基于学习解耦几何布局对应的跨视角地理定位

摘要

跨视角地理定位旨在通过将查询地面全景图像与参考的带位置信息的航拍图像数据库进行匹配，来估计查询图像的位置。作为一项极具挑战性的任务，其困难在于两个视角之间巨大的视角变化以及拍摄时间的差异。尽管困难重重，最近的研究在跨视角地理定位基准上取得了显著进展。然而，现有方法在跨区域基准 (即训练和测试数据来自两个不同数据集) 上仍然表现不佳。zhang 和 li 等人 (2023) [1] 将这种不足归因于缺乏提取视觉特征布局的空间配置能力以及模型对训练集中低级细节的过度拟合。因此 zhang 和 li 等人 [1] (2023) 提出了 GeoDTR 模型，该方法通过一种新式几何布局提取器模块明确地将几何信息从原始特征中分离出来，并学习来自航拍图像和地面全景图像对的视觉特征之间的空间相关性。该模块生成一组几何布局描述符并调整原始特征产生高质量的潜在表示。此外，他们详细阐述了两类数据增强方法，(i) 布局模拟，可以改变空间配置同时保持低级细节不变。(ii) 语义增强，可以改变低级细节并鼓励模型捕获空间配置。这些增强有助于改善跨视角地理定位模型的性能，特别是在跨区域基准上。此外，zhang 和 li 等人 [1] 提出了一种基于反事实的学习过程，以使几何布局提取器能够探索空间信息。通过大量的实验验证 GeoDTR 不仅仅取得了很好的检索效果，也显著提升了在跨区域基准上的性能。

关键词：跨视角地理定位，图像检索，几何布局，反事实学习

1 引言

跨视角地理定位被定义为从一组带位置信息的航拍图像 (也称为参考图像) 中估计地面全景图像 (也称为查询图像) 的位置。地面全景图像通常由安装在车辆上的摄像头或者由行人拍摄。跨视角地理定位可应用于不同领域，如自动驾驶 (Kim 和 Walter，2017 [2])、无人机导航 (Shetty 和 Gao，2019 [3]) 和增强现实 (Chiu 等，2018 [4])。大多数现有的跨视角地理定位方法都将问题框架化为检索任务，如：Shi 等，2019 [5]，2020a [6]；Yang，Lu 和 Zhu，2021 [7]；Hu 等，2018 [8]；Vo 和 Hays，2016 [9]；Workman，Souvenir 和 Jacobs，2015 [10]；Toker 等，2021 [11]；Shi 等，2020b [12]；Liu 和 Li，2019 [13]；Wang 等，2021 [14]；Zheng，Wei 和 Yang，2020 [15]。这些方法通常训练模型，将对应的航拍图像和地面全景图像对 (也称为地空图像对) 在潜在空间中靠近，并将不匹配的图像对彼此推开。在部署阶段，具有最高相似度的航拍图像的位置即为给定查询地面全景图像的预测位置。

跨视角地理定位被认为是一个极具挑战性的问题，原因在于：1) 视角发生了剧烈的变化，2) 拍摄时间存在差异，3) 地面全景图像和航拍图像的分辨率不同 (Wilson 等，2021 [16])。解决这些挑战需要对图像内容和视觉特征的空间配置有着深入的了解，例如建筑物和道路等。大多

数现有方法 (shi 等, 2019 [5], 2020b [12]; Hu 等, 2018 [8]; Lin 等, 2015 [17]; Cai 等, 2019 [18]; Vo 和 Hays, 2016) [9]) 通过利用从卷积神经网络 (Convolutional Neural Networks,CNNs) 中提取的特征来匹配地面全景图像和航拍图像。例如, Shi 等人 (2019 [5]) 通过使用空间感知位置嵌入 (Spatial-aware Position Embedding,SPE) 模块直接对物体特征之间的相对位置进行编码。然而, 因为它是一个全局属性, 这些方法在探索视觉特征的空间配置上存在局限性。最近, 随着 Transformer(Vaswani 等, 2017 [19]) 的进展, 一些方法 (Yang, Lu 和 Zhu, 2021 [7]; Zhu, Shah 和 Chen, 2022 [20]) 探索从全局上下文信息中提取潜在特征。然而, 这些方法仅仅依赖多头注意力机制来隐式探索输入特征之间的相关性。因此, 这些方法中的相关性不可避免地被纠缠在一起。

本文介绍的 GeoDTR, 它分别处理了视觉特征的低级特征和空间配置。为了捕获空间配置, zhang 和 li 等人 [1] 提出了一种新颖的几何布局提取子模块。该子模块生成一组几何布局描述符, 反映图像中视觉特征之间的全局上下文信息。通过布局模拟、语义增强 (Layout simulation, Semantic augmentation, LS) 和反事实 (Counterfactual, CF) 训练模式, 增强了几何布局描述符的质量。布局模拟通过扰动低级细节为航空和地面全景图像对生成不同的布局, 提高了训练数据的多样性。与现有的跨视角地理定位数据增强方法不同, 布局模拟在训练过程中保持了几何/空间对应关系。因此, 它可以普遍应用于任何地理定位方法。此外, zhang 和 li 等人 [1] 观察到布局模拟通过正则化提高了在跨区域实验中的性能。因此他们引入了一种新颖的基于距离的反事实 (CF) 训练模式, 以加强对提取描述符的学习。具体来说, 它为几何布局提取器提供辅助监督, 以完善全局上下文信息。他们提出的 GeoDTR 模型在常见的跨视角地理定位数据集 CVUSA(Workman, Souvenir 和 Jacobs, 2015 [10]) 和 CVACT(Liu 和 Li, 2019) [13] 上相比其他算法显著提高了性能, 并达到了最先进水平。

2 相关工作

2.1 跨视角地理定位

2.1.1 基于特征的跨视角地理定位

基于特征的地理定位方法使用 CNNs 从局部信息中提取航拍图和地面全景图像的潜在表示 (Lin, Belongie 和 Hays, 2013 [21]; Lin 等, 2015 [17]; Workman, Souvenir 和 Jacobs, 2015 [10])。现有工作研究了不同的聚合策略 (Hu 等, 2018 [8])、训练范式 (Vo 和 Hays, 2016 [9])、损失函数 (例如 HER(Cai 等, 2019 [18])) 和 SEH(Guo 等, 2022 [22]), 以及特征转换 (例如特征融合 (Regmi 和 Shah, 2019 [23]) 和跨视角特征传输 (Cross-View Feature Transport,CVFT)(Shi 等, 2020b [12]))。上述基于特征的方法未能充分探索空间信息的有效性, 这是因为 CNN 的局部性质导致其无法探索全局相关性。通过利用 Transformer 捕获全局上下文信息的能力, GeoDTR 通过基于 Transformer 的子模块学习了地面全景图像和航拍图像之间的几何对应关系, 从而实现了更好的性能表现。

2.1.2 基于几何的跨视角地理定位

最近,学习在航拍图像和地面全景视图之间匹配几何对应关系变得热门起来。Liu 和 Li(2019) [13] 提出在航拍图像和地面全景图像中编码摄像机方向来训练模型。Shi 等人 (2019) [5] 提出

了 SAFA，通过从地面全景图像和极坐标转换的航拍图像中学到的几何对应关系来聚合特征，后来，她们又提出了动态相似性匹配 (DSM)(Shi 等，2020a [6])，通过类似滑动窗口的算法对有限视野的地面全景图像进行地理定位。CDE(Toker 等，2021) 将 GAN(Goodfellow 等，2014) 和 SAFA(Shi 等，2019) 结合起来，同时学习跨视角地理定位任务中地面全景图像生成。尽管这些基于几何的方法取得了显著的效果，但它们受到 CNN 探索像素之间局部相关性的特性限制。但是 GeoDTR 不仅明确建模了局部相关性，还通过基于 Transformer 的子模块探索了全局上下文信息。这种全局上下文信息的质量通过 CF 学习模式和 LS 技术得到进一步加强。并且得益于模型的架构设计，GeoDTR 并不完全依赖于极坐标转换的航拍图像。

近期的研究 (Yang，Lu 和 Zhu，2021 [7]；Zhu，Shah 和 Chen，2022 [20]) 也在探索捕获图像中的非局部相关性。L2LTR(Yang，Lu 和 Zhu，2021 [7]) 研究了基于 ViT(Dosovitskiy 等，2020 [24]) 的混合方法，而 TransGeo(Zhu，Shah 和 Chen，2022 [20]) 则提出了一种纯 Transformer 模型。上述方法由于完全依赖 Transformer，因此隐含地从原始特征中建模了空间信息。然而，GeoDTR 明确地将低级细节和空间信息从原始特征中解耦。此外，GeoDTR 的可训练参数比 L2LTR(Yang，Lu 和 Zhu，2021 [7]) 少，并且不需要像 TransGeo(Zhu，Shah 和 Chen，2022 [20]) 中提出的两阶段训练范式。

2.1.3 跨视角地理定位中的数据增强

数据增强在计算机视觉中很受欢迎。然而，由于航拍图像和地面全景图像之间空间对应关系的脆弱性，跨视角地理定位中的数据增强非常有限，很容易被轻微干扰破坏。例如，大多数现有方法 (Liu 和 Li，2019 [13]；Rodrigues 和 Tani，2022 [25]；Vo 和 Hays，2016 [9]；Cai 等，2019 [18]) 在随机旋转或平移一个视图的同时固定另一个视图。另一方面，Rodrigues 和 Tani(2021) 会根据街景图像中的分割随机遮挡地面物体。在 GeoDTR 模型中，zhang 和 li 等人 [1] 提出了 LS 技术，在训练阶段保持了两个视图图像之间的几何对应关系，同时变化几何布局和视觉特征。同时 zhang 和 li 等人 [1] 进行的大量实验表明，LS 不仅可以显著提高 GeoDTR 在跨区域数据集上的性能，而且可以普遍应用于其他现有方法。

2.2 2.2 反事实学习

反事实在因果推断 (Pearl，2009 [26]) 中的理念已成功应用于多个研究领域，如可解释人工智能 (Byrne，2019 [27])、视觉问答 (Abbasnejad 等，2020 [28])、物理模拟 (Baradel 等，2019 [29]) 和强化学习 (Wang 等，2019 [30])。在 DeoGTR 模型的训练中，zhang 和 li 等人 [1] 提出了一种新颖的基于距离的反事实 (CF) 学习模式，用于增强 GeoDTR 学到的几何描述符的质量。实验证明，其所提出的 CF 学习模式提高了 GeoDTR 的性能。

3 方法

3.1 问题定义

对于一组地面-空中图像对 $(I_i^g, I_i^a), i = 1, ..., N$，其中上标 g 和 a 分别代表地面全景图像和航拍图像，N 是图像对的数量。每个图像对都与一个独特的地理位置相关联。在跨视角地

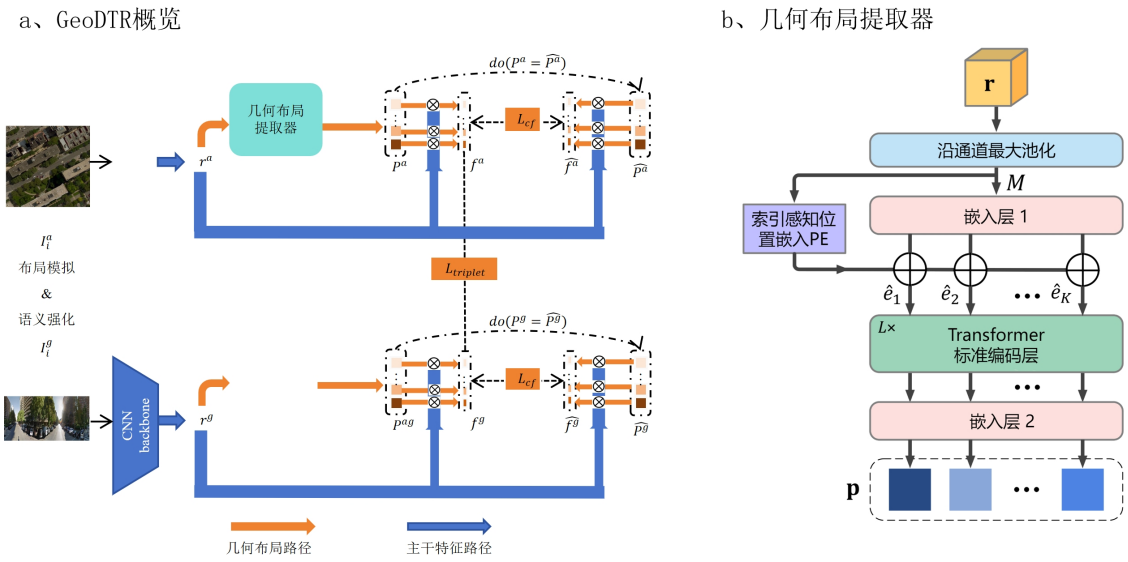

图 1. (a) GeoDTR 模型的总体流程概览。(b) 几何布局提取器的结构图

理定位任务中，给定查询地面全景图像 I_q^g，其中 q 是索引，我们要寻找最佳匹配的参考航拍图像 I_b^a，其中 $b \in 1, ..., N$。

为了进行可行的地面全景图像和航拍图像之间的比较，我们寻求图像的有区别的潜在表示 f^g 和 f^a。这些表示预计能够捕捉到视角变化以及丰富的低层次细节，如文本模式。然后，图像检索任务可以明确表示为：

$$b = \underset{i \in \{1,...,N\}}{arg\ mind}(f_q^g, f_i^a) \tag{1}$$

其中 $d(\cdot, \cdot)$ 表示 L2 距离。为了方便表示，我们使用上标 v 来表示同时适用于地面 (g) 和航拍 (a) 视角的情况。

3.2 几何布局调制表示

为了生成高质量的跨视角地理定位潜在表示，zhang 和 li 等人 [1] 强调了视觉特征的空间配置和低层次特征。空间配置不仅反映了图像中视觉特征的位置，还反映了它们之间的全局上下文信息。人们可以期望这样的几何信息在视角变化过程中保持稳定。同时，颜色和纹理等低层次特征有助于在不同视角中识别视觉特征。

具体而言，zhang 和 li 等人 [1] 提出了以下潜在表示的分解形式：

$$f^v = \mathbf{p}^v \circ \mathbf{r}^v \tag{2}$$

其中，$\mathbf{p}^v = \{p_m^v\}_{m=1,...,K}$ 是聚合了视觉特征的空间配置的 K 个几何布局描述符的集合，$\mathbf{r}^v = \{r_j^v\}_{j=1,...,C}$ 表示由任何主干编码器生成的 C 个通道的原始潜在表示。p_m^v 和 r_j^v 都是 $\mathbb{R}^{H \times W}$ 中的向量，其中 H 和 W 分别是原始潜在表示的高度和宽度。调制操作 $\mathbf{p}^v \circ \mathbf{r}^v$ 可以展开为：

$$(< p_1^v, r_1^v >, ..., < p_1^v, r_C^v >, ..., < p_K^v, r_1^v >, ..., < p_K^v, r_C^v >) \tag{3}$$

其中 $< p_m^v, r_j^v >$ 表示 p_m^v 和 r_j^v 的 Frobenius 内积。在这个意义上，得到的 $f^v \in \mathbb{R}^{CK}$ 被称为几何布局调制表示，并将被输入到公式 (1) 中，以检索最佳匹配的空中图像。

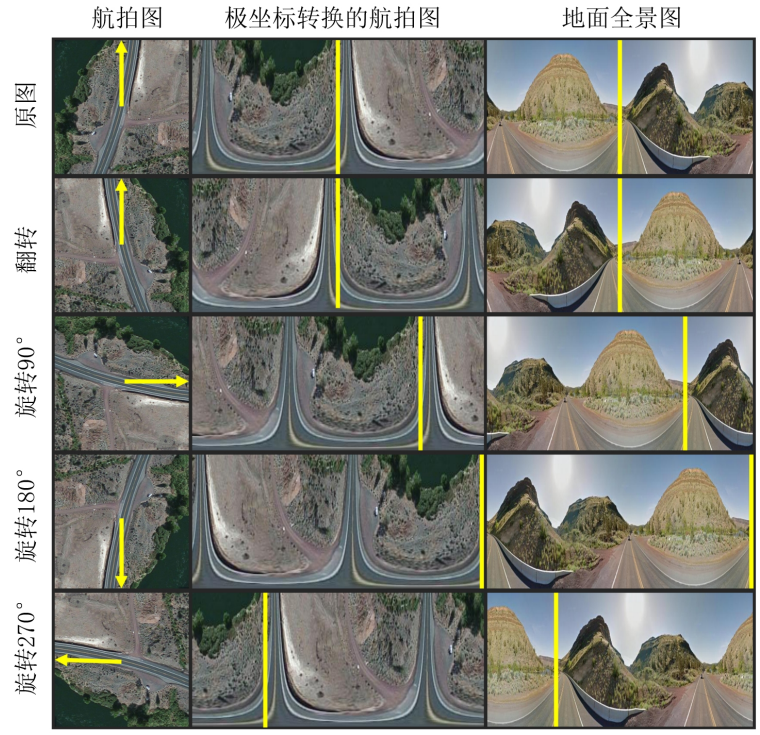

图 2. 布局模拟的示意图。从左到右依次是航拍图像、极坐标变换后的航拍图像和地面全景图像。黄色箭头和线表示北方方向。

3.3 模型概览

GeoDTR(见图1(a)) 是一个包括地面视角和空中视角两个分支的孪生神经网络。在每个分支内部，有两个不同的处理路径，即主干特征路径和几何布局路径。

在主干特征路径中，一个 CNN 主干编码器处理输入图像，生成原始潜在表示 $\mathbf{r}^v$，其中 $v = g$ 或者 $v = a$。由于 CNN 主干的特性，这些表示携带了位置信息和低层次特征信息。

几何布局路径致力于探索视觉特征之间的全局上下文信息。该路径包括一个核心子模块，即几何布局提取器，它基于原始潜在表示 $\mathbf{r}^v$ 生成一组几何布局描述符 $\mathbf{p}^v$。这些描述符将调制 $\mathbf{r}^v$，将其中的几何布局信息整合进来。通过对几何布局的独立处理，可以避免引入来自不同视觉特征的低层次特征之间的不必要的相关性。接下来，我将详细描述 GeoDTR 的核心组成部分。

3.4 布局模拟和语义强化

zhang 和 li 等人 [1] 精心设计了两类增强方法，即布局模拟和语义增强 (LS)，旨在提高提取的布局描述符的质量，以及改善跨视角地理定位模型的泛化能力。

3.4.1 布局模拟

布局模拟采用随机翻转和随机旋转 (90°、180° 或 270°) 的组合方式，同时应用于地面真实航拍图像和地面全景图像。通过这种方式，保留了低级细节，但几何布局被修改了。如图2所示，布局模拟可以生成具有不同几何布局的匹配的地空对图像对。

3.4.2 语义强化

语义增强则是分别对航拍图像和地面全景图像中的低级特征进行随机修改。zhang 和 li 等人 [1] 采用颜色抖动来改变图像的亮度、对比度和饱和度。此外，他们还随机应用高斯模糊，并将图像随机转换为灰度图像或分色调图像。

需要注意的是，与先前的数据增强方法不同，布局模拟并不会破坏两个视图中视觉特征之间的几何对应关系。实验证明，布局模拟在跨区域性能上极大地提升了 GeoDTR 的表现，同时几乎不削弱同区域的性能。将布局模拟应用于现有方法也提升了它们在跨区域实验中的性能。

3.5 几何布局提取器

该子模块用于挖掘视觉特征之间的全局上下文信息，并生成有效的几何布局描述符。尽管视角发生变化，视觉特征对齐基本保持不变。因此，将几何布局信息整合到潜在表示 f^v 中将提高其在跨视角地理定位中的区分能力。需要注意的是，几何布局是一种全局属性，它捕捉了地面全景图像和航拍图像中不同位置的多个视觉特征的空间配置。例如，单个视觉特征可以跨越整个图像，比如道路。因此，几何布局描述符应该能够掌握视觉特征之间的全局相关性。

为了实现这个目标，几何布局提取器是在标准 Transformer 的基础上构建的。如图1b 所示，网络由一个沿通道的最大池化层、一个 Transformer 模块和两个嵌入层组成，这两个嵌入层分别位于 Transformer 之前和之后。最大池化层产生一个显著性图 $M \in \mathbb{R}^{H \times W}$。然后，$M$ 通过第一个嵌入层进行处理，投影到 K 个嵌入向量 $E = [e_1, e_2, ..., e_k]$，其中每个嵌入向量的维度为 $(H \times W)/2$。在标准的可学习位置嵌入 (Positional Embedding,PE)E_{pe}(Dosovitskiy et al. 2021 [24]) 的基础上，zhang 和 li 等人 [1] 引入了一个额外的索引感知位置嵌入添加到 E 中：

$$\hat{E} = E + \text{HardTanh}(E_{pe} + W_{LN} M^{idx}) \tag{4}$$

其中 $M^{idx} mj = \frac{1}{C} \arg\max c \in 1, ..., C r^c mj$，$WLN$ 是一个可学习的线性变换，将 M^{idx} 映射到 K 个不同的子空间，$W_{LN} \in \mathbb{R}^{(H \times W) \times (K \times \frac{H \times W}{2})}$。Transformer 模块探索 $\hat{E}$ 之间的相关性。在经过第二个嵌入层的投影后，几何布局提取器生成一组 K 个几何布局描述符 $\mathbf{p}$。

3.6 反常学习策略

由于缺乏地面真值的几何布局描述符，子模块 $G^v(\cdot)$ 在训练过程中只能接收间接且不充分的监督信号。受 (Rao et al. 2021) 的启发，zhang 和 li 等人 [1] 提出了一种基于对立的 (CF-based) 学习过程。具体而言，他们在公式 (2) 中应用了一种干预操作 $do(\mathbf{p}^v = \hat{\mathbf{p}}^v)$，将 $\mathbf{p}^v$ 替换为一组虚构的布局描述符 $\hat{\mathbf{p}}^v$。这样就得到了一个虚构的表示 $\hat{f}^v$。$\hat{\mathbf{p}}^v$ 的元素来自均匀分布 $U[-1, 1]$。为了惩罚 $\hat{\mathbf{p}}^v$，并鼓励 $\mathbf{p}^v$ 捕捉更多独特的几何线索，他们通过最小化提出的对立损失来最大化 f^v 和 $\hat{f}^v$ 之间的距离：

$$L_{cf}^v = \log(1 + e^{-v[d(f^v, \hat{f}^v)]}) \tag{5}$$

其中 ^{v} 是一个参数，用于调节收敛速度。对立损失通过惩罚虚构描述符 $\hat{p}$，为几何布局描述符 p 提供了一种弱监督信号。通过这种方式，模型可以远离明显的“错误”解，并学习

到更好的潜在特征表示。除了对立损失，我们还采用了加权软边缘三元组损失，将匹配的对更接近，不匹配的对彼此远离：

$$L_{triplet} = \log(1 + e^{[d(f_m^g, f_m^a) - d(f_m^g, f_n^a)]}) \tag{6}$$

其中 是一个超参数，控制训练的收敛性。m, n $1, 2, ..., N$ 并且 $m \neq n$。最终的损失为：

$$L = L_{triplet} + L_{cf}^a + L_{cf}^g \tag{7}$$

4 复现细节

4.1 与已有开源代码对比

zhang 和 li 等人 [1] 在他们实验室的代码库中提供了相关工作的代码，由于他们没有详细说明他们文章中使用的骨干网络的设置，我的复现工作参考了提供的模型文件，节省推测他们骨干网络相关参数的时间。

此外，我使用了 zhang 和 li 等人 [1] 建议的 MD5 代码的方式清洗了 CVUSA 数据集，以删除一些重复的数据。

我还根据他们的工作思考并设计了 LS 模块，旨在获得具有更好泛化能力的模型。具体的改进将在“主要贡献”(第 4.6 节) 中详细描述。

4.2 数据集

为了评估 GeoDTR 的有效性，zhang 和 li 等人 [1] 在 CVUSA(Workman, Souvenir, and Jacobs 2015 [10]) 和 CVACT(Liu and Li 2019 [13]) 两个数据集上进行了大量实验，因此我也在这两个数据集上进行了实验。CVUSA 和 CVACT 都包含 35,532 对训练样本。CVUSA 提供了 8,884 对测试样本，CVACT 的验证集 (CVACT val) 中也有相同数量的样本对。此外，CVACT 还提供了一个具有挑战性的大规模测试集 (CVACT test)，其中包含 92,802 对样本。在 CVUSA 中，zhang 和 li 等人 [1] 在原始的训练集和测试集中分别找到了 762 对和 43 对重复的样本对。他们从训练集中移除了这些重复的样本对，但保持测试集不变以进行公平比较。

4.3 评估指标

与现有方法 (Shi 等人，2019 [5]；Toker 等人，2021 [11]；Yang，Lu 和 Zhu，2021 [7]；Hu 等人，2018 [8]；Liu 和 Li，2019 [13]；Shi 等人，2020a [6]) 类似，zhang 和 li 等人 [1] 选择使用前 K 个预测的召回准确率 (R@K) 作为评估指标。R@K 衡量了给定查询图像，在前 K 个预测中排名靠前的真实航拍图像的概率。在接下来的实验中，我将评估所有方法在 R@1、R@5、R@10 和 R@1% 上的性能。

4.4 实验环境搭建

我在 Ubuntu 18.4 上设置了实验环境，受硬件资源限制，最高版本的 CUDA 是 10.1，因此使用的 pytorch 版本是 1.8.1，torchvision 版本是 0.9.1，配置命令可以在 pytorch 官方网站上找到。这个配置与 zhang 和 li 等人 [1] 要求的 pytorch 版本高于 1.11 和 torch 版本高于 0.11

不匹配，这可能是后续实验结果不匹配的原因之一，所以如果您有足够的资源，请按照 zhang 和 li 等人 [1] 的要求进行配置。还有一些将被使用的库，如 tqdm、scipy、PIL 和 OpenCV，可以直接使用 pip 命令安装。

4.5 实验细节

我使用 PyTorch(Paszke 等人，2019) 实现了模型。地面全景图像和航拍图像分别调整为 122×671 和 256×256 的大小。与先前的方法 (Shi 等人，2019 [5], 2020a [6]；Yang，Lu 和 Zhu，2021 [7]；Toker 等人，2021 [11]；Hu 等人，2018 [8]) 类似，我将批大小设置为 32，并将 设置为 10 作为软边界三元组损失的超参数。在每个批次中，使用穷举式的小批量策略 (Vo 和 Hays，2016 [9]) 构建三元组对。我使用 AdamW(Loshchilov 和 Hutter，2017 [31]) 在四个 GTX 1080 GPU 上进行了 200 个 epoch 的训练，权重衰减为 0.03。学习率选择为 10^{-4}，并采用余弦学习率调度。反事实损失中的 设为 5。骨干网络使用 ResNet34(He 等人，2016 [32])，在 ImageNet(Deng 等人，2009 [33]) 上进行了预训练。我采用了一个具有 4 个头的 2 层 Transformer 作为几何布局提取器，并随机初始化。几何描述符的数量 K 设置为 8。

4.5.1 语义增强

我使用 TorchVision 库实现了语义增强。我使用 ColorJitter 随机调整亮度、对比度和饱和度。在这个函数中，我将每个参数设置为 0.3。在训练过程中，我使用 RandomGrayscale 和 RandomPosterize 以 0.2 的概率随机应用图像灰度化和图像分级。最后，我使用从 {1, 3, 5} 中随机选择的卷积核大小和在 [0.1, 5] 范围内随机选择的 σ 对图像进行高斯模糊。

4.5.2 几何布局提取器

几何布局提取器是基于标准 Transformer 构建的。在实现中，我采用了 PyTorch 官方的 Transformer 实现。对于每个 Transformer 编码器层，我将潜在维度设置为 168，前馈层维度设置为 2048，丢弃概率设置为 0.3。我使用层归一化和 GeLU 激活函数作为每一层的标准化和激活函数。最后，在生成的几何布局描述符的输出之前，应用了 HardTanh 激活函数。

4.6 创新点

许多现有的 CVGL 方法探索数据挖掘技术以提高性能。批内挖掘方法 [18,22] 利用 HER 和 SEH 等损失函数，在单个批次中强调困难样本。全局挖掘方法 [34,35] 维护全局挖掘池，从整个训练数据集中选择困难样本构建训练批次。然而，批内挖掘方法受限于训练批次内样本多样性的缺乏，而全局挖掘策略则需要额外的内存和计算资源来维护挖掘池。

在这项工作中，zhang 和 li 等人 [1] 提出了一个几何布局提取器 (Geometric Layout Extractor，GLE) 模块来捕捉空间配置，以及布局模拟和语义增强 (LS) 技术来增加训练数据的多样性。此外，还引入了反事实 (CF) 学习模式，为几何布局提取器模块提供额外的监督。然而，所提出的 GLE 模块仅在学习的子空间内探索几何相关性，这可能导致信息的丢失。此外，所提出的 LS 技术在增强跨区域性能方面的有效性仍然有限，因为在这项工作中它们只是以数据增强的方式使用。

为了解决上述限制，我在数据增强方面进行了扩展。在这项工作中，我采用 LS 技术作为特殊的数据增强，隐式地为布局和语义特征引入了批内对比信号。受交叉视图地理定位中的困难样本挖掘策略的启发 [34]，我将 LS 技术扩展到显式地包含批内对比信号。通过这种方式，期望模型能够区分单个批次中这些“生成的难样本“。我将这个过程称为“难样本增强”过程。

与先前的数据增强方法不同，新的 LS 方法不会破坏两个视图中视觉特征的几何对应关系。因此我在训练的时候对每一个图像对进行布局模拟和语义强化。对每一个输入的图像对进行 2 次随机的 LS 操作，因此生成的航拍图像和地面全景图想在视觉特征上相互对应着相同的原始视觉特征，互为难样本。这样可以有效地促进模型学习到更多图像的布局信息，增强其泛化能力与健壮性。

5 实验结果分析

5.1 复现结果

在相同数据集上进行训练和测试的实验结果分别如图3和图4所示。在 CVUSA 数据集中无极坐标变换的实验结果中可以看到，训练 200epochs 的实验结果时比训练 300 个 epochs 的实验结果要低近 1 个百分点，也就说训练的轮数越多学习到的几何布局信息提取能力就越强。但是也在某一种程度上说明了模型学习率和学习率衰减时存在了问题，耗费大量的资源只能得到很小的回报，这是不正常的。

Method	R@1	R@5	R@10	R@1%	Size
Author without polar@200	93.76%	98.47%	99.22%	99.85%	
Ours without polar@200	90.62%$^{-3.14}$	97.49%$^{-0.98}$	98.61%$^{-0.61}$	99.79%$^{-0.06}$	47.52MB
Ours without polar@300	91.56%$^{+0.94}$	97.72%$^{+0.23}$	98.60%$^{-0.01}$	99.77%$^{-0.02}$	47.52MB
Author with polar@200	95.43%	98.86%	99.34%	99.86%	
Ours with polar@200	94.51%$^{-0.92}$	98.58%$^{-0.28}$	99.23%$^{-0.11}$	99.82%$^{-0.04}$	48.51MB

图 3. 在 CVUSA 数据集上训练和验证 200peochs 有极坐标变换和无极坐标变换以及 300epochs 有极坐标变换

Method	R@1	R@5	R@10	R@1%	Size
Author without polar@200	85.43%	94.81%	96.11%	98.26%	
Ours without polar@200	83.1%$^{-2.33}$	93.63%$^{-1.18}$	95.15%$^{-0.94}$	98.50%$^{+0.24}$	47.52MB
Author with polar@200	86.21%	95.44%	96.72%	98.77%	
Ours with polar@300	85.96%$^{-0.25}$	95.23%$^{-0.21}$	96.57%$^{-0.15}$	98.71%$^{-0.06}$	48.51MB

图 4. 在 CVACT 数据集上训练和验证 200peochs 无极坐标变换和 300epochs 有极坐标变换

在不同数据集上训练测试实验结果分别如图5和图6所示。根据在 CVUAS 数据集上训练验证得到的结果，在跨区域测试基准实验中，无极坐标变换的 epoch 数改为 300，可以看见

在 CVUSA 上训练并在 CVACT 上测试的结果与原作者论文中的结果要优于在 CVACT 数据集上训练并在 CVUSA 上测试的结果。我猜测是由于 CVACT 数据集是在一个城市的密集采集，因此数据之间的相关性会高于 CVUSA 数据集，因此造成在 CVACT 上训练的时候可能存在过拟合的情况。当然 CVUAS 的数据集范围要比 CVACT 数据集的范围更广，这也是在不同的跨区域基准数据集上测试结果的误差。

Method	R@1	R@5	R@10	R@1%	Size
Author without polar@200	43.72%	66.99%	74.61%	91.83%	
Ours without polar@200	38.20%$^{-5.52}$	61.48%$^{-5.51}$	69.77%$^{-4.84}$	89.80%$^{-2.03}$	47.52MB
Ours without polar@300	39.12%$^{+0.92}$	62.53%$^{+0.95}$	70.59%$^{+0.82}$	90.13%$^{+0.33}$	47.52MB
Author with polar@200	53.14%	75.62%	81.90%	93.80%	
Ours with polar@200	47.80%$^{-5.34}$	70.84%$^{-4.78}$	78.38%$^{-3.52}$	93.15%$^{-0.65}$	48.51MB

图 5. 在 CVUSA 上训练，200epochs 无极坐标变换和 300epochs 有极坐标变换，并在 CVACT 上验证

Method	R@1	R@5	R@10	R@1%	Size
Author without polar@200	29.85%	49.25%	57.11%	82.47%	
Ours without polar@200	22.79%$^{-7.06}$	39.89%$^{-9.36}$	48.12%$^{-8.99}$	74.64%$^{-7.83}$	47.52MB
Author with polar@200	44.07%	64.66%	72.08%	90.09%	
Ours with polar@300	40.86%$^{-3.21}$	60.85%$^{-3.81}$	68.44%$^{-3.64}$	87.17%$^{-2.94}$	48.51MB

图 6. 在 CVACT 上训练，200epoch 有极坐标变换和无极坐标变换以及 300epoc 有极坐标变换，并在 CVUSA 上验证

6 总结与展望

GeoDTR 将几何布局从原始输入特征中分离出来，更好地探索了视觉特征之间的空间相关性。并且还引入了布局模拟和语义增强技术，提高了 GeoDTR 和其他现有的跨视图地理定位模型的泛化能力。此外，还引入了一种新的基于反事实的学习方法来训练 GeoDTR。大量的实验证明了 GeoDTR 在标准、细粒度和跨区域交叉视图地理定位任务上的优势。但是目前对几何布局描述符的解释在 GeoDTR 中还没有得到充分的探索。在未来的研究中，考虑到所提出的 GeoDTR 的泛化能力，研究 GeoDTR 在不同环境下，如弱光场景或不同季节下的性能，也将继续研究它们的性质，并朝着更易于解释的模型努力。

参考文献

[1] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3480–3488, 2023.

[2] Dong-Ki Kim and Matthew R Walter. Satellite image-based localization via learned embeddings. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2073–2080. IEEE, 2017.

[3] Akshay Shetty and Grace Xingxin Gao. Uav pose estimation using cross-view geolocalization with satellite imagery. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1827–1833. IEEE, 2019.

[4] Han-Pang Chiu, Varun Murali, Ryan Villamil, G Drew Kessler, Supun Samarasekera, and Rakesh Kumar. Augmented reality driving using semantic geo-registration. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 423–430. IEEE, 2018.

[5] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020.

[7] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34:29009–29020, 2021.

[8] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018.

[9] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 494–509. Springer, 2016.

[10] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015.

[11] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021.

[12] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11990–11997, 2020.

[13] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633, 2019.

[14] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):867–879, 2021.

[15] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 1395–1403, 2020.

[16] Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Visual and object geo-localization: A comprehensive survey. *arXiv preprint arXiv:2112.15202*, 2021.

[17] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5007–5015, 2015.

[18] Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8391–8400, 2019.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[20] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022.

[21] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013.

[22] Yulan Guo, Michael Choi, Kunhong Li, Farid Boussaid, and Mohammed Bennamoun. Soft exemplar highlighting for cross-view image-based geo-localization. *IEEE transactions on image processing*, 31:2094–2105, 2022.

[23] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 470–479, 2019.

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly,

et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[25] Royston Rodrigues and Masahiro Tani. Global assists local: Effective aerial representations for field of view constrained image geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3871–3879, 2022.

[26] Judea Pearl. Causal inference in statistics: An overview. 2009.

[27] Ruth MJ Byrne. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJCAI*, pages 6276–6282, 2019.

[28] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10044–10054, 2020.

[29] Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics. *arXiv preprint arXiv:1909.12000*, 2019.

[30] Yue Wang, Yao Wan, Chenwei Zhang, Lu Bai, Lixin Cui, and Philip Yu. Competitive multi-agent deep reinforcement learning with counterfactual thinking. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1366–1371. IEEE, 2019.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[34] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021.

[35] Sijie Zhu, Taojiannan Yang, and Chen Chen. Revisiting street-to-aerial view image geo-localization and orientation estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 756–765, 2021.