度和多层次的拍摄操作，因此该数据集具有平移、旋转和缩放等特性。实验数据的具体分布如表 2 所列。

表 2 实验数据统计

Table 2 Experimental data statistics

训练集		
类别	像数目	大学数量
无人机	37 854	701
卫星	701	701
测试集		
类别	图像数目	大学数量
卫星(图库)	951	951
无人机(图库)	51 355	951
无人机(查询集)	37 854	701
卫星(查询集)	701	701

University-1652 数据集包含无人机视角定位和无人机导航两个跨视角地理定位任务。第一种任务是无人机视角定位，旨在给定一张无人机视角的图像，寻找与其最匹配的卫星图像，以便利用卫星图像的 GPS 信息对无人机视角中的目标进行地理定位。在此任务中，查询集包含了 37 854 张无人机图像，而图库中包含了 951 张相应的卫星图像。第二种任务是无人机导航，该任务要求基于给定的卫星图像，无人机能够检索到其先前拍摄的相应无人机视角图像，从而实现导航的目的，即返回到图像的拍摄地点。在此任务中，查询集由 701 张卫星图像组成，而图库则包含了 51 355 张对应的无人机图像。

4.3 评价指标

本文使用召回精度 R@K 和平均精度(AP)来评价模型的性能。@K 表示从查询集(Query)中选择一张查询图像放入参考图库集(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。召回分数越高表明网络的性能越好，具体计算过程如式(12)所示：

$$Recall@K=\sum_{i=1}^{n}S_{i,k}/n \tag{12}$$

其中，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。R@K 值越大，检索算法的性能越好。

平均精度(AP)由 Precision-Recall 曲线下的面积构成。精确率 Precision 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。精确率的计算过程如式(13)所示：

$$Precision=\frac{TP}{TP+FP}=\frac{TP}{N} \tag{13}$$

其中，TP 代表正类被预测为正类的数量；TN 代表负类被预测为负类的数量；FP 代表负类被预测为正类的数量；FN 代表正类被预测为负类的数量；N 代表识别为正样本的数量。

4.4 实验结果

4.4.1 与现有方法的比较

将本文方法与近几年的最新方法进行了比较，结果如表 3 所列。当输入图像尺寸为 256×256 时，在无人机视角定位任务(无人机→卫星)中实现了 90.24%的 R@1 精度，超过了当前最先进模型 0.96%；在无人机导航任务(卫星 → 无人机)中实现了 90.85%的 AP 精度，超过了当前最先进模型 1.56%。当输入尺寸调整为 384×384 时，较大的图像尺寸拥有更多的像素点，可以捕捉到更丰富的细节和上下文信息，

性能仍得到全面增强。在无人机目标定位任务（无人机→卫星）中，R@1 精度进一步达到了 91.37%，超过当前最先进模型 1.75%。在无人机导航任务（卫星→无人机）中，AP 精度进一步达到了 91.75%，超过当前最先进模型 2.36%，实现了指标最大增幅2.64%。由于 University-1652 数据集是一个涵盖多视角、多数据源的综合数据集，具有平移、旋转和缩放的特征，因此，较优的实验结果表明本文方法在面对复杂、变化的跨视角地理图像时具有良好的鲁棒性。

表 3 在 University-1652 数据集上与主流方法的比较

Table 3 Comparison with state-of-the-art methods on University-1652

方法	发表	骨干网络	输入图尺寸	无人机→卫星		卫星→无人机	
				R@1/%	*AP*/%	*R*@1/%	*AP*/%
Baseline[4]	ACM MM 2020	ResNet-50	256×256	58.49	63.31	71.18	58.74
LPN[11]	TCSVT 2021	ResNet-50	256×256	74.18	77.39	85.16	73.68
RK-Net(USAM)[2]	TIP 2022	ResNet-50	256×256	77.07	80.09	85.16	74.06
FSRA[40]	TCSVT 2022	Vit-Transformer	256×256	84.51	86.71	88.45	83.37
SGM[30]	IEEE Access 2022	Swin-Tiny	256×256	82.14	84.72	88.16	81.81
MSBA[29]	Remote Sens 2021	ResNet-50	256×256	82.33	84.78	90.58	81.61
PAAN[41]	JRNAL 2022	SE-ResNet-50	256×256	84.51	86.78	91.01	82.28
DWDR[42]	—	Swin-B	256×256	86.41	88.41	91.30	86.02
MCCG[2]	TCSVT 2023	ConvNeXt	256×256	89.28	91.01	94.29	89.29
Ours	—	ConvNeXt	256×256	**90.24** ↑0.96	**91.80** ↑0.79	**94.58** ↑0.29	**90.85** ↑1.56
PCL[43]	TCSVT 2021	ResNet-50	384×384	81.63	85.46	89.73	80.84
FSRA[40]	TCSVT 2022	Vit-Transformer	384×384	84.82	87.03	87.59	83.37
SDPL[44]	—	ResNet-50	384×384	82.57	85.08	88.73	80.79
OSNet[45]	LGRS 2023	—	384×384	84.79	87.07	93.15	84.93
MCCG[2]	TCSVT 2023	ConvNeXt	384×384	89.64	91.32	94.30	89.39
Ours	—	ConvNeXt	384×384	**91.39** ↑1.75	**92.78** ↑1.46	**95.29** ↑0.99	**91.75** ↑2.36

注：加粗字体表示精度最优的方法，下划线字体表示精度次优的方法。

综上所述，在无人机视角定位和无人机导航两个跨视角地理定位任务中，本文方法实现了显著的性能提升，在不同尺度下均超过了近年来大多数有竞争力的方法。

4.4.2 消融实验

为了验证所提方法的有效性，本文（TIM）从网络结构、分支作用和图像大小 3 个方面对三维交互机制进行了消融实验分析。

1）不同网络中三维交互机制（TIM）的效果

为了验证 TIM 的有效性，本文分别使用了 ResNet-101，ConvNeXt-Tiny 和 ConvNeXt-Small 作为主干网络进行实验，结果如表 4 所列。

表 4 三维交互机制在不同网络上的精度结果

Table 4 Results of TIM on different backbones

(%)

骨干网络类型	无人机→卫星		卫星→无人机	
	R@1	*AP*	*R*@1	*AP*
ResNet-101	72.36	76.10	86.88	72.70
ResNet-101+TIM	75.01	78.73	88.72	74.33
ConvNeXt-Tiny	89.28	91.01	94.29	89.29
ConvNeXt-Tiny+TIM	90.24	91.80	94.58	90.85
ConvNeXt-Small	89.40	91.01	95.01	89.93
ConvNeXt-Small+TIM	89.84	91.40	95.01	90.57

我们发现，不同的网络使用了三维交互机制后，两个跨视角任务的 R@1 和 AP 精度均得到了提升。其中，TIM 在 ResNet 网络上提升效果最大，在无人机视角定位任务中 R@1 精度和 AP 精度分别提升了 2.65%和 2.63%。在无人机导航任务中，R@1 精度和 AP 精度分别提升了 1.84%和1.63%；TIM 在 ConvNeXt-Tiny 网络上提升性能最优，在无人机视角定位任务中 R@1 精度和 AP 精度分别提升了 0.96%和 0.79%。在无人机导航任务中，R@1 精度和 AP 精度分别提升了0.29%和 1.56%。因此，为了获得跨视角地理定位任务的最优性能，本文方法选用 ConvNeXt-Tiny 网络作为主干网络。其次，实验结果显示 TIM 在多种类型骨干网中均有稳定的提升效果，表明该机制具有良好的模型鲁棒性。

为了进一步对比不同网络在不同指标下的增长幅度，本文针对 TIM 在不同网络上的性能进行了可视化，如图 3 所示。从图中可以看出，ConvNeXt-Tiny 网络的各项精度均优于 ResNet 网络，并且 TIM 机制对两种网络均有提升效果。这说明三维交互机制（TIM）是一个适用于多种网络的即插即用的特征注意力模块。

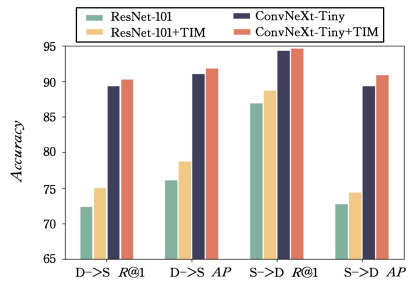


图 3 三维交互机制在不同网络上的精度

Fig. 3 Accuracy of TIM on different backbones

2）三维交互机制（TIM）中各分支的影响和作用

现有的实验结果表明，本文提出的三维交互机制可有效提高模型的性能，但 TIM 中不同分支的贡献度有所不同，因此本文进一步对 TIM 的 3 个分支进行消融实验。如表 5 所列，去除 SE 分支后，TIM 在无人机视角定位任务中 R@1 精度下降了 1.42%，AP 精度下降了 1.29%；在无人机导航任务中 R@1 精度下降了 0.57%，AP 精度下降了 1.85%。该数据表明，在翻转通道中使用 SE 注意力促进了模型关注信息量较大的特征并抑制了不重要的特征。由于该操作对翻转后

的新特征进行了特征重校准，因此该分支有效地增强了三维交互机制的全局特征提取能力。去除 SK 分支后，TIM 在无人机视角定位任务中，R@1 精度下降了 1.11%，AP 精度下降了 0.94%；同时在无人机导航任务中 R@1 精度下降了 0.43%，AP 精度下降了 1.3%。该数据表明，在通道中使用 SK 注意力处理不翻转的特征图，利用自适应的卷积核对原始特征进行特征提取，可以有效地促使该分支关注原始图像中的关键区域。同时去除 SE 和 SK 分支后，TIM 在无人机视角定位任务中 R@1 精度下降了 1.08%，AP 精度下降了 0.97%，在无人机导航任务中 R@1 精度下降了 1.14%，AP 精度下降了 0.48%。从上述数据可以看出，每个分支都有效提升了 TIM 的性能，而 SE 分支对 TIM 的贡献度较大，SK 分支对 TIM 的贡献度略小。整个三维交互机制中，使用 SE 分支对两个翻转后的新特征进行特征重校准，以及使用 SK 分支对不翻转的原始特征进行注意力重分配，有效地增强了不同分支中的通道特征；而三维交互机制间各分支之间完成空间上的特征交互，最终提升了跨视角地理定位任务的整体精度。

表 5 三维交互机制的消融实验结果

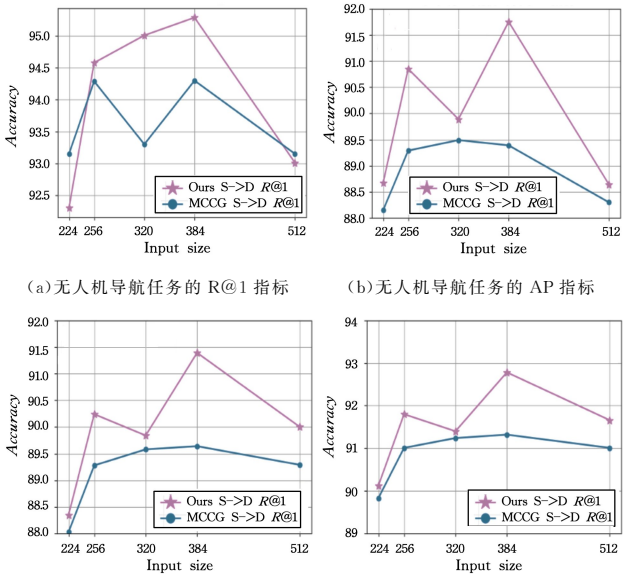
Table 5 Ablation experimental results of TIM

(%)

方法	无人机→卫星		卫星→无人机	
	R@1	*AP*	*R@1*	*AP*
TIM	90.24	91.80	94.58	90.85
TIM w/o SE	88.82	90.51	94.01	89.00
TIM w/o SK	89.13	90.86	94.15	89.55
TIM w/o SE and SK	89.16	90.83	93.44	90.37

3)输入图像大小的影响

为了研究不同输入大小对模型性能的影响，本文选取 5 种输入尺寸进行实验并与目前的最新精度进行比较，实验结果如图 4 所示。



(a)无人机导航任务的 R@1 指标 (b)无人机导航任务的 AP 指标

(c)无人机定位任务的 R@1 指标 (d)无人机定位任务的 AP 指标

图 4 与最新方法在不同输入尺寸下的比较

Fig. 4 Comparison with state-of-the-art methods with different input sizes

在无人机导航(见图 4(a)和图 4(b))和无人机视角定位(见图 4(c)和图 4(d))两个任务中，通过改变输入尺寸大小，指标性能均发生了变化，总体呈现先下降后上升的趋势；并且该变化是具有弹性的，最大输入尺度的图像效果并非最优。总体来说，从 4 幅图中可以看出，不同尺寸条件下本文方法的精度几乎均优于当前最新方法[2]。当输入图像尺寸为 384×384 时，较大的图像尺寸不仅拥有更丰富的特征信息，而且能避免模型出现过拟合问题，因此各类指标的精度在该尺寸下均达到最高。

4)损失函数的有效性实验

为了逐步验证联合损失函数的有效性，我们分别独立使用交叉熵损失和三元组损失指导模型训练，实验结果如表 6 所列。将跨视角地理定位问题独立视为一个图像分类问题时，在无人机视角定位任务中 R@1 精度下降了 22.28%，AP 精度下降了 19.62%；在无人机导航任务中 R@1 精度下降了 15.58%，AP 精度下降了 24.2%。实验结果表明，仅依赖于使用交叉熵损失对模型进行训练，将图像分类至相应标签的效果较差。因此，我们推测跨视角地理定位问题的本质仍然属于图像检索问题。将跨视角地理定位问题独立视为一个图像检索问题时，在无人机视角定位任务中 R@1 精度下降了 4.21%，AP 精度下降了 3.73%；在无人机导航任务中 R@1 精度下降了 2.24%，AP 精度下降了 3.89%。实验结果表明，单独使用三元组损失训练模型，随机选择一张图像作为锚点，然后拉近同类标签图像、推远不同标签图像的方法略优于单独使用交叉熵损失训练的方法，但是整体效果仍然较差。因此，使用交叉熵损失和三元组损失联合指导模型训练的方法更有效。

表 6 损失函数的消融实验

Table 6 Ablation experiments of loss functions

(%)

Loss	无人机→卫星		卫星→无人机	
	R@1	*AP*	*R@1*	*AP*
Loss w/o Triple	67.99	72.18	79.03	66.65
Loss w/o CE	86.03	88.17	92.44	86.96
Total Loss	90.24	91.80	94.58	90.85

4.4.3 定性结果的可视化

此外，图 5 展示了本文方法在 University-1652 数据集中两个跨视角地理定位任务上的可视化检索结果。黄色框表示正确匹配的图像，蓝色框表示错误匹配的图像。在无人机导航任务(卫星→无人机)中，从测试集中随机选择 3 张卫星图像，对于每个卫星视角图像，模型从图库集中返回前 5 个类似的无人机图像；在无人机视角定位任务(无人机→卫星)中，从测试集中随机选择 3 张无人机图像，对于每个无人机视角图像，模型从图库集中返回前 5 个类似的卫星图像，结果如图 5 所示。可以观察到，即使对于随机选择的具有视角差异的图像，本文方法也取得了较好的匹配结果。尤其在无人机视角定位任务(无人机→卫星)中，可以发现本文方法能够有效地区分与正例样本非常相似的负例样本。

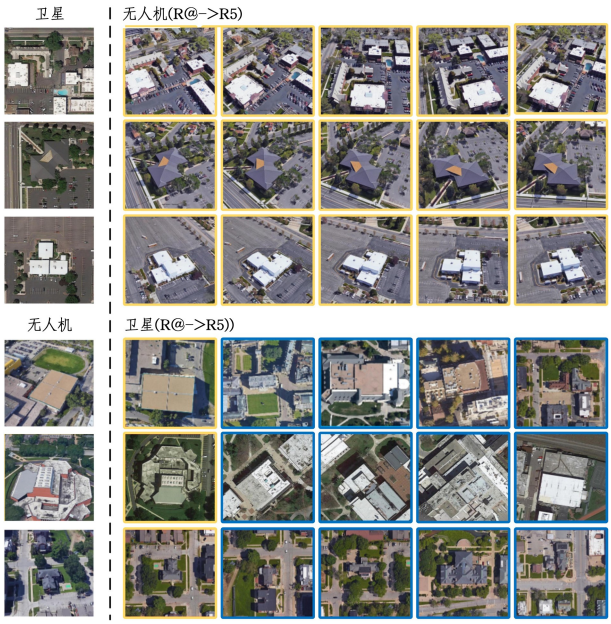


图 5 跨视角地理定位检索结果(电子版为彩图)

Fig. 5 Cross-view image retrieval results

结束语 本文研究了无人机视图和卫星视图之间的跨视角匹配任务，设计了一种跨视角地理定位三维交互机制。与现有方法不同，本文关注图像的完整结构和特征更加丰富的三维语义信息，并利用完整的空间交互实现了更精确的特征提取。该方法中包含了一个即插即用的三维交互机制，在3个通道中使用了不同的注意力使模型对跨视角图像的平移、缩放、旋转具有鲁棒性。实验结果表明，本文方法在University-1652公共数据集上实现了无人机视角定位和无人机导航两个任务的最优性能。虽然本文方法提升了跨视角地理定位的精度，但没有进一步利用局部特征。因此，我们将在后续工作中进一步关注细粒度特征提取和多模态信息使用问题[46]，将其运用在跨视角地理定位领域中。

参 考 文 献

[1] LIN J, ZHENG Z, ZHONG Z, et al. Joint representation learning and keypoint detection for cross-view geo-localization[J]. IEEE Transactions on Image Processing, 2022, 31: 3780-3792.

[2] SHEN T R, WEI Y M, KANG L, et al. Mccg: a ConvNeXt-based multiple-classifier method for cross-view geo-localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(3): 1456-1468.

[3] ZHU S, SHAH M, CHEN C. Transgeo: Transformer is all you need for cross-view image geo-localization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 1162-1171.

[4] 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.

[5] CHEN D, KRÄHENBüHL P. Learning from all vehicles[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 17222-17231.

[6] PENG T, LI Q, ZHU P. Rgb-t crowd counting from drone: a benchmark and mmccn network[C]// Proceedings of the Asian Conference on Computer Vision. 2020.

[7] LUO H Y, CHEN T X, LI X J, et al. Keepedge: a knowledge distillation empowered edge intelligence framework for visual assisted positioning in UAV delivery[J]. IEEE Transactions on Mobile Computing, 2022, 22(8): 4729-4741.

[8] SHUGAEV M, SEMENOV I, ASHLEY K, et al. Arcgeo: localizing limited field-of-view images using cross-view matching [C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 209-218.

[9] 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.

[10] SUN Y X, YE Y M, KANG J, et al. Cross-view object geo-localization in a local region with satellite imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-16.

[11] WANG T, ZHENG Z, YAN C, et al. Each part matters: local patterns facilitate cross-view geo-localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(2): 867-879.

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

[13] LOWE D G. Object recognition from local scale-invariant features[C]// Proceedings of the seventh IEEE International Conference on Computer Vision. IEEE, 1999, 2: 1150-1157.

[14] TIAN Y, CHEN C, SHAH M. Cross-view image matching for geo-localization in urban environments[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3608-3616.

[15] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.

[16] LIN T Y, CUI Y, BELONGIE S, et al. Learning deep representations for ground-to-aerial geolocalization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5007-5015.

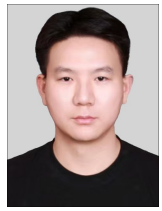
[17] HU S, FENG M, NGUYEN R M H, et al. Cvm-net: cross-view matching network for image-based ground-to-aerial geo-localization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7258-7267.

[18] 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: 11990-11997.

[19] YANG H, LU X, ZHU Y. Cross-view geo-localization with layer-to-layer Transformer[J]. Advances in Neural Information Processing Systems, 2021, 34: 29009-29020.

[20] ZHU R, YANG M, YIN L, et al. UAV's status is worth considering: A fusion representations matching method for geo-localization[J]. Sensors, 2023, 23(2): 720.

[21] HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks[C]// Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11—14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016: 630-645.
[22] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv: 2010. 11929, 2020.
[23] ZHANG X, LI X, SULTANI W, et al. Cross-view geo-localization via learning disentangled geometric layout correspondence [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023: 3480-3488.
[24] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv: 1409. 0473, 2014.
[25] LI S, BAK S, CARR P, et al. Diversity regularized spatiotemporal attention for video-based person re-identification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 369-378.
[26] XU J, ZHAO R, ZHU F, et al. Attention-aware compositional network for person re-identification [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2119-2128.
[27] FU J, ZHENG H, MEI T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4438-4446.
[28] 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.
[29] ZHUANG J, DAI M, CHEN X, et al. A faster and more effective cross-view matching method of UAV and satellite images for UAV geolocalization[J]. Remote Sensing, 2021, 13(19): 3979.
[30] ZHUANG J, CHEN X, DAI M, et al. A semantic guidance and Transformer-based matching method for UAVs and satellite Images for UAV Geo-Localization[J]. IEEE Access, 2022, 10: 34277-34287.
[31] ZHU Y, YANG H, LU Y, et al. Simple, effective and general: a new backbone for cross-view image geo-localization[J]. arXiv: 2302. 01572, 2023.
[32] LIU Z, MAO H, WU C Y, et al. A convnet for the 2020s[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11976-11986.
[33] 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.
[34] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.
[35] MISRA D, NALAMADA T, ARASANIPALAI A U, et al. Rotate to attend: convolutional triplet attention module[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 3139-3148.
[36] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.
[37] LI X, WANG W, HU X, et al. Selective kernel networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 510-519.
[38] ZENG W, WANG T, CAO J, et al. Clustering-guided pairwise metric triplet loss for person reidentification[J]. IEEE Internet of Things Journal, 2022, 9(16): 15150-15160.
[39] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification [C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1026-1034.
[40] DAI M, HU J, ZHUANG J, et al. A Transformer-based feature segmentation and region alignment method for UAV-view geo-localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(7): 4376-4389.
[41] BUI D V, KUBO M, SATO H. A part-aware attention neural network for cross-view geo-localization between UAV and satellite[J]. Journal of Robotics, Networking and Artificial Life, 2022, 9(3): 275-284.
[42] WANG T, ZHENG Z, ZHU Z, et al. Learning cross-view geo-localization embeddings via dynamic weighted decorrelation regularization[J]. arXiv: 2211. 05296, 2022.
[43] TIAN X, SHAO J, OUYANG D, et al. UAV-satellite view synthesis for cross-view geo-localization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(7): 4804-4815.
[44] CHEN Q, WANG T, YANG Z, et al. Sdpl: shifting-dense partition learning for UAV-view geo-localization[J]. arXiv: 2403. 04172, 2024.
[45] SONG H, WANG Z, LEI Y, et al. Learning visual representation clusters for cross-view geo-location[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 1-5.
[46] WANG Y P, LI Y, WANG J B, et al. A robust lightweight deep learning method for remote sensing scene image classification and retrieval under label noise[J]. Journal of Image and Graphics, 2021, 26(12): 2991-3004.



ZHOU Bowen, born in 1999, postgraduate. His main research interests include deep learning and cross-view geo-localization.

LI Yang, born in 1984, Ph. D, associate professor, is a senior member of CCF (No. D24215). His main research interests include computer vision and image processing.

（责任编辑：何杨）