GeoMix: Descriptor-Free Visual Localization via Global Context and Multi-Detector Training

Yejun Zhang[1], Xinjue Wang[1], Zihan Wang[1], Esa Rahtu[2], and Juho Kannala[1,3]

[1] Aalto University, Espoo, Finland
{yejun.zhang,xinjue.wang,zihan.1.wang,juho.kannala}@aalto.fi
[2] Tampere University, Tampere, Finland
esa.rahtu@tuni.fi
[3] University of Oulu, Oulu, Finland

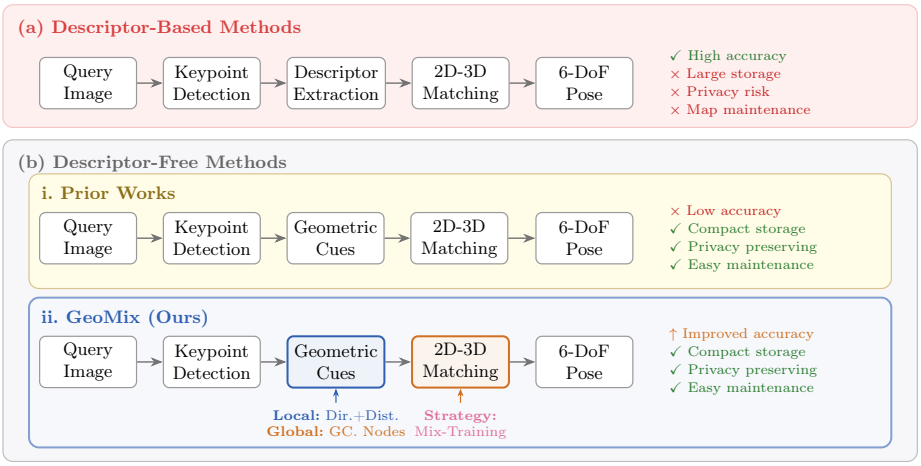

Fig. 1: Comparison of visual localization pipelines. (a) Descriptor-based methods achieve high accuracy but require large storage, raise privacy concerns, and incur costly map maintenance. (b-i) Descriptor-free methods offer compact, privacy-preserving localization yet lag in accuracy. (b-ii) GeoMix bridges this gap with directional and distance-aware geometric embeddings, global context nodes, and a Mix-Training strategy. Dir.: direction; Dist.: distance; GC: global context.

Abstract. Descriptor-free visual localization eliminates high-dimensional descriptor storage, preserves scene privacy, and simplifies map maintenance, yet its accuracy still lags far behind descriptor-based pipelines. We identify this gap to insufficient geometric discriminability in geometry-only matching. Without visual appearance, current methods underutilize local geometry cues, lack the global context among keypoints, and overfit to a single keypoint detector. We further observe that descriptor-free matching naturally enables multi-detector training, as heterogeneous keypoints can be optimized in a shared geometry-only space without aligning descriptor spaces. Building on these insights, we propose

GeoMix, a descriptor-free 2D-3D matching framework that strengthens geometric discriminability at three levels. Locally, directional and distance-aware embeddings enrich neighborhood aggregation with fine-grained spatial structure. Globally, learnable context nodes aggregate and redistribute scene-wide information via cross-attention to resolve ambiguities beyond local receptive fields. At the training level, Mix-Training exploits this detector-agnostic geometry space to learn representations across multiple keypoint detectors. Extensive experiments on MegaDepth, Cambridge Landmarks, 7Scenes, and Aachen Day-Night show that GeoMix sets a new state of the art among descriptor-free methods, reducing 75th-percentile rotation error by 89% and translation error by up to 90% over the previous best, while generalizing zero-shot to unseen detectors and narrowing the gap to descriptor-based pipelines. Code is available at `https://github.com/YejunZhang/Geomix`.

Keywords: Visual Localization · Descriptor-Free Matching · Graph Neural Network

1 Introduction

Accurate camera localization underpins autonomous navigation, augmented reality, and large-scale mapping [2, 5, 22, 42]. The dominant paradigm achieves high accuracy by matching local image features to a pre-built 3D map via visual descriptors and recovering pose with a PnP-RANSAC solver [50, 53, 54, 57, 77].

Yet this success comes at a cost: storing high-dimensional descriptors for millions of 3D points consumes substantial memory, descriptors can be inverted to reveal sensitive scene content [9, 43], and map updates require recomputing descriptors whenever new images are observed [17, 69, 76]. These constraints motivate descriptor-free localization [6, 69, 76], which represents keypoints through geometric cues such as keypoint positions and spatial relations, enabling compact and privacy-preserving maps. Despite these practical advantages, descriptor-free methods still lag significantly behind descriptor-based counterparts in accuracy, preventing adoption in precision-critical systems.

The accuracy gap reflects a fundamental challenge. Without visual appearance, the network must extract sufficient discriminative signal from geometry alone, and current methods [69, 75, 76] fall short at every stage of this process. At the local level, neighborhood aggregation in graph neural networks (GNNs) relies on basic spatial relations, overlooking directional and metric cues that could sharpen per-point discriminability. At the global scale, local graph convolution operators have limited receptive fields; repetitive structures such as corridors or building facades are geometrically indistinguishable locally and therefore produce ambiguous correspondences. At the training level, existing methods train with a single keypoint detector, causing the network to absorb detector-specific patterns rather than learning geometry that transfers across conditions.

We observe, however, that the descriptor-free formulation itself conceals an untapped advantage. In descriptor-based pipelines, each keypoint detector is

tightly coupled with its own descriptor space, so switching detectors requires reconciling incompatible representations [11, 17, 27]. The descriptor-free setting removes this coupling entirely: since matching relies purely on geometric relations, keypoints from *any* detector are interchangeable. This detector-agnostic property has not been exploited by prior methods, yet it opens a natural path to stronger generalization—training over diverse detectors simultaneously to force the network to learn geometry that transcends any single detector's characteristics.

Building on this observation, we propose **GeoMix**, a descriptor-free 2D–3D matching framework that strengthens **Geo**metric discriminability at all three levels through enriched local–global representations and **Mix**-Training over multiple detectors. At the local level, directional and distance-aware embeddings enrich neighborhood aggregation with fine-grained spatial structure. At the global level, learnable global context nodes aggregate and redistribute global information via cross-attention, enabling each feature to resolve ambiguities beyond its local receptive field. At the training level, our Mix-Training strategy jointly optimizes over multiple keypoint detectors, leveraging the detector-agnostic property unique to the descriptor-free setting to learn robust geometric representations. Extensive experiments on MegaDepth, Cambridge Landmarks, 7Scenes, and Aachen Day-Night show that GeoMix sets a new state of the art among descriptor-free methods, reducing 75th-percentile rotation error by 89% and translation error by up to 90% over the previous best, substantially narrowing the gap to descriptor-based pipelines.

Our contributions are as follows:

- We introduce GeoMix, a descriptor-free 2D–3D matching framework that strengthens geometric discriminability at both local and global levels: directional and distance-aware embeddings sharpen per-point discrimination in neighborhood aggregation, while learnable global context nodes aggregate and redistribute scene-wide information via cross-attention to disambiguate repetitive structures.
- We propose Mix-Training, a multi-detector training strategy uniquely enabled by the descriptor-free setting: since keypoints from any detector are interchangeable without descriptors, joint optimization forces the network to learn geometry that generalizes to entirely unseen detectors at test time.
- Extensive experiments across four benchmarks show that GeoMix reduces 75th-percentile rotation error by 89% and translation error by up to 90% over the best prior descriptor-free method, while generalizing zero-shot to unseen detectors.

2 Related work

Structure-Based Localization. Structure-based visual localization achieves high accuracy by establishing 2D–3D correspondences against a pre-built 3D map. Classical methods [53, 54] relied on Structure-from-Motion (SfM) models to match features between query images and 3D points. Learning-based advances

have significantly improved image retrieval [1,24,25,48], with recent works [26,31] leveraging powerful pretrained models to further boost retrieval accuracy. Feature extraction methods [13, 18, 49, 66] extract robust visual representations for reliable keypoint detection and description. For keypoint matching, SuperGlue [51] employs graph neural networks to refine correspondences, while recent advances [29, 38, 59, 74] continue to push the boundaries of matching accuracy and efficiency. Detector-free methods [10, 65, 72] further improve accuracy by matching dense image pairs directly, though at higher computational cost. More recently, end-to-end regression methods such as Reloc3r [16] directly predict relative camera poses from image pairs, offering fast inference but requiring large-scale training data and substantial model capacity.

Scene Compression. Scene compression reduces storage via map pruning [7, 8, 19, 34, 36, 73], descriptor compression [15, 28, 30, 34, 52, 68, 73], or hybrid approaches [7, 34, 73]. Scene coordinate regression [3, 4, 35, 67, 70, 71] learns compact representations but struggles to generalize to new scenes. While these techniques save storage, they do not fully address privacy concerns or descriptor maintenance. Our descriptor-free method is complementary and can be combined with scene compression to further reduce model size.

Privacy-Preserving Visual Localization. As demonstrated in [47], image details can be recovered from descriptors, raising data privacy concerns. To address this, privacy-preserving methods [40, 41, 43, 60, 63, 64] obscure spatial details by lifting 2D/3D points to geometric primitives such as lines, ray clouds [40], and sphere clouds [41], but generally suffer from accuracy degradation or require additional sensor inputs. Other methods adopt alternative strategies, including scene landmarks [14], semantic segmentation [46], Gaussian splatting feature fields [44], and NeRF-based representations [45], yet they rely on extensive priors or labels and limit generalizability.

Cross-Detector Matching. Standard localization pipelines assume the same feature extractor for mapping and querying. To enable cross-detector map reuse, prior works learn pairwise descriptor translations [17], project features into shared embedding spaces [27], or augment descriptors with geometric context [11], all requiring descriptor storage or alignment. However, these methods must be retrained or fine-tuned for every new detector pair, limiting scalability as the number of detectors grows. In contrast, our descriptor-free formulation sidesteps the cross-detector problem entirely: since matching relies solely on geometric relations, keypoints from any detector are directly interchangeable without additional training or adaptation.

3 Method

3.1 Problem Setting and Keypoint Representation

Problem Formulation. Given a query image I, we first retrieve its top-k database images and collect the 3D points observed in them to form a candidate point

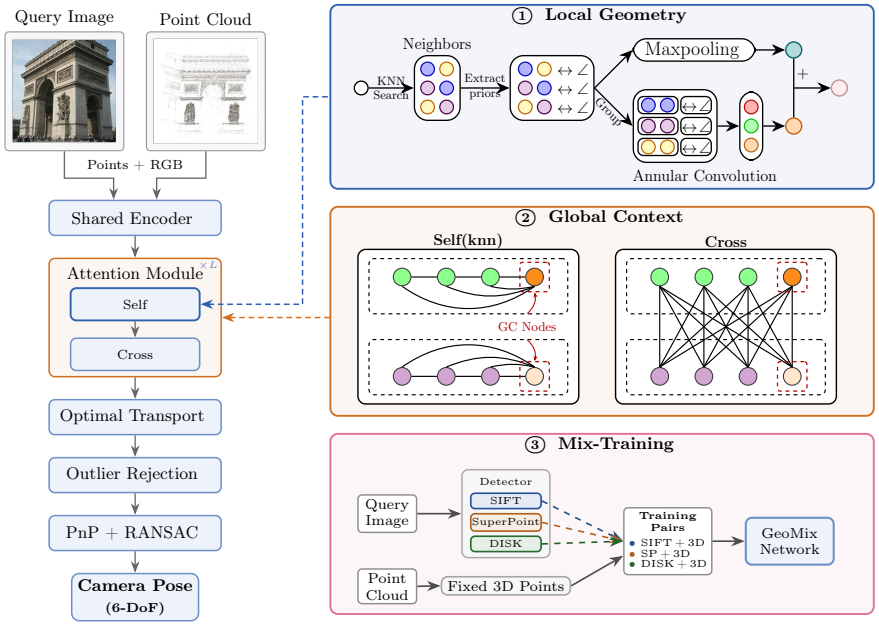

Fig. 2: Overview of GeoMix. A shared encoder lifts 2D keypoints and 3D points to bearing vectors with RGB colors. The attention module alternates self-attention and cross-attention over L layers: self-attention aggregates local geometry via annular convolution and max-pooling enriched by directional and distance embeddings, while learnable global context (GC) nodes broadcast scene-wide information through cross-attention. Optimal transport with outlier rejection produces 2D–3D correspondences for pose estimation. During training, Mix-Training varies the query-side keypoint detector while keeping the 3D map fixed, encouraging detector-agnostic geometric learning.

cloud $\mathbf{Q} = \{\mathbf{q}_j \in \mathbb{R}^3\}_{j=1}^{N}$. Because the retrieved database images are registered in the scene, their known camera poses provide the extrinsic parameters used below. We estimate the 6-DoF camera pose $(\mathbf{R}, \mathbf{t})$ of the query image by establishing 2D–3D correspondences between its M keypoints $\mathbf{P} = \{\mathbf{p}_i \in \mathbb{R}^2\}_{i=1}^{M}$ and $\mathbf{Q}$.

Keypoint Representation. Since 2D keypoints are defined in pixel space while 3D scene points reside in world coordinates, directly comparing them is challenging without a common representation. Following [69,75,76], we adopt bearing vectors to bridge this modality gap: by back-projecting both onto the normalized image plane, 2D and 3D keypoints become geometrically comparable. For a 2D keypoint $\mathbf{p}_i = (u, v)$, we back-project it through the camera intrinsics $\mathbf{K}$ to obtain the bearing vector $\mathbf{b}_{p_i}$:

$$[\mathbf{b}_{p_i}^\top, 1]^\top = \mathbf{K}^{-1}[u, v, 1]^\top.$$

For a 3D scene point $\mathbf{q}_j$ and the world-to-database-image transformation with rotation $\mathbf{R}$ and translation $\mathbf{t}$, we compute $\mathbf{q}_j^c = \mathbf{R}\mathbf{q}_j + \mathbf{t}$, where $\mathbf{q}_j^c$ represents $\mathbf{q}_j$ in the camera's frame of reference. We then obtain the bearing vector $\mathbf{b}_{q_j}$ by

projecting onto the normalized image plane:

$$[\mathbf{b}_{q_j}^\top, 1]^\top = \frac{\mathbf{q}_j^c}{q_{j,z}^c},$$

where $q_{j,z}^c$ denotes the z-coordinate of $\mathbf{q}_j^c$.

3.2 Network Architecture

Fig. 2 shows our architecture, which builds upon the descriptor-free A2-GNN backbone [75]. The model consists of a bearing/RGB feature encoder, a geometric aggregation GNN with angle–annular convolution and max-pooling for local structure modeling, a global context mechanism, an optimal transport layer for differentiable correspondence estimation, and an outlier rejection module. While inheriting the encoder, local aggregation, optimal transport, and rejection modules from A2-GNN, our GeoMix introduces three key components in Fig. 1(b-ii): directional and distance edge embeddings, learnable global context nodes, and a Mix-Training strategy. These components enhance local and global geometric embedding and improve robustness.

Feature Encoder. We use ResNet-style encoders to project inputs into a high-dimensional feature space. For each keypoint with bearing vector $\mathbf{b}_i \in \{\mathbf{b}_{p_i}, \mathbf{b}_{q_j}\}$ and RGB color $\mathbf{c}_i$, we process geometric and photometric cues through separate branches and sum their embeddings to generate the initial feature vector $\mathbf{f}_i \in \mathbb{R}^d$:

$$\mathbf{f}_i = \mathcal{F}_b(\mathbf{b}_i) + \mathcal{F}_c(\mathbf{c}_i), \tag{1}$$

where $\mathcal{F}_b$ and $\mathcal{F}_c$ denote the bearing vector and color encoders.

Geometric Graph Neural Network. We employ a GNN where each node represents a 2D keypoint or 3D point. The network alternates between self-attention and cross-attention to progressively refine features. Self-attention layers extract local geometric context via dual-branch aggregation, while cross-attention layers incorporate global scene context using global context nodes.

Local Geometry. We construct a local graph by connecting each point to its k nearest neighbors. For central node i with neighbor set $\mathcal{N}_i$, self-attention uses dual-branch aggregation: geometry-guided max-pooling extracts salient metric cues with permutation invariance, while annular convolution partitions neighbors into distance-based groups to encode topological structure. For each neighbor $j \in \mathcal{N}_i$, we compute relative displacement $\Delta\mathbf{p}_{ij} = \mathbf{p}_j - \mathbf{p}_i$ and form a unified geometric embedding shared by both branches by concatenating the relative displacement with its unit direction:

$$\mathbf{e}_{ij}^{geo} = \left[\Delta\mathbf{p}_{ij}\,;\, \frac{\Delta\mathbf{p}_{ij}}{\|\Delta\mathbf{p}_{ij}\|_2}\right] \in \mathbb{R}^4. \tag{2}$$

This embedding is subsequently projected and aggregated via a max-pooling operation:

$$\mathbf{h}_i^{max} = \max_{j \in \mathcal{N}_i} \left(\phi_{feat}([\mathbf{f}_i \,;\, \mathbf{f}_j - \mathbf{f}_i]) + \phi_{geo}(\mathbf{e}_{ij}^{geo}) \right), \tag{3}$$

yielding a geometry-aware representation $\mathbf{h}_i^{max}$, where ϕ_{feat} and ϕ_{geo} denote MLPs with instance normalization and LeakyReLU.

Simultaneously, the annular convolution branch partitions the k nearest neighbors into g groups by ranking Euclidean distances to the central node, so that each group contains k/g neighbors at a similar radius. Two successive 1D convolutions with kernels of size $1 \times (k/g)$ and $1 \times g$ are applied: the first aggregates features within each group, and the second aggregates across groups. We apply this operation to both the concatenated features and the geometric embedding in parallel:

$$\mathbf{h}_i^{ann} = \phi_{ann}^{feat}([\mathbf{f}_i \,;\, \mathbf{f}_j - \mathbf{f}_i]) + \phi_{ann}^{geo}(\mathbf{e}_{ij}^{geo}), \tag{4}$$

where ϕ_{ann}^{feat} and ϕ_{ann}^{geo} denote the annular convolutions applied to neighbor features and geometric embeddings, respectively, each followed by batch normalization and ReLU.

Finally, we aggregate local features from two complementary paths: first, an annular convolution path $\mathbf{h}_i^{ann}$ with annular geometric embeddings, and second, a maxpooling path $\mathbf{h}_i^{max}$ with geometric embeddings. These are fused via element-wise summation: $\mathbf{f}_i^{local} = \mathbf{h}_i^{ann} + \mathbf{h}_i^{max}$, which then feeds into the global context module.

Global Context Nodes. To capture long-range dependencies beyond the local KNN graph, we introduce N_g learnable global context nodes $\{\mathbf{g}_k \in \mathbb{R}^d\}_{k=1}^{N_g}$, maintained independently for each 2D and 3D point set. These nodes aggregate information from all local features, exchange context among themselves, and broadcast back via cross-attention:

$$\mathbf{g}_k' = \mathbf{g}_k + \phi_{agg}\left([\mathbf{g}_k \,;\, \mathrm{CA}(\mathbf{g}_k, \mathcal{F})]\right), \tag{5}$$

$$\mathbf{g}_k'' = \mathbf{g}_k' + \mathrm{SA}(\mathbf{g}_k', \mathcal{G}'), \tag{6}$$

$$\mathbf{f}_i^{self} = \mathbf{f}_i^{local} + \phi_{broad}\left(\left[\mathbf{f}_i^{local} \,;\, \mathrm{CA}(\mathbf{f}_i^{local}, \mathcal{G}'')\right]\right), \tag{7}$$

where $\mathrm{SA}(\cdot)$ and $\mathrm{CA}(\cdot)$ denote Self-Attention and Cross-Attention mechanisms respectively. $\mathcal{G}'$ and $\mathcal{G}''$ denote the sets of intermediate global context features, $[\cdot\,;\cdot]$ denotes concatenation, and ϕ_{agg} and ϕ_{broad} are MLP blocks for aggregation and broadcasting, respectively. The final output $\mathbf{f}_i^{self}$ corresponds to the locally updated feature for node i enriched with global context.

Optimal Transport and Outlier Rejection. To establish the initial set of correspondences $\mathcal{M}_{init}$, we employ a differentiable Optimal Transport layer. By applying the Sinkhorn algorithm [12, 62] to the score matrix derived from the enhanced features, we compute an optimal assignment matrix that encapsulates the soft matching probabilities between 2D query keypoints and 3D scene points.

Fig. 3: Keypoint distributions of SIFT, SuperPoint, and DISK on the same image. Each detector produces a distinct spatial layout with few shared keypoints (rightmost), motivating Mix-Training to learn detector-agnostic geometric representations.

Subsequently, following A2-GNN [75], we utilize a learning-based outlier rejection module to refine these initial matches. This classifier estimates a confidence score for each putative match, indicating its probability of being an inlier. The final set of robust correspondences is obtained by retaining only those matches where the predicted confidence exceeds a fixed threshold τ.

3.3 Mix-Training Strategy

Different keypoint detectors adopt different detection strategies and produce markedly different keypoint distributions. As shown in Fig. 3, SIFT [39] favors blob-like structures, SuperPoint [13] prefers corner regions, and DISK [66] targets repeatable local features, resulting in few shared keypoints on the same image. However, existing descriptor-free methods are typically trained with a single detector, causing the network to overfit to that detector's keypoint distribution and generalize poorly to others.

To address this, we introduce Mix-Training, a multi-detector training strategy uniquely enabled by the descriptor-free setting. In descriptor-based methods, each detector is tightly coupled with its own descriptor space, making it impractical to mix detectors without reconciling incompatible representations. The descriptor-free formulation removes this constraint: since matching is based purely on geometric relationships between 2D-3D correspondences, keypoints from any detector can serve as input without modifying the matching pipeline.

Specifically, we maintain a pool of L detectors $\mathcal{S} = \{D_l\}_{l=1}^{L}$ (L=3 in practice). For each query image, we apply all L detectors independently to generate L distinct sets of 2D keypoints. Each set is paired with the same 3D scene points to produce L training samples per image. Note that we fix the 3D side and vary only the 2D query-side detector, since the 3D maps are reconstructed via SfM with SIFT and regenerating them for each detector would be prohibitively expensive. By exposing the network to diverse keypoint distributions against shared 3D

geometry, Mix-Training encourages detector-invariant geometric reasoning and improves generalization to unseen detectors at test time. At inference a single detector still yields 1,024 keypoints per query, identical to single-detector training, so the gains stem from detector diversity rather than a larger keypoint budget.

3.4 Training Loss

Following GoMatch [76], the loss function $\mathcal{L}$ consists of a matching loss $\mathcal{L}_m$ and an outlier rejection loss $\mathcal{L}_{or}$. The matching loss minimizes the negative log-likelihood of matching scores:

$$\mathcal{L}_m = -\frac{1}{N_m}\Big(\sum_{(i,j)\in\mathcal{M}_{gt}} \log \tilde{\mathrm{P}}_{ij} + \sum_{i\in\mathcal{U}_q} \log \tilde{\mathrm{P}}_{i(N+1)} + \sum_{j\in\mathcal{U}_d} \log \tilde{\mathrm{P}}_{(M+1)j}\Big), \quad (8)$$

where $\tilde{\mathrm{P}}$ denotes the matching score matrix, $\mathcal{M}_{gt}$ the set of ground-truth matches, $\mathcal{U}_q$ the unmatched query keypoints, and $\mathcal{U}_d$ the unmatched database 3D points. N_m is the total number of ground-truth, unmatched query, and unmatched database keypoints.

The outlier rejection loss $\mathcal{L}_{or}$ enhances robustness against outliers:

$$\mathcal{L}_{or} = -\frac{1}{N_c}\sum_{i=1}^{N_c} w_i \left(y_i \log p_i + (1-y_i)\log(1-p_i)\right), \quad (9)$$

where N_c denotes the total number of initial correspondences, p_i is the predicted inlier probability, y_i the ground-truth label, and w_i the class-balancing weight. The total training loss is $\mathcal{L} = \mathcal{L}_m + \mathcal{L}_{or}$.

4 Experiments

Training. We train our model on the MegaDepth dataset [37] for 50 epochs using the Adam optimizer [33] with a learning rate of 10^{-3} and a batch size of 16. We use N_g=4 global context nodes, 9 nearest neighbors for local graph construction, and a group size of 3 for neighborhood aggregation, following the same settings as A2-GNN [75]. Training runs on a single NVIDIA RTX 4090 GPU with 24 GB of VRAM and converges in approximately 22 hours.

Datasets. Following prior work [76], we train and evaluate on MegaDepth [37], a large-scale outdoor dataset with SfM-based 3D models. We follow GoMatch [76] and use the official test split of 53 sequences, with the remaining sequences divided into 99 training, 16 validation, and 49 test scenes after query sampling and co-visibility filtering. During training, we extract 2D keypoints with SIFT, SuperPoint, and DISK, and establish 2D-3D correspondences against the original SIFT-based SfM point clouds. For visual localization, we evaluate on Cambridge Landmarks [32], a medium-scale outdoor dataset with four scenes and SfM-based

Table 1: Matching results on MegaDepth under two training protocols. *Single-training* uses SIFT only; *Mix-training* jointly trains with SIFT, SuperPoint, and DISK. The upper section evaluates detectors seen during training, while the lower section tests zero-shot transfer to unseen detectors (DeDoDe v2 and R2D2). Best results are in **bold**.

Method	Detector	Reproj. AUC (%) @1 / 5 / 10px (↑)	Rotation (°) Quantile @25 / 50 / 75% (↓)	Translation
Single-training	SIFT	18.68 / 58.34 / 66.19	0.06 / 0.15 / 2.26	0.00 / 0.01 / 0.20
	SuperPoint	0.04 / 8.55 / 17.50	1.67 / 7.14 / 27.40	0.15 / 0.68 / 2.94
	DISK	0.02 / 17.70 / 32.00	0.74 / 2.82 / 16.06	0.06 / 0.25 / 1.69
Mix-training	SIFT	**21.51** / **65.61** / **73.76**	**0.05** / **0.11** / **0.52**	**0.00** / **0.01** / **0.05**
	SuperPoint	**0.15** / **20.84** / **36.89**	**0.60** / **1.57** / **7.57**	**0.05** / **0.14** / **0.77**
	DISK	**0.04** / **30.30** / **48.80**	**0.33** / **0.96** / **4.22**	**0.03** / **0.08** / **0.41**
Single-training	DeDoDe v2	7.87 / 47.27 / 57.31	0.11 / 0.39 / 4.19	0.01 / 0.03 / 0.44
	R2D2	0.29 / 27.29 / 37.62	0.29 / 1.94 / 17.07	0.03 / 0.18 / 1.89
Mix-training	DeDoDe v2	**11.16** / **60.30** / **70.56**	**0.07** / **0.17** / **0.83**	**0.01** / **0.02** / **0.07**
	R2D2	**0.62** / **46.87** / **60.22**	**0.13** / **0.34** / **1.95**	**0.01** / **0.03** / **0.19**

Table 2: Visual localization on Cambridge Landmarks. Median translation (cm) and rotation (°) errors are reported. MB denotes total map size. Best results per group are in **bold**.

Methods		**Cambridge-Landmarks** (cm, °)					
		King's	Hospital	Shop	St. Mary	Avg.	**MB**
E2E	MS-Trans. [58]	83 / 1.47	181 / 2.39	86 / 3.07	162 / 3.99	128 / 2.73	**71**
	DSAC* [4]	**15** / **0.30**	21 / 0.40	**5** / **0.30**	13 / 0.40	14 / 0.35	112
	HSCNet [35]	18 / **0.30**	**19** / **0.30**	6 / **0.30**	**9** / **0.30**	**13** / **0.30**	592
	Reloc3r [16]	42 / 0.36	62 / 0.55	13 / 0.58	34 / 0.58	38 / 0.52	10294
DB	HybridSC [7]	81 / 0.59	75 / 1.01	19 / 0.54	50 / 0.49	56 / 0.66	**3**
	AS [55]	13 / 0.22	20 / 0.36	**4** / **0.21**	8 / 0.25	11 / 0.26	813
	SP [13]+SG [51]	**12** / **0.20**	**15** / **0.30**	**4** / **0.20**	**7** / **0.21**	**10** / **0.23**	3215
DF	GoMatch [76]	25 / 0.64	283 / 8.14	48 / 4.77	335 / 9.94	173 / 5.87	**48**
	DGC-GNN [69]	18 / 0.47	75 / 2.83	15 / 1.57	106 / 4.03	54 / 2.23	69
	A2-GNN [75]	15 / 0.39	59 / 1.74	12 / 1.16	76 / 2.65	41 / 1.49	69
	GeoMix (Ours)	**14** / **0.37**	**27** / **0.82**	**6** / **0.73**	**23** / **0.72**	**18** / **0.66**	69

3D reconstructions; 7Scenes [61], a small-scale indoor dataset with seven RGB-D scenes and SLAM-derived ground-truth poses; and Aachen Day-Night [56], which requires localizing 922 queries against a 3D model built from 4,328 daytime database images under drastic day-night illumination changes. Detailed descriptions of the correspondence retrieval strategy and the Aachen Day-Night retriangulation procedure are provided in the supplementary material.

Table 3: Visual localization on 7Scenes. Median translation (cm) and rotation (°) errors are reported. MB denotes total map size. Best results per group are in **bold**.

	Methods	**7Scenes** (cm, °) (SIFT)								
		Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Avg.	**MB**
E2E	MS-Trans. [58]	11/4.66	24/9.60	14/12.2	17/5.66	18/4.44	17/5.94	26/ 8.45	18/7.28	**71**
	DSAC* [4]	**2**/1.10	**2**/1.24	**1**/ 1.82	**3**/1.15	**4**/1.34	**4**/1.68	**3**/ 1.16	**3**/1.36	196
	HSCNet [35]	**2**/**0.70**	**2**/**0.90**	**1**/ **0.90**	**3**/**0.80**	**4**/**1.00**	**4**/**1.20**	**3**/ **0.80**	**3**/**0.90**	1036
	Reloc3r [16]	3/0.88	3/0.81	**1**/ 0.95	4/0.88	6/1.10	4/1.26	7/ 1.26	4/1.02	10268
DB	AS [55]	3/0.87	**2**/1.01	**1**/ 0.82	4/1.15	7/1.69	5/1.72	**4**/ **1.01**	4/1.18	-
	SP [13]+SG [51]	**2**/**0.85**	**2**/**0.94**	**1**/ **0.75**	**3**/**0.92**	**5**/**1.30**	**4**/**1.40**	5/ 1.47	**3**/**1.09**	22977
DF	GoMatch [76]	4/1.65	13/3.86	9/ 5.17	11/2.48	16/3.32	13/2.84	89/21.1	22/5.77	**302**
	DGC-GNN [69]	**3**/1.41	5/1.81	4/ 3.13	7/1.66	8/2.03	8/2.14	83/21.5	17/4.81	355
	A2-GNN [75]	**3**/1.37	5/1.78	4/ 2.70	6/1.56	**7**/1.86	**7**/2.00	72/17.0	15/4.04	355
	GeoMix (Ours)	**3**/**1.31**	**4**/**1.52**	**3**/ **2.30**	**5**/**1.46**	**7**/**1.74**	**7**/**1.88**	**45**/**11.1**	**11**/**3.04**	355

Keypoint Detectors. Unless otherwise specified, our main model uses mix-training on MegaDepth with three 2D query-side keypoint detectors: SIFT [39], SuperPoint [13], and DISK [66], covering both classical hand-crafted and modern learning-based approaches. Note that the 3D point clouds in MegaDepth are reconstructed with SIFT, so training with SuperPoint or DISK introduces a cross-detector setting between the 2D and 3D sides. To isolate the effect of detector diversity, Tab. 1 also includes a single-training baseline trained with SIFT only. For evaluation on MegaDepth, we test with SIFT, SuperPoint, and DISK, and further evaluate on R2D2 [49] and DeDoDe-v2 [20] as zero-shot detectors to test generalizability to unseen feature extractors. For visual localization, Cambridge Landmarks uses SuperPoint and 7Scenes uses SIFT, while Aachen Day-Night is evaluated with retriangulated 3D models from multiple detectors, following the configurations described above.

Evaluation Metrics. Following [69, 75, 76], we adopt different metrics for matching and visual localization tasks. For matching evaluation on MegaDepth, we report the Area Under the Curve (AUC) of mean reprojection error at 1, 5, and 10 pixel thresholds, along with translation and rotation error quantiles at 25%, 50%, and 75%. For visual localization on Cambridge Landmarks and 7Scenes, we report the median translation (cm) and rotation (°) errors per scene. All camera poses are estimated using PnP-RANSAC [21, 23] on the established correspondences. We also provide Oracle results, which use ground truth matches as an upper bound for reference.

Matching and Localization Results. Tab. 1 compares single-training (SIFT only) with mix-training (SIFT, SuperPoint, DISK) under different test-time detectors. Single-training suffers severe cross-detector degradation, revealing a strong detector-specific bias. Mix-training resolves this: it not only narrows the cross-detector gap dramatically, but also improves same-detector performance, sug-

Table 4: Visual localization on Aachen Day-Night [56]. Percentage of correctly localized queries at (0.25m, 2°) / (0.5m, 5°) / (5m, 10°) thresholds. For descriptor-free methods, the 2D and 3D sides use the same detector type; SP+SG is a descriptor-based reference evaluated on the same retriangulated map. Matching time per image excludes PnP-RANSAC; Params is model size in millions. Best results among descriptor-free methods are in **bold**.

Method	Detector	Day	Night	Time (s)	Params (M)
Descriptor-based reference					
SP [13]+SG [51]	SuperPoint	85.1 / 91.6 / 96.0	77.6 / 86.7 / 94.9	0.138	12.04
Descriptor-free					
GoMatch [76]	SIFT	0.5 / 1.2 / 6.9	0.0 / 0.0 / 1.0	**0.077**	**1.32**
	SuperPoint	2.2 / 4.7 / 18.4	1.0 / 2.0 / 10.2		
	R2D2	2.3 / 4.2 / 10.2	1.0 / 3.1 / 3.1		
	DISK	4.2 / 7.9 / 26.3	0.0 / 1.0 / 4.1		
DGC-GNN [69]	SIFT	2.8 / 6.3 / 15.9	0.0 / 0.0 / 1.0	0.284	5.65
	SuperPoint	8.1 / 13.3 / 31.1	0.0 / 2.0 / 7.1		
	R2D2	4.6 / 8.1 / 16.0	1.0 / 1.0 / 3.1		
	DISK	11.7 / 20.6 / 38.5	1.0 / 2.0 / 9.2		
A2-GNN [75]	SIFT	3.8 / 7.5 / 16.7	0.0 / 0.0 / 3.1	0.108	2.67
	SuperPoint	11.2 / 20.3 / 36.9	1.0 / 3.1 / 11.2		
	R2D2	6.4 / 11.4 / 19.5	1.0 / 3.1 / 4.1		
	DISK	16.1 / 24.6 / 41.7	0.0 / 1.0 / 7.1		
GeoMix (Ours)	SIFT	**7.9** / **15.9** / **32.6**	**2.0** / **4.1** / **5.1**	0.115	3.50
	SuperPoint	**25.6** / **37.4** / **58.1**	**9.2** / **16.3** / **27.6**		
	R2D2	**11.0** / **18.4** / **33.1**	**3.1** / **4.1** / **7.1**		
	DISK	**27.8** / **42.5** / **61.4**	**4.1** / **8.2** / **12.2**		

gesting that diverse detector exposure acts as a regularizer that strengthens geometric reasoning.

We further compare against prior descriptor-free methods on MegaDepth with k=10 retrieved images in Tab. 5. GeoMix outperforms all baselines by a substantial margin, improving AUC@5/10px over A2-GNN by +11.20/+11.52%. The pose estimation gains are equally striking: the 75%-quantile rotation error drops from 4.60° to 0.52° and the translation error from 0.48 to 0.05, both roughly 90% reductions. To assess matcher quality independently of PnP-RANSAC, we further report matcher-only metrics (precision, recall, F1) on MegaDepth in the supplementary, where GeoMix attains the best scores on all three, consistent with the pose-based ranking.

To verify cross-dataset transfer, we evaluate on Cambridge Landmarks and 7Scenes as shown in Tabs. 2 and 3. GeoMix achieves state-of-the-art results among descriptor-free methods on all scenes, with the largest gains on challenging cases such as Old Hospital, where translation error drops to 27 cm, 54% lower than A2-GNN, and Stairs, where GeoMix achieves 45 cm / 11.1° vs. 72 cm / 17.0° for A2-GNN. These results substantially narrow the gap with descriptor-based

Table 5: Comparison on MegaDepth (k=10 retrieved images). Best results are in **bold**.

Method	Reproj. AUC (%) @1 / 5 / 10px (↑)	Rotation (°) Quantile @25 / 50 / 75% (↓)	Translation Quantile @25 / 50 / 75% (↓)
Oracle	34.59 / 85.02 / 92.02	0.04 / 0.06 / 0.12	0.00 / 0.01 / 0.01
GoMatch [76]	18.90 / 35.67 / 44.99	0.18 / 1.29 / 16.65	0.02 / 0.12 / 1.92
DGC-GNN [69]	15.30 / 51.70 / 60.01	0.07 / 0.26 / 5.41	0.01 / 0.02 / 0.57
A2-GNN [75]	17.29 / 54.41 / 62.24	0.06 / 0.19 / 4.60	0.01 / 0.02 / 0.48
GeoMix (Ours)	**21.51 / 65.61 / 73.76**	**0.05 / 0.11 / 0.52**	**0.00 / 0.01 / 0.05**

Table 6: Ablation study on MegaDepth evaluating the contribution of each proposed component: GC nodes, local geometry embeddings, and Mix-Training. Best results are in **bold**.

Methods	GC Node	Local Geometry: Annular Conv.	Local Geometry: Maxpooling	Mix Train	Reproj. AUC (%) @1 / 5 / 10px (↑)	Rotation (°) Quantile@25 / 50 / 75% (↓)	Translation Quantile@25 / 50 / 75% (↓)
Baseline					16.57 / 52.87 / 60.66	0.06 / 0.22 / 6.04	0.01 / 0.02 / 0.64
Variants	✓				17.24 / 54.59 / 62.36	0.06 / 0.19 / 4.47	0.01 / 0.02 / 0.47
	✓	✓			17.53 / 55.53 / 63.39	0.06 / 0.18 / 3.80	0.01 / 0.02 / 0.38
	✓	✓	✓		18.68 / 58.34 / 66.19	0.06 / 0.15 / 2.26	0.00 / 0.01 / 0.20
GeoMix	✓	✓	✓	✓	**21.51 / 65.61 / 73.76**	**0.05 / 0.11 / 0.52**	**0.00 / 0.01 / 0.05**

methods such as SuperPoint+SuperGlue, while requiring only 69 MB storage versus 3215 MB and preserving privacy by design. Scene-coordinate regressors such as DSAC* are also compact, but must be retrained per scene (hours each), whereas GeoMix is trained once on MegaDepth and reused zero-shot; even 32-d compact descriptors would still need roughly 6× more storage than our geometry-only map.

We further stress-test robustness under drastic appearance changes on Aachen Day-Night in Tab. 4, where both the 2D and 3D sides use the same detector type via retriangulation. GeoMix consistently outperforms all baselines across all four detectors, with especially large gains under the challenging nighttime setting, demonstrating strong robustness to illumination variation.

Generalizability. We evaluate generalizability along two axes. For cross-dataset transfer, the mix-trained model is directly evaluated on 7Scenes and Cambridge Landmarks without finetuning, maintaining strong accuracy despite the domain shift, as discussed above. For cross-detector transfer, we test on two detectors entirely unseen during training: R2D2 [49] and DeDoDe-v2 [20]. As shown in Tab. 1, mix-training outperforms single-training by a large margin on both, with the median rotation error dropping from 1.94° to 0.34° for R2D2 and from 0.39° to 0.17° for DeDoDe-v2, confirming that mix-training encourages reliance on geometric structure rather than detector-specific patterns.

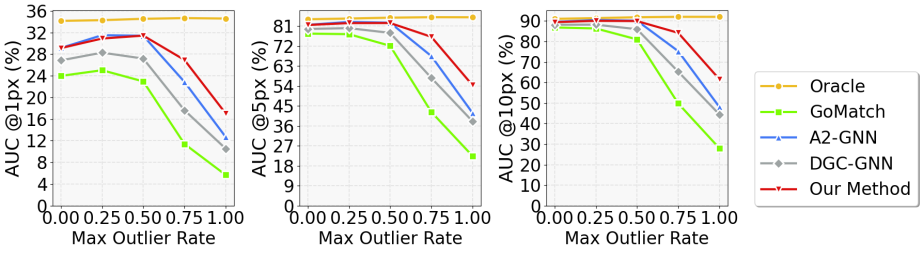

Fig. 4: Sensitivity to outlier ratio. Reprojection AUC at 1, 5, and 10 px thresholds for prior descriptor-free methods and GeoMix as the outlier ratio increases from 0 to 1. Oracle denotes the upper bound obtained with ground-truth matches.

Ablation Study. Tab. 6 presents a progressive ablation on MegaDepth with $k = 10$ retrieved images, starting from the A2-GNN [75] baseline. We first add the learnable global context (GC) nodes, which aggregate long-range dependencies via cross-attention; this consistently improves all AUC thresholds, confirming that global aggregation helps disambiguate correspondences. We then inject directional and distance-aware embeddings into the two local geometry branches (annular convolution and maxpooling), with the maxpooling branch contributing a notably larger boost, indicating that geometry guidance is most effective at the feature aggregation stage. Finally, mix-training contributes the single largest gain, improving AUC@5/10px by over 7%, consistent with our finding that detector diversity acts as a strong regularizer. The full model improves AUC by +4.94/+12.74/+13.10% over the baseline and reduces the 75%-quantile rotation error from 6.04° to 0.52°.

Sensitivity to Keypoint Outliers. Following [69,75,76], we evaluate robustness by varying the outlier ratio, defined as the fraction of keypoints without a ground-truth match among all input keypoints. At ratio 0, only ground-truth inliers are provided; at ratio 1, all detected 2D keypoints and top-k retrieved 3D points are used without filtering, reflecting the realistic deployment setting. As shown in Fig. 4, GeoMix consistently outperforms all baselines across the full range. Below 50%, GeoMix closely approaches the Oracle upper bound. At ratio 1.0, GeoMix still achieves the best performance, demonstrating strong robustness to keypoint outliers in practical conditions.

Matching Efficiency. Tab. 4 reports the matching time of descriptor-free methods on Aachen Day-Night, excluding PnP-RANSAC. Despite the additional global context nodes and geometry injection modules, our GeoMix runs at 0.115s per image pair, comparable to A2-GNN at 0.108s and only 1.5× slower than GoMatch at 0.077s, the lightest baseline, while achieving significantly higher localization accuracy. Compared to DGC-GNN, which requires 0.284s per image pair and is 2.5× slower than ours, GeoMix offers both better accuracy and higher throughput, striking a favorable balance between efficiency and performance. In

terms of model size, GeoMix contains 3.50M parameters, comparable to A2-GNN (2.67M) and GoMatch (1.32M), while being notably smaller than DGC-GNN (5.65M). This confirms that the accuracy gains of GeoMix come without significant parameter overhead. We also compare end-to-end query cost against on-the-fly descriptor matching on Aachen (top-5, 1,024 keypoints): GeoMix matches the query directly to the descriptor-free map in 111.3 ms/query, whereas an on-the-fly hierarchical baseline (SP+SG matched to the 5 retrieved database images, then chained to 3D via their known associations) takes 138.2 ms. SP+SG stays more accurate, but GeoMix offers a competitive descriptor-free alternative for storage- and privacy-sensitive deployments.

5 Conclusion and Limitations

We present GeoMix, a descriptor-free 2D–3D matching framework that enhances geometric discriminability at three complementary levels: locally via directional and distance-aware embeddings, globally via learnable context nodes that aggregate and redistribute scene-wide information, and at the training level via Mix-Training that exploits the detector-agnostic property unique to the descriptor-free setting to learn geometry generalizing across and even beyond seen detectors. Extensive experiments on four benchmarks show that GeoMix sets a new state of the art among descriptor-free methods, substantially narrowing the gap to descriptor-based pipelines while preserving compact storage, privacy, and zero descriptor maintenance overhead. Despite these advances, a performance gap to descriptor-based methods remains, especially under high outlier ratios where purely geometric cues alone lack sufficient discriminability. Moreover, the 3D maps are reconstructed with a fixed detector, potentially introducing residual bias toward specific point distributions. The map still extends incrementally: new images are added with only their pose and 3D-point associations, and any descriptors for triangulating new geometry are used transiently and discarded. Future work could jointly diversify the 3D reconstruction side and incorporate lightweight appearance cues to further close the gap without compromising the benefits of the descriptor-free paradigm.

Acknowledgements

This work was supported by the Research Council of Finland (projects 352788, 362407, 373999, 373780, 353139, 362409, 373997, 373778) and the Finnish Doctoral Program Network in Artificial Intelligence, AI-DOC (decision number VN/3137/2024-OKM-6). We acknowledge the computational resources provided by the CSC-IT Center for Science, Finland.

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307 (2016)
2. Billinghurst, M., Clark, A., Lee, G., et al.: A survey of augmented reality. Foundations and Trends® in Human–Computer Interaction **8**(2-3), 73–272 (2015)
3. Brachmann, E., Cavallari, T., Prisacariu, V.A.: Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5044–5053 (2023)
4. Brachmann, E., Rother, C.: Visual camera re-localization from rgb and rgb-d images using dsac. IEEE transactions on pattern analysis and machine intelligence **44**(9), 5847–5865 (2021)
5. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on robotics **32**(6), 1309–1332 (2016)
6. Campbell, D., Liu, L., Gould, S.: Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 244–261. Springer (2020)
7. Camposeco, F., Cohen, A., Pollefeys, M., Sattler, T.: Hybrid scene compression for visual localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7653–7662 (2019)
8. Cao, S., Snavely, N.: Minimal scene descriptions from structure from motion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 461–468 (2014)
9. Chelani, K., Kahl, F., Sattler, T.: How privacy-preserving are line clouds? recovering scene details from 3d lines. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15668–15678 (2021)
10. Chen, H., Luo, Z., Zhou, L., Tian, Y., Zhen, M., Fang, T., Mckinnon, D., Tsin, Y., Quan, L.: Aspanformer: Detector-free image matching with adaptive span transformer. In: European Conference on Computer Vision. pp. 20–36. Springer (2022)
11. Cubero, P.C., Gálvez, A.J., Mateus, A., Araújo, J., Jensfelt, P.: Matcha: Cross-algorithm matching with feature augmentation. arXiv preprint arXiv:2506.22336 (2025)
12. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
13. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 224–236 (2018)
14. Do, T., Miksik, O., DeGol, J., Park, H.S., Sinha, S.N.: Learning to detect scene landmarks for camera localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11132–11142 (2022)
15. Dong, H., Chen, X., Dusmanu, M., Larsson, V., Pollefeys, M., Stachniss, C.: Learning-based dimensionality reduction for computing compact and effective local feature descriptors. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 6189–6195. IEEE (2023)

16. Dong, S., Wang, S., Liu, S., Cai, L., Fan, Q., Kannala, J., Yang, Y.: Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. arXiv preprint arXiv:2412.08376 (2025)
17. Dusmanu, M., Miksik, O., Schönberger, J.L., Pollefeys, M.: Cross-descriptor visual localization and mapping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6058–6067 (2021)
18. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint description and detection of local features. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 8092–8101 (2019)
19. Dymczyk, M., Lynen, S., Cieslewski, T., Bosse, M., Siegwart, R., Furgale, P.: The gist of maps-summarizing experience for lifelong localization. In: 2015 IEEE international conference on robotics and automation (ICRA). pp. 2767–2773. IEEE (2015)
20. Edstedt, J., Bökman, G., Zhao, Z.: DeDoDe v2: Analyzing and Improving the DeDoDe Keypoint Detector . In: IEEE/CVF Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2024)
21. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
22. Fuentes-Pacheco, J., Ruiz-Ascencio, J., Rendón-Mancha, J.M.: Visual simultaneous localization and mapping: a survey. Artificial intelligence review **43**, 55–81 (2015)
23. Gao, X.S., Hou, X.R., Tang, J., Cheng, H.F.: Complete solution classification for the perspective-three-point problem. IEEE transactions on pattern analysis and machine intelligence **25**(8), 930–943 (2003)
24. Ge, Y., Wang, H., Zhu, F., Zhao, R., Li, H.: Self-supervising fine-grained region similarities for large-scale image localization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 369–386. Springer (2020)
25. Gordo, A., Almazan, J., Revaud, J., Larlus, D.: End-to-end learning of deep visual representations for image retrieval. International Journal of Computer Vision **124**(2), 237–254 (2017)
26. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2024)
27. Jaenal, A., Cubero, P.C., Araújo, J., Mateus, A.: Towards visual localization interoperability: Cross-feature for collaborative visual localization and mapping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26783–26792 (2025)
28. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE transactions on pattern analysis and machine intelligence **33**(1), 117–128 (2010)
29. Jiang, H., Karpur, A., Cao, B., Huang, Q., Araujo, A.: Omniglue: Generalizable feature matching with foundation model guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19865–19875 (2024)
30. Ke, Y., Sukthankar, R.: Pca-sift: A more distinctive representation for local image descriptors. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. vol. 2, pp. II–II. IEEE (2004)

31. Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K.M., Scherer, S., Krishna, M., Garg, S.: Anyloc: Towards universal visual place recognition. IEEE Robotics and Automation Letters (2023)
32. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: Proceedings of the IEEE international conference on computer vision. pp. 2938–2946 (2015)
33. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
34. Laskar, Z., Melekhov, I., Benbihi, A., Wang, S., Kannala, J.: Differentiable product quantization for memory efficient camera relocalization. arXiv preprint arXiv:2407.15540 (2024)
35. Li, X., Wang, S., Zhao, Y., Verbeek, J., Kannala, J.: Hierarchical scene coordinate classification and regression for visual localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11983–11992 (2020)
36. Li, Y., Snavely, N., Huttenlocher, D.P.: Location recognition using prioritized feature matching. In: Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11. pp. 791–804. Springer (2010)
37. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2041–2050 (2018)
38. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17627–17638 (2023)
39. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**, 91–110 (2004)
40. Moon, H., Lee, C., Hong, J.H.: Efficient privacy-preserving visual localization using 3d ray clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9773–9783 (2024)
41. Moon, H., Lee, J., Kim, J., Hong, J.H.: Depth-guided privacy-preserving visual localization using 3d sphere clouds. In: Proceedings of the British Machine Vision Conference (2024)
42. Mur-Artal, R., Tardós, J.D.: Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE transactions on robotics **33**(5), 1255–1262 (2017)
43. Pan, L., Schönberger, J.L., Larsson, V., Pollefeys, M.: Privacy preserving localization via coordinate permutations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18174–18183 (2023)
44. Pietrantoni, M., Csurka, G., Sattler, T.: Gaussian splatting feature fields for (privacy-preserving) visual localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
45. Pietrantoni, M., Humenberger, M., Csurka, G., Sattler, T.: Can we make nerf-based visual localization privacy-preserving? arXiv preprint arXiv:2508.18971 (2025)
46. Pietrantoni, M., Humenberger, M., Sattler, T., Csurka, G.: Segloc: Learning segmentation-based representations for privacy-preserving visual localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15380–15391 (2023)
47. Pittaluga, F., Koppal, S.J., Kang, S.B., Sinha, S.N.: Revealing scenes by inverting structure from motion reconstructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 145–154 (2019)

48. Revaud, J., Almazán, J., Rezende, R.S., Souza, C.R.d.: Learning with average precision: Training image retrieval with a listwise loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5107–5116 (2019)
49. Revaud, J., De Souza, C., Humenberger, M., Weinzaepfel, P.: R2d2: Reliable and repeatable detector and descriptor. Advances in neural information processing systems **32** (2019)
50. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12716–12725 (2019)
51. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020)
52. Sattler, T., Havlena, M., Radenovic, F., Schindler, K., Pollefeys, M.: Hyperpoints and fine vocabularies for large-scale location recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2102–2110 (2015)
53. Sattler, T., Leibe, B., Kobbelt, L.: Fast image-based localization using direct 2d-to-3d matching. In: 2011 International Conference on Computer Vision. pp. 667–674. IEEE (2011)
54. Sattler, T., Leibe, B., Kobbelt, L.: Improving image-based localization by active correspondence search. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12. pp. 752–765. Springer (2012)
55. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & effective prioritized matching for large-scale image-based localization. IEEE transactions on pattern analysis and machine intelligence **39**(9), 1744–1756 (2016)
56. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., Pajdla, T.: Benchmarking 6dof outdoor visual localization in changing conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8601–8610 (2018)
57. Sattler, T., Torii, A., Sivic, J., Pollefeys, M., Taira, H., Okutomi, M., Pajdla, T.: Are large-scale 3d models really necessary for accurate visual localization? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1637–1646 (2017)
58. Shavit, Y., Ferens, R., Keller, Y.: Learning multi-scene absolute pose regression with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2733–2742 (2021)
59. Shi, Y., Cai, J.X., Shavit, Y., Mu, T.J., Feng, W., Zhang, K.: Clustergnn: Cluster-based coarse-to-fine graph neural network for efficient feature matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12517–12526 (2022)
60. Shibuya, M., Sumikura, S., Sakurada, K.: Privacy preserving visual slam. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 102–118. Springer (2020)
61. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in rgb-d images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2930–2937 (2013)
62. Sinkhorn, R., Knopp, P.: Concerning nonnegative matrices and doubly stochastic matrices. Pacific Journal of Mathematics **21**(2), 343–348 (1967)

63. Speciale, P., Schonberger, J.L., Kang, S.B., Sinha, S.N., Pollefeys, M.: Privacy preserving image-based localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5493–5503 (2019)
64. Speciale, P., Schonberger, J.L., Sinha, S.N., Pollefeys, M.: Privacy preserving image queries for camera localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1486–1496 (2019)
65. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8922–8931 (2021)
66. Tyszkiewicz, M., Fua, P., Trulls, E.: Disk: Learning local features with policy gradient. Advances in Neural Information Processing Systems **33**, 14254–14265 (2020)
67. Wang, F., Jiang, X., Galliani, S., Vogel, C., Pollefeys, M.: Glace: Global local accelerated coordinate encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21562–21571 (2024)
68. Wang, Q.: Mad-dr: Map compression for visual localization with matchness aware descriptor dimension reduction. In: European Conference on Computer Vision (ECCV). pp. 261–278. Springer (2025)
69. Wang, S., Kannala, J., Barath, D.: DGC-GNN: Leveraging geometry and color cues for visual descriptor-free 2d-3d matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20881–20891 (2024)
70. Wang, S., Laskar, Z., Melekhov, I., Li, X., Kannala, J.: Continual learning for image-based camera localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3252–3262 (2021)
71. Wang, S., Laskar, Z., Melekhov, I., Li, X., Zhao, Y., Tolias, G., Kannala, J.: Hscnet++: Hierarchical scene coordinate classification and regression for visual localization with transformer. International Journal of Computer Vision pp. 1–21 (2024)
72. Wang, Y., He, X., Peng, S., Tan, D., Zhou, X.: Efficient loftr: Semi-dense local feature matching with sparse-like speed. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21666–21675 (2024)
73. Yang, L., Shrestha, R., Li, W., Liu, S., Zhang, G., Cui, Z., Tan, P.: Scenesqueezer: Learning to compress scene for camera relocalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8259–8268 (2022)
74. Zhang, J., Xia, Z., Dong, M., Shen, S., Yue, L., Zheng, X.: Comatcher: Multi-view collaborative feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21970–21980 (2025)
75. Zhang, Y., Wang, S., Kannala, J.: A2-gnn: Angle-annular gnn for visual descriptor-free camera relocalization (2025), `https://arxiv.org/abs/2502.20036`
76. Zhou, Q., Agostinho, S., Ošep, A., Leal-Taixé, L.: Is geometry enough for matching in visual localization? In: European Conference on Computer Vision. pp. 407–425. Springer (2022)
77. Zuliani, M.: Ransac for dummies. Vision Research Lab, University of California, Santa Barbara (2009)