Learning to Produce Semi-dense Correspondences for Visual Localization

Khang Truong Giang[1] Soohwan Song[2]* Sungho Jo[1]*
[1] School of Computing, KAIST, Daejeon, Republic of Korea
[2] College of AI Convergence, Dongguk University, Seoul, Republic of Korea
*corresponding authors

Abstract

This study addresses the challenge of performing visual localization in demanding conditions such as nighttime scenarios, adverse weather, and seasonal changes. While many prior studies have focused on improving image matching performance to facilitate reliable dense keypoint matching between images, existing methods often heavily rely on predefined feature points on a reconstructed 3D model. Consequently, they tend to overlook unobserved keypoints during the matching process. Therefore, dense keypoint matches are not fully exploited, leading to a notable reduction in accuracy, particularly in noisy scenes. To tackle this issue, we propose a novel localization method that extracts reliable semi-dense 2D-3D matching points based on dense keypoint matches. This approach involves regressing semi-dense 2D keypoints into 3D scene coordinates using a point inference network. The network utilizes both geometric and visual cues to effectively infer 3D coordinates for unobserved keypoints from the observed ones. The abundance of matching information significantly enhances the accuracy of camera pose estimation, even in scenarios involving noisy or sparse 3D models. Comprehensive evaluations demonstrate that the proposed method outperforms other methods in challenging scenes and achieves competitive results in large-scale visual localization benchmarks. The code will be available at https://github.com/TruongKhang/DeViLoc.

1. Introduction

Visual localization is the process of determining the 6 degrees of freedom (DoF) camera pose for a given query image within a known scene. This fundamental task in computer vision is critical for applications such as robot navigation [33] and virtual or augmented reality [35, 37]. Most leading studies primarily employ a structure-based approach [8, 41, 43, 47, 68], consistently exhibiting high localization performance across diverse challenging conditions [31, 42, 54, 58].

Traditionally, structure-based methods heavily rely on feature matching (FM) [11, 46, 51, 55, 64]. These methods establish sparse correspondences between 3D points and 2D pixel-level keypoints in images, followed by estimating camera poses using RANSAC-based Perspective-n-Point (PnP). The recent advancements in FM-based methods [41, 42, 53] have shown outstanding performance across various benchmarks, particularly in large-scale scenes. However, despite these achievements, FM-based methods encounter substantial challenges in practical scenarios, including dealing with complex lighting conditions, seasonal variations, and changes in perspectives.

Addressing these challenges necessitates a more robust and informative feature-matching approach incorporating detailed 3D points. Current methods, including those achieving semi-dense 2D-2D correspondences through detector-free image matching [15, 24, 53], are limited by relying solely on matched sparse features. This limitation persists when considering semi-dense matches, which only account for keypoints observed in a 3D model, neglecting valuable information from unobserved keypoints. Furthermore, FM-based methods demand a detailed 3D point cloud map for accurate localization. Continuously refining 3D feature points for new image inputs is time-intensive, and in some scenarios, localization must be performed with a noisy or sparse 3D point cloud. Overreliance on predetermined 2D and 3D points in the database can lead to a degradation in pose accuracy, especially in noisy cases with texture-less surfaces or repetitive patterns.

Recent studies introduce scene coordinate regression (SCR) methods [6, 8, 30, 56], aiming to achieve dense 2D-3D correspondences. Unlike FM-based approaches, SCR methods use an implicit representation of scenes as a learnable function, predicting the dense 3D scene coordinates of a query image. SCR methods excel in compact and stable settings, eliminating the need for storing 3D models; however, they face challenges in dynamic environments and adapting to new viewpoints, limiting their applicability in large-scale scenes. Therefore, there is a need for an alternative method capable of finding dense and accurate 2D-3D

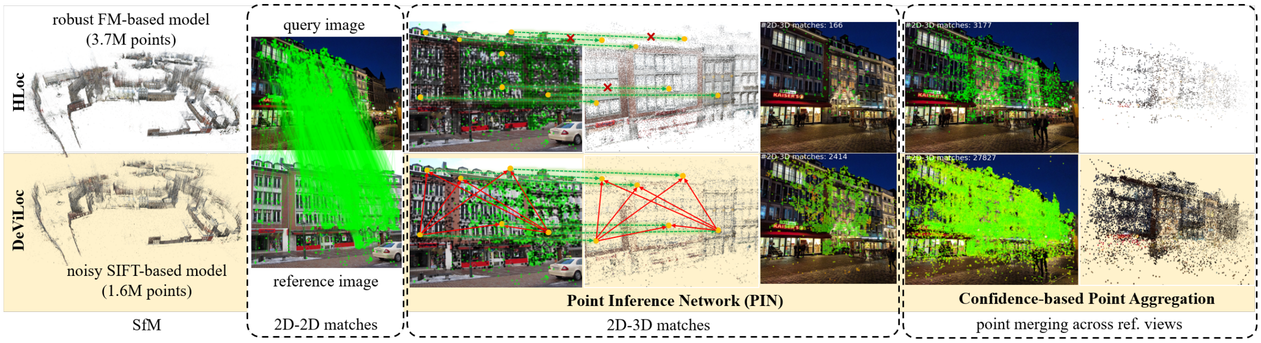

Figure 1. The comparison of the 2D-3D correspondence finding process in our method (DeViLoc) and an existing method (HLoc [41]). HLoc heavily relies on a robust 3D point cloud but discards many detected 2D keypoints (depicted in green) during the 2D-3D matching process. In contrast, our method efficiently handles a noisy point cloud through the point inference process of PIN. PIN transforms the entire set of 2D-2D matches into 2D-3D matches. Our method then produces numerous accurate 2D-3D matches across multiple views using a confidence-based aggregation module. These abundant matches significantly enhance localization performance, particularly in scenarios characterized by noisy or sparse 3D point clouds.

correspondences for visual localization.

Therefore, this study proposes a novel FM-based method, semi-**De**nse **Vi**sual **Loc**alization (DeViLoc), aiming to predict dense 2D-3D correspondences for robust and accurate localization. In contrast to existing FM-based methods relying solely on pre-existing 3D points, our method directly converts semi-dense 2D-2D matches into 2D-3D matches. This abundant 2D-3D match information significantly enhances the precision of camera pose estimation, even when dealing with noisy or sparse 3D models.

This method comprises two main components: 1) the *Point Inference Network* (PIN) and 2) the *Confidence-based Point Aggregation* (CPA) module. PIN plays a crucial role in our method by converting semi-dense 2D-2D matches into 2D-3D matches. It achieves this by directly regressing all 2D keypoints, both observed and unobserved, into 3D scene coordinates. The process involves encoding scene geometry from observed points into latent vectors and propagating 3D information to unobserved positions through attention layers. Next, the CPA module aggregates 2D-3D matches from multiple query-reference pairs, identifying consistent and highly confident 3D points corresponding to the same 2D keypoints in multiple matching views. This step effectively removes outliers from dense matches, and the filtered 2D-3D matches expedite RANSAC-based pose estimation. Ultimately, DeViLoc significantly increases the number of accurate 2D-3D matches for localization.

Fig. 1 illustrates the 2D-3D matching process of our method and an existing one (HLoc+LoFTR [41, 53]). The existing method, despite having a dense and precise 3D model, rejects many important points during 2D-3D estimation. In contrast, our method confidently transforms 2D-2D matches into 2D-3D matches, generating numerous matches even in the presence of noisy 3D input and night-time conditions. Consequently, the method yields robust and accurate localization results based on dense matching information, especially in challenging conditions like night-time scenarios, adverse weather, and seasonal changes.

This paper makes the following contributions:

- We propose a novel visual localization method that leverages rich matching information by directly converting semi-dense 2D-2D matches into 2D-3D matches. This method significantly improves the accuracy of camera pose estimation, particularly in scenarios with noisy or sparse 3D models.
- We introduce a network architecture, Point Inference Network (PIN), designed to directly regress 2D keypoints into 3D points. This network effectively exploits geometric and visual connections between unobserved and observed keypoints, ensuring accurate estimation of 3D information along with associated uncertainties.
- We conducted a comprehensive evaluation of our method across diverse datasets. The results indicate that our proposed approach outperforms other state-of-the-art methods in challenging scenes and achieves competitive performance in large-scale visual localization benchmarks. The source code is publicly available.

2. Related Works

Visual localization, which involves estimating camera poses from visual inputs, has been a subject of study for decades [8, 28, 31, 39, 49, 54]. Early approaches [1, 49, 59] primarily relied on an image retrieval strategy to directly estimate camera poses from the most similar images in database. This approach is intuitive and efficient, but its performance is significantly influenced by the density of images in the database. To overcome this limitation, an alternative approach learns a neural network to directly predict absolute camera poses from the input images [26, 27, 61].

Many studies [8, 9, 11, 25, 41, 47, 51, 54, 55, 68] have shifted their focus towards structure-based approach due to its stability and scalability in diverse scenes. The structure-based methods estimate camera pose by establishing a set of 2D-3D correspondences between image pixels and 3D coordinates of the scene. Leveraging this precise correspondence set, a camera pose can be accurately computed using a PnP solver [22, 29, 40] within a RANSAC paradigm [3, 4, 21]. The structure-based methods can be broadly categorized into two main groups: feature matching (FM) and scene coordinate regression (SCR).

Feature Matching. FM-based methods [11, 31, 41, 46, 51] initially reconstruct a 3D model of the environment from database images using Structure-from-Motion (SfM) [50, 63]. Each 3D point in the model is associated with one or several feature descriptors for localization. When a localization request for a query image is made, these methods detect a set of 2D keypoints along with their descriptors and proceed to match them with the 3D points. Several works generate the 2D-3D matches by examining all points in the 3D model [31, 44]; however, they face difficulty when dealing with large scenes.

To address this challenge, recent studies [41, 57, 64] have introduced a coarse-to-fine strategy. They initially identify a set of reference images in a database through image retrieval. They then establish 2D-3D matches based on 2D-2D matches between the query and reference images. Subsequent works have focused on enhancing the performance of image matching by incorporating transformers [15, 24, 42, 53] or semantic information [65]. These approaches produce robust and accurate 2D-2D matches, contributing to the robust construction of 3D model and the accurate generation of 2D-3D matches during the localization step. Consequently, FM-based methods have achieved state-of-the-art performance.

However, these methods exhibit inflexibility due to heavily relying on high-fidelity point cloud reconstruction. This time-consuming step is not suitable for online applications such as SLAM or robot navigation, where the 3D point cloud is constructed on-the-fly from a sequence of images. In this situation, the point cloud might be noisy or incomplete, thus degrading localization performance. Furthermore, existing FM-based methods only utilize observed 3D points in the database to generate 2D-3D matches, discarding numerous unrelated 2D-2D matches. This process might compromise the performance in the presence of noisy 3D inputs. On the other hand, our proposed framework aims to adapt to various kinds of 3D inputs and predict semi-dense 2D-3D correspondences.

Scene Coordinate Regression. In contrast to the explicit utilization of 3D models seen in FM-based methods, SCR [6, 8, 9, 13, 14, 52, 60] employs an implicit representation of scenes in a form of a machine learning model. This model predicts dense 3D scene coordinates for an input query image. While SCR-based methods offer a concise representation of scenes, they encounter challenges when adapting to large-scale scenes, novel scenes, or challenging conditions. Several approaches have been proposed to address these challenges by predicting a scene part-by-part [7, 30] or employing a coarse-to-fine prediction [56, 66]. However, these methods still exhibit lower performance compared to FM-based methods.

Drawing inspiration from SCR, our method aims to predict 3D coordinates for all keypoints in the query image, identified using a detector-free image-matching model. However, in contrast to the dense prediction in SCR, our semi-dense approach reduces the possibility of incorrect 3D prediction by focusing only on regions of interest between the query and reference images. Consequently, our method achieves more accurate performance than SCR-based methods in outdoor scenes.

3D Scene Representation. FM-based methods [24, 41, 42, 53] have proven to be effective, but they come with a drawback of requiring substantial storage capacity to store both the 3D coordinates and the associated visual features [41, 42]. Consequently, recent studies [10, 12, 16, 38, 67, 70] have shifted their focus toward discovering more space-efficient representations of the scene. For example, NeuMap [57] proposed encoding a 3D point cloud into a set of latent codes and then regressing 3D coordinates based on these codes. While these methods successfully reduce storage demands, they discard crucial scene information, thereby limiting their performance. In contrast, this study focuses on improving performance when dealing with challenging point cloud inputs.

3. Proposed Method

We follow the coarse-to-fine localization paradigm [41] that first retrieves a set of reference images via image retrieval, and then estimates a camera pose by finding 2D-3D correspondences between image pixels and 3D coordinates of the scene. The proposed method aims to address the task of 2D-3D prediction, while applying image retrieval and a PnP pose solver, similar to [41].

3.1. Overview

Given a query image I^q and a set of reference images $\{I^{r,1}...I^{r,N_v}\}$ retrieved from a database, our method establishes a set of 2D-3D correspondences $M = \{(k_i, p_i)\}$, where k_i is a 2D keypoint in the query image and p_i represents the corresponding 3D coordinate. Fig. 2 describes the overall architecture of the proposed method, DeViLoc. For each pair of query I^q and reference I^r, DeViLoc first extracts local 2D-2D feature correspondences $\{(k^q_i, k^r_i)\}$ using an image-matching network. Here, we employed a

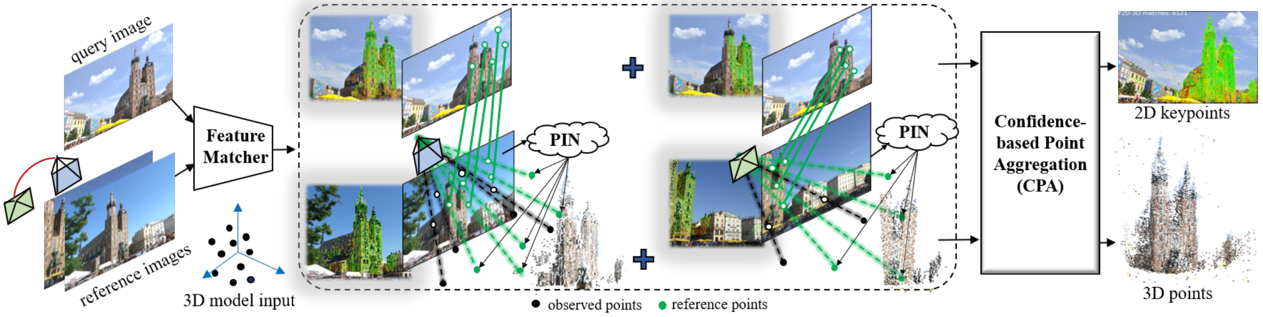

Figure 2. Overview of DeViLoc. First, a feature matcher is employed to detect 2D-2D matches for each pair of query-reference images. Subsequently, the PIN module infers a set of 3D coordinates for all detected 2D keypoints based on the observed data in the reference image. Finally, the CPA module integrates all 2D-3D matches obtained across all query-reference pairs.

detector-free image-matching model [24] to produce semi-dense correspondences.

The Point Inference Network (PIN) then converts detected 2D keypoints of the reference image into 3D points (Section 3.2). It takes reference image I^r with known camera parameters, along with a sparse set of observed 3D points as inputs, and predicts a set of 3D scene coordinates, $\{p_i\}$, corresponding to all keypoints $\{k_i^r\}$. Based on predicted 3D points $\{p_i\}$ and 2D-2D matches $\{(k_i^q, k_i^r)\}$, we could produce dense 2D-3D matches, $\{(k_i^q, p_i)\}$, for the query image. The basic concept of PIN is inspired by depth completion [62, 69], where a dense depth map is reconstructed from a sparse set of observed depth measurements and an input image. However, unlike depth completion, PIN aims to infer a discrete set of depth points, leading to reduced computational costs. Due to this discrete depth prediction, PIN utilizes attention [19] and MLP layers without the need for complex CNNs, as commonly found in conventional depth completion methods.

Next, the Confidence-based Point Aggregation (CPA) module integrates all 2D-3D matches from multiple query-reference image pairs (Section 3.3). CPA integrates 2D-3D matches with small distances of keypoints into a representative match. It effectively removes outliers by considering the confidence information of each match during the integration process.

3.2. Point Inference Network

For reference image I^r, let $O^r = \{o_i^r\}$ be the set of 3D points observed from the constructed 3D model. PIN predicts 3D points $P^r = \{p_i^r\}$ corresponding to detected 2D keypoints $K^r = \{k_i^r\}$ as:

$$P^r = \texttt{PIN}(K^r, O^r, F^r) \tag{1}$$

where $K^r \in \mathbb{R}^{N^r_{points} \times 2}$, $O^r \in \mathbb{R}^{N^o_{points} \times 3}$, and F^r is a feature map extracted from image I^r using convolutional networks ($F^r \in \mathbb{R}^{h \times w \times c}$). We directly utilized the output features of the image-matching network [24] as F^r.

The method begins with a preprocessing step where keypoints K^r and observed 3D points O^r are transformed into the same camera-coordinate system using the known camera parameters. Subsequently, we decompose the observed 3D points into observed keypoints $K^o \in \mathbb{R}^{N^o_{points} \times 2}$ and depth values $D^o \in \mathbb{R}^{N^o_{points} \times 1}$. The PIN network then aims to estimate depths D^r for reference keypoints K^r, utilizing observed keypoints K^o, observed depths D^o, and the feature map F^r. The network architecture of PIN is illustrated in Fig. 3, comprising two fundamental steps to leverage both spatial information and visual similarity for depth estimation: geometric guidance and visual guidance.

To implement geometric guidance, we initially employ two MLP-based encoders to learn embeddings for all 2D keypoint coordinates $\{K^r, K^o\}$, as well as for the observed depths, D^o. Additionally, we encode scene geometry from the observed depths by utilizing a self-attention layer.

$$D^o_{emb} = \texttt{DepthEnc}(D^o) \tag{2}$$

$$[K^r_{emb}, K^o_{emb}] = \texttt{KeypointEnc}([K^r, K^o]) \tag{3}$$

The relative position between the observed and unobserved keypoints serves as a primary geometric cue for propagating depth information from observed to unobserved positions. Therefore, we represent the observed positions with latent codes, P^o_{lc}, combined from keypoint embeddings K^o_{emb} and depth features D^o_{emb}. These latent codes are then passed into a cross-attention layer to generate latent codes P^r_{lc} related to reference keypoints K^r_{emb}:

$$P^r_{lc} = \texttt{CrsAtt}(K^r_{emb}, P^o_{lc}) \tag{4}$$

In Eq. 4, K^r_{emb} is the query, and P^o_{lc} represents the key and value for the cross-attention function, $\texttt{CrsAtt}(.)$.

The geometric guidance mentioned above may lack robustness in estimating accurate depths due to the sparsity of

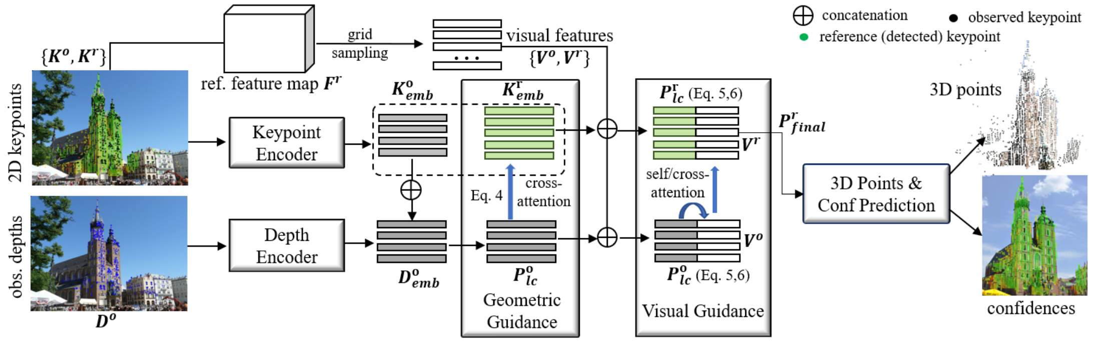

Figure 3. Point Inference Network (PIN). The network begins by learning embeddings for all keypoints (K^o_{emb}, K^r_{emb}) and observed depths (D^o_{emb}). Subsequently, attention layers are employed for both geometric and visual guidance. Finally, the learned latent codes (P^r_{lc}) are utilized to perform regression for the 3D points along with confidence values.

observed data points. Therefore, we also incorporate visual features for more detailed guidance. Considering the feature map F^r in Eq. 1, we utilize a bilinear grid sampling to extract sets of visual features V^r and V^o for both K^r and K^o. These visual features are appended to the latent codes, P^r_{lc} and P^o_{lc}. Subsequently, we employ multiple self/cross-attention layers to facilitate visual guidance:

$$P^r_{lc} = \texttt{Concat}(P^r_{lc}, V^r), P^o_{lc} = \texttt{Concat}(P^o_{lc}, V^o) \quad (5)$$

$$P^o_{lc} = \texttt{SelfAtt}(P^o_{lc}), P^r_{final} = \texttt{CrsAtt}(P^r_{lc}, P^o_{lc}) \quad (6)$$

After learning the final latent features, P^r_{final}, corresponding to reference keypoints K^r, we use two MLP-based networks to predict the depths and confidences:

$$D^r = \texttt{MLP}(P^r_{final}), \quad C^r = \texttt{MLP}([P^r_{final}, D^r]) \quad \text{(7a,b)}$$

Finally, the 3D scene coordinates for the reference keypoints are estimated as follows:

$$P^r = (T^{-1})_{cam \rightarrow scene}(D^r[\hat{K}^r, 1]^T) \quad (8)$$

where T^{-1} transforms the 3D points from camera-coordinate to scene-coordinate system. The confidences C^r represent the uncertainty associated with the predicted 3D coordinates, P^r.

3.3. Confidence-based Point Aggregation

PIN produced the 2D-3D correspondences, $(K^q, P) = \{(k^q_i, p_i)\}$, along with their associated confidence values $C^r = \{c_i\}$ for every query-reference pair (I^q, I^r). Subsequently, we aggregated the 2D-3D matches $\{(k^q_i, p_i)^n\}^{i=1...N_m}_{n=1...N_v}$ across N_v sets of matches, with N_m being the number of matches per set. The goal of CPA is to eliminate outliers from the aggregated matches and determine the final matches that exhibit high confidence and consistency. To accomplish this, we started by discarding matches with low confidence through a threshold τ. Next, we grouped adjacent matches using a keypoint quantization step. If the coordinates of two keypoints are closer than s pixels, these keypoints are assigned to the same group. This process is represented by a function Q_s, where s denotes the quantization size ($s \in \{2, 4\}$ in our experiments). Once matches within the same group were identified, we merged them using a confidence-based averaging operation.

Let k_j be a quantized keypoint, the formulas for the aggregated 3D point p^{agg}_j and the corresponding confidence c^{agg}_j can be written as follows:

$$p^{agg}_j = \frac{\sum_{i,n} \mathbf{1}(Q_s(k^n_i) = k_j)c^n_i p^n_i}{\sum_{i,n} \mathbf{1}(Q_s(k^n_i) = k_j)c^n_i} \quad (9)$$

$$c^{agg}_j = \frac{\sum_{i,n} \mathbf{1}(Q_s(k^n_i) = k_j)c^n_i}{N_{k_j}} \quad (10)$$

Here, $\mathbf{1}(Q_s(k^n_i) = k_j)$ is a binary indicator, and $N_{k_j} = \sum_{i,n} \mathbf{1}(Q_s(k^n_i) = k_j)$ represents the number of keypoints quantized into k_j.

3.4. Loss Functions

In summary, the proposed approach generates a set of 2D-3D matches, denoted as $M = \{(k_j, p^{agg}_j, c^{agg}_j)\}$, for query image I^q. Utilizing ground-truth camera matrices (C^q, T^q) and available depth information $D^q = \{d^q_j\}$, we define two functions, incorporating point-matching loss and confidence loss, to train the proposed network.

Point-matching loss. We computed the ground-truth 3D coordinates $\{p^{gt}_j\}$ for the keypoints $\{k_j\}$ using the depths $\{d^q_j\}$ and camera matrices (C^q, T^q). Then, we employed the $L1$ function to calculate the loss between the ground-truth 3D points and predicted 3D points:

$$L^q_{point} = \frac{1}{|M|} \sum_j ||p^{agg}_j - p^{gt}_j|| \quad (11)$$

Confidence loss. To train the confidences, we projected the 3D points $\{p^{agg}_j\}$ onto the image plane and assigned a

	Methods	7scenes (indoor)							Cambridge landmarks (outdoor)				
		Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Court	King's	Hospital	Shop	St. Mary's
D	HSCNet [30]	2/0.7	2/0.9	1/0.9	3/0.8	**4/1.0**	4/1.2	**3/0.8**	28/0.2	18/0.3	19/0.3	6/0.3	9/0.3
	DSAC++ [6]	**2/0.5**	2/0.9	1/0.8	**3/0.7**	4/1.1	4/1.1	9/2.6	40/0.2	18/0.3	20/0.3	6/0.3	13/0.4
	DSAC* [8]	2/1.10	2/1.24	1/1.82	3/1.15	4/1.34	4/1.68	3/1.16	49/0.3	15/0.3	21/0.4	5/0.3	13/0.4
	SANet [66]	3/0.88	3/1.08	2/1.48	3/1.00	5/1.32	4/1.40	16/4.59	328/1.95	32/0.54	32/0.53	10/0.47	16/0.57
	DSM [56]	2/0.68	2/0.80	1/0.80	3/0.78	4/1.11	**3/1.11**	4/1.16	43/0.19	19/0.35	23/0.38	6/0.30	11/0.34
S	AS [47]	3/0.87	2/1.01	1/0.82	4/1.15	7/1.69	5/1.72	4/1.01	24/0.13	13/0.22	20/0.36	4/0.21	8/0.25
	InLoc [55]	3/1.05	3/1.07	2/1.16	3/1.05	5/1.55	4/1.31	9/2.47	-	-	-	-	-
	PixLoc [43]	2/0.80	**2/0.73**	1/0.82	3/0.82	4/1.21	3/1.20	5/1.30	30/0.14	14/0.24	16/0.32	5/0.23	10/0.34
	HLoc[SP+SG]	2/0.84	2/0.93	1/0.74	3/0.92	5/1.27	4/1.40	5/1.47	16/0.11	**12/0.20**	15/0.30	4/0.20	**7/0.21**
	HLoc[LoFTR]	3/0.93	2/0.87	1/0.73	4/1.02	5/1.24	5/1.48	6/1.47	19/0.11	16/0.26	16/0.29	4/0.21	9/0.26
SD	NeuMap [57]	2/0.81	3/1.11	2/1.17	3/0.98	4/1.11	4/1.33	4/1.12	**6/0.10**	14/0.19	19/0.36	6/0.25	17/0.53
	DeViLoc (Ours)	2/0.78	**2/0.74**	**1/0.65**	3/0.82	**4/1.02**	3/1.19	4/1.12	18/0.11	**12/0.21**	**13/0.28**	**4/0.18**	7/0.23

Table 1. Evaluation on 7scenes and Cambridge landmarks. The metrics are the median translation (cm) and rotation (o) errors. The SCR-based methods highlighted in red were trained per scene. The best and second-best results are marked in **bold** and cyan. DeViLoc outperforms the other methods in overall, despite being trained only on MegaDepth.

label to each point based on pixel error. If the error between the projected 2D point and the query keypoint k_j is less than θ pixels ($\theta = 8$ in our experiments), the corresponding confidence c_j^{agg} is labeled as $l_j = 1$, and vice versa. Consequently, the confidence loss, denoted as L^q_{conf}, is established through the binary cross-entropy function:

$$L^q_{conf} = \frac{1}{|M|}\sum_j (l_j \log c_j^{agg} + (1-l_j)\log(1-c_j^{agg})) \quad (12)$$

The final loss results from the combination of these two loss terms:

$$L^q = L^q_{point} + \lambda * L^q_{conf} \quad (13)$$

Here, λ was set to 0.25 in our experiments.

4. Experiments

4.1. Implementation Details

We conducted experiments on various datasets, including 7scenes [52], Cambridge [27], Aachen Day-Night [45, 48], RobotCar Seasons [36, 48], and Extended CMU Seasons [2, 58].

Detailed pipeline. To demonstrate the adaptability of the proposed method to noisy and sparse 3D inputs, we utilized 3D point clouds generated through SIFT-based SfM in COLMAP [34, 50]. These 3D point clouds are consistently available across all datasets. We employed DenseVLad [59], NetVLad [1], or CosPlace [5] to retrieve the top-k reference images. Subsequently, the proposed framework was applied to predict 2D-3D matches from these inputs. During the prediction, we utilized the efficient feature-matching model, TopicFM [24], to generate semi-dense 2D-2D matches. Finally, the camera pose was computed from the 2D-3D matches using PnP functions in COLMAP.

Training. Our network was trained on MegaDepth [32], comprising outdoor scenes from various locations. The trained model was directly used for evaluation across all datasets, eliminating the need for finetuning or retraining.

4.2. Evaluation on Cambridge and 7scenes

We conducted a comparative analysis of our method (**DeViLoc**) against various state-of-the-art structure-based methods. Based on the characteristics outlined in the related works (Section 2), we primarily organized these methods into three categories: dense (D), sparse (S), and semi-dense (SD) methods. In this classification, SCR-based methods like HSCNet [30], SANet [66], DSAC* [8], DSAC++ [6], and DSM [56] were placed in the dense group (D) due to their strategy of making dense 2D-3D predictions. Conversely, methods such as Active Search [47], InLoc [55], and HLoc[SP+SG] [18, 41, 42] executed a sparse matching process between 2D keypoints and the 3D point cloud, classifying them in the sparse group (S). Despite PixLoc [43] not directly detecting 2D keypoints in the query image, its effective utilization of 3D points from a sparse point cloud to find corresponding 2D positions in the query image led us to categorize PixLoc in group S as well. In contrast to both groups (D and S), our method predicts 3D coordinates for all detected keypoints without any point rejection. To the best of our knowledge, NeuMap [57] is the most similar work to ours. Therefore, we assigned DeViLoc and NeuMap to the semi-dense group (SD).

We compared the estimated camera poses of all methods to the ground-truth poses, calculating translation (in cm) and rotation (in degrees) errors [27] and presenting the median errors for each scene in Table 1. The results highlight the effectiveness of dense methods, following the SCR approach, in indoor scenes. This success is attributed to their scene-specific training (e.g., HSCNet, DSAC++, DSAC*) or training on extensive indoor datasets (e.g., DSM, trained on ScanNet [17]). However, these dense methods face challenges in generalizing to outdoor scenes in Cambridge. In contrast, despite being trained on outdoor scenes, our method demonstrates superior performance compared to the dense methods on the 7scenes dataset. The method secures first-place rankings in three scenes (Fire, Heads, Pumpkin)

	Methods	Aachen Day-Night		RobotCar-Seasons		Extended CMU-Seasons		
		Day	Night	Day-all	Night-all	Urban	Suburban	Park
D	ESAC [7]	42.6 / 59.6 / 75.5	6.1 / 10.2 / 18.4	-	-	-	-	-
S	AS [47]	85.3 / 92.2 / 97.9	39.8 / 49.0 / 64.3	50.9 / 80.2 / 96.6	6.9 / 15.6 / 31.7	81.0 / 87.3 / 92.4	62.6 / 70.9 / 81.0	45.5 / 51.6 / 62.0
	D2Net [20]	84.8 / 92.6 / 97.5	84.7 / 90.8 / 96.9	54.5 / 80.0 / 95.3	20.4 / 40.1 / 55.0	94.0 / 97.7 / 99.1	93.0 / 95.7 / 98.3	89.2 / 93.2 / 95.0
	S2DNet [23]	84.5 / 90.3 / 95.3	74.5 / 82.7 / 94.9	53.9 / 80.6 / 95.8	14.5 / 40.2 / 69.7	-	-	-
	HLoc[SP] [18, 41]	80.5 / 87.4 / 94.2	68.4 / 77.6 / 88.8	53.1 / 79.1 / 95.5	7.2 / 17.4 / 34.4	89.5 / 94.2 / 97.9	76.5 / 82.7 / 92.7	57.4 / 64.4 / 80.4
	PixLoc [43]	64.3 / 69.3 / 77.4	51.0 / 55.1 / 67.3	52.7 / 77.5 / 93.9	12.0 / 20.7 / 45.4	88.3 / 90.4 / 93.7	79.6 / 81.1 / 85.2	61.0 / 62.5 / 69.4
	HLoc[SP+SG]	**89.6** / 95.4 / **98.8**	86.7 / **93.9** / **100.**	**56.9** / 81.7 / **98.1**	33.3 / 65.9 / 88.8	95.5 / 98.6 / **99.3**	90.9 / 94.2 / 97.1	85.7 / 89.0 / 91.6
	LBR [64]	88.3 / **95.6** / **98.8**	84.7 / **93.9** / **100.**	56.7 / 81.7 / **98.2**	24.9 / 62.3 / 86.1	-	-	-
	HLoc+PixLoc	84.7 / 94.2 / **98.8**	81.6 / **93.9** / **100.**	**56.9** / **82.0** / **98.1**	**34.9** / 67.7 / 89.5	**96.9** / **98.9** / **99.3**	93.3 / 95.4 / 97.1	87.0 / 89.5 / 91.6
	HLoc[TopicFM][24]	88.8 / 94.7 / 97.9	86.7 / 92.9 / **100.**	-	-	-	-	-
SD	NeuMap [57]	80.8 / 90.9 / 95.6	48.0 / 67.3 / 87.8	-	-	-	-	-
	DeViLoc (Ours)	87.4 / 94.8 / 98.2	**87.8** / **93.9** / **100.**	**56.9** / 81.8 / **98.0**	31.3 / **68.9** / **92.4**	95.7 / 98.4 / **99.2**	**97.1** / **98.3** / **99.4**	**92.1** / **95.1** / **96.3**

Table 2. Evaluated results on the long-term benchmark [48] using the recall metrics at thresholds of $\{(25cm, 2^o), (50cm, 5^o), (5m, 10^o)\}$. We compare with various complex baselines that integrate robust FM models into HLoc [41] or use PixLoc to refine HLoc's poses. Our method achieves state-of-the-art performance, especially in the highly challenging localization in the CMU dataset (marked in **bold red**).

and second-place ranking in one scene (Kitchen). It also outperforms dense methods on the Cambridge dataset.

In comparison to methods in group S, DeViLoc consistently delivers superior performance. Among these, HLoc[SP+SG] is the only method achieving competitive results with our approach on the Cambridge dataset. Notably, HLoc employs a complex pipeline involving re-triangulating SIFT-based point clouds using robust local features and matches detected by SuperPoint (SP) [18] and SuperGlue (SG) [42]. In contrast, our method directly employs the noisy SIFT-based inputs and generally outperforms HLoc on both the 7scenes and Cambridge datasets.

4.3. Evaluation on large-scale challenging scenes

We compared DeViLoc with several contemporary image-matching methods including D2Net [20], SP+SG [18, 42], and TopicFM [24], which are frequently incorporated into the HLoc pipeline as robust feature matchers. When using these methods, HLoc requires the extraction of new local features and matches from the database to recalculate the 3D point clouds. Furthermore, we also compared DeViLoc to other methods such as PixLoc [43], ESAC [7], and NeuMap [57], which are not based on the HLoc pipeline. Table 2 presents the results of all these methods.

Aachen. DeViLoc performed comparably with recent FM-based baselines. Additionally, when compared to HLoc[TopicFM], which utilized the same feature matcher, our method exhibited an overall superior performance, demonstrating the effectiveness of the proposed pipeline.

RobotCar. As illustrated in Table 2, our method surpassed other methods and demonstrated competitive performance compared to HLoc[SP+SG] for the day-time queries. Particularly noteworthy is that DeViLoc significantly outperformed HLoc[SP+SG] for the night-time queries. It is important to highlight that both SP and SG were trained using multiple datasets, while DeViLoc was exclusively trained on MegaDepth [32].

CMU. Our method significantly outperformed the state-of-the-art pipeline HLoc[SP+SG] with a large margin. Compared to the complex pipeline HLoc+PixLoc, which uses PixLoc to refine the estimated poses of HLoc[SP+SG], DeViLoc improved accuracy by up to 5.1% on the scene "Park". This demonstrates the effectiveness and stability of our approach in difficult localization conditions.

4.4. Ablation Study

Evaluating the performance with noisy and sparse inputs. To assess how effectively DeViLoc handles noisy and sparse inputs, we conducted an experiment using the Aachen Day-Night dataset. Our pipeline utilized 3D point cloud inputs generated by different image-matching models, including SIFT, SP+SG, and LoFTR. As illustrated in Fig. 4, the 3D point cloud generated by SIFT exhibits significantly more noise and sparsity compared to those produced by the SP+SG and LoFTR models. The results for each input model are presented in Table 3. We observe that the precise and dense 3D inputs from SP+SG or LoFTR only slightly improve performance. This highlights the adaptability of our approach in handling various types of 3D inputs, proving effective even in the presence of noisy and sparse data.

Visualization of semi-dense matching. Fig. 5 illustrates the detected 2D keypoints and the corresponding 3D points produced by DeViLoc. We visualized multiple pairs of query and reference views. Despite a higher number of detected keypoints compared to the observed keypoints (especially in reference view 1), our method is capable of effectively estimating 3D points along with their uncertainties. Notably, points in the sky or near-edge regions of the scene tend to exhibit lower confidence. Ultimately, our method yields a significant number of 2D-3D matches after the point aggregation step, substantially improving performance, particularly in challenging scenes.

Effectiveness of proposed modules. We implemented three models to measure the contributions of the proposed modules, PIN (Section 3.2) and CPA (Section 3.3):

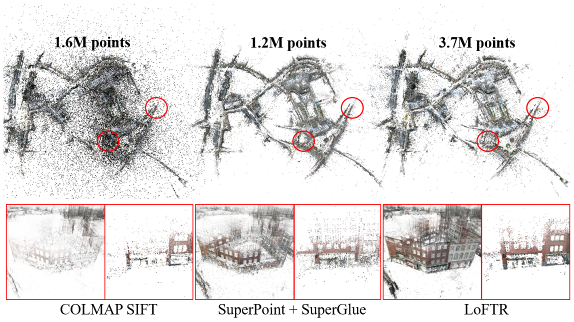

Figure 4. Comparison between point clouds built from traditional FM (SIFT [34]), sparse FM (SP+SG [18, 42]), and detector-free FM (LoFTR [53]). DeViLoc can handle well the noisy SIFT-based input to achieve competitive performance compared to the precise (SP+SG) or dense (LoFTR) inputs (shown in Table 3).

Models	Day	Night
	(0.25m,2^o) / (0.5m,5^o) / (5.0m,10^o)	
DeViLoc[half-SIFT]	87.5 / 94.1 / 97.9	86.7 / 92.9 / **100.**
DeViLoc[SIFT]	87.4 / 94.8 / 98.2	87.8 / **93.9** / **100.**
DeViLoc[SP+SG]	87.3 / **95.3** / **98.3**	**88.8** / 92.9 / **100.**
DeViLoc[LoFTR]	**87.9** / 94.7 / 98.2	**88.8** / 92.9 / **100.**

Table 3. Ablation study of DeViLoc on Aachen Day-Night when using different point cloud inputs.

- Model-A: This model executes the standard 2D-3D matching process without utilizing the PIN and CPA modules, akin to existing pipelines like HLoc [41].
- Model-B: This model only integrates the PIN module, generating semi-dense matches. All these matches are directly fed into the PnP solver for pose estimation.
- Model-C: This model combines both PIN and CPA, including the entire process of the proposed method.

Table 4 presents the results of pose estimation using AUC with thresholds of $\{2^o, 5^o, 10^o\}$ [42]. The findings indicate that the proposed modules (Model-B and Model-C) significantly enhance performance as compared to the baseline (Model-A). Notably, these modules do not impose a substantial runtime burden, requiring only about $100ms$ to generate a more extensive set of 2D-3D matches, ultimately enhancing overall performance.

Impact of the top-k reference images. Table 4 shows that employing more reference images can enhance performance. However, this improvement comes at the expense of significantly increased runtime and the number of matches.

Using different feature matcher. We evaluated the performance by substituting the detector-free TopicFM [24] with the detector-based SP+SG. The AUC metrics did not decrease significantly, as indicated in Table 4.

5. Conclusions and Limitations

This study introduces a robust structure-based framework for visual localization that minimizes reliance on the precise reconstruction of 3D point clouds. Our approach exhibits stable performance even when confronted with sparse and noisy 3D inputs. To achieve this, we present two novel modules: the Point Inference Network and the Confidence-based Point Aggregation. Consequently, the method generates numerous 2D-3D correspondences, leading to significant enhancements in challenging conditions, including textureless scenes, large-scale environments, and variations in weather and seasons. However, the computational efficiency of our proposed method has limitations. The runtime experiences a slowdown as the number of matching pairs between query and reference images increases. Addressing this limitation will be a focus of our future work.

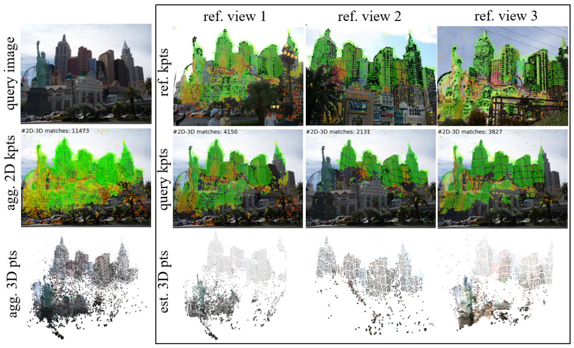

Figure 5. Illustration of 2D-3D correspondences estimated by DeViLoc for several pairs of images. The observed 2D keypoints are marked in black, while the reference keypoints are represented in orange (low confidence) or green (high confidence).

Model	AUC ($2^o/5^o/10^o$)	Time (s)	#P $\times 10^6$	#M
A[SIFT] (A ⇔ HLoc)	72.4 / 85.5 / 91.1	0.49	11.6	42
A[SP+SG]	72.8 / 86.4 / 92.2	0.53	11.6	54
A[LoFTR]	74.7 / 87.7 / 93.2	0.51	11.6	297
B (A+PIN)	78.8 / 90.0/ 94.7	0.56	15.9	4743
C (A+PIN+CPA)⇔DeViLoc	**79.6 / 90.5 / 94.9**	0.57	15.9	4246
C (TopicFM→SP+SG)	78.2 / 89.9 / 94.7	1.02	17.6	993
DeViLoc (top-5)	81.0 / 91.4 / 95.5	0.90	15.9	7668
DeViLoc (top-10)	**82.6 / 92.2 / 96.0**	1.92	15.9	13970

Table 4. Effectiveness of the proposed PIN and CPA (top), impact of the feature matcher (middle), and ablation of the top-k image retrieval (bottom) on MegaDepth [32]. The top-3 retrieval is used by default. Except for Model-A, the others were tested with SIFT inputs. **P** is the model parameters and **M** is the 2D-3D matches.

6. Acknowledgments

This work was supported by the Ministry of Trade, Industry and Energy (MOTIE, Korea) under Grant 20007058, and by the National Research Foundation of Korea (NRF) funded by the Korean Government (MSIT) under Grant RS-2023-00208052.

References

[1] 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*, pages 5297–5307, 2016. 2, 6

[2] Hernán Badino, Daniel Huber, and Takeo Kanade. Visual topometric localization. In *2011 IEEE Intelligent vehicles symposium (IV)*, pages 794–799. IEEE, 2011. 6

[3] Daniel Barath and Jiří Matas. Graph-cut ransac. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6733–6741, 2018. 3

[4] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. Magsac++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1304–1312, 2020. 3

[5] 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*, pages 4878–4888, 2022. 6

[6] Eric Brachmann and Carsten Rother. Learning less is more-6d camera localization via 3d surface regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4654–4662, 2018. 1, 3, 6

[7] Eric Brachmann and Carsten Rother. Expert sample consensus applied to camera re-localization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7524–7533, 2019. 3, 7

[8] Eric Brachmann and Carsten Rother. 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. 1, 2, 3, 6

[9] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6684–6692, 2017. 3

[10] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. 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*, pages 5044–5053, 2023. 3

[11] Jan Brejcha, Michal Lukáč, Yannick Hold-Geoffroy, Oliver Wang, and Martin Čadík. Landscapear: Large scale outdoor augmented reality by matching photographs with terrain models using learned descriptors. In *European Conference on Computer Vision*, pages 295–312. Springer, 2020. 1, 3

[12] Federico Camposeco, Andrea Cohen, Marc Pollefeys, and Torsten Sattler. Hybrid scene compression for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7653–7662, 2019. 3

[13] Tommaso Cavallari, Luca Bertinetto, Jishnu Mukhoti, Philip Torr, and Stuart Golodetz. Let's take this online: Adapting scene coordinate regression network predictions for online rgb-d camera relocalisation. In *2019 International Conference on 3D Vision (3DV)*, pages 564–573. IEEE, 2019. 3

[14] Tommaso Cavallari, Stuart Golodetz, Nicholas A Lord, Julien Valentin, Victor A Prisacariu, Luigi Di Stefano, and Philip HS Torr. Real-time rgb-d camera pose estimation in novel scenes using a relocalisation cascade. *IEEE transactions on pattern analysis and machine intelligence*, 42(10): 2465–2477, 2019. 3

[15] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022. 1, 3

[16] Wentao Cheng, Weisi Lin, Kan Chen, and Xinfeng Zhang. Cascaded parallel filtering for memory-efficient image-based localization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1032–1041, 2019. 3

[17] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6

[18] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 6, 7, 8

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[20] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*, 2019. 7

[21] Martin A Fischler and Robert C Bolles. A paradigm for model fitting with applications to image analysis and automated cartography (reprinted in readings in computer vision, ed. ma fischler. *Comm. ACM*, 24(6):381–395, 1981. 3

[22] Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang, and Hang-Fei Cheng. Complete solution classification for the perspective-three-point problem. *IEEE transactions on pattern analysis and machine intelligence*, 25(8):930–943, 2003. 3

[23] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. S2dnet: Learning image features for accurate sparse-to-dense matching. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 626–643. Springer, 2020. 7

[24] Khang Truong Giang, Soohwan Song, and Sungho Jo. Topicfm: Robust and interpretable topic-assisted feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2447–2455, 2023. 1, 3, 4, 6, 7, 8

[25] Martin Humenberger, Yohann Cabon, Nicolas Guerin, Julien Morat, Vincent Leroy, Jérôme Revaud, Philippe Rerole, Noé

Pion, Cesar de Souza, and Gabriela Csurka. Robust image retrieval-based visual localization using kapture. *arXiv preprint arXiv:2007.13867*, 2020. 3

[26] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5974–5983, 2017. 3

[27] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 3, 6

[28] Minjung Kim, Junseo Koo, and Gunhee Kim. Ep2p-loc: End-to-end 3d point to 2d pixel localization for large-scale visual localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21527–21537, 2023. 2

[29] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *CVPR 2011*, pages 2969–2976. IEEE, 2011. 3

[30] Xiaotian Li, Shuzhe Wang, Yi Zhao, Jakob J. Verbeek, and Juho Kannala. Hierarchical scene coordinate classification and regression for visual localization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11980–11989, 2019. 1, 3, 6

[31] Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua. Worldwide pose estimation using 3d point clouds. In *European conference on computer vision*, pages 15–29. Springer, 2012. 1, 2, 3

[32] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 6, 7, 8

[33] Hyon Lim, Sudipta N Sinha, Michael F Cohen, and Matthew Uyttendaele. Real-time image-based 6-dof localization in large-scale environments. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1043–1050. IEEE, 2012. 1

[34] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 6, 8

[35] Simon Lynen, Torsten Sattler, Michael Bosse, Joel A Hesch, Marc Pollefeys, and Roland Siegwart. Get out of my lab: Large-scale, real-time visual-inertial localization. In *Robotics: Science and Systems*, page 1, 2015. 1

[36] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 6

[37] Sven Middelberg, Torsten Sattler, Ole Untzelmann, and Leif Kobbelt. Scalable 6-dof localization on mobile devices. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 268–283. Springer, 2014. 1

[38] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. Meshloc: Mesh-based visual localization. In *European Conference on Computer Vision*, pages 589–609. Springer, 2022. 3

[39] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. Visual localization using imperfect 3d models from the internet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13175–13186, 2023. 2

[40] Mikael Persson and Klas Nordberg. Lambda twist: An accurate fast robust perspective three point (p3p) solver. In *Proceedings of the European conference on computer vision (ECCV)*, pages 318–332, 2018. 3

[41] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 1, 2, 3, 6, 7, 8

[42] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 3, 6, 7, 8

[43] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3247–3257, 2021. 1, 6, 7

[44] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. 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*, pages 752–765. Springer, 2012. 3

[45] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *BMVC*, page 4, 2012. 6

[46] Torsten Sattler, Michal Havlena, Filip Radenovic, Konrad Schindler, and Marc Pollefeys. Hyperpoints and fine vocabularies for large-scale location recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2102–2110, 2015. 1, 3

[47] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016. 1, 3, 6, 7

[48] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8601–8610, 2018. 6, 7

[49] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2007. 2

[50] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 3, 6

[51] Johannes L Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. Semantic visual localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6896–6906, 2018. 1, 3

[52] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 3, 6

[53] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 1, 2, 3, 8

[54] Linus Svärm, Olof Enqvist, Fredrik Kahl, and Magnus Oskarsson. City-scale localization for cameras with known vertical direction. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1455–1461, 2016. 1, 2, 3

[55] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 1, 3, 6

[56] Shitao Tang, Chengzhou Tang, Rui Huang, Siyu Zhu, and Ping Tan. Learning camera localization via dense scene matching. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1831–1841, 2021. 1, 3, 6

[57] Shitao Tang, Sicong Tang, Andrea Tagliasacchi, Ping Tan, and Yasutaka Furukawa. Neumap: Neural coordinate mapping by auto-transdecoder for camera localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 929–939, 2023. 3, 6, 7

[58] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, et al. Long-term visual localization revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2074–2088, 2020. 1, 6

[59] 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*, pages 1808–1817, 2015. 2, 6

[60] Julien P. C. Valentin, Angela Dai, Matthias Nießner, Pushmeet Kohli, Philip H. S. Torr, Shahram Izadi, and Cem Keskin. Learning to navigate the energy landscape. *2016 Fourth International Conference on 3D Vision (3DV)*, pages 323–332, 2016. 3

[61] Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation. In *Proceedings of the IEEE international conference on computer vision*, pages 627–637, 2017. 3

[62] Yufei Wang, Bo Li, Ge Zhang, Qi Liu, Tao Gao, and Yuchao Dai. Lrru: Long-short range recurrent updating networks for depth completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9422–9432, 2023. 4

[63] Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, pages 127–134. IEEE, 2013. 3

[64] Fei Xue, Ignas Budvytis, Daniel Olmeda Reino, and Roberto Cipolla. Efficient large-scale localization by global instance recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17348–17357, 2022. 1, 3, 7

[65] Fei Xue, Ignas Budvytis, and Roberto Cipolla. Sfd2: Semantic-guided feature detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5216, 2023. 3

[66] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. Sanet: Scene agnostic network for camera localization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 42–51, 2019. 3, 6

[67] Luwei Yang, Rakesh Shrestha, Wenbo Li, Shuaicheng Liu, Guofeng Zhang, Zhaopeng Cui, and Ping Tan. Scenesqueezer: Learning to compress scene for camera relocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8259–8268, 2022. 3

[68] Bernhard Zeisl, Torsten Sattler, and Marc Pollefeys. Camera pose voting for large-scale image-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2704–2712, 2015. 1, 3

[69] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2023. 4

[70] Qunjie Zhou, Sérgio Agostinho, Aljoša Ošep, and Laura Leal-Taixé. Is geometry enough for matching in visual localization? In *European Conference on Computer Vision*, pages 407–425. Springer, 2022. 3