DSFormer: 一种用于视觉位置识别的双尺度交叉学习 Transformer

Haiyang Jiang[1], Songhao Piao[1*], Chao Gao[2*], Lei Yu[3], Liguo Chen[4],

***Abstract*—视觉地点识别（VPR）对于鲁棒的移动机器人定位至关重要，但在不同环境条件和视角下保持可靠性能方面面临重大挑战。为了解决这个问题，我们提出了一种新颖的框架，该框架集成了双尺度变换器（DSFormer），一个基于 Transformer 的跨学习模块，与一种创新的块聚类策略。DSFormer 通过在从最后两个 CNN 层提取的双尺度特征之间实现双向信息传递来增强特征表示，通过自身注意力捕捉每个尺度内的长程依赖关系以及共享的跨注意力进行跨尺度学习，捕获语义丰富性和空间细节。补充这一点的是，我们的块聚类策略从多个不同的视角对广泛使用的旧金山超大型（SF-XL）训练数据集重新划分，以优化数据组织，进一步增强对视点变化的鲁棒性。这些创新结合在一起，不仅产生了能够适应环境变化的鲁棒全局嵌入，还将所需的训练数据量减少了约 30 %，相较于之前的划分方法。综合实验表明，我们的方法在大多数基准数据集上实现了最先进的性能，作为使用 512 维全局描述符的全局检索解决方案，超越了先进的重排序方法如 DELG、Patch-NetVLAD、TransVPR 和 R2Former，并显著提高了计算效率。源代码将在 https://github.com/aurorawhisper/dsformer.git 发布。**

***Index Terms*—Visual Place Recognition, Transformer, Block Clustering, Dual Scale Cross-Learning**

I. 介绍

V 精确率 是机器人系统中的一项基本功能，使机器人能够通过将视觉输入与预先存在的带有地理标签的数据库匹配，在环境中大致定位自身，这对于机器人执行大规模地理定位任务至关重要。然而，在动态的现实世界中实现强大且高效的识别仍面临若干显著挑战。诸如条件变化（例如。，光照、天气、季节变化、长期变化等）、视角变化和感知别名 [1] （即不同位置展示出重复或结构相似的模式）等挑战降低了特征的一致性和独特性。此外，实际的机器人系统通常要求在严格的内存和计算约束下实时性能，这些挑战进一步恶化。

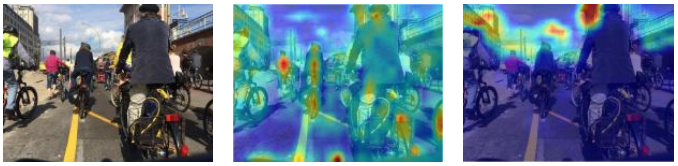
传统的 VPR 方法 [2]–[5] 通常依赖于在 ImageNet [8] 上预训练的 CNN 骨干网（如 例如，, ResNet [6] 和 VGG [7] 等）来提取局部特征。这些特征随后通过诸如 GeM [9] 或 NetVLAD [10] 等技术聚合成紧凑的描述符，所得模型在特定于 VPR 的数据集上进行微调，例如 MSLS [11] 和 Pittsburgh30k [12] 。虽然这些低内存和低延迟的紧凑表示能够支持高效的大规模检索任务，但它们在复杂环境中通常表现出有限的鲁棒性。这是由于 ImageNet 预训练骨干网的有限泛化能力，其偏向于对象中心的分布，以及训练数据集多样性的不足，未能捕捉现实世界中条件变化、视点和场景结构全谱的变化。一种有前景的解决方案涉及两阶段方法 [13]–[17] ，其中全局特征首先用于从数据库中检索前 k 个候选，然后通过局部特征匹配对这些候选进行重新排序。然而，与局部特征匹配相关的大量内存使用和高延迟对实际应用构成重大挑战，强调了对鲁棒轻量级全局描述符的需求。

最近的进展 [18]–[20] 利用 DINOv2 [21] 的能力，这是一个基于变压器的模型，训练于 LVD-142M 数据集。这些方法使用 DINOv2 作为提取特征的骨干，一种常用的 CNN 和经过 ImageNet 训练的 ViTs 的替代方案，实现了鲁棒的高性能结果。同时，CosPlace [22] 和 EigenPlaces [23] 引入了一个大规模的 VPR 密集训练数据集，称为 SF-XL [24] ，在训练过程中消除了挖掘负例的需求，并能够从广泛的数据集中进行鲁棒学习。然而，他们基于网格的分区策略导致聚类效率低下，造成类别间的数据不平衡，并减少了有效利用。最近的进展 [18]–[20] 利用 DINOv2 [21] 的能力，这是一个基于变压器的模型，训练于 LVD-142M 数据集。这些方法使用 DINOv2 作为提取特征的骨干，一种常用的 CNN 和经过 ImageNet 训练的 ViTs 的替代方案，实现了鲁棒的高性能结果。同时，CosPlace [22] 和 EigenPlaces [23] 引入了一个大规模的 VPR 密集训练数据集，称为 SF-XL [24] ，在训练过程中消除了挖掘负例的需求，并能够从广泛的数据集中进行鲁棒学习。然而，他们基于网格的分区策略导致聚类效率低下，造成类别间的数据不平衡，并减少了有效利用。

在这项研究中，我们提出了一种新颖的基于 Transformer 的模型 Dual-Scale-Former（DSFormer），旨在整合来自最终两个 CNN 层的双尺度特征。与之前的将多层特征进行简单连接且层间交互最少的方法不同，DSFormer 运用了自注意力机制来捕捉每个尺度内的长距离依赖关系，并通过交叉注意力模块动态融合跨尺度相关性并分配权重，从而生成与语义和结构线索相协调的鲁棒全局嵌入，如图 1 所示。同时，我们引入了一种基于 HDBSCAN [25] 的块聚类方法以重新划分 SF-XL。这种划分方法减轻了类间数据分布显著不平衡的问题，并减少了冗余，从而提高了数据利用的整体效率。我们的贡献可以总结为如下几点：

- 我们提出了 DSFormer，这是一种基于 Transformer 的模块，通过双尺度特征交叉学习生成辨别性的全局描述符，从而增强了 VPR 中对环境和视点变化的鲁棒性。
- 我们提出了一种针对 SF-XL 的块聚类策略，与 EigenPlaces 相比，该策略优化了数据利用效率，并将体积减少了约 30 %，同时实现了更优异的性能。

*Corresponding author: S. Piao. and C. Gao.
[1] H. Jiang and S. Piao are with the Multi-agent Robot Research Center, Department of Faculty of Computing, Harbin Institute of Technology, Harbin, 150001 China. E-mail: jianghy.hgd@stu.hit.edu.cn, piaosh@hit.edu.cn.
[2] C. Gao is with the Institute for AI Industry Research (AIR), Tsinghua University, BeiJing, 100084, China. Email: chao.gao@cantab.net
[3] L. Yu is with the School of Artificial Intelligence, Wuhan University, Wuhan, 430072, China. E-mail: ly.wd@whu.edu.cn.
[4] L. Chen is with the Soochow University, Soochow, 215031, China. E-mail: chenliguo@suda.edu.cn.
Sponsored by Xinchen Qihang Inc.



(a) Original image (b) ResNet50 (c) Our work

Fig. 1. 对比预训练基础模型（ResNet-50）和我们方法的特征热图可视化。预训练模型强调无关的区域，如动态的行人，而我们的方法则瞄准辨别性区域，如建筑物，这些区域与地点识别更相关。

传统的 VPR 方法主要依赖于聚合技术，这些技术将手工设计特征的统计属性汇总成紧凑的表示。VPR 任务可以通过仅专注于图像内容来设计，以解决图像检索任务，从而催生出许多高性能的基于学习的解决方案。随着深度学习的兴起，现代 VPR 方法已转向利用预训练的 CNN 和 ViT 骨干网络，并结合先进的聚合策略以生成强大的全局描述符。三种突出的基于学习的聚合方法 NetVLAD、GeM 和 MixVPR 体现了这一趋势，每种方法都提供了增强特征表示的不同机制。NetVLAD，一种开创性的方法，用一个可学习的聚类层取代了传统的 VLAD 手工聚类，从而实现了端到端的训练和适应性的特征分组。其灵活性和强大的性能激发了众多变体，例如 CRN [5]，其优化上下文关系，TransVLAD [29]，融合了基于 Transformer 的增强，以及 MultiRes-NetVLAD [2]，通过集成多分辨率特征来提高尺度间的鲁棒性。与之相比，GeM 引入了一个可学习参数以动态调整对特征大小的敏感性。这种适应性使其成为一个广泛采用的基线，近期一些框架如 CosPlace [22] 和 EigenPlaces [23] 利用 GeM 通过优化在如 SF-XL 等密集数据集上的训练，在大规模 VPR 中实现了显著的性能提升。最近，MixVPR [28] 运用多层感知器（MLPs）来混合深度特征，生成具有高度辨别力的全局描述符，这些描述符能够捕捉复杂模式。尽管这些基于学习的聚合方法在计算效率上表现出色，并在受控或简单场景中表现稳定，其局限性在更具挑战性的真实环境中显现。

为了解决全球检索方法在复杂环境中的局限性，两阶段方法逐渐兴起，通过对初始全球阶段检索到的前 k 个候选进行重新排序来细化 VPR 性能。这些方法引入了一个随后重新排序阶段，利用局部特征的空间一致性匹配来提高检索精度。总体而言，两阶段技术可以分为基于 RANSAC 和基于学习的范式，各自通过不同的策略解决几何验证问题。基于 RANSAC 的方法，例如 Patch-NetVLAD [14]、DELG [13]、TransVPR [15] 和 EUPN [17]，通过估算查询和参考图像之间的单应性来识别局部特征中的匹配内点。然后，基于内点的数量对候选进行重新排序，有效地结合了空间关系，以提高对视点变化和环境变化的鲁棒性。Patch-NetVLAD 通过补丁级匹配来优化 NetVLAD 描述符，而 DELG 则在统一框架中结合了全局和局部特征，展示了在各种数据集上的多样性。相反，基于学习的方法，如 R ^{2} Former [16] 采用基于 Transformer 的匹配网络来替代计算密集的 RANSAC 过程，直接计算局部特征间的匹配分数，同时利用注意力机制来建模复杂的空间依赖性，并简化重新排序。通过整合来自局部特征的几何信息，两阶段方法显著优于单独的全球检索技术，尤其是在光照变化剧烈、存在遮挡和感知混淆的场景中。然而，由于局部特征存储导致的高内存开销和几何验证带来的时间延迟，限制了两阶段方法的实用性。

尽管通过先进的重排序技术提高了检索精度，VPR 方法的性能依赖于预训练的骨干网络和采用的训练数据集。通过 DINOv2 [21]，一种基于 Transformer 的模型，在广泛的 LVD-142M 数据集上的预训练，骨干网络泛化方面取得了重大进展。这个大规模训练赋予 DINOv2 以强大的可泛化特征，超越了在 ImageNet 对象中心分布上训练的传统 CNN 和 ViTs 的能力。AnyLoc [30] 首次利用这一点，使用预训练的 DINOv2 骨干网络提取多层特征，通过 VLAD 集中，实现有效的跨域 VPR，且对任务特定注释的依赖最小。后续的基于 DINOv2 的 VPR 方法 [18]–[20] 进一步证明了，由于 DINOv2 本身的泛化能力，该骨干网络在极端光照变化、视点转换等具有挑战性的场景中达到了最先进的性能。对于传统训练数据集，如 MSLS [11] 和 Pittsburgh30k [12]，环境条件（如照明、天气、季节变化和视点）有限的变化，使得模型很难泛化到多样化和未见过的测试域。为了解决这个问题，CosPlace [22] 引入了经过处理的 SF-XL 数据集，这是一个大规模、密集采样的集合，通过广泛的真实世界覆盖提高鲁棒性而不进行负样本挖掘。EigenPlaces [23] 在此基础上重新处理 SF-XL，通过细化的分区来更好地处理视点变化。然而，这些努力受到次优的数据分类的阻碍，固定的网格划分误判地理上邻近的位置，导致数据利用效率低下。

II. 方法论

为了解决 VPR 任务中的内在挑战，我们提出了两种协同策略，旨在提高特征表示和数据利用效率。首先，我们引入了 DSFormer，一种新型的基于 Transformer 的架构，它有效地捕获和整合双尺度特征，显著提高了特征表示的判别能力。其次，我们开发了一种先进的块聚类技术，利用 HDBSCAN [25] 来预处理密集的 SF-XL 数据集 [24]，通过确保地理位置接近的地点具有一致的类别分配，优化训练数据的利用效率。

A. DSFormer

我们的模型仍然采用 ResNet-50 作为特征提取的主干网络，利用它来编码丰富的视觉信息。与仅依赖最终层或倒数第二层特征的先前方法不同，我们利用从最后两层提取的双尺度特征来构建更健壮的全局描述符。虽然 ResNet-50 提供了强大的局部归纳偏差，但它可能在整体上下文意识上有所欠缺。为了解决这个问题，我们引入了 DSFormer，通过自注意力和交叉注意力机制的协同结合，实现了两层之间的跨层学习，有效地捕捉了长程依赖和尺度间的相关性。最后，应用 GeM 池化模块将双尺度特征聚合成一个紧凑的全局描述符。图 2 展示了该过程的完整框架。

1) Transformer 编码器： Dosovitskiy 等人提出的 Vision Transformer 通过将图像视为适合 Transformer 架构的嵌入序列来重新定义视觉处理，在我们的改编中，我们用从预训练的主干网获得的特征块替代了原始图像块。给定不同尺度的特征图 $f_i \in \mathbb{R}^{C_i \times H_i \times W_i}, i = 1, 2$，我们将它们划分为 $N_i = H_i \times W_i$ 个块。f_i 被展平成嵌入：

$$z_0 = [x_f^1; x_f^2; \cdots; x_f^N] + E_{pos} \tag{1}$$

其中 x_f^j 表示 f_i 的第 j 个特征补丁，E_{pos} 提供位置编码以保持空间关系。这些嵌入随后通过利用多头注意力（MHA）的 Transformer 编码层进行处理：

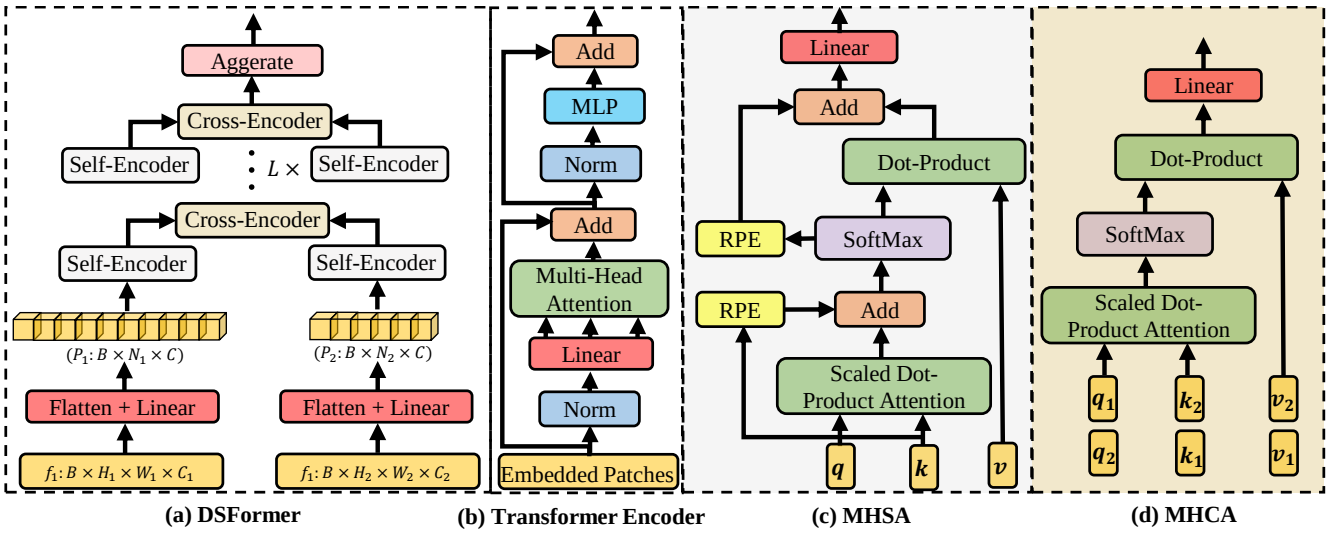


Fig. 2. （a）DSFormer 模块：输入由 ResNet-50 的双尺度特征图组成，线性映射为两个维度相同的嵌入补丁，通过包含自编码器和共享交叉编码器的 L 编码块进行处理。(b) Transformer 编码器层。(c) 自编码器块采用带有相对位置编码的多头自注意力（MHSA）。(d) 交叉编码器利用多头交叉注意力（MHCA）并操作双尺度补丁嵌入作为输入。

$$z_t = \text{MHA}(\text{LN}(z_{l-1})) + z_{l-1}, z_l = \text{FFN}(\text{LN}(z_t)) + z_t \quad (2)$$

$$\text{MHA}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) = \text{Concat}(\text{head}_1, \cdots, \text{head}_h)W^O \quad (3)$$

$$\text{head}_i = \text{Attention}(\boldsymbol{q}W_i^{\boldsymbol{q}}, \boldsymbol{k}W_i^{\boldsymbol{k}}, \boldsymbol{v}W_i^{\boldsymbol{v}}) \quad (4)$$

这个基于补丁的 Transformer 框架旨在处理不带分类标记的特征输入，是我们 DSFormer 的基础，该模型利用这些原理来整合用于 VPR 任务的双尺度特征。

2) 多头自注意力： 为了有效地建模特征块序列中的空间关系，我们通过结合改进的相对位置编码 (IRPE) [32] 模块来采用定制的自注意力机制。三个可学习的函数，记为 $\text{rpe}_{\boldsymbol{q}}$ 、$\text{rpe}_{\boldsymbol{k}}$ 、$\text{rpe}_{\boldsymbol{v}}$ ，用于将相对距离映射到相应的编码值，如下所示：

$$\text{Attention}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) = \text{rpe}_{\boldsymbol{v}}(\text{softmax}(\frac{\boldsymbol{q} \cdot \boldsymbol{k}^T}{\sqrt{d_k}} + \text{rpe}_{\boldsymbol{q}}(\boldsymbol{q}) + \text{rpe}_{\boldsymbol{k}}(\boldsymbol{k}))\boldsymbol{v}) \quad (5)$$

通过结合 IRPE，DSFormer 能够有效捕捉远距离依赖关系和空间连贯性，从而提高特征的可区分性，并在不同的环境条件和视点变化中增强 VPR 的鲁棒性。

3) MHCA： 为了促进 DSFormer 中双尺度特征之间的有效跨学习，我们引入了一个专门的共享交叉注意力模块，该模块整合了 ResNet-50 骨干网络最后两层的特征，实现了粗粒度和细粒度表示的融合，从而提高了视觉位置识别任务中全局描述符的鲁棒性和辨识性。给定一组特征补丁嵌入：$\boldsymbol{z_i} \in \mathbb{R}^{N_i \times C}$ 。交叉注意力定义为：

$$\text{Attention}(\boldsymbol{q}_i, \boldsymbol{k}_m, \boldsymbol{v}_m) = \text{softmax}\left(\frac{\boldsymbol{q}_i \cdot {\boldsymbol{k}_m}^T}{\sqrt{d_{\boldsymbol{k}_m}}}\right)\boldsymbol{v}_m \quad (6)$$

$$\boldsymbol{q}_i, \boldsymbol{k}_i, \boldsymbol{v}_i = \boldsymbol{z}_i(W^{\boldsymbol{q}_i}, W^{\boldsymbol{k}_i}, W^{\boldsymbol{v}_i}), i = 1, 2.$$

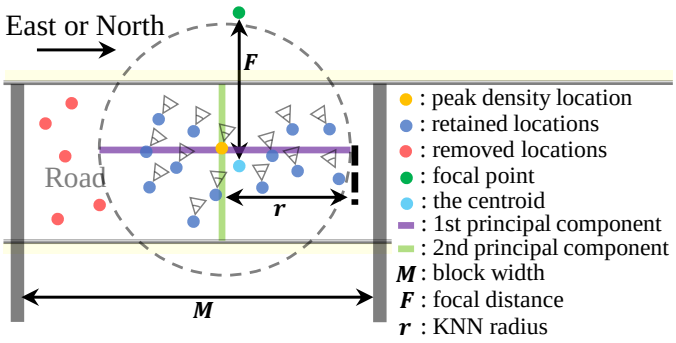


Fig. 3. 分块聚类原理概述。最初，该区域沿东或北方向分割成宽度为 M 米的块，其中被街道边界截断的采样位置组合为一个单独的类别。KNN 方法保留了以峰值密度位置为中心、半径为 r 以内的采样位置，其他则被舍弃。SVD 用于识别两个主要方向，从每个质心沿这些方向定义一个焦点，距离为 F ，而采样位置的图像则指向该焦点。

在 Attention 中 $i \neq m$ 是

为了提高数据的利用效率，我们引入了一种块聚类方法来重新划分 SF-XL 数据集，使用 UTM 坐标 $\{east, north\}$ 。相同的过程用于两个方向的坐标值，以东方作为代表性例子。

给定一个 UTM 坐标 $x_{i,j} = (x_e^i, x_n^i)$ ，表示分配给 j 组 $\mathcal{G}_e^j$ 的第 i 个位置，其中 (x_e^i, x_n^i) 分别表示东坐标和北坐标。而，

$$j = \left\lfloor \frac{(x_e^i - \min(\{x_e^i\}_{i=1}^I)) \mod (M \cdot N)}{M} \right\rfloor + 1 \quad (7)$$

这里，M 是每个块的宽度，N 表示组的总数，I 是位置的总数。随后，组 $\mathcal{G}_e^j$ 被 HDBSCAN [25] 聚类为 K 类，利用其层次方法识别具有抗噪能力的不同形状和大小的基于密度的簇。然而，类内距离的不确定性可能导致同一类内的采样位置相距甚远，从而降低模型对空间接近性的敏感度。为了解决这个问题，我们采用 k -最近邻（KNN）方法，仅保留在峰值密度位置半径 r 米范围内的采样点，如图 3 所示。这有效地限制了类内距离并减少了数据冗余，提高

TABLE I
DSFormer 和最新的 VPR 方法在多个数据集上的性能比较，以 R@1（Recall@1）和 R@5（Recall@5）为指标。† 表示在我们新的 SFXL 分区进行的训练。

Method	Training Set	MSLS Val		MSLS Chal		Pitts30k		Tokyo24/7		Nordland		Average	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Two-Stage Methods													
DELG [13]	GLD [33]	84.5	90.4	53.8	62.5	89.6	95.3	87.6	93.3	67.9	77.8	76.7	83.9
Patch-NetVLAD [14]	MSLS	79.5	86.2	48.1	57.6	88.7	94.5	85.1	87.0	62.6	70.8	72.8	79.2
SP-SuperGLUE [34], [35]	SS [34]	84.5	88.8	50.6	56.9	88.7	95.1	88.3	89.2	65.3	70.4	75.5	80.1
TransVPR [15]	MSLS	86.5	91.1	63.9	73.2	89.0	94.9	77.8	81.6	73.4	81.4	78.1	84.4
R ^{2} Former [16]	MSLS	89.7	95.0	73.0	85.9	91.1	95.2	86.3	90.8	77.1	89.0	83.3	91.2
Retrieval Baselines on a ResNet-50 backbone													
NetVLAD [10]	MSLS	78.2	87.1	55.8	71.5	74.8	88.4	48.9	67.3	53.0	75.4	62.1	77.9
GeM [9]	MSLS	75.9	88.9	53.6	70.5	77.1	90.6	39.7	63.5	43.9	68.8	58.0	76.5
CosPlace [22]	SF-XL	85.7	91.2	63.0	74.6	90.4	95.4	80.0	88.6	68.3	86.7	77.5	87.3
EigenPlaces [23]	SF-XL	85.9	91.1	62.6	74.8	91.0	95.7	78.1	88.9	71.9	88.9	77.9	87.1
MixVPR [28]	GSV-Cities [36]	83.6	91.5	59.2	72.3	90.6	95.5	80.0	89.5	70.8	86.9	76.8	87.1
BoQ + PCA [37]	GSV-Cities	85.8	90.8	63.2	74.2	90.2	94.7	80.3	88.9	73.6	89.5	78.6	87.6
GeM † (ResNet-50)	SF-XL	87.6	91.9	65.3	76.0	91.9	96.0	82.9	92.1	75.5	91.1	80.6	89.4
DSFormer (ResNet-50)	SF-XL	88.9	93.2	68.1	77.6	91.9	96.4	88.6	93.7	81.5	94.1	83.8	91.0
Retrieval Baselines on a DINOv2 backbone													
CricaVPR + PCA [19]	GSV-Cities	86.9	93.6	69.3	81.9	92.3	96.8	87.6	94.9	88.1	97.5	84.8	92.9
SALAD + PCA [20]	GSV-Cities	90.3	96.2	73.2	87.0	91.2	95.8	92.4	97.1	84.5	95.7	86.3	94.4
BoQ + PCA	GSV-Cities	89.6	95.0	73.8	84.7	94.0	98.6	89.2	94.0	84.8	97.0	86.3	93.9
GeM † (DINOv2)	SF-XL	90.3	95.9	74.6	87.3	93.5	97.0	95.9	97.8	87.7	97.4	88.4	95.1
DSFormer (DINOv2)	SF-XL	92.7	96.2	75.8	87.4	93.6	97.0	95.9	97.8	88.4	97.2	89.3	95.1

了所得数据集的质量。同时，为了防止类间可见区域重叠，我们排除密度峰值位置间距离小于 l 米的类，进一步减小了数据集的大小。对于北向，我们使用北坐标 x_n^j 执行相同操作以获得组的集合 $\{\mathcal{G}_n^j\}_{j=1}^N$ 。进一步合并 $\{\mathcal{G}_e^j\}_{j=1}^N$ 和 $\{\mathcal{G}_n^j\}_{j=1}^N$ ：

$$\{\mathcal{G}_j\}_{j=1}^N = \{\mathcal{G}_e^j \cup \mathcal{G}_n^j\}_{j=1}^N \tag{8}$$

为了防止合并后的聚类中不同类别之间出现重叠区域，对于其峰值密度位置距离低于 l 米的任何类别，将再次删除。我们的聚类确保沿同一条街道的采样位置在各自的类别中均匀分布，防止由于不均匀划分而导致的类内样本不足，从而保持更有效的类别。

为了提高对视点变化的鲁棒性，我们基于 EigenPlaces [23] 中提出的方法实施了一种图像选择策略。对于每个保留的类别，由其 UTM 坐标 $X_k \in \mathbb{R}^{n\times 2}$ 表示，其中 n 是 k 类别中采样位置的数量，我们执行奇异值分解 (SVD) 以提取中心矩阵 $\widehat{X_k} = X_k - E[X_k]$ 的两个主方向：一个通常与捕获图像中车辆行驶的道路对齐，另一个垂直于它，代表路边。然后我们计算每个方向对应的焦点 c_k ，方法如下：

$$c_k = E[X_k] + F \cdot V \tag{9}$$

其中 F 表示焦距，V 代表从 SVD 中导出的方向向量，$E[X_k]$ 是质心的 UTM 坐标。每组根据两个方向分为两组，最终得到 $2N$ 组。

III. 评估

在本节中，我们评估了所提出的 DSFormer 的性能。评估分为三个关键部分：（1）实验设置，其中概述了实现细节、数据集、评估指标和对比方法；（2）结果与分析，展示了定量结果以及与当前最先进方法的比较见解；（3）消融实验，研究了我们方法中各个组成部分的贡献。

A. 实验设置

1) 实现细节： 我们的模型利用了预训练的 ResNet-50 骨干网络 [6] 进行特征提取，截断在最后两层以保留语义丰富的局部特征表示。除最后两个残差块外的所有残差块均被冻结，对通道维度应用线性投影以标准化，使其与交叉注意机制兼容。该实现是用 PyTorch 构建的。输入图像被调整为 320×320 ，并随机缩放并进行颜色抖动增强。我们在处理过的大规模 SF-XL 数据集 [24] 上训练我们的模型，该数据集包含约 4.2 百万张图像，分为 10 组，捕捉到各种极端条件。训练遵循类似于 EigenPlaces [38] 的框架，采用大边距余弦损失（LMCL）[39] 作为损失函数。模型使用 Adam 优化器 [40] 以学习率 1×10^{-5} 进行优化，而分类器使用一个单独的 Adam 优化器，学习率为 0.01。训练在具有 16 GB 内存的 NVIDIA RTX 4080 GPU 上运行长达 40 个 epoch，批量大小为 32 张图像，每个 epoch 定义为对一组进行 10000 次迭代。DSFormer 模块由 3 层组成，注意机制使用 16 个共享参数头。对于块聚类，我们设置块宽度为 $M = 10$ 米，组数为 $N = 5$ （总计 10 组方向），KNN 半径为 $r = 7.5$ 米，峰密度位置之间的最小距离为 $l = 40$ 米，焦距为 $F = 15$ 米。此外，我们还评估了配备 DINOv2（ViT-B）骨干的 DSFormer。该模型使用分辨率为 322×322 的输入图像进行训练。在训练期间，DINOv2 的最后两层被微调。该架构仅包含一层 DSFormer 层，采用 12 个具有共享参数的注意头。

2) 数据集和指标： 我们在五个广泛采用的 VPR 基准数据集和两个大规模数据集上评估我们的方法：MSLS 验证集 [11] 、MSLS 挑战集 [11] 、Pittsburgh30k [12] 、Tokyo24/7 [41] 、Nordland [42] （使用 [43] 的冬夏分区分别作为查询集和参考集），SF-XL 测试数据集 [22], [44]

TABLE II
在 MSLS 验证集上使用 NVIDIA RTX 3060 评估 DSFormer 与双阶段 VPR 方法在延迟和内存使用方面的比较。

Method	Des. Size	Memory (GB)	Latency(s)/query	
			Extract	Rerank
DELG	$1000 \times 128 + 2048$	7.8	0.096	35.8
Patch-NetVLAD	$2826 \times 4096 + 4096$	908.3	1.006	16.88
SP-SuperGLUE	$2048 \times 256 + 4096$	41.6	0.092	12.381
TransVPR	$1200 \times 256 + 256$	24.2	0.017	2.118
R ^{2} Former	$500 \times 131 + 256$	5.2	0.025	0.423
DSFormer (Ours)	1×512	0.039	0.024	N/A

。对于 Nordland，性能的测量误差容忍度为 ±2 帧。所有其他数据集使用 Recall@N（R@N）作为指标，其中成功的条件是前 N 个检索结果中至少有一个在 25 米的位置误差容忍度内匹配到真实值。

我们将我们的 DSFormer 与几种最先进的（SOTA）VPR 方法进行了比较，这些方法分为两类。检索基线包括 NetVLAD [10]、GeM [9]、CosPlace [22]、EigenPlaces [23]、MixVPR [28] 和 BoQ [37]。为了公平比较，这些方法使用 320×320 的输入图像大小，并使用 ResNet-50 作为骨干网络，生成 512 维的全局描述符（对于 BoQ，应用 PCA 进行降维）。此外，我们还包括了一些最近采用 DINOv2（ViT-B）骨干网的检索方法，即 CricaVPR [19]、SALAD [20] 和 BoQ。这些方法使用 322×322 的输入图像分辨率，通过 PCA 将描述符降低至 512 维，以确保公平比较。为了在评估中保持一致性，我们重新实现了大多数检索方法，以确保在训练和测试期间输入图像分辨率保持不变。两阶段方法包括 DELG [13]、Patch-NetVLAD [14]、TransVPR [15] 和 R ^{2} Former [16]，它们首先利用全局描述符检索前 100 名候选者，然后进行基于局部描述符的重新排序，输入图像大小为 640×480。这里报告的所有比较结果均来自我们对各方法的再现。

B. 结果与分析

我们在标准基准数据集上评估 DSFormer 与最新的 VPR 方法的表现，如表 I 所示。DSFormer（ResNet-50）在平均指标上表现出色，平均而言，在五个数据集中，R@1 上超越了 EigenPlaces 5.9 % 和 R ^{2} Former 0.5 %。尽管 DSFormer 在 MSLS 上的表现略低于 R ^{2} Former，但在其余数据集上展示了优势。值得注意的是，R2Former 的训练数据集来源与 MSLS 的评估集相同，并且虽然 R2Former 使用了比我们模型的 320×320 配置更高的 480×640 输入分辨率，这可能导致我们的方法在 MSLS 中的评估性能略低于 R2Former 的结果。此外，我们将我们方法的计算成本与两阶段方法进行了比较。DSFormer 实现的内存占用比 R2Former 小 132 倍，同时还表现出比所有基于重排序方法显著更高的计算效率，因为它消除了重排序过程的必要性，如表 II 所示。DSFormer（DINOv2）在基于 DINOv2 的方法中取得了最佳性能，平均 R@1 上比 SALAD 高出 3.0 %。这一结果证实了即使采用 DINOv2 骨干时，DSFormer 依然有效。

我们评估了 GeM †（ResNet-50）模型，该模型使用我们的块聚类策略在 SF-XL 数据集上进行了训练，与 CosPlace 和 EigenPlaces 进行比较——它们都使用 GeM 模型在 SF-XL 的不同分区上进行训练，平均在 R@1 指标上分别超过 3.1 % 和 2.7 % 的性能提升。这一改进归功于我们的块聚类方法，它不仅提高了检索精度，还与 CosPlace 和 EigenPlaces 使用的分区技术相比，将数据集大小分别减少了约 25 %（从 5.6M 减少到 4.2M）和 32 %（从 6.2M 减少到 4.2M）。此外，DSFormer 平均在 R@1 上比 GeM † 高 3.2 %（ResNet-50 骨干），这证明了 DSFormer 模块的有效性。这一优势在具有挑战性的 Tokyo24/7 和 Nordland 数据集上尤为明显，DSFormer 分别实现了 5.7 % 和 6.0 % 的性能提升，进一步验证了其在处理严重外观变化时的鲁棒性。此外，以 DINOv2 为主干的 GeM † 比其 ResNet-50 在平均 R@1 上提高了 7.8 %，这表明 DINOv2（ViT-B）具有显著更强的泛化能力。

TABLE III
在大规模 SF-XL 测试集上，补充性能比较 DSFormer 和最新 VPR 方法的 Recall@1 指标。

Method	v1	v2	Night	Occlusion	Average
DELG	85.0	93.5	28.3	30.3	59.3
Patch-NetVLAD	37.9	77.4	12.2	13.2	35.2
SP-SuperGLUE	41.6	78.9	12.2	13.2	36.5
TransVPR	37.0	68.4	4.9	15.8	31.5
R ^{2} Former	45.5	74.9	11.6	22.4	38.6
NetVLAD	19.7	41.6	3.2	6.6	17.8
GeM	17.6	38.3	3.2	9.2	17.1
CosPlace	75.8	85.6	26.8	36.8	56.3
EigenPlaces	80.0	90.3	25.3	32.9	57.1
MixVPR	61.6	85.5	13.1	25.0	46.3
BoQ + PCA	60.0	82.9	15.0	25.0	45.7
GeM † (ResNet-50)	82.1	89.5	29.2	38.2	59.8
DSFormer (ResNet-50)	85.7	91.3	31.5	42.1	62.7
CricaVPR + PCA	77.5	88.8	31.1	47.4	61.2
SALAD + PCA	80.4	92.1	41.8	38.2	63.1
BoQ + PCA	73.6	87.3	32.4	39.5	58.2
GeM † (DINOv2)	93.5	94.3	52.1	47.4	71.8
DSFormer (DINOv2)	93.7	94.8	54.1	52.6	73.8

在现实场景中，数据库通常由大量的参考图像组成，而查询图像可能并不一定来自于与参考图像相同的数据集。为了评估在这些具有挑战性的场景中的 VPR 方法的性能，我们新增了一个大型的基准测试，SF-XL，它包括多个查询子集（v1、v2、夜晚和遮挡）。该基准测试的数据库包含大量参考图像（约 280 万），而大多数查询子集的查询图像不来自 SF-XL，并涵盖复杂的条件，如夜间环境和重大遮挡。我们在这个基准上评估了全局检索方法和两阶段方法的性能，结果总结在表 III 中。我们提出的 DSFormer（ResNet-50）实现了最高的准确率，在 R@1 的平均精度上比 EigenPlaces 高出 5.6 %。两阶段方法的次优结果可以归因于它们对全局检索精度的严重依赖。此外，我们的 DSFormer（DINOv2）仍然表现出色，比 SALAD 高出 10.7 %。

C. 消融

1) DSFormer 上的消融研究： 在消融研究中，我们系统性地分析了 DSFormer 层数及其核心组件对模型性能的影响。结果如表 IV 所示，分为两个主要部分。第一部分探讨了改变 DSFormer 层数的效果。当所有 DSFormer 层被移除时，模型在外观变化严重的数据集（如 Tokyo24/7 和 Nordland 数据集）上表现出明显的性能下降。在 MSLS 验证集和 Tokyo24/7 上，采用三个 DSFormer 层可实现最佳结果。然而，将层数增加到四层并未带来显著的益处，并且在某些情况下还导致轻微的性能下降。因此，我们采

TABLE IV
关于 DSFormer 层数以及 IRPE、自编码器 (SE)、交叉编码器 (CE) 和区块聚类 (BC) 影响的消融研究，展示了在多个数据集上 512 维输出的 Recall@1 结果。

Ablated Versions	MSLS Val	Pitts30K	Tokyo24/7	Nordland
Three layers	88.9	91.9	88.6	81.5
Zero layer	87.2	91.5	81.9	77.7
One layer	87.6	91.7	85.1	81.0
Two layers	88.6	92.2	86.7	80.8
Four Layers	87.7	92.0	87.6	82.8
Remove IRPE	88.4	91.9	88.3	80.8
Remove SE	87.8	91.6	87.0	81.9
Remove CE	87.7	91.9	86.0	83.0
Remove BC	87.4	91.9	88.3	77.2

用三层 DSFormer 配置作为基准，它在性能和效率之间提供了更好的平衡。第二部分考察了单个 DSFormer 组件的贡献。我们独立移除了 IRPE、自编码器（SE）和交叉编码器（CE），每次移除都导致多个数据集上的性能下降。这些结果强调了 DSFormer 模块在提高模型鲁棒性和整体效果中的重要性。最后一部分展示了所提出的块聚类策略的消融研究（移除聚类等同于在 SF-XL 数据集上使用 EigenPlaces 中的数据分配方案），可以观察到多个评估数据集上的性能下降，从而验证了聚类设计的有效性。

TABLE V
焦距消融研究，展示了在 MSLS Val 和 Pitts30K 数据集上，具有 ResNet-50 骨干和 512 维输出的 GeM 模型的 Recall@1 结果。

Focal Distance(m)	MSLS Val			Pitts30K		
	R@1	R@5	R@10	R@1	R@5	R@10
5	85.7	91.1	92.4	91.2	96.1	97.0
10	87.4	92.0	93.1	91.5	96.3	97.3
15	87.6	91.9	93.2	91.9	96.0	97.1
20	87.4	91.8	93.2	91.6	96.2	97.2
25	87.4	92.6	93.8	92.0	96.3	97.3
30	87.3	92.0	93.6	91.7	96.0	97.2

在本节中，我们进行了消融实验，以研究块聚类中一个重要参数的影响：焦距，定义为从类中心到焦点的距离。该参数影响采样图像的方向和视角，可能会影响模型在视觉定位任务中的性能。为了评估其影响，我们在 MSLS Val 和 Pitts30K 数据集上测试了 GeM † （ResNet-50）模型在各种焦距下的表现，如表 V 所示。在较短距离（例如 5 米）时，有限的视野可能限制模型捕捉全面场景信息的能力，导致精度降低。当焦距增加到 15 米时，更宽的视角增强了模型对环境的理解，从而提高了性能。然而，过大的距离（例如 30 米）会带来收益递减，因为更广泛的视角可能引入无关细节或噪声，导致性能出现小幅波动。基于这些发现，我们选择 15 米作为默认焦距，因为它在捕捉足够的背景信息与避免过多噪声之间取得了平衡。

IV. 结论

本研究介绍了 DSFormer，这是一个先进的模型，旨在解决视觉位置识别（VPR）中的基本挑战，特别是视点和外观的变化。我们通过引入一种新颖的区块聚类策略来优化 SFXL 数据集中的数据分区，增强训练效率并减少数据冗余。此外，所提出的双尺度 Transformer（DSFormer）模块通过利用双尺度交叉学习机制改善了特征表示。在多个基准数据集上进行实验评估，结果表明 DSFormer 有效地捕获了鲁棒特征表示，使其能够处理各种挑战，并优于现有的最先进的方法。

References

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," IEEE Transactions on Robotics , vol. 32, no. 1, pp. 1–19, 2015.

[2] A. Khaliq, M. Milford, and S. Garg, "Multires-netvlad: Augmenting place recognition training with low-resolution imagery," IEEE Robotics and Automation Letters , vol. 7, no. 2, pp. 3882–3889, 2022.

[3] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in CVPR , 2019.

[4] J. Revaud, J. Almazan, R. Rezende, and C. de Souza, "Learning with average precision: Training image retrieval with a listwise loss," in ICCV , 2019.

[5] H. J. Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization." in CVPR , vol. 1, no. 2, 2017, p. 3.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , 2016, pp. 770–778.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556 , 2014.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition . Ieee, 2009, pp. 248–255.

[9] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," IEEE Transactions on Pattern Analysis and Machine Intelligence , vol. 41, no. 7, pp. 1655–1668, 2018.

[10] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , 2016, pp. 5297–5307.

[11] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 2020, pp. 2626–2635.

[12] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in Proceedings of the IEEE conference on computer vision and pattern recognition , 2013, pp. 883–890.

[13] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in European Conference on Computer Vision . Springer, 2020, pp. 726–743.

[14] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 2021, pp. 14 141–14 152.

[15] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 2022, pp. 13 648–13 657.

[16] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, "R2former: Unified retrieval and reranking transformer for place recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 2023, pp. 19 370–19 380.

[17] H. Jiang, S. Piao, H. Yu, W. Li, and L. Yu, "Robust visual place recognition for severe appearance changes," IEEE Robotics and Automation Letters , 2024.

[18] F. Lu, L. Zhang, X. Lan, S. Dong, Y. Wang, and C. Yuan, "Towards seamless adaptation of pre-trained models for visual place recognition," in The Twelfth International Conference on Learning Representations , 2024. [Online]. Available: https://openreview.net/forum?id=TVg6hlfsKa

[19] F. Lu, X. Lan, L. Zhang, D. Jiang, Y. Wang, and C. Yuan, "Cricavpr: Cross-image correlation-aware representation learning for visual place recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 2024, pp. 16 772–16 782.

[20] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 2024, pp. 17 658–17 668.

[21] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby et al. , "Dinov2: Learning robust visual features without supervision," arXiv preprint arXiv:2304.07193 , 2023.

[22] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 2022, pp. 4878–4888.

[23] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision , 2023, pp. 11 080–11 090.

[24] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys et al. , "City-scale landmark identification on mobile devices," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) . IEEE, 2011, pp. 737–744.

[25] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," The Journal of Open Source Software , vol. 2, no. 11, p. 205, 2017.

[26] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," IEEE Transactions on Robotics , vol. 28, no. 5, pp. 1188–1197, 2012.

[27] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition . IEEE, 2010, pp. 3304–3311.

[28] A. Ali-bey, B. Chaib-draa, and P. Giguère, "Mixvpr: Feature mixing for visual place recognition," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) , January 2023, pp. 2998–3007.

[29] Y. Xu, P. Shamsolmoali, E. Granger, C. Nicodeme, L. Gardes, and J. Yang, "Transvlad: Multi-scale attention-based global descriptors for visual geo-localization," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision , 2023, pp. 2840–2849.

[30] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," IEEE Robotics and Automation Letters , vol. 9, no. 2, pp. 1286–1293, 2023.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations , 2021.

[32] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, "Rethinking and improving relative position encoding for vision transformer," in Proceedings of the IEEE/CVF International Conference on Computer Vision , 2021, pp. 10 033–10 041.

[33] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in Proceedings of the IEEE International Conference on Computer Vision , 2017, pp. 3456–3465.

[34] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , 2018, pp. 224–236.

[35] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 2020, pp. 4938–4947.

[36] A. Ali-bey, B. Chaib-draa, and P. Giguere, "Gsv-cities: Toward appropriate supervised visual place recognition," Neurocomputing , vol. 513, pp. 194–203, 2022.

[37] A. Ali-bey, B. Chaib-draa, and P. Giguère, "BoQ: A place is worth a bag of learnable queries," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , June 2024, pp. 17 794–17 803.

[38] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , October 2023, pp. 11 080–11 090.

[39] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , 2018, pp. 5265–5274.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in 3rd International Conference on Learning Representations, ICLR , 2015.

[41] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , 2015, pp. 1808–1817.

[42] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA) , p. 2013, 01 2013.

[43] D. Olid, J. M. Fácil, and J. Civera, "Single-view place recognition under seasonal changes," in PPNIV Workshop at IROS 2018 , 2018.

[44] G. Barbarani, M. Mostafa, H. Bayramov, G. Trivigno, G. Berton, C. Masone, and B. Caputo, "Are local features all you need for cross-domain visual place recognition?" in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 2023, pp. 6154–6164.