FROM TOPOLOGY TO METRIC: GEOMETRY-GROUNDED REFINEMENT FOR METRIC PLACE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual Place Recognition (VPR) aims to localize query images by retrieving visually matched geo-tagged database images. Existing works mainly achieve Topological Place Recognition, where the coordinate of the retrieved database image is directly assigned to the query location, overlooking the spatial offset between matched images and only getting the coarse geographical location. To address this issue, we introduce Metric Place Recognition and propose GeoMPR, a geometry-grounded framework that extends topological retrieval toward precise geographic position estimation. GeoMPR first retrieves a top-k set of geo-tagged candidates and filters the candidates to form a locally coherent reference set. Given the query and these references, VGGT estimates their relative camera geometry. The Spatial Offset Estimation Module (SOEP) first establishes the transformation mapping between the predicted relative camera geometry and the spatial offset using these reference images, and subsequently leverages this mapping to derive the spatial offset estimation of the query image relative to the anchor, and refines the estimation using residual compensation and global correction. Moreover, we introduce a Geometry-aware Spatial Offset Loss to explicitly supervise the spatial displacement between query and anchor images for accurate geographic position estimation. Extensive experiments on several challenging VPR benchmarks demonstrate that GeoMPR substantially improves place recognition accuracy under strict distance thresholds. Notably, GeoMPR improves the 5-meter recognition accuracy of the SOTA method on Tokyo24/7 by **42.5 pp** (36.8% $\rightarrow$ 79.3%), demonstrating the effectiveness of geometry-grounded refinement. Codes will be publicly available.

1 INTRODUCTION

Visual Place Recognition (VPR) estimates the location of a query image by retrieving visually similar images from a geo-tagged database (Liu et al., 2019; Berton et al., 2022b). Most VPR methods represent images with global or local descriptors and return the coordinate of a retrieved database image as the query location (Arandjelovic et al., 2016; Cao et al., 2020). We refer to this paradigm as **Topological Place Recognition (TPR)**: it identifies the most compatible database place and assigns its geographic position to the query image, but does not explicitly estimate the precise geographic position of query image. This coordinate assignment of TPR can be inaccurate even when retrieval succeeds: database images are sampled at discrete, often sparse locations along trajectories, so the database rarely contains an image captured at the exact position of query image. The retrieved image may therefore be several meters or tens of meters away from the query. Furthermore, TPR typically assigns the geographic coordinates of the visually most similar candidate image to the query image. However, prior work Izquierdo & Civera (2024a) has revealed that visual similarity does not reliably reflect geographic proximity, particularly over short distances, so the geographically closest reference may not rank first. Therefore, the geographic coordinate estimation in TPR inherently exhibits an offset from the ground truth, which is unacceptable for high-precision localization tasks.

Eliminating this offset requires geometric refinement beyond visual similarity. Two-stage VPR methods use local matching and geometric verification mainly to re-rank candidates, and generally still return a selected database coordinate (Cao et al., 2020; Hausler et al., 2021; Wang et al., 2022). Vision geometry foundation models such as DUSt3R (Wang et al., 2024) and VGGT (Wang et al., 2025) offer a direct geometric cue by predicting relative camera geometry, depth, and camera centers. However, their predictions lie in local coordinate systems with arbitrary scale and orien-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tation and cannot be directly interpreted as geographic coordinates. To address the spatial offset inherent in TPR, we first introduce **Metric Place Recognition (MPR)**, which extends TPR from assigning the query image the geographic coordinate of a retrieved database image to estimating the precise geographic coordinate of query image.

To realize MPR, we propose GeoMPR, a novel geometry-grounded refinement metric place recognition framework. We first use a conventional VPR model to retrieve a top-k set of geo-tagged candidates, whose top-ranked image serves as the anchor image. Candidates are then filtered by their geographic distance from the anchor to form a locally coherent reference set. Then, VGGT jointly processes these filtered references and the query image to obtain anchor-centered relative camera geometry. Next, the Spatial Offset Estimation Module (SOEP) obtains a transformation matrix from the relative camera geometry to the spatial offset through confidence-weighted differentiable ridge regression, and based on this matrix, produces the initial spatial offset estimation for query image. We further refine the initial offset estimation through residual compensation and global correction. Finally, we obtain the precise coordinates of the query image by combining the geographic coordinate of the anchor image and the spatial offset of the query image relative to the anchor image. During the training process, we propose Geometry-aware Spatial Offset Loss that combines anchor-relative offset regression and Euclidean metric loss to explicitly supervise the spatial displacement between query and anchor images for accurate geographic position estimation. Our main contributions are summarized as follows:

- We introduce Metric Place Recognition (MPR), a new formulation that extends conventional retrieval-based Topological Place Recognition (TPR) from discrete reference-image identification to precise geographic position estimation. MPR explicitly addresses the spatial displacement caused by discrete and sparse database sampling.
- We propose GeoMPR, a geometry-grounded coarse-to-fine metric place recognition framework designed to achieve geographic position estimation for query images. Within this framework, we introduce the Spatial Offset Estimation Module (SOEP) to estimate the spatial offset between the query image and reference images.
- We design Geometry-aware Spatial Offset Loss that jointly supervises anchor-relative offset regression and Euclidean metric error. This objective aligns model optimization with distance-thresholded metric place recognition.
- We evaluate GeoMPR on several challenging VPR benchmarks using strict 3-meter, 5-meter, and 10-meter localization accuracy. GeoMPR consistently outperforms retrieval-only TPR, with particularly substantial improvements under the 5-meter criterion.

2 Related Work

2.1 Visual Place Recognition

Visual Place Recognition (VPR) is commonly formulated as retrieving the most similar geo-tagged image for a query by using global features. Early approaches relied on hand-crafted features (Bay et al., 2008), extracting global descriptors via aggregation algorithms such as Bag-of-Words (Angeli et al., 2008) or VLAD (Hervé et al., 2010). With the progress of deep learning technology, Numerous works (Sünderhauf et al., 2015; Arandjelovic et al., 2016; Chen et al., 2017; Naseer et al., 2017) have leveraged various deep features to enhance robustness against viewpoint, appearance, and environmental variations in the VPR task. Representative methods include NetVLAD(Arandjelovic et al., 2016), CosPlace(Berton et al., 2022a), MixVPR(Ali-bey et al., 2023), and EigenPlaces(Berton et al., 2023). However, global aggregation may discard spatial structure and remains vulnerable to perceptual aliasing. Two-stage visual place recognition methods (Cao et al., 2020; Hausler & Milford, 2020; Hausler et al., 2021; Wang et al., 2022; Berton et al., 2021; Shen et al., 2022; Zhu et al., 2023) reorder the top-k retrieved candidates using local descriptors or geometric verification, thereby partially alleviating this issue. Although these methods improve candidate ranking, their final prediction is generally still the coordinate of a selected database image. Recent foundation-model-based methods, such as AnyLoc(Keetha et al., 2023) and SelaVPR(Lu et al., 2024b), further exploit transferable visual representations for retrieval and re-ranking. Moreover, the work(Izquierdo & Civera, 2024a) shows that embedding distances may poorly reflect geographic distances at short ranges, resulting in the geographically closest reference image often not being the top-ranked image. We refer to

these retrieval-centered approaches as Topological Place Recognition. In contrast, our Metric Place Recognition retains retrieval as a global prior and uses multi-view geometry to refine the retrieved reference into a precise geographic coordinate.

2.2 Feed-Forward 3D Foundation Models

Building on conventional 3D reconstruction, recent end-to-end learning-based methods (Long et al., 2024; Lu et al., 2025; Murai et al., 2025; Reizenstein et al., 2021; Smart et al., 2024; Szymanowicz et al., 2025; Yang et al., 2025; Zhang et al., 2025; 2024) encode scene priors via neural networks, significantly boosting robustness and cross-dataset generalization. DUST3R(Wang et al., 2024) pioneered regressing view-consistent 3D point maps from two RGB images without camera calibration, while its successor MASt3R(Murai et al., 2025) refines metric scale estimation via confidence-weighted losses. For single-pass dense-view interaction, Fast3R(Yang et al., 2025) extends to N-view processing with Flash-Attention and parallel fusion, supporting 1000+ image inference and SOTA accuracy in seconds. State-of-the-art VGGT (Wang et al., 2025) jointly predicts camera parameters, depth, and point maps in one forward pass, outperforming task-specific methods. Nevertheless, the camera centres and reconstructed points produced by feed-forward multi-view models are expressed in local coordinate systems with potentially arbitrary scales and orientations. Thus, they cannot directly provide real-world geographic coordinates. In our work, we refer to the top-ranked reference image as the anchor image. VGGT serves as the geometric backbone of GeoMPR to obtain the camera geometry of the reference images and the query image relative to the anchor image. Then, the spatial offset of the query image relative to the anchor image is predicted using SOEP. Finally, the real-world coordinates of the query image are obtained by adding the predicted spatial deviation to the real-world coordinates of the anchor image. The recovered geometry is therefore used as an intermediate representation for metric refinement.

3 Method

3.1 Overview

Fig. 1 illustrates the proposed GeoMPR framework. Given a query image, GeoMPR first retrieves a set of geo-tagged candidate images. The top-ranked image is selected as the anchor image. Next, geographic filtering retains reference images within a predefined geographic coordinate distance threshold of the anchor image to form a geographically closest reference set (Sec. 3.2). Then, VGGT is used to recover the relative camera geometry between the query image and the geographically consistent reference images with respect to the anchor image, and simultaneously, we compute the spatial offset based on the known geographic coordinates of the reference images (Sec. 3.3). Based on these geometric representations, the proposed Spatial Offset Estimation Module (SOEP) predicts the spatial offset from the query image to the anchor image (Sec.3.4). Finally, the query coordinate is recovered by combining the anchor image coordinate and the predicted spatial offset (Sec.3.5). Furthermore, we present the loss for fine-tuning (Sec.3.6).

3.2 Geographic Anchor-based Reference Construction

Although retrieval-based VPR provides an effective place-level hypothesis, the retrieved candidates may not form a geographically consistent neighborhood due to perceptual aliasing and visual similarity between different locations. Directly using all retrieved images for geometry-grounded refinement may introduce noisy correspondences and degrade the estimation of spatial displacement. Therefore, we construct a geographically coherent reference set based on the anchor image.

Given a geo-tagged database $\mathcal{G} = \{(I_j, g_j)\}_{j=1}^{M}$, where I_j denotes the database image and $g_j \in \mathbb{R}^2$ represents its geographic coordinate, a VPR model first retrieves an ordered top-k candidate set for a query image I_q: $\mathcal{D}_k(I_q) = \{(I_0, g_0), (I_1, g_1), \cdots, (I_{k-1}, g_{k-1})\}$. The top-ranked retrieval (I_0, g_0) is selected as the anchor image. To provide reliable geometric constraints for subsequent spatial offset estimation, we further filter the retrieved candidates according to their geographic proximity to the anchor. Specifically, we retain candidates whose distance from the anchor is smaller than a predefined threshold τ:

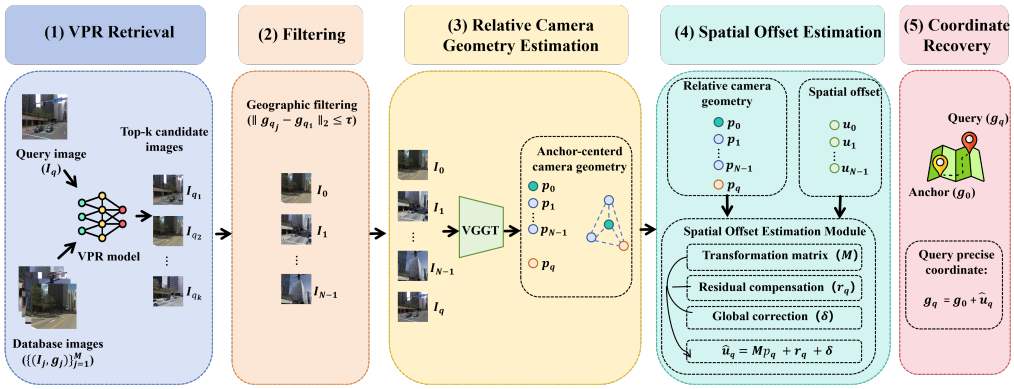

Figure 1: Overview of GeoMPR. (1) VPR model retrieves top-k geo-tagged candidate images for the query image. (2) Geographic filtering selects a locally coherent reference set based on the geographic distance from the anchor. (3) VGGT estimates the query and reference images' camera centers in an anchor-centered local coordinate system, while the reference images provide their anchor-relative spatial offsets. (4) The SOEP maps the query's relative camera geometry to its spatial offset, then refines the estimate through residual compensation and global correction. (5) The query's geographic coordinate is recovered by adding the predicted offset to the anchor coordinate.

$$\mathcal{R}(I_q) = \{(I_i, g_i) \in \mathcal{D}_k(I_q) \mid \|g_i - g_0\|_2 < \tau\}. \tag{1}$$

This operation transforms visually ranked retrieval results into a local geographic neighborhood, providing more reliable information for subsequent estimation of spatial offset.

3.3 Relative Camera Geometry Estimation and Spatial Offset computation

In this section, we employ VGGT to estimate the relative camera geometry between the query and reference images. VGGT jointly processes the selected reference images and the query image, and predicts their camera geometry in a common local coordinate system with the anchor image as the origin. Subsequently, we derive the camera center for each image based on the output of VGGT. Let p_q and p_i denote the camera centers of the query image and the i-th reference image, respectively, expressed relative to the anchor. Since the geographic coordinates of the reference images are known, their each anchor-relative spatial offset $\mathbf{u}_i$ are computed as:

$$\mathbf{u}_i = g_i - g_0. \tag{2}$$

where i denotes the i-th reference image. Based on these camera centers and spatial offsets, feature representations for each reference image are constructed to facilitate the estimation of the spatial offset for the query image.

3.4 Spatial Offset Estimation Module

Based on the previously obtained relative camera geometry and relative spatial offset, the Spatial Offset Estimation Module (SOEP) estimates the spatial offset between the query image and the anchor image. Fig. 2 illustrates the proposed Spatial Offset Estimation Module framework.

3.4.1 Preliminary

Based on previous relative camera centers and spatial offsets, we construct a reference image feature ϕ_i for each reference image.

$$\phi_i = [p_i,\ u_i,\ p_q - p_i,\ \|p_q - p_i\|_2,\ \|p_i\|_2,\ \|u_i\|,\ p_q,\ p_i^T p_q,\ \cos(p_q, p_i)]. \tag{3}$$

where $\cos(p_q, p_i)$ represents the cosine similarity between p_q and p_i. Then, an MLP followed by a Transformer(Lee et al., 2019) generates dense reference image features:

$$\{x_i\}_{i=0}^{N-1} = \text{Transformer}(\{\text{MLP}(\phi_i)\}_{i=0}^{N-1}). \tag{4}$$

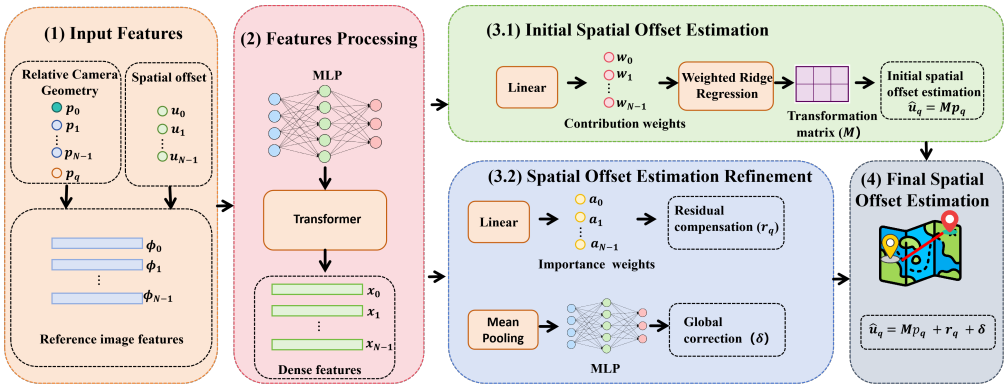

Figure 2: Framework of the Spatial Offset Estimation Module (SOEP). (1) **Input Features**: anchor-relative camera geometry, geographic offsets, and query-reference relationships are combined to construct reference image features. (2) **Feature Processing**: an MLP and a Transformer encode the reference features into dense representations. (3.1) **Initial Offset Estimation**: a prediction head estimates the contribution weights of each reference image, and weighted ridge regression learns the mapping from relative camera geometry to the spatial offset. (3.2) **Offset Refinement**: a prediction head estimates residual compensation, while mean-pooled reference features are used to predict a global correction by an MLP. (4) **Final Offset Estimation**: the initial offset, residual compensation, and global correction are combined to obtain the query's spatial offset relative to the anchor.

where N denotes the number of reference images, x_i represents the dense feature of the i-th reference image.

3.4.2 Initial Spatial Offset Estimation

The geometry predicted by VGGT is defined in an arbitrary local coordinate system and cannot be directly converted into geographic displacement. Therefore, we learn a geometry-to-offset transformation using weighted ridge regression. The Spatial Offset Estimation Estimation (SOEP) estimates the contribution weight w_i of each reference image through a reliability prediction head h_w:

$$w_i = \frac{\exp(h_w(x_i))}{\sum_{j=0}^{N-1} \exp(h_w(x_j))}. \tag{5}$$

Using the predicted contribution weights, the SOEP estimates the transformation matrix M from relative camera center to spatial offset via differentiable weighted ridge regression. Based on the transformation matrix M, the initial spatial offset $\hat{u}_q$ between the query image and the anchor can be obtained as:

$$\hat{u}_q = Mp_q. \tag{6}$$

3.4.3 Spatial Offset Estimation Refinement

Although the transformation matrix M provides an initial spatial offset estimate, VGGT predictions may contain local geometric drift, scale distortion, and pose noise. To compensate for these errors, SOEP further refines the initial spatial offset estimation $\hat{u}_q$ using residual compensation and global correction.

For each reference image, the residual between the ground-truth geographic offset and the estimated offset is computed as:

$$r_i = u_i - Mp_i. \tag{7}$$

Through a prediction head h_a, we derive the importance weights a_i for the offsets of each reference image, which are subsequently used to calculate the residual compensation r_q:

$$r_q = \sum_{i=0}^{N-1} a_i r_i, \quad a_i = \frac{\exp(h_a(x_i))}{\sum_{j=0}^{N-1} \exp(h_a(x_j))}. \tag{8}$$

Meanwhile, a global correction prediction head outputs a global correction term δ from the aggregated reference representation. The final predicted spatial offset of the query image relative to the anchor image is:

$$\hat{u}_q = Mp_q + r_q + \delta. \tag{9}$$

3.5 Geographic Coordinate Recovery

The Spatial Offset Estimation Module predicts anchor-relative spatial offset $\hat{u}_q \in \mathbb{R}^2$ of the query image. Let $g_0 \in \mathbb{R}^2$ denote the geo-tagged coordinate of the anchor image. The final metric geographic coordinate of the query image is recovered as:

$$\hat{g}_q = g_0 + \hat{u}_q. \tag{10}$$

Therefore, VGGT does not directly predict an absolute geographic coordinate. Its local camera geometry is used as an intermediate representation for estimating the spatial offset.

3.6 Geometry-aware Spatial Offset Loss

Existing VPR objectives mainly optimize visual similarity, while Metric Place Recognition requires accurate estimation of the spatial displacement between the query image and anchor images. Therefore, we formulate Geometry-aware Spatial Offset Loss to optimize the trainable components of GeoMPR.

Let $\hat{u}_q$ denote the predicted anchor-relative spatial offset from SOEP, and u_q^* denote the ground-truth offset computed from the query and anchor coordinates. We first apply an offset regression loss L_{offset}:

$$L_{offset} = \frac{1}{B}\sum_{b=1}^{B}\sum_{d=1}^{2}\mathrm{SmoothL1}_{\beta}\left(\hat{u}_{q,d}^{(b)} - u_{q,d}^{*(b)}\right). \tag{11}$$

where B is the batch size, b indexes a training sample, $d \in \{1, 2\}$ indexes the spatial dimensions, and β controls the transition point of the Smooth-L_1 function. We further introduce a Euclidean metric loss L_{metric}:

$$L_{metric} = \frac{1}{B}\sum_{b=1}^{B}\left\|\hat{u}_q^{(b)} - u_q^{*(b)}\right\|_2. \tag{12}$$

This metric loss directly reflects the evaluation objective of Metric Place Recognition, where the prediction accuracy is determined by whether the estimated coordinate falls within a specific geographic distance threshold. The Geometry-aware Spatial Offset Loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{offset} + \lambda_{metric}\mathcal{L}_{metric}. \tag{13}$$

where λ_{metric} balances the contribution of the metric supervision. By jointly optimizing the offset regression loss and the Euclidean metric loss, the proposed objective encourages SOEP to learn spatially meaningful geometric relationships, enabling more accurate metric place recognition.

4 Experiments

4.1 Datasets and Performance Evaluation

Datasets. We evaluate our method on several widely used VPR benchmark datasets, which encompass diverse challenges arising from viewpoint and environmental condition variations. A detailed overview of the datasets used for evaluation is provided in Table 1. Specifically, **MSLS** (Warburg et al., 2020) is a challenging large-scale dataset comprising images collected across urban, suburban, and natural environments, with substantial variations in visual appearance. **Pitts30k** (Torii et al., 2013), derived from Google Street View, primarily features significant viewpoint variations. In contrast, **Tokyo24/7** (Torii et al., 2015) is characterized by pronounced changes in illumination conditions, making it particularly suitable for evaluating robustness against appearance variations.

Table 1: Summary of the evaluation datasets.

Dataset	Description	Number	
		Database	Queries
Pitts30k	urban, panorama	10,000	6,816
MSLS-val	urban, suburban	18,871	740
Tokyo24/7	urban, day/night	75,984	315
SVOX	urban, multi-domain	17166	14278

Baidu (Sun et al., 2017) was captured in a modern indoor shopping mall, mainly presenting viewpoint variations. **SVOX** (Moreno Berton et al., 2021) mainly presents severe domain shifts caused by variations in weather, time, and viewpoint. More details can be found in App. C. **Performance Evaluation**. Since GeoMPR aims to estimate the metric geographic position of a query image, we evaluate the place recognition accuracy based on the distance between the predicted coordinate and the ground-truth coordinate. Specifically, a prediction is considered correct if the geographic distance between them is within a predefined threshold. Following previous VPR evaluation protocols, we report the percentage of correctly localized queries under three strict distance thresholds: 5 meters, 10 meters, and 15 meters.

4.2 Implementation Details

All experiments are implemented in PyTorch and conducted on an NVIDIA RTX A6000 GPU. UTM coordinates are employed as the standard geographic representation for all images in the subsequent experiments. We fine-tune the Spatial Offset Estimation Module and the last four transformer blocks of VGGT Wang et al. (2025) with the previous layers frozen. We employ the loss in Section 3.6 for training on the GSV-Cities (Ali-bey et al., 2022) dataset. Since the original labels in the GSV-Cities dataset are in latitude and longitude, we convert them into UTM coordinates before training. Each training sample comprises a query image and four randomly chosen database images whose coordinate distance to the query is less than 20 meters. The model is trained using the Adam optimizer with an initial learning rate of 1e-4. Batch size is 8. Each training batch contains. Besides, we set the epochs to 10. More implementation details are provided in the Appendix B.

4.3 Comparisons with TPR Approaches

In Table 2, compares GeoMPR with retrieval-based Topological Place Recognition (TPR) methods, including NetVLAD (Arandjelovic et al., 2016), BoQ (Ali-Bey et al., 2024), EDTformer (Jin et al., 2025), SALAD (Izquierdo & Civera, 2024b), SuperVLAD (Lu et al., 2024c), SFRS (Ge et al., 2020), MixVPR (Ali-bey et al., 2023), CosPlace (Berton et al., 2022a), CricaVPR (Lu et al., 2024a), EigenPlaces (Berton et al., 2023), and ImAge (Lu et al., 2026), under strict distance thresholds. GeoMPR uses ImAge as the retrieval backbone. The corresponding TPR baseline assigns the coordinate of the top-ranked retrieved image to the query. We report the percentage of queries whose predicted coordinates fall within 5, 10, and 15 meters of the ground-truth coordinates. GeoMPR achieves the highest or tied-highest accuracy across the evaluated datasets. The most notable improvements over the strongest competing TPR method occur on Tokyo24/7 at 5 m, where the accuracy increases by 35.8 percentage points, Pitts30k at 5 m and 10 m, with gains of 19.4 and 8.9 points, respectively, and SVOX at 5 m, with a gain of 16.0 points. The largest gain is obtained on Tokyo24/7 under the strict 5-meter threshold. Its substantial day–night appearance changes may cause retrieval methods to select a visually correct image whose database coordinate is spatially displaced from the query. By using the retrieved image as an anchor and estimating the anchor-relative offset from nearby references, GeoMPR can correct this displacement and achieve more accurate metric localization.

Meanwhile, we evaluate the performance gains over retrieval-only baselines under various distance thresholds. As shown in Fig. 3, GeoMPR achieves limited improvements under relaxed distance thresholds. However, the performance gains become increasingly significant as the threshold becomes stricter, since directly assigning the retrieved image coordinate cannot capture the spatial displacement between the query and reference. These results demonstrate that geometry-grounded offset estimation effectively complements different VPR retrieval models for fine-grained metric place recognition.

Table 2: Comparison to the TPR methods on benchmark datasets. The best is highlighted in **bold** and the second is underlined. Notably, GeoMPR adopts ImAge as the VPR model.

Method	Tokyo24/7			SVOX			Pitts30k			MSLS-val		
	5m	10m	15m	5m	10m	15m	5m	10m	15m	5m	10m	15m
NetVLAD	17.8	42.2	48.6	61.8	73.2	77.9	35.4	68.6	77.6	13.9	32.4	43.2
SFRS	27.3	57.1	64.4	67.4	78.9	81.4	38.6	74.9	83.4	15.4	38.4	51.4
MixVPR	37.1	79.0	83.5	79.8	94.0	96.8	42.1	80.7	89.2	24.1	55.1	71.8
CosPlace	29.2	61.3	69.8	77.7	92.0	94.7	40.1	77.6	86.1	23.8	52.2	69.9
EigenPlaces	31.1	67.6	76.2	80.1	94.1	96.5	40.7	78.9	87.2	24.7	54.7	72.4
CricaVPR	34.9	76.8	88.3	75.8	91.6	95.4	42.9	81.9	87.6	25.9	56.9	74.3
SALAD	36.8	82.5	90.8	79.1	94.1	97.2	40.7	78.7	88.7	26.9	59.5	77.6
BoQ	43.5	87.6	95.2	80.5	95.2	98.1	42.5	81.3	90.8	26.9	59.7	79.1
SuperVLAD	39.7	81.9	89.2	78.4	93.6	96.9	41.5	79.3	88.9	27.4	58.9	76.6
EDTformer	40.6	86.0	94.9	78.7	93.9	97.3	40.9	79.5	89.4	27.0	59.5	78.1
ImAge	36.8	86.0	95.2	78.8	94.4	97.8	41.5	80.2	90.3	28.8	62.8	80.9
GeoMPR (ours)	**79.3**	**92.3**	**95.8**	**96.5**	**98.3**	**98.5**	**62.3**	**90.8**	**91.6**	**34.0**	**69.3**	**86.0**

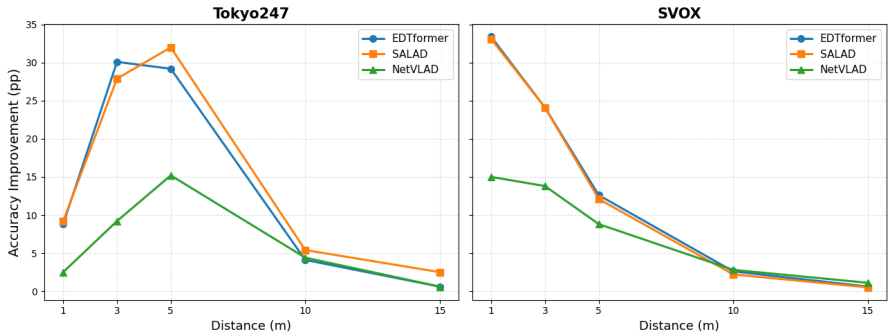

Figure 3: Performance gains of GeoMPR with different VPR retrieval models on Tokyo24/7 and MSLS-val under various localization distance thresholds. Here, pp stands for percentage points.

4.4 Ablation Studies

Effectiveness with Different VPR Retrieval Models. To investigate whether GeoMPR can generalize across different retrieval-based VPR models, we replace the VPR component in GeoMPR with several representative retrieval methods and compare their performance with the corresponding retrieval-only baselines. Specifically, each VPR model first retrieves the top-ranked reference image and directly assigns its coordinate as the predicted location, while GeoMPR further estimates the spatial offset between the query and the retrieved anchor to obtain a refined metric position. As shown in Table 3, integrating GeoMPR consistently improves the localization accuracy of different VPR models across all evaluated benchmarks. These results demonstrate that GeoMPR is not dependent on a specific retrieval architecture; instead, it can effectively leverage different retrieval models as place-level anchors and improve their localization accuracy through geometry-grounded spatial offset estimation. More results are provided in the App. B.

Effect of VGGT Fine-tuning Strategies. To investigate whether adapting the geometry foundation model can further improve Metric Place Recognition, we evaluate GeoMPR with different VGGT fine-tuning strategies. Specifically, we compare the original frozen VGGT with partially fine-tuned variants, where different blocks of VGGT are updated during training while the remaining components are kept fixed. As shown in Table 4, fine-tuning VGGT consistently improves the localization accuracy of GeoMPR compared with using a fully frozen geometry model. This indicates that adapting the geometric representations to the target VPR scenarios can provide more accurate spatial cues for the Spatial Offset Estimation Module (SOEP). The results demonstrate that although pretrained vision geometry models already provide strong geometric priors, task-specific adaptation is beneficial for achieving more accurate metric place recognition.

Effect of Spatial Offset Refinement Components. To investigate the contribution of the offset refinement components in SOEP, we evaluate the impact of residual compensation and global correc-

Table 3: Effectiveness of GeoMPR with Different VPR Retrieval Models.

Method	Pitts30k		MSLS-val		Tokyo24/7		SVOX		Baidu	
	5m	10m	5m	10m	5m	10m	5m	10m	5m	10m
NetVLAD	35.4	68.6	13.9	32.4	17.8	42.2	61.8	73.2	31.3	44.2
GeoMPR (ours)	**46.0**	**76.6**	**19.0**	**37.5**	**33.0**	**46.6**	**70.6**	**76.0**	**34.3**	**47.5**
BoQ	42.5	81.3	26.9	59.7	43.5	87.6	80.5	95.2	47.0	69.7
GeoMPR (ours)	**53.3**	**86.5**	**30.1**	**64.1**	**70.7**	**92.0**	**91.9**	**97.1**	**56.2**	**71.1**
EDTformer	40.9	79.5	27.0	59.5	40.6	86.0	78.7	93.9	45.4	71.1
GeoMPR (ours)	**53.3**	**85.6**	**30.0**	**63.1**	**69.8**	**90.1**	**91.3**	**96.5**	**58.6**	**75.2**
SALAD	40.7	78.7	26.9	59.5	36.8	82.5	79.1	94.1	40.3	62.1
GeoMPR (ours)	**52.3**	**85.0**	**30.0**	**62.7**	**68.8**	**87.9**	**91.2**	**96.3**	**49.3**	**64.3**

Table 4: Comparison of different VGGT fine-tuning strategies for GeoMPR.

Method	Pitts30k		MSLS-val		Tokyo24/7		SVOX		Baidu	
	5m	10m	5m	10m	5m	10m	5m	10m	5m	10m
w/o training	53.0	86.6	30.9	67.2	73.9	92.3	91.7	96.9	44.6	57.7
w/ trianing (ours)	**62.3**	**90.8**	**34.0**	**69.3**	**79.3**	**92.3**	**96.5**	**98.3**	**52.4**	**59.8**

tion on GeoMPR performance. After obtaining the initial spatial offset through geometry-to-offset alignment, SOEP further introduces two complementary correction mechanisms: residual compensation reduces reference-specific geometric estimation errors, while global correction learns a global adjustment to compensate for systematic offset bias. As shown in Table 5, removing either residual compensation or global correction degrades the localization accuracy, demonstrating that both components provide complementary benefits for metric place recognition. Residual compensation improves the adaptation to local geometric variations among reference images, whereas global correction further refines the overall offset estimation. By combining both components, SOEP achieves the most accurate spatial offset prediction and improves the final GeoMPR localization performance.

Table 5: Ablation study of spatial offset refinement components in SOEP.

Method	Pitts30k		MSLS-val		Tokyo24/7		SVOX		Baidu	
	5m	10m	5m	10m	5m	10m	5m	10m	5m	10m
w/o Residual Compensation	52.6	87.2	32.1	67.1	69.2	91.1	93.2	97.1	49.0	59.2
w/o Global Correction	53.5	86.6	30.8	67.4	73.9	92.3	92.3	96.9	44.6	57.7
GeoMPR (Full)	**62.3**	**90.8**	**34.0**	**69.3**	**79.3**	**92.3**	**96.5**	**98.3**	**52.4**	**59.8**

5 CONCLUSION

In this paper, we introduced Metric Place Recognition and proposed GeoMPR, a geometry-grounded framework that extends conventional retrieval-based Topological Place Recognition toward fine-grained geographic position estimation. By treating the retrieved image as a reliable geographic anchor rather than the final prediction, GeoMPR combines visual retrieval with geometry-based spatial offset estimation to recover the displacement between the query and reference images. Furthermore, we designed a geometry-aware metric learning objective to optimize spatial offset estimation and improve metric localization accuracy. Extensive experiments on five challenging VPR benchmarks demonstrate that GeoMPR significantly improves place recognition performance under strict distance thresholds, showing the effectiveness of geometry-grounded refinement.

AI USE STATEMENT

Generative AI tools were employed exclusively for supportive purposes, including language polishing, enhancing the clarity and organization of the manuscript, and assisting with the design of figures and tables. All core contributions—encompassing research conceptualization, methodological decisions, implementation, experimentation, analysis, and conclusions—were independently conducted and rigorously verified by the authors. The authors assume full responsibility for the integrity and accuracy of this work, including all AI-assisted text and artifacts.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we provide comprehensive descriptions of the proposed method, model architecture, training objectives, and inference procedures in the main text. Detailed implementation specifics are documented in the Appendix. All datasets utilized in our experiments are publicly accessible via the official sources cited in the references. Furthermore, we include thorough descriptions of the datasets and data preprocessing pipelines in the Appendix to facilitate consistent evaluation. To fully support reproducibility, the source code and configuration files will be made publicly available upon publication.

ETHICS STATEMENT

This study does not involve human subjects, nor does it utilize any personal or sensitive data. All experiments are conducted exclusively on publicly available benchmark datasets and publicly released pretrained models. The proposed methods and released artifacts do not present any foreseeable risks of harmful application, and the authors declare no conflicts of interest. This work strictly adheres to the ICLR Code of Ethics.

REFERENCES

Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.09.127.

Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguere. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2998–3007, 2023.

Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Boq: A place is worth a bag of learnable queries. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17794–17803. IEEE, 2024.

Adrien Angeli, David Filliat, Stéphane Doncieux, and Jean-Arcady Meyer. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE transactions on robotics*, 24 (5):1027–1037, 2008.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. doi: 10.1109/CVPR.2016.572.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

Gabriele Berton, Carlo Masone, Valerio Paolicelli, and Barbara Caputo. Viewpoint invariant dense matching for visual geolocalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12149–12158. IEEE, 2021.

Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4878–4888, 2022a. doi: 10.1109/CVPR52688.2022.00483.

Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5386–5397. IEEE, 2022b.

Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11080–11090, 2023. doi: 10.1109/ICCV51070.2023.01017.

Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *European conference on computer vision*, pp. 726–743. Springer, 2020.

Zetao Chen, Adam Jacobson, Niko Sünderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford. Deep learning features at scale for visual place recognition. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3223–3230. IEEE, 2017.

Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *European conference on computer vision*, pp. 369–386. Springer, 2020.

Stephen Hausler and Michael Milford. Hierarchical multi-process fusion for visual place recognition. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3327–3333. IEEE, 2020.

Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias 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*, pp. 14141–14152, 2021. doi: 10.1109/CVPR46437.2021.01392.

Jégou Hervé, Douze Matthijs, S Cordelia, and P Patrick. Aggregating local descriptors into a compact image representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3304–3311, 2010.

Sergio Izquierdo and Javier Civera. Close, but not there: Boosting geographic distance sensitivity in visual place recognition. In *European Conference on Computer Vision*, pp. 240–257. Springer, 2024a.

Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17658–17668. IEEE, 2024b.

Tong Jin, Feng Lu, Shuyu Hu, Chun Yuan, and Yunpeng Liu. Edtformer: An efficient decoder transformer for visual place recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):8835–8848, 2025.

Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 9(2):1286–1293, 2023.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Liu Liu, Hongdong Li, and Yuchao Dai. Stochastic attraction-repulsion embedding for large scale image localization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2570–2579. IEEE, 2019.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9970–9980. IEEE, 2024.

Feng Lu, Xiangyuan Lan, Lijun Zhang, Dongmei Jiang, Yaowei Wang, and Chun Yuan. Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16772–16782. IEEE, 2024a.

Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. In *International conference on learning representations*, volume 2024, pp. 14314–14335, 2024b.

Feng Lu, Xinyao Zhang, Canming Ye, Shuting Dong, Lijun Zhang, Xiangyuan Lan, and Chun Yuan. Supervlad: Compact and robust image descriptors for visual place recognition. *Advances in Neural Information Processing Systems*, 37:5789–5816, 2024c.

Feng Lu, Tong Jin, Canming Ye, Xiangyuan Lan, Yunpeng Liu, and Chun Yuan. Towards implicit aggregation: Robust image representation for place recognition in the transformer era. *Advances in Neural Information Processing Systems*, 38:85096–85124, 2026.

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22820–22830. IEEE, 2025.

Gabriele Moreno Berton, Valerio Paolicelli, Carlo Masone, and Barbara Caputo. Adaptive-attentive geolocalization from few queries: a hybrid approach. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2917–2926, 2021. doi: 10.1109/WACV48630.2021.00296.

Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16695–16705. IEEE, 2025.

Tayyab Naseer, Gabriel L Oliveira, Thomas Brox, and Wolfram Burgard. Semantics-aware visual localization under challenging perceptual conditions. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2614–2620. IEEE, 2017.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10881–10891. IEEE, 2021.

Yanqing Shen, Ruotong Wang, Weiliang Zuo, and Nanning Zheng. Tcl: Tightly coupled learning strategy for weakly supervised hierarchical place recognition. *IEEE Robotics and Automation Letters*, 7(2):2684–2691, 2022.

Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: zero-shot gaussian splatting from uncalibrated image pairs (2024). *URL https://arxiv. org/abs/2408.13912*, 1:2, 2024.

Xun Sun, Yuanfan Xie, Pei Luo, and Liang Wang. A dataset for benchmarking image-based localization. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5641–5649. IEEE, 2017.

Niko Sünderhauf, Sareh Shirazi, Feras Dayoub, Ben Upcroft, and Michael Milford. On the performance of convnet features for place recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 4297–4304. IEEE, 2015.

Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, Joao F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. In *2025 International Conference on 3D Vision (3DV)*, pp. 670–681. IEEE, 2025.

Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025. doi: 10.1109/CVPR52734.2025.00499.

Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. Transvpr: Transformer-based place recognition with multi-level attention aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13648–13657, 2022. doi: 10.1109/CVPR52688.2022.01328.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024. doi: 10.1109/CVPR52733.2024.01956.

Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935. IEEE, 2025.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Forrester Cole, Deqing Sun, Ming-Hsuan Yang, et al. Monst3r: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations*, volume 2025, pp. 82863–82886, 2025.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024.

Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. $r\hat{}\{2\}$ former: Unified retrieval and reranking transformer for place recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19370–19380. IEEE, 2023.

A More method implementation details

It is worth noting that the relative camera geometry and geographic coordinates mentioned in the main text refer to the 3D relative camera center and the 2D UTM coordinates, respectively.

A.1 Spatial Offset Estimation Module

A.1.1 Relative Camera Geometry Estimation and Spatial Offset computation

The selected reference images and the query image are arranged as

$$\mathcal{I} = [I_0, I_1, \ldots, I_{N-1}, I_q] \tag{14}$$

and jointly input to VGGT. The anchor image I_0 is placed at the first position so that the predicted camera geometry is expressed relative to the anchor image. For each input image, VGGT predicts a pose encoding $\mathbf{e}_i$:

$$\mathbf{e}_i = [\mathbf{t}_i, \mathbf{q}_i, \mathbf{f}_i], \tag{15}$$

where $\mathbf{t}_i \in \mathbb{R}^3$ is the translation vector, $\mathbf{q}_i \in \mathbb{R}^4$ is the rotation quaternion, and $\mathbf{f}_i \in \mathbb{R}^2$ represents the field-of-view parameters. After converting $\mathbf{q}_i$ into a rotation matrix $\mathbf{R}_i$, the camera centre is calculated as

$$\mathbf{c}_i = -\mathbf{R}_i^{\mathsf{T}} \mathbf{t}_i. \tag{16}$$

The reference and query camera centres are expressed relative to the anchor:

$$\mathbf{p}_i = \mathbf{c}_i - \mathbf{c}_0, \qquad i = 0, \ldots, N-1, \tag{17}$$

$$\mathbf{p}_q = \mathbf{c}_q - \mathbf{c}_0. \tag{18}$$

Similarly, the geographic coordinates of reference images are converted into spatial offset:

$$\mathbf{u}_i = \mathbf{g}_i - \mathbf{g}_0, \qquad i = 0, \ldots, N-1. \tag{19}$$

By construction:

$$\mathbf{p}_0 = \mathbf{0}, \qquad \mathbf{u}_0 = \mathbf{0}. \tag{20}$$

Because both coordinate systems share the same anchor origin, no independent translation term is required. The spatial offset problem can therefore be formulated as

$$\mathbf{u}_i \approx \mathbf{M}\mathbf{p}_i, \qquad \mathbf{M} \in \mathbb{R}^{2\times 3}, \tag{21}$$

where $\mathbf{M}$ maps the local 3D camera coordinate system onto the 2D UTM plane.

A.1.2 Scale Normalization

The local 3D reconstruction produced by VGGT does not necessarily have a fixed metric scale. To reduce the influence of scene-dependent scale variations, the 3D camera centres and UTM coordinates are normalized independently. The corresponding RMS scales are:

$$s_3 = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} \|\mathbf{p}_i\|_2^2}, \tag{22}$$

$$s_2 = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} \|\mathbf{u}_i\|_2^2}. \tag{23}$$

The normalized coordinates are defined as

$$\widetilde{\mathbf{p}}_i = \frac{\mathbf{p}_i}{\max(s_3, \epsilon)}, \qquad \widetilde{\mathbf{p}}_q = \frac{\mathbf{p}_q}{\max(s_3, \epsilon)}, \qquad \widetilde{\mathbf{u}}_i = \frac{\mathbf{u}_i}{\max(s_2, \epsilon)}, \tag{24}$$

where ϵ is a small positive constant used to prevent numerical instability. By default, p_i, p_q, and u_i are normalized in the main text.

A.1.3 Weighted Geometric Alignment

The encoded features are used to predict a reliability score for each reference image. The normalized regression weight of the i-th reference is defined as

$$w_i = \frac{\exp(\ell_i)}{\sum_{j=0}^{N-1} \exp(\ell_j)}, \qquad \ell_i = h_w(\mathbf{x}_i), \qquad i = 0, \ldots, N-1, \tag{25}$$

where $h_w(\cdot)$ is a linear prediction head and N is the number of reference images retained after geographic filtering. The normalized weights satisfy

$$w_i \geq 0, \qquad \sum_{i=0}^{N-1} w_i = 1. \tag{26}$$

The learned weights allow SOEP to adaptively control the contribution of each reference image. References that are more geometrically consistent with the remaining images can be assigned larger weights, whereas references affected by retrieval errors or inaccurate VGGT pose estimation can be assigned smaller weights.

Based on the predicted weights, the weighted normal matrix $\mathbf{A} \in \mathbb{R}^{3\times3}$ and cross-covariance matrix $\mathbf{C} \in \mathbb{R}^{2\times3}$ are calculated as

$$\mathbf{A} = \sum_{i=0}^{N-1} w_i \widetilde{\mathbf{p}}_i \widetilde{\mathbf{p}}_i^{\mathsf{T}}, \tag{27}$$

$$\mathbf{C} = \sum_{i=0}^{N-1} w_i \widetilde{\mathbf{u}}_i \widetilde{\mathbf{p}}_i^{\mathsf{T}}. \tag{28}$$

The 3D-to-2D mapping matrix is obtained through differentiable weighted ridge regression:

$$\mathbf{M} = \mathbf{C}\left(\mathbf{A} + \lambda \mathbf{I}_3\right)^{-1}, \tag{29}$$

where $\mathbf{I}_3$ is the 3×3 identity matrix and λ is the ridge regularization coefficient. To ensure that λ remains positive, it is parameterized as

$$\lambda = \lambda_{\text{floor}} + \text{softplus}(\rho), \tag{30}$$

where $\lambda_{\text{floor}} > 0$ is a fixed lower bound and ρ is a learnable scalar.

Rather than explicitly computing the matrix inverse in Eq. equation 29, the mapping matrix is obtained by solving the corresponding linear system, which provides better numerical stability and remains fully differentiable.

The initial spatial offset prediction of the query location is then

$$\widetilde{\mathbf{u}}_{q,\text{base}} = \mathbf{M}\widetilde{\mathbf{p}}_q. \tag{31}$$

A.1.4 Residual Refinement

A single global linear mapping may not fully describe the relationship between the VGGT reconstruction and the UTM plane because the estimated camera geometry may contain local drift, scale distortion, or pose noise. To compensate for these local residuals, the fitting residual of each reference image is calculated as

$$\mathbf{r}_i = \widetilde{\mathbf{u}}_i - \mathbf{M}\widetilde{\mathbf{p}}_i, \qquad i = 0, \ldots, N-1. \tag{32}$$

A second prediction head estimates an interpolation score for each reference. The normalized interpolation weights are defined as

$$a_i = \frac{\exp(\alpha_i)}{\sum_{j=0}^{N-1} \exp(\alpha_j)}, \qquad \alpha_i = h_a(\mathbf{x}_i), \qquad i = 0, \ldots, N-1, \tag{33}$$

where $h_a(\cdot)$ is a linear prediction head. These weights satisfy

$$a_i \geq 0, \qquad \sum_{i=0}^{N-1} a_i = 1. \tag{34}$$

The residual compensation term for the query image is calculated as

$$\mathbf{r}_q = \sum_{i=0}^{N-1} a_i \mathbf{r}_i = \sum_{i=0}^{N-1} a_i \left(\widetilde{\mathbf{u}}_i - \mathbf{M}\widetilde{\mathbf{p}}_i\right). \tag{35}$$

In parallel, the reference features are aggregated by mean pooling:

$$\overline{\mathbf{x}} = \frac{1}{N} \sum_{i=0}^{N-1} \mathbf{x}_i. \tag{36}$$

The pooled reference feature is concatenated with the normalized query camera center and the logarithm of the number of reference images. An MLP then produces a global learned correction:

$$\boldsymbol{\delta} = s_{\text{res}} \, \text{MLP}_{\delta} \left([\overline{\mathbf{x}}, \widetilde{\mathbf{p}}_q, \log N] \right), \tag{37}$$

where s_{res} controls the magnitude of the learned correction. The final spatial offset of query is

$$\widehat{\widetilde{\mathbf{u}}}_q = \underbrace{\mathbf{M}\widetilde{\mathbf{p}}_q}_{\text{global geometric alignment}} + \underbrace{\sum_{i=0}^{N-1} a_i \left(\widetilde{\mathbf{u}}_i - \mathbf{M}\widetilde{\mathbf{p}}_i \right)}_{\text{local residual interpolation}} + \underbrace{\boldsymbol{\delta}}_{\text{learned correction}}. \tag{38}$$

The anchor-relative spatial offset in meters is recovered by restoring the UTM scale:

$$\widehat{\mathbf{u}}_q = s_2 \widehat{\widetilde{\mathbf{u}}}_q. \tag{39}$$

Finally, the absolute UTM coordinate of the query image is obtained by adding the known anchor coordinate:

$$\widehat{\mathbf{g}}_q = \mathbf{g}_0 + \widehat{\mathbf{u}}_q. \tag{40}$$

A.2 Training Objective

During training, the ground-truth query coordinate is also expressed relative to the anchor:

$$\mathbf{u}_q^* = \mathbf{g}_q - \mathbf{g}_0. \tag{41}$$

Let $\widehat{\mathbf{u}}_q \in \mathbb{R}^2$ and $\mathbf{u}_q^* \in \mathbb{R}^2$ denote the predicted and ground-truth anchor-relative UTM coordinates, respectively. The offset regression loss L_{offset} is defined as:

$$L_{offset} = \frac{1}{B} \sum_{b=1}^{B} \sum_{d=1}^{2} \text{SmoothL1}_{\beta} \left(\frac{\widehat{u}_q^{(b)} - u_q^{*(b)}}{s_c} \right), \tag{42}$$

where B is the batch size and β is the transition point between the quadratic and linear regions of the Smooth-L_1 function. The Euclidean metric loss L_{metric}:

$$L_{metric} = \frac{1}{B} \sum_{b=1}^{B} \left\| \hat{u}_q^{(b)} - u_q^{*(b)} \right\|_2. \tag{43}$$

The total training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{offset}} + \lambda \frac{\mathcal{L}_{\text{metric}}}{s_c}, \tag{44}$$

where λ controls the contribution of the metric distance term.

B Additional experimental results

B.1 Effectiveness with Different VPR Retrieval Models

In the main paper, we demonstrate that GeoMPR can consistently improve different retrieval-based VPR models by introducing geometry-grounded spatial offset estimation. Here, we provide additional results with more retrieval models and datasets. As shown in Table 6, GeoMPR consistently improves the localization accuracy of various VPR retrieval models compared with directly assigning the coordinate of the retrieved image. These results further demonstrate that GeoMPR does not rely on a specific retrieval architecture. Instead, different VPR models can serve as effective place-level anchors, while the proposed spatial offset estimation corrects the displacement between the query image and the retrieved reference. This verifies the general applicability of geometry-grounded refinement for bridging topological retrieval and metric place recognition.

Table 6: Additional evaluation of GeoMPR with different VPR retrieval models. GeoMPR consistently improves retrieval-only baselines by estimating the spatial offset between the query and retrieved anchor.

Method	Pitts30k		MSLS-val		Tokyo24/7		SVOX		Baidu	
	5m	10m	5m	10m	5m	10m	5m	10m	5m	10m
SuperVLAD	41.5	79.3	27.4	58.9	39.7	81.9	78.4	93.6	40.9	62.6
GeoMPR (Ours)	**52.7**	**85.3**	**29.7**	**63.9**	**66.6**	**85.3**	**90.6**	**96.0**	**49.2**	**64.6**
SFRS	38.6	74.9	15.4	38.4	27.3	57.1	67.4	78.9	34.6	46.4
GeoMPR (Ours)	**49.2**	**80.7**	**22.9**	**42.5**	**44.1**	**63.1**	**76.4**	**81.2**	**35.3**	**47.1**
MixVPR	42.1	80.7	24.1	55.1	37.1	79.0	79.8	94.0	43.1	62.8
GeoMPR (Ours)	**52.2**	**85.1**	**29.3**	**60.2**	**64.4**	**81.5**	**90.9**	**95.8**	**49.1**	**63.7**
CosPlace	40.1	77.6	23.8	52.2	29.2	61.3	77.7	92.0	25.8	35.5
GeoMPR (Ours)	**50.7**	**83.0**	**28.2**	**58.7**	**52.0**	**68.5**	**88.9**	**94.0**	**26.8**	**37.0**
CricaVPR	42.9	81.9	25.9	56.9	34.9	76.8	75.8	**91.6**	37.0	**58.3**
GeoMPR (Ours)	**50.0**	**82.5**	**27.0**	**55.8**	**63.1**	**80.0**	**84.4**	91.3	**38.5**	53.6
EigenPlaces	40.7	78.9	24.7	54.7	31.1	67.6	80.1	94.1	36.1	50.7
GeoMPR (Ours)	**51.0**	**83.5**	**28.9**	**61.4**	**57.7**	**73.9**	**90.7**	**95.8**	**40.5**	**52.7**

B.2 Performance under Stricter Localization Thresholds

The main experiments evaluate GeoMPR under 5-meter, 10-meter, and 15-meter distance thresholds. To further investigate the capability of GeoMPR for fine-grained metric place recognition, we provide additional results under an even stricter 3-meter localization threshold. As shown in Table 7, retrieval-based Topological Place Recognition (TPR) methods suffer from considerable performance degradation under the strict 3-meter criterion. This is because TPR directly assigns the coordinate of the retrieved database image as the query location and cannot account for the spatial displacement between the query and retrieved reference. In contrast, GeoMPR explicitly estimates the anchor-relative spatial offset through geometry-grounded reasoning and recovers a more accurate geographic position. The results demonstrate that GeoMPR maintains a clear advantage under extremely strict localization requirements. These findings further validate that the proposed Metric Place Recognition framework effectively bridges the gap between topological retrieval and precise geographic position estimation.

C Dataset Details

GSV-Cities (Ali-bey et al., 2022). GSV-Cities is a large-scale dataset developed specifically for supervised visual place recognition (VPR). It contains approximately 560,000 street-view images covering 67,000 geographically distinct locations across more than 40 cities worldwide. With accurate geographic annotations and substantial variations in capture time and visual appearance, the dataset provides a well-suited benchmark for supervised VPR training. In this work, we employ GSV-Cities as one of the main datasets for outdoor training.

Pitts30k (Torii et al., 2013). Pitts30k-test is a widely used subset of the Pittsburgh 250k dataset, which is collected from GPS-tagged Google Street View panoramas in Pittsburgh, Pennsylvania, USA. It contains large-scale urban street-view images covering various city scenes, including roads, bridges, and buildings. The dataset provides multiple viewpoints for each location and exhibits significant variations in viewpoint and illumination, making it a representative benchmark for evaluating VPR methods in urban environments.

MSLS-val (Warburg et al., 2020). MSLS-val is the validation split of the Mapillary Street-Level Sequences (MSLS) dataset, consisting of street-level image sequences collected worldwide using different imaging devices, such as smartphones, dash cameras, and mapping systems. It includes diverse geographic regions and environmental conditions, with variations in cities, seasons, weather,

Table 7: Comparison of GeoMPR and retrieval-based TPR methods under the strict 3-meter localization threshold. GeoMPR achieves more accurate metric place recognition by correcting the spatial offset between retrieved anchors and query images.

Method	Pitts30k	MSLS-val	Tokyo24/7	SVOX	Baidu
NetVLAD	16.2	7.6	5.4	47.4	22.4
SFRS	17.4	8.9	9.5	51.4	25.8
MixVPR	18.7	11.2	12.1	<u>59.4</u>	29.1
CosPlace	18.0	12.0	8.3	57.1	18.5
EigenPlaces	18.5	12.7	11.1	14.1	29.0
CricaVPR	18.1	11.8	10.5	22.5	23.8
SALAD	18.2	11.8	11.4	58.6	25.9
BoQ	<u>19.2</u>	<u>13.1</u>	12.7	32.0	<u>30.8</u>
SuperVLAD	18.4	12.3	11.7	58.2	26.8
EDTformer	18.5	13.0	12.1	58.4	29.0
ImAge	18.6	12.7	<u>13.0</u>	58.3	26.4
GeoMPR (ours)	**21.9**	**15.4**	**44.1**	**84.0**	**31.9**

and acquisition platforms. Due to its large-scale and highly diverse scenarios, MSLS-val is widely adopted for evaluating the robustness and generalization capability of VPR models.

SVOX (Moreno Berton et al., 2021). SVOX is an urban visual place recognition benchmark specifically designed to evaluate model robustness under cross-domain and environmental variations. Its gallery is composed of Google Street View imagery covering the city of Oxford, while the query set includes images captured under diverse weather and illumination conditions. In the standard test split, the shared gallery consists of 17,166 images. We conduct evaluations on the standard SVOX query.

Baidu Mall (Sun et al., 2017). Baidu Mall is an indoor localization dataset collected in a large-scale shopping mall, featuring significant viewpoint changes and pronounced perceptual aliasing. Following the MegaLoc work, we utilize 689 database images and 2,292 query images. The reference and query images are acquired under diverse conditions and contain challenging repetitive patterns, along with dynamic distractors such as pedestrians.

Tokyo24/7 (Torii et al., 2015). Tokyo24/7 is an urban VPR benchmark composed of GPS-tagged street-view images collected in Tokyo, Japan, mainly from Google Street View panoramas. The dataset contains images captured at different times of day, including daytime and nighttime scenes, to simulate long-term appearance changes. With dense urban areas containing buildings, roads, and open spaces, as well as significant variations in viewpoint, illumination, and dynamic objects, Tokyo24/7 provides a challenging testbed for evaluating VPR methods under complex real-world conditions.

D Details of comparison method

NetVLAD (Arandjelovic et al., 2016) is a pioneering deep VPR method that introduces a learnable VLAD aggregation layer into CNN-based image retrieval frameworks. Built upon a VGG-16 backbone and trained with weakly supervised ranking loss on Pitts30k, NetVLAD learns compact global descriptors that significantly outperform traditional hand-crafted feature based approaches.

SFRS (Ge et al., 2020) improves NetVLAD-based retrieval by addressing the impact of inaccurate GPS annotations in VPR datasets. It introduces a self-supervised fine-grained region similarity strategy to mine reliable hard positive samples through multi-generation training. Based on a VGG-16 backbone and trained on Pitts30k, SFRS achieves improved retrieval robustness over previous methods.

CosPlace (Berton et al., 2022a) tackles the scalability issue of large-scale VPR training by reformulating place recognition as a classification problem. It constructs the SF-XL dataset and introduces geographic grouping strategies to enable efficient learning of place-discriminative representations. Using VGG-16 or ResNet backbones, CosPlace produces compact 512-dimensional descriptors with low memory consumption, making it suitable for city-scale visual localization.

MixVPR (Ali-bey et al., 2023) proposes a holistic feature aggregation strategy that improves global representation learning without relying on local matching or multi-scale aggregation. It introduces stacked Feature-Mixer blocks, which are composed of lightweight MLP layers, to progressively incorporate global relationships into feature maps. Trained on large-scale datasets such as GSV-Cities with ResNet-based backbones, MixVPR achieves strong retrieval performance with fewer parameters compared with previous aggregation methods.

EigenPlaces (Berton et al., 2023) improves viewpoint robustness by explicitly exposing the model to diverse viewpoints of the same place during training. It constructs viewpoint-aware training samples through clustering without requiring additional supervision. Combined with VGG-16 or ResNet backbones and GeM pooling, EigenPlaces achieves competitive performance on multiple benchmarks while reducing training memory requirements and descriptor dimensions.

CricaVPR (Lu et al., 2024a) introduces cross-image correlation-aware representation learning for VPR. By exploiting feature correlations among images within a batch, it enables each image representation to benefit from contextual information from other samples. Moreover, it proposes a multi-scale convolution enhanced adaptation strategy for efficiently adapting frozen vision foundation models, such as DINOv2, leading to improved robustness and reduced training cost.

SALAD (Izquierdo & Civera, 2024b) revisits VLAD-style aggregation by formulating the soft assignment process as an optimal transport problem. It considers bidirectional relationships between local features and cluster representations and introduces a dustbin cluster to remove non-informative features. With DINOv2 as the backbone, SALAD achieves strong performance through efficient fine-tuning and provides fast single-stage retrieval without re-ranking.

BoQ (Ali-Bey et al., 2024) proposes a query-based aggregation framework for global feature learning. It learns a set of global queries and utilizes cross-attention to extract consistent information from image features. Supporting both CNN and Vision Transformer backbones, BoQ achieves competitive retrieval accuracy as a single-stage method without additional re-ranking, while maintaining efficient inference.

SuperVLAD (Lu et al., 2024c) improves VLAD-based aggregation by removing explicit cluster centers and introducing a simplified cluster representation mechanism. It further proposes a low-dimensional 1-Cluster VLAD variant with ghost clusters, enabling compact descriptors while preserving discriminative ability. Using transformer-based backbones such as DINOv2, SuperVLAD achieves strong cross-domain generalization with lower-dimensional representations.

EDTFormer (Jin et al., 2025) introduces a lightweight transformer decoder architecture for VPR, where learnable queries are used to efficiently aggregate discriminative information from image features. It further enhances DINOv2 adaptation through a Low-rank Parallel Adaptation (LoPA) strategy, achieving effective parameter-efficient fine-tuning with reduced training overhead.

ImAge (Lu et al., 2026) proposes an implicit aggregation paradigm that eliminates explicit feature aggregation modules in VPR. Instead of using hand-designed aggregation operations, ImAge introduces learnable aggregation tokens before transformer blocks and exploits self-attention to implicitly capture global contextual information. The aggregation tokens are initialized through K-means clustering, improving representation robustness.

E Limitations Discussion Future Work

E.1 Limitations Discussion

Although GeoMPR extends retrieval-based Topological Place Recognition toward Metric Place Recognition, it still relies on the quality of the retrieved reference images. Since the proposed framework uses the top-ranked retrieval result as the geographic anchor and estimates the query position through the spatial offset relative to this anchor, the quality and distribution of database im-

ages can affect the final localization accuracy. In particular, sparse database sampling or significant differences between database and query trajectories may introduce larger spatial gaps that make accurate offset estimation more challenging. Moreover, GeoMPR assumes that the retrieval stage can provide a reasonable place-level hypothesis. When the top-ranked retrieved image corresponds to an incorrect location due to severe perceptual aliasing or retrieval failure, the subsequent geometry-grounded refinement may not recover the correct geographic position, since the estimated spatial offset is still based on an inaccurate anchor. Therefore, improving the robustness of retrieval and jointly modeling retrieval uncertainty with geometric refinement remain important directions.

E.2 Future Work

In the current framework, retrieval, geometric reasoning, and spatial offset estimation are optimized separately. Although this design provides a flexible way to integrate existing VPR models with vision geometry foundation models, an end-to-end training framework that jointly optimizes retrieval and offset estimation could further improve the consistency between place recognition and metric localization. Furthermore, GeoMPR currently exploits only a subset of the geometric outputs provided by VGGT, mainly using relative camera geometry for spatial offset estimation. Future work could investigate how to incorporate additional geometric cues, such as depth, point maps, or richer scene-level representations, to provide stronger geometric constraints and achieve more accurate metric place recognition.