

复现论文：Transformer Is All You Need for Cross-view Image Geo-localization

摘要

本文研究了在跨视角情况下图像的检索问题，复现了在 2022 年 CVPR 发表的论文 [19]，并在此基础上提出了自己的改进。具体来说，给定一张由地面水平视角拍摄的街景图像，需要在数据库已有的一系列从空中垂直视角拍摄的航拍图像中找出与这张街景图像对应的航拍图像。由于剧烈的视角变化，导致街景图像与航拍图像之间存在巨大的语义鸿沟，使得跨视角图像检索任务极具挑战性。本文复现了 [19] 的研究工作，并在此基础上提出在图像输入阶段将航拍图像进行极坐标变换后再送入模型进行第一阶段训练；紧接着在模型训练的第二阶段对极坐标变换后的航拍图像使用 [19] 提出的非均匀裁剪策略。此外，不同于以往研究，本文提出使用基于对比学习的 InfoNCE 损失而不是基于度量学习的三元组损失或者三元组的各种变体损失来训练模型。在主流的 CVUSA 数据集的实验结果表明，本文所提出的极坐标变换策略和 InfoNCE 损失函数是有所成效的。

关键词：图像检索；极坐标变换；对比学习

1 引言

跨视角图像检索指的是给定一张由地面水平视角拍摄的街景图像，需要在数据库已有的一系列从空中垂直视角拍摄的航拍图像中找出与这张街景图像对应的航拍图像，这是一个具有挑战性且意义重大的任务。图 1 形象地展示了街景图像和航拍图像这两种跨视角图像的产生过程。之所以说具有挑战性，是因为两种交叉视角的图像是正交的，可共享的信息很少；同一个场景在不同视角下的几何外观和空间布局具有非常大的差异；跨视角图像的采集时间往往是不同的，这也会带来同一个场景在不同视角的图像中具有明暗、亮度等方面差异。总体来说，这种挑战性源于跨视角图像中存在的巨大的语义鸿沟。但是，这种极具挑战的任务有着十分重要的研究意义。航拍图像通常由卫星拍摄而来，这种卫星拍摄的图像附有 GPS 坐标信息。一旦检索出街景图像所对应的航拍图像，就可以根据航拍图像定位到街景图像的空间地理位置。因此，跨视角图像检索任务在自动驾驶、无人机导航、机器导航、3D 重建、地理位置标记、辅助导航等方面具有重要的应用价值。

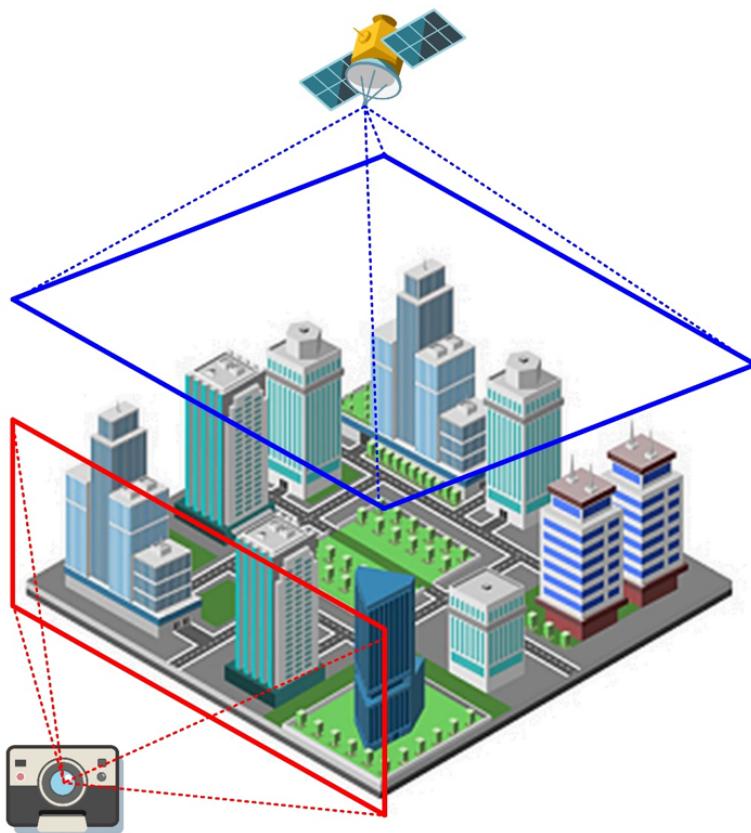


图 1. 跨视角图像的产生过程

近来，随着深度学习的不断发展，在跨视角图像检索方面的研究成果层出不穷。主流的跨视角图像检索模型主要分为两大类，如图 2 所示：一是基于 CNN 的孪生/双流/双塔网络架构，例如：CVM-Net [6]，SAFA [11]，L2LTR [17] 等；二是基于 Transformer 的双分支网络架构，例如：TransGeo [19] 等。值得一提的是，Shi 等人 [11] 在 2019 年提出的极坐标变换的图像预处理方式取得了显著的实验效果，并且在许多其他模型中都有效果。受此启发，本文基于 Transformer 架构，在 [19] 研究的基础上，提出结合 [11] 的极坐标变换策略，即在航拍图像输入之前，先将航拍图像经过极坐标变换，变成和街景图像等宽等高的航拍图，再送入模型进行第一阶段训练；紧接着在第二阶段训练中利用 [19] 提出的非均匀裁剪策略对变换后的航拍图进行裁剪，以降低计算复杂度，提高训练速度。此外，在跨图像检索/地理定位的经典架构中，不管是基于 CNN 的模型架构，还是基于 CNN+Transformer 的模型架构，亦或是基于 Vision Transformer 的模型架构，几乎所有研究人员都采用三元组（及其各种变体）损失函数来训练模型，而并未考虑采用其他损失函数来训练模型。本文跳脱出三元组损失函数的范畴，注意到 InfoNCE 损失在多模态预训练中被证明是有用的 [9]，可以用来弥合模态之间的差异，于是受此启发，通过分析 InfoNCE 损失与三元组损失之间的差异，决定舍弃过往研究中所采用的基于度量学习的三元组损失函数，而是采用基于对比学习的 InfoNCE 损失函数来训练模型，取得良好的成效。本文所提出的优化模型架构如图 3 所示。

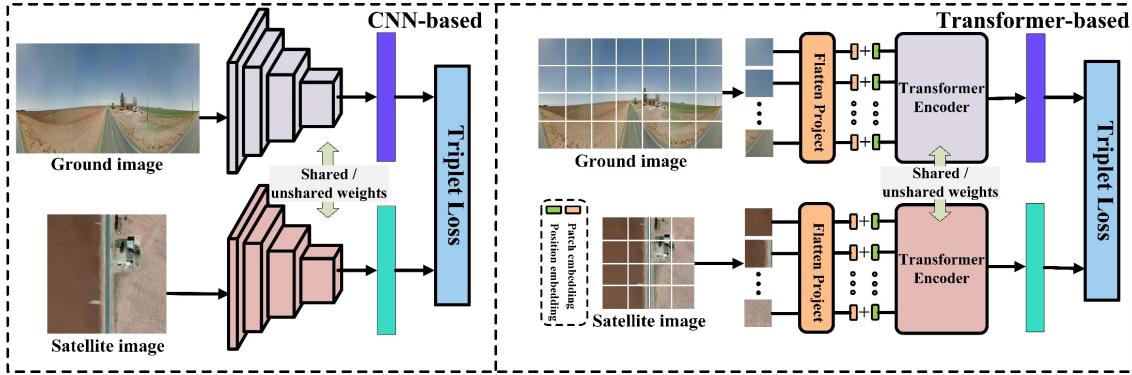


图 2. 主流的跨视角图像检索架构

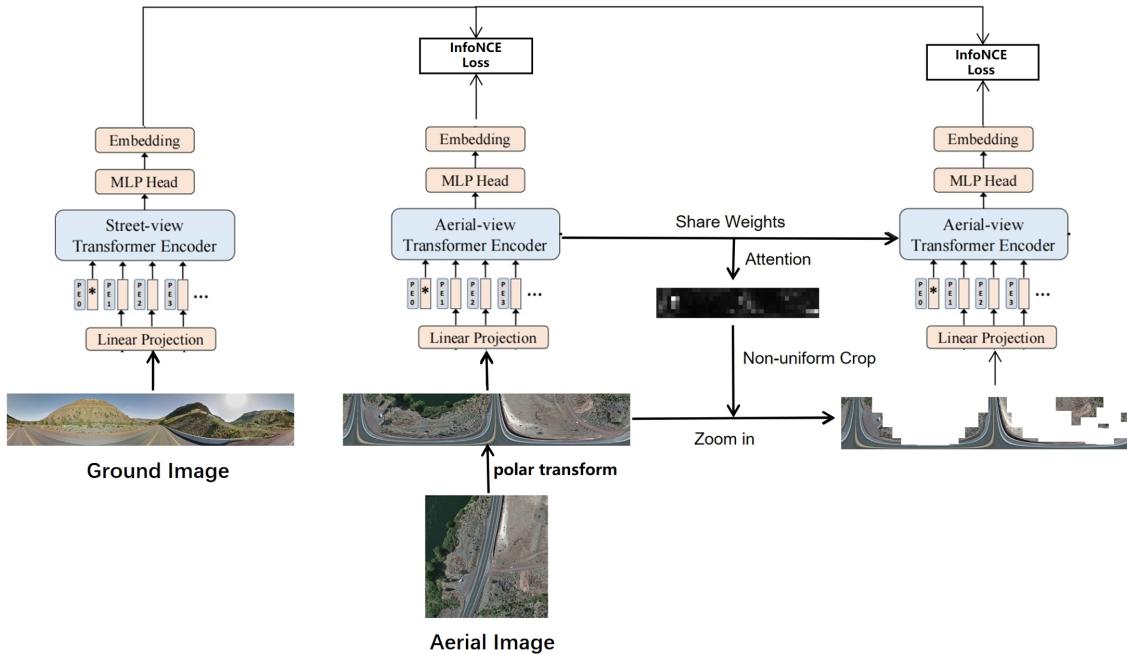


图 3. 本文的模型总览图

总的来说，本文的贡献主要在于两个方面：

- (1) 在 [19] 的研究基础之上结合极坐标变换策略，进一步提高了模型的性能，证明极坐标变换这一预处理技术在跨视角图像检索领域不同的模型架构中依然有较好的泛化表现。
- (2) 在跨图像检索/地理定位领域提出使用损失函数 InfoNCE 以取代三元组（及其各种变体）损失函数来训练模型，实验结果证明 InfoNCE 损失函数能够有效提高 [19] 的模型性能，初步证明 InfoNCE Loss 的有效性。

2 相关工作

Workman 等人 [16] 最早的一项研究表明，CNN 提取的特征远优于传统手工提取的特征。在他们的工作中，他们还提出了 CVUSA 数据集，这是当今跨视角图像检索/地理定位任务的主要基准数据集之一。随后，许多跨视角图像检索工作的研究者纷纷开始采用双分支的 CNN 架构 [1, 6, 8, 12, 15]，对两个视图分别提取特征，通过监督学习使模型学习一个嵌入空间，将街景图像与航拍图像在嵌入空间中进行相似度度量，将相似度得分最高的航拍图像作为模型

的检索结果。但是，这种方式无法对两种不同视角图像显著的外观差异进行建模，导致模型的检索性能较差。

后来，Shi 等人 [11] 基于两种视图的几何先验知识设计了一种极坐标变换方法，使变换后的航拍图像与街景图像具有相似的几何布局，极大地减小了跨视角图像之间的域差异，取得了显著的实验成果。Regmi 等人 [10] 通过添加额外的生成模块 GAN [4]，利用街景图像生成航拍图像，从一种视图转换为另一种视图来减小跨视图之间的域差异。Toker 等人 [13] 在极坐标变换的基础上，进一步训练生成网络，使生成的图像更加真实以便于模型训练。随后，L2LTR [17] 在 ResNet [5] 的基础上采用 vanilla ViT [3]，形成了一种 CNN+Transformer 的混合架构。但是，由于采用 CNN 作为特征提取器，自注意力和位置编码只在 CNN 高层特征上使用，没有充分利用第一层的全局建模能力和位置编码信息。

相比之下，TransGeo [19] 采用基于纯粹的 Transformer 模型而非 CNN+Transformer 的混合架构，充分利用了 Transformer 的全局建模能力和位置编码信息，将局部信息整合到全局嵌入向量中，有助于模型理解整个场景。在 TransGeo 中，作者使用 Transformer 架构中的注意力图执行一个额外的缩放步骤，即模拟人眼的“关注和放大”机制，可以以更高的分辨率观看较小的对象。为了提高方法的通用性，TransGeo 还使用了自适应锐度感知最小化 (ASAM) [7] 来平滑损失面，提高模型的性能。

纵观跨视角图像检索/地理定位研究工作的历程，几乎所有的模型，包括 CNN、CNN 与 Transformer 的混合架构、Vision Transformer 在内，都使用三元组损失、加权软边界三元组损失、加权软边界排序三元组损失 [6] 等损失函数来训练模型，并未跳脱出三元组损失函数的范畴。

本文在 [19] 的工作基础上，使用 [11] 提出的图像预处理技术，对航拍图像进行极坐标变换，并在训练的第二阶段沿用 [19] 提出的非均匀裁剪策略和“关注放大”机制对极坐标变换后的航拍图像进行裁剪和分辨率的调整。此外，本文不同于先前研究所采用的三元组损失函数及其一系列损失函数变体，而是引入基于对比学习的 InfoNCE 损失函数，进一步优化了 [19] 提出的模型性能。实验结果表明，极坐标变换这一预处理技术在跨视角图像检索领域不同的模型中有较好的泛化性，InfoNCE 损失函数可以作为三元组损失函数及其变体的替代，能够提高模型的性能。

3 本文方法

3.1 本文方法概述

本文提出的优化模型总览图如图 3 所示。[19] 的模型总览图如图 4 所示。可以看到，本文的模型和 [19] 的模型在结构上大体相同，因为本文的工作正是在其基础上的进一步改进和优化。

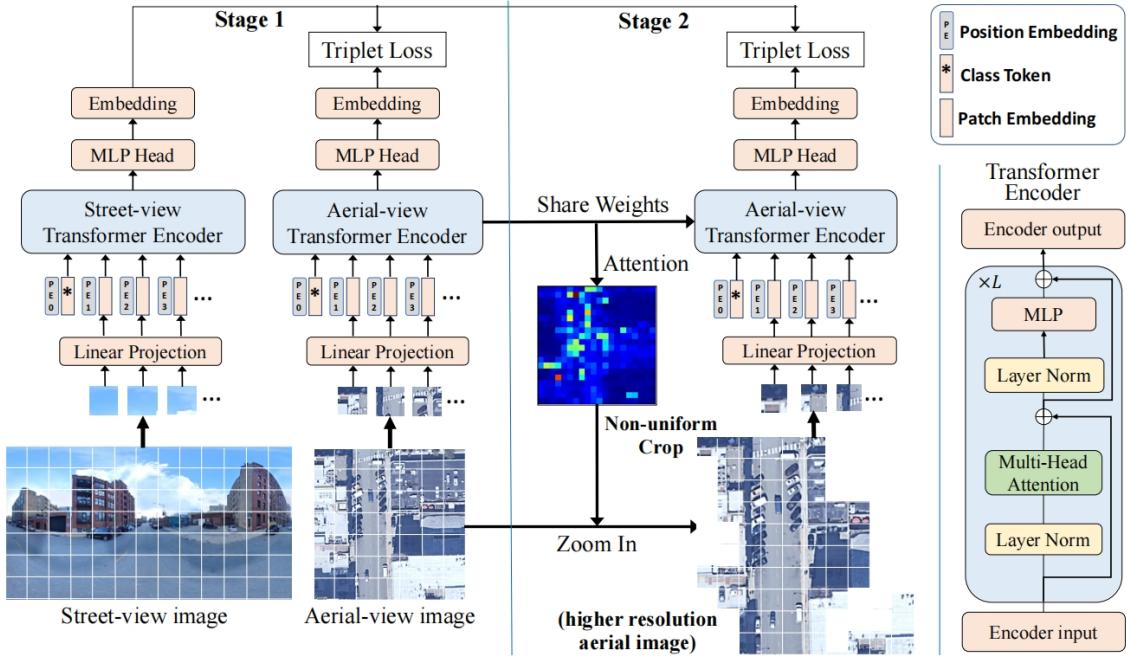


图 4. TransGeo 模型总览图, 图引自 [19]

3.2 极坐标变换

极坐标变换最初由 Shi 等人 [11] 提出, 具体的坐标变换过程由下面的公式给出。

$$\begin{aligned} x_i^s &= \frac{A_a}{2} + \frac{A_a}{2} * \frac{y_i^t}{H_g} * \sin \frac{2\pi}{W_g} \\ y_i^s &= \frac{A_a}{2} - \frac{A_a}{2} * \frac{y_i^t}{H_g} * \cos \frac{2\pi}{W_g} \end{aligned} \quad (1)$$

其中, A_a 表示原始航拍图像的边长, W_g , H_g 分别表示街景图像的宽度和高度, 表示原始航拍图像的像素点, 表示变换后航拍图像的像素点。那么, 为什么极坐标变换在跨视角图像检索中卓有成效呢? Shi 等人 [11] 指出, 航拍图像经过极坐标变换后 (如图 5 所示), 街景图像和航拍图像在几何外观和空间布局上的差异被极大地缩小, 两种跨视角图像之间的特征被很好地对齐, 这种对齐效果非常有利于模型进行训练和学习, 从而提高模型的跨视角图像检索性能。此外, Shi 等人 [11] 强调, 对航拍图像应用极坐标变换策略不可避免会带来局部图像畸变, 但是这种畸变带来的影响是可以被缓解的。Shi 等人 [11] 的做法是提出学习多个鲁棒的空间特征描述符并将多个描述符聚合, 即使用 SAFA 模块来克服极坐标变换图像畸变问题。在本文中, 同样引入极坐标变换策略, 但是并不采用 SAFA 模块来克服极坐标变换带来的图像畸变, 而是利用 Transformer 的多头注意力机制——学习图像的多个全局特征并将其聚合, 作为图像的全局特征描述符。因此, Transformer 本身就具备应用极坐标变换策略的良好特性。以上便是本文引入极坐标变换优化策略的动机——极坐标变换可以极大减小跨视角图像的域差异以及 Transformer 对极坐标变换的良好特性。

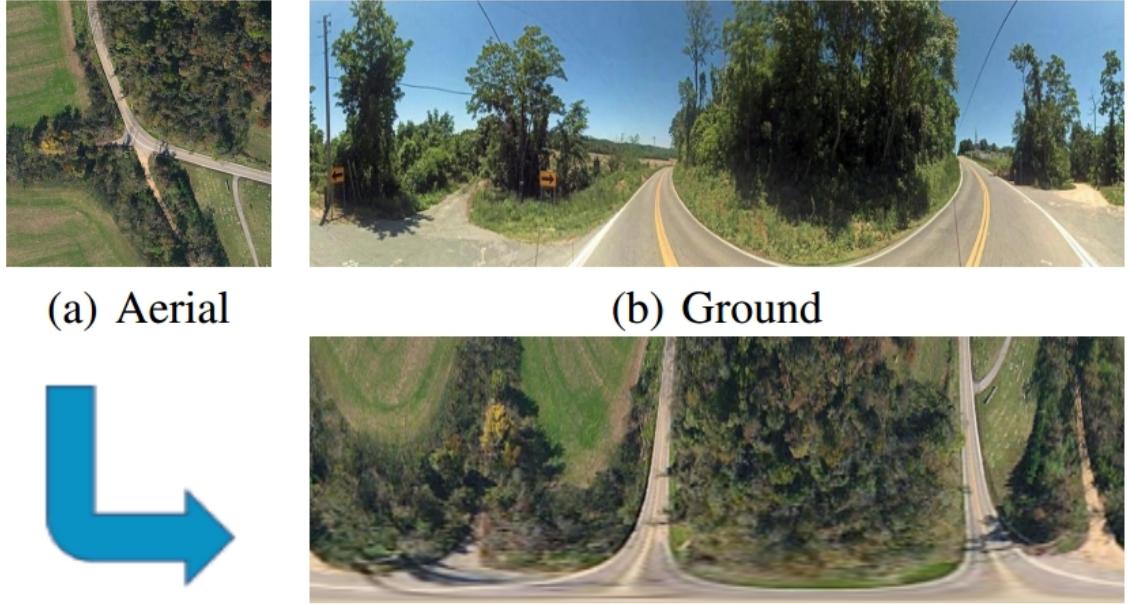


图 5. 极坐标变换后的航拍图像，图引自 [11]

3.3 InfoNCE Loss

本文使用基于对比学习的 InfoNCE 损失来训练模型。InfoNCE 损失和三元组损失分别由以下两个公式给出：

$$\mathcal{L}(q, R)_{InfoNCE} = -\log \frac{\exp(q \cdot r_+ / \tau)}{\sum_{i=0}^R \exp(q \cdot r_i / \tau)} \quad (2)$$

$$\mathcal{L}_{triplet} = -\log(1 + e^{\alpha(d_{pos} - d_{neg})}) \quad (3)$$

其中， q 表示街景图像的编码向量， r_+ 表示正样本的编码向量， R 表示负样本数量， τ 表示温度系数； d_{pos} 表示 query 和正样本之间的距离， d_{neg} 表示 query 和负样本之间的距离。三元组损失（及其各种变体）通过从训练数据中选择三元组（query、正样本和负样本）来构建训练样本，目标是使得 query 与正样本之间的距离尽可能地小，同时与负样本之间的距离尽可能地大；而对比学习 InfoNCE 损失基于样本的相似性（如余弦相似度）来计算正样本对的互信息，并通过最大化这个互信息值、同时尽可能地减小负样本对之间的相似性来训练模型。从两者的公式表达和原理表述来看，三元组（及其变体）损失每次只关注一个 query 和一个正样本以及一个 query 和一个负样本之间的距离，是一种一对一的训练方式；而 InfoNCE 损失每次关注的是一个 query 和一个正样本以及一个 query 和多个负样本之间的相似度，是一种一对多的训练方式。针对跨视角图像检索这种需要处理大规模地理空间数据的任务来说，显然 InfoNCE 一对多的训练方式更加有效。并且，InfoNCE 损失在多模态预训练中被证明是有用的 [9]，可以用来弥合模态之间的差异。本文通过如上分析 InfoNCE 损失与三元组损失之间的差异，决定引入 InfoNCE 损失函数来训练模型。

4 复现细节

4.1 与已有开源代码对比

本文在已有的开源代码基础上，主要进行以下几个方面的工作：

(1) 将已有的开源代码进行重构，以适应不同的 cuda 版本。

(2) 重新设计一套 DDP 分布式训练框架，以使用多 GPU 来训练模型。具体来说：由于本文当中分布式训练只需要使用一台机器，不必进行不同机器的不同进程通信，只需要进行同一台机器的不同进程间通信。因此无需使用 tcp 协议来完成集群中的不同节点之间的数据传输和通信，只需要使用本地会话来完成同一节点的数据传输和通信即可。相应的，DDP 分布式训练代码仅指定通信主机的 ip 地址以及任意可用的端口号即可：“os.environ[“MASTER_ADDR”] = ‘localhost’”、“os.environ[“MASTER_PORT”] = ‘12345’”。

(3) 参考开源代码，自己完整地将 ViT 模型框架代码写出来。具体来说，不同于已有的开源代码的 ViT 模型框架，本文实现的 ViT 模型框架进行了精心设计和调整以适配后续的改进策略。

(4) 添加极坐标变换策略的代码。本文的一个核心创新点是：在图像预处理阶段使用极坐标变换策略来提高模型的性能。因此本文在前三个方面调整了开源代码之后，接着添加了极坐标变换的代码。

(5) 添加基于对比学习 InfoNCE 损失函数代码。本文的另一个核心创新点是：使用基于对比学习的 InfoNCE 损失函数策略来提高模型的性能。因此本文在前三个方面调整了开源代码之后，接着添加了 InfoNCE 损失函数的代码。

4.2 实验环境搭建

本文的实验环境如下：

- Python == 3.9, PyTorch == 1.11.0, torchvision == 0.12.0

同时务必保证环境中已经安装以下几个库：

- numpy, matplotlib, pillow, ptflops, timm

4.3 创新点

本文的核心创新点有两个：

(1) 在 [19] 的研究基础之上结合极坐标变换策略，进一步提高了模型的性能，证明极坐标变换这一预处理技术在跨视角图像检索领域不同的模型架构中依然有较好的泛化表现。

(2) 在跨图像检索/地理定位领域提出使用损失函数 InfoNCE 以取代三元组（及其各种变体）损失函数来训练模型，实验结果证明 InfoNCE 损失函数能够有效提高 [19] 的模型性能，初步证明 InfoNCE Loss 的有效性。

5 实验结果分析

5.1 数据集和评价指标

本文采用主流的基准数据集 CVUSA [18] 进行实验。CVUSA 数据集采集于美国郊区，包含 35532 个用于训练的图像对和 8884 个用于测试的图像对，数据集中的街景图像全部是全景图。本文采用跨视角图像检索/地理定位研究中常用的召回率作为评价指标，并报告了本文的优化模型在 CVUSA 数据集上前 1、前 5、前 10 和前 1% 的召回率结果。

5.2 模型参数细节

本文的方法使用 pytorch 来实现。对于 CVUSA 数据集，全景图像被调整为 112×616 大小，航拍图像在预处理阶段经过极坐标变换后也被调整为 112×616 大小，然后以 32 大小的 batch-size 输入本文的模型。本文使用 12 个 Transformer 编码器，每个多头注意模块有 6 个头。模型使用在 ImageNet-1K [2] 上现成的预训练权值 [14] 进行初始化。InfoNCE 损失函数计算图像相似度分数使用的缩放因子使用在 ImageNet-22K 上的预训练模型中设置的缩放因子 14.286。本文模型最终输出特征的维数为 1000。

5.3 原论文复现结果

表 1 报告了本文对原论文 [19] 的复现情况。蓝色加粗字体表示当前指标下模型性能更优秀的一方。从实验结果可见，本文复现的实验结果稍差于原论文 [19] 的实验结果，这可能是由于超参数设置或者硬件性能等因素导致的差异。本文将在下面几个小节中接着报告加入本文提出的优化策略（创新点）后，模型性能的提升情况。

表 1. 使用极坐标变换策略在 CVUSA 数据集测试改进效果（不同机器）

Method	R@1	R@5	R@10	R@1%
TransGeo	94.08	98.36	99.04	99.77
ours	92.74	98.06	98.77	99.80

5.4 极坐标变换优化策略（消融实验）

本文首先单独使用极坐标变换对模型进行优化。具体来说，本文在第一阶段将输入的航拍图变换为和全景图像等宽等高的图像。图 6 展示了实验中几组航拍图像的变换效果。

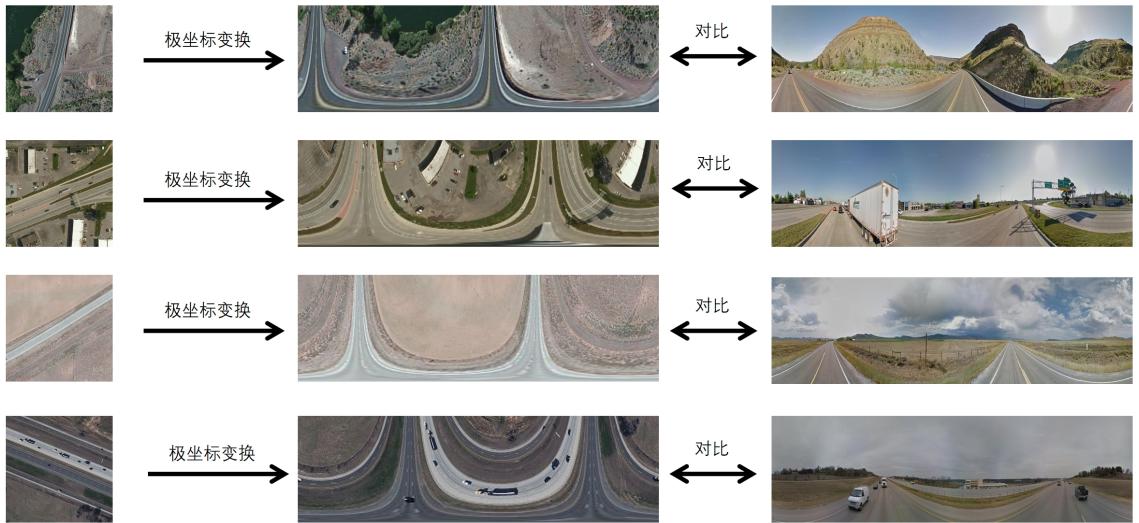


图 6. 原始航拍图（左）经过极坐标变换（中）与全景图（右）的对比效果

从图 6 中可以看出，航拍图像经过极坐标变换之后的图像与全景图像在几何外观和空间布局上面已经有了明显的对齐，同时变换后的图像不可避免有轻微的图像畸变现象。下面将模型训练过程中使用极坐标变换和未使用极坐标变换分别关注到的语义信息进行灰度可视化展示，如图 7 所示。

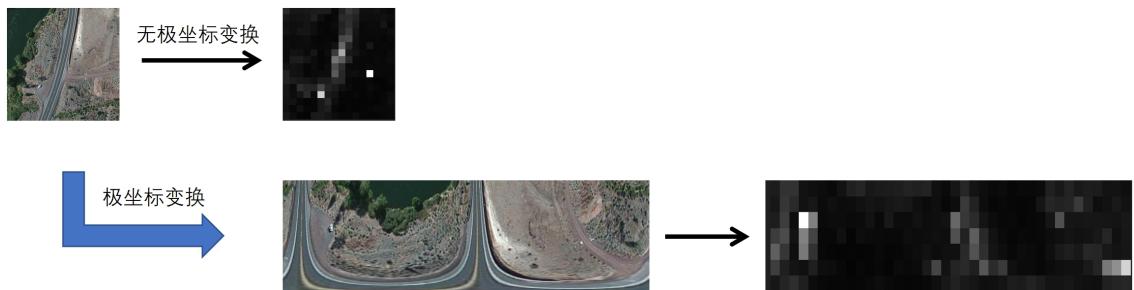


图 7. 无极坐标变换（右上）和有极坐标变换（右下）模型关注的语义信息对比

从图 7 中可以看出，无论使用还是不使用极坐标变换，模型都主要关注道路信息。但是不同的是，相比于不使用极坐标变换的情况，模型在对航拍图进行极坐标变换之后关注的信息明显更加丰富——除了道路信息以外，模型还会关注一些道路两侧蕴含丰富特征信息的区域。

本文在 CVUSA 数据集上单独测试极坐标变换的实验效果。表 2 和表 3 分别报告了使用极坐标变换后，模型的性能与原论文 [19] 模型 TransGeo 在不同机器（原论文实验结果）和相同机器（本文对原论文的复现结果）的对比结果。蓝色加粗字体表示当前指标下模型性能更优秀的一方。

表 2. 使用极坐标变换策略在 CVUSA 数据集测试改进效果（不同机器）

Method	R@1	R@5	R@10	R@1%
TransGeo	94.08	98.36	99.04	99.77
ours	93.53	98.28	99.02	99.72

表 3. 使用极坐标变换策略在 CVUSA 数据集测试改进效果（相同机器）

Method	R@1	R@5	R@10	R@1%
TransGeo	92.74	98.06	98.77	99.80
ours	93.53	98.28	99.02	99.72

从表 2 的实验结果可以看出，引入极坐标变换优化策略后，模型的性能几乎和原论文 [19] 模型性能相当。从表 3 的实验结果可以看出，在消除了超参数和硬件差异等因素的影响之后，模型性能有了提升，说明在相同的硬件条件下，本文提出的结合极坐标变换优化策略能够小幅度提高模型的性能，证明 [11] 提出的极坐标变换方法在如今跨视角图像检索/地理定位研究领域不同的模型中依然有较好的泛化性。

5.5 InfoNCE 损失函数优化策略（消融实验）

本文紧接着单独使用 InfoNCE 损失函数训练模型，测试 InfoNCE 损失函数对模型的优化效果。表 4 和表 5 分别报告了在不同机器和相同机器上 [19] 的模型 TransGeo 和本文使用 InfoNCE 损失函数优化后的模型性能对比。蓝色加粗字体表示当前指标下模型性能更优秀的一方。

表 4. 使用 InfoNCE 损失函数在 CVUSA 数据集测试改进效果（不同机器）

Method	R@1	R@5	R@10	R@1%
TransGeo	94.08	98.36	99.04	99.77
ours	94.36	98.29	98.96	99.57

表 5. 使用 InfoNCE 损失函数在 CVUSA 数据集测试改进效果（相同机器）

Method	R@1	R@5	R@10	R@1%
TransGeo	92.74	98.06	98.77	99.80
ours	94.36	98.29	98.96	99.57

从表 4 和表 5 可见，不管是在相同机器还是不同机器上，模型使用 InfoNCE 损失函数进行训练后，性能在 R@1 指标上都有了提升，初步证明 InfoNCE 损失函数可以作为三元组（及其各种变体）损失函数的替代给模型带来更好的性能。

5.6 综合优化策略（主实验）

本文最后综合使用两种优化策略对模型进行进一步优化。表 6 和表 7 分别报告了在不同机器和相同机器上 [19] 的模型 TransGeo 和本文使用 InfoNCE 损失函数优化后的模型性能对比。蓝色加粗字体表示当前指标下模型性能更优秀的一方。

表 6. 使用两种优化策略在 CVUSA 数据集测试改进效果（不同机器）

Method	R@1	R@5	R@10	R@1%
TransGeo	94.08	98.36	99.04	99.77
ours	94.62	98.28	98.84	99.49

表 7. 使用两种优化策略在 CVUSA 数据集测试改进效果（相同机器）

Method	R@1	R@5	R@10	R@1%
TransGeo	92.74	98.06	98.77	99.80
ours	94.62	98.28	98.84	99.49

从表 6 和表 7 可见，不管是在相同机器还是不同机器上，模型综合使用两种优化进行训练后，性能在 R@1 指标上有了进一步提升。

6 总结与展望

本文在 [19] 的研究基础上，提出结合 [11] 的极坐标变换方法对图像进行预处理，并舍弃以往跨视角图像检索/地理定位研究中主流的三元组（及其各种变体）损失函数，而是采用 InfoNCE 损失函数来训练模型。本文在 CVUSA 数据集上的实验结果说明本文优化后的模型比 [19] 提出的模型性能更佳，证明本文提出的这两种优化策略的有效性。本文初步证明：(1) 极坐标变换在如今跨视角图像检索/地理定位研究的不同模型中依然有较好的泛化性；(2) 跨视角图像检索/地理定位研究可以考虑采用 InfoNCE 损失函数来提高模型的性能。

参考文献

- [1] Sudong Cai, Yulan Guo, Salman H. Khan, Jiwei Hu, and Gongjian Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8390–8399. IEEE, 2019.
- [2] 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 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.
- [3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

- [4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.
- [6] Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7258–7267. Computer Vision Foundation / IEEE Computer Society, 2018.
- [7] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5905–5914. PMLR, 2021.
- [8] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5624–5633. Computer Vision Foundation / IEEE, 2019.
- [9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.
- [10] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 470–479. IEEE, 2019.
- [11] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information*

Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 10090–10100, 2019.

- [12] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1998–2006. IEEE Computer Society, 2017.
- [13] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 6488–6497. Computer Vision Foundation / IEEE, 2021.
- [14] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 2021.
- [15] Nam N. Vo and James Hays. Localizing and orienting street views using overhead imagery. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 494–509. Springer, 2016.
- [16] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocation with aerial reference imagery. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3961–3969. IEEE Computer Society, 2015.
- [17] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. In Marc’Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 29009–29020, 2021.
- [18] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4132–4140. IEEE Computer Society, 2017.
- [19] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1152–1161, 2022.