MULTILOC: LOOK AROUND AS YOU LOCALIZE FOR FAST AND ROBUST VISUAL RE-LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Relative camera pose is a basic geometric cue for visual re-localization and scene understanding. When using images alone, visual evidence may not sufficiently support reliable estimation of 3D-consistent motion. The challenge is especially acute in unfamiliar and ever-changing environments for real-time applications, where a system must infer spatial structure under tight time constraints from what it sees rather than rely on scene-specific reconstruction or training. This raises a fundamental question: how can live 3D camera motion be estimated accurately for re-localization in unseen environments while drawing on enough scene context to resolve pose ambiguity? We introduce MultiLoc, a multi-view-guided relative pose regressor trained at scale to achieve spatial and geometric consistent representations for robust visual re-localization in unseen scenarios with high inference speed. Specifically, MultiLoc creates a minimal 3D-sub-scene representation of the environment and efficiently transforms it into 3D spatially and geometrically consistent features in a single forward pass, enabling precise pose estimates with high inference speed gains. Across diverse indoor, outdoor, and in-the-wild visual re-localization benchmarks—Indoor6, Cambridge Landmarks and WaySpots—MultiLoc consistently outperforms state-of-the-art relative pose regression methods. We also highlight that MultiLoc, a pose regressor, also performs competitively with inference-heavy and structure-based approaches while retaining sub-second inference and generalizing to unseen scenes. Given a small posed support set, it also surpasses relative pose regression, feature-matching, and non-regression methods on several relative camera pose benchmarks. Code will be released.

1 INTRODUCTION

Visual Re-localization estimates an agent's position and orientation within an environment from visual observations alone. The said environment is represented by reference images and their camera poses. The poses, obtained through Structure-from-Motion, 3D scanners, or geo-tagging, provide a global anchor that corrects drift inherent in odometry and incremental SLAM. It also supports accurate pose estimation, system initialization, and loop closure, especially in GNSS-denied environments. Such capabilities make visual re-localization important for autonomous driving Liu et al. (2021b), navigation Suomela et al. (2023), augmented reality Castle et al. (2008), and other robotic applications.

Traditionally, structure-based methods Sarlin et al. (2021); Sattler et al. (2016); Taira et al. (2018) build a 3D map, match image features to map points, and estimate camera pose with PnP Gao et al. (2003) and RANSAC Fischler & Bolles (1981). They require high mapping time, substantial descriptor storage, and slow inference. Recent 3D foundation models like MASt3R Leroy et al. (2024) achieve state-of-the-art re-localization without storing descriptors but they still require substantial mapping time and memory for Structure-from-Motion (SfM) reconstruction Bruns et al. (2025), while remaining significantly slower during inference than relative pose regression (RPR) methods. Conversely, scene coordinate regression (SCR) methods Brachmann & Rother (2021); Li et al. (2020); Nguyen et al. (2024); Wang et al. (2024b); Brachmann et al. (2023) implicitly encode scenes to directly predict 2D-3D correspondences. Recent variants like the ACE family Brachmann et al. (2023); Wang et al. (2024a); Bruns et al. (2025) even enhance mapping efficiency, scene scalabil-

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

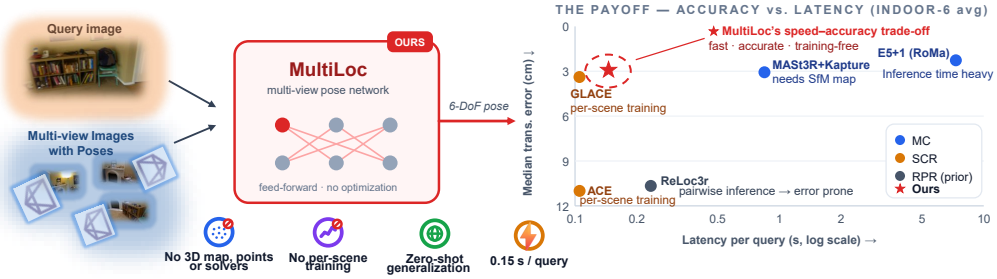


Figure 1: **One forward pass, no scene-training or map.** Given a query and K reference images with 6-DoF poses—a camera constellation, or 3D sub-scene (left)—MultiLoc achieves sub-scene consistent spatial and geometric representations by jointly fusing multi-view images and the available poses. By doing so, just with a single forward pass, MultiLoc matches the accuracy of feature matching (MC) and scene-coordinate regression (SCR) methods while offering one of the fastest inference speed across all the baselines on Indoor-6 (right).

ity, and robustness to domain shifts. However, both paradigms fundamentally necessitate per-scene optimization and an inference-time pose solver.

Interestingly, a third line of work under structureless feature-matching forgoes the 3D model entirely. Semi-generalized E5+1 solver Zheng & Wu (2015); Panek et al. (2026), for example, determines relative pose using five correspondences to a primary retrieved image and resolves absolute scale via one correspondence to a second, bypassing both explicit 3D maps and per-scene training. However, such methods typically operate on pair of images for 2D-2D matches and rely on geometric solvers while trading off their accuracy with inference speed compared to pose regression.

For real-time applications, pose regression methods are generally sought that encode scene representation in network weights and predict camera pose. Relative Pose Regression (RPR) methods, such as Dong et al. (2025); Li et al. (2025); Chen et al. (2021a); Balntas et al. (2018) are trained on image pairs to predict the relative camera pose between the two images. For visual localization, during inference, one image is the query image while another typically comes from the top-k retrieved geo-tagged reference images of the environment. The predicted relative poses are then utilized with ground-truth poses of the reference (or database) images to recover the absolute pose using methods like Umeyama alignment Umeyama (2002) or motion-averaging module Dong et al. (2025). Recently, Dong et al. (2025); Li et al. (2025) has shown the generalizability and faster inference speed capabilities of such feed-forward RPR based methods when trained on a large-scale of data. Despite achieving high performance and inference speed, pose regression methods often exhibit inferior performance compared to other paradigms. We attribute this limitation to a lack of global spatial awareness; by training on image pairs, these models merely approximate poses from pairwise relationships Arnold et al. (2022) rather than capturing the broader scene context. Furthermore, such works typically estimate relative camera poses without 3D-aware constraints, neglecting the geometric and structural consistency between pairs.

To address these limitations, we introduce **MultiLoc**, a novel large-scale trained relative pose regressor. It comprises of a multi-view guided pose regressor that aggregates holistic spatial cues from k reference images and their corresponding ground-truth camera poses in a single forward pass. Unlike existing RPR methods that rely solely on pairwise visual appearance, MultiLoc integrates multi-view inputs and poses to achieve superior spatial and geometric grounding as shown in Figure 1. By incorporating multiple reference views and poses directly into the regression phase, our model anchors the query image within a known 3D context, or what we refer as *3D sub-scene context*, providing highly accurate estimates for the camera pose estimation task. For visual relocalization, we follow the two stage pipeline that includes image retrieval followed by our pose regressor. It is worth noting that our work for re-localization is independent of any camera calibration parameters, scene-specific training and post-hoc geometric solver or optimization for metric scale recovery. Thus, MultiLoc departs from existing RPR methods, which regress relative poses from isolated image pairs and afterwards attain global metric poses through geometric solvers that inherit each pairwise estimate's errors. Finally, we demonstrate the zero-shot capability of our proposed method, MultiLoc, by benchmarking on multiple datasets. To the best of our knowledge,

MultiLoc is the first pose regression work to achieve visual re-localization performance comparable to other non-regression methods like scene coordinate regression or feature matching. Therefore, the main contributions of this work are as follows:

1. We propose a new approach for relative pose regression by recasting it to *anchor the query in a 3D sub-scene context during estimation for re-localization.* This helps grounding query pose prediction with multi-view spatial and geometric consistent features.
2. For effective 3D-sub-scene representation learning and grounding the query in such, we introduce a novel multi-view and geometry-aware relative pose regressor.
3. We present MultiLoc, a two-stage re-localization framework, based on our proposed regressor, attaining robust re-localization performance in unseen environments with only one forward pass and real-time inference speed without scene-specific optimizations.
4. Finally, we investigate effectiveness of the newly proposed paradigm by benchmarking on relative camera pose estimation, where our regressor outperforms both non-pose-regression and pose regression approaches; and on visual re-localization where MultiLoc achieves state-of-the-art within RPR and comparable performance with non-regression methods.

2 Related Work

Typically, a common body of work addresses both relative pose estimation and visual re-localization:

Structure-based Methods. These methods estimate pose from 2D–3D correspondences and typically require scene-specific mapping or training as discussed in Section 1, though ACE variants train in minutes Brachmann et al. (2023); Wang et al. (2024a); Bruns et al. (2025). Scene-agnostic regressors Revaud et al. (2024); Tang et al. (2021) avoid retraining but add inference costs through cost volumes or sparse SfM correspondences. Most recently, FastForward Barroso-Laguna et al. (2026) reduces mapping to retrieval and localizes in one feed-forward pass using ray-encoded features sampled from posed images instead of an explicit 3D model. However, it still regresses scene coordinates and uses PnP-RANSAC, requiring known camera intrinsics. MultiLoc directly regresses pose and avoids these requirements. Although FastForward advances visual re-localization, its retrieval GeM-AP checkpoint and source code are unavailable, preventing a fair direct benchmark. Nonetheless, we report its performance in the appendix for completeness.

Pose Regression Methods. Such regressors forgo intermediate correspondences and predict pose directly, either in a scene's absolute frame, called Absolute Pose Regression (APR), Kendall et al. (2015); Brahmbhatt et al. (2018); Shavit et al. (2021); Chen et al. (2021b) or relative to another image, termed as Relative Pose Regression (RPR), Dong et al. (2025); Li et al. (2025); Balntas et al. (2018); Chen et al. (2021a); Zhou et al. (2020); Winkelbauer et al. (2021); both offer real-time inference, with APR trailing RPR in cross-scene generalization. Bridging APR and structure-based methods, marepo Chen et al. (2024) decouples pose regression into a per-scene geometry network that predicts scene coordinates and a scene-agnostic transformer that regresses pose directly from them, attaining accuracy competitive with structure-based methods while replacing the explicit solver with a learned forward pass; its geometry network, however, is per-scene trained, so marepo still incurs mapping time, unlike the fully feed-forward RPR paradigm we adopt. Map-free re-localization Arnold et al. (2022) instead bridges RPR and feature matching, regressing metric relative pose to a single reference image via monocular depth and explicit matching, though it depends on known camera intrinsics for geometric back-projection. MultiLoc is likewise an RPR method, combining high accuracy and inference speed with no scene-specific training; accordingly, our localization benchmarks focus on comparison against such methods.

3D Foundation Models. Recent works use large-scale pre-training to learn generalizable 3D representations. DUSt3R Wang et al. (2024c), building upon CroCo Weinzaepfel et al. (2023), pioneered this shift with a pointmap representation that establishes dense 2D-3D correspondences from a pair of images. CUT3R Wang et al. (2025b) extends this approach to image streams through recurrent state updates. To reason over all views jointly, VGGT Wang et al. (2025a) introduced a unified transformer that processes multiple images simultaneously, capturing richer spatial context than independent pairs. A common limitation of the image-only variants is their inability to exploit auxiliary priors—such as known camera poses—that are readily available in visual re-localization and can offer tight geometric constraints. Pow3R Jang et al. (2025) and G-CUT3R Khafizov et al.

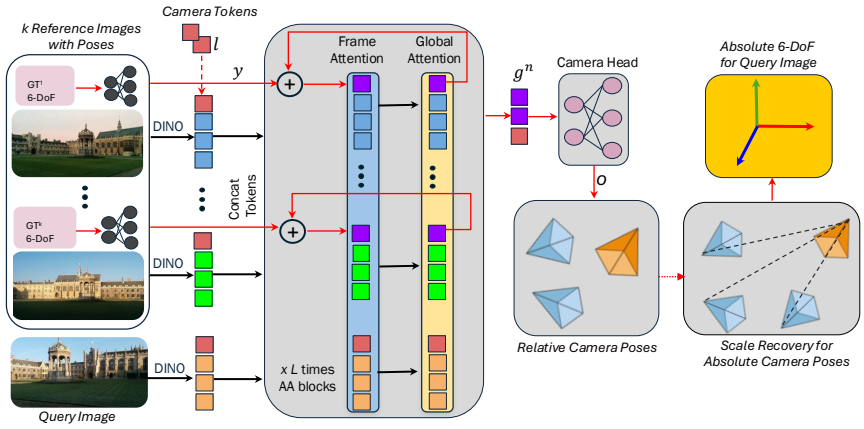


Figure 2: **MultiLoc Architecture overview.** The retrieved images' extrinsics are embedded via our camera encoders into camera tokens y and added with learnable camera tokens l, while all images pass through DINO-v2 to obtain patch tokens. All the tokens are then concatenated and processed by alternating-attention (AA) transformer blocks. Finally, camera tokens are used to predict relative poses between query and reference images using the camera head and metric scale is recovered by the pre-computed scene-scale s. The red arrows denote the flow of camera tokens.

(2025) address this for local paradigms: Pow3R conditions DUSt3R on camera intrinsics, poses, or depth, while G-CUT3R injects the same modalities into CUT3R's recurrent decoder via zero-convolution. However, both remain bounded by *local* processing—pairwise for Pow3R, sequential for G-CUT3R—so scene-level consistency across many views must still be recovered through post-hoc alignment or accumulated in a compressed state prone to drift. Concurrent to our work, MapAnything Keetha et al. (2026) generalizes this joint paradigm into a universal reconstruction backbone that accepts a variable number of images and optional geometric inputs such as camera intrinsics, poses, or depth. Therefore, we include MapAnything as a strong baseline in our work.

Visual Re-Localization. Building on these foundation models, ReLoc3r Dong et al. (2025) specialized DUSt3R for visual localization, achieving state-of-the-art relative pose regression through large-scale supervision. Re-localization, however, has a specific structure that these methods do not fully exploit. The task pairs a query image with a set of reference frames whose camera poses are known. We therefore propose to leverage a joint multi-view transformer, as in VGGT Wang et al. (2025a), conditioned on the poses of the reference images. We hypothesize that jointly processing all available views together with the reference camera poses facilitates a more holistic and better scene-grounded understanding of spatial context, leading to more robust pose estimation.

3 METHOD

Problem Statement. Given a database of k images, $\mathbf{I} = \{I_i\}_{i=1}^{k}, \mid I_i \in \mathbb{R}^{H\times W\times 3}$, from a scene with their absolute camera poses, $P = \{p_1, p_2, \ldots, p_k\} \mid p_i \in \mathrm{SE}(3)$, in a common world coordinate system, visual re-localization aims to find the absolute camera pose of a query image I_q.

Method Overview. We present MultiLoc, a robust relative pose regression framework designed for visual re-localization and camera pose estimation task. Unlike existing methods such as Dong et al. (2025); Li et al. (2025), which operate on isolated image pairs, MultiLoc ingests a sequence of $(k+1)$ images. By utilizing the first k relative camera poses as geometric priors along with the images, our model achieves a more comprehensive spatial understanding of the scene, leading to significantly enhanced relative camera pose estimation accuracy for the query frame.

3.1 ARCHITECTURE

We begin by describing the architecture that is common for both camera pose estimation and visual re-localization and as shown in Figure 2. Input to our model at a time are *k+1* images and k corresponding relative camera poses. The poses are input as members of $SE(3)$ manifold which are flattened to $\mathbb{R}^{12}$. A small multi-layer perceptron (MLP) is then applied to encode each of the

camera pose as: MLP : $p_i \in \mathbb{R}^{12} \mapsto y_i \in \mathbb{R}^d$. The poses are normalized by the average distance, $s = \frac{1}{k}\sum_{i=1}^{k} \|c_i\|$ where c_i is the translation component of a camera pose.

The images, $I = \{I_1, \ldots, I_k, I_q\}$, are tokenized in a patch of size 14 and fed into the DINO-v2 Oquab et al. (2023) model to obtain per-patch features $f \in \mathbb{R}^d$. Learnable camera tokens, $l = \{l_i\}_{i=1}^{k+1}$ are then added with ground-truth camera tokens, y, to form final pose tokens as: $g = l + y$. For the query image, we use $y_q = \mathbf{0} \in \mathbf{R}^d$. Finally, pose tokens are concatenated with register tokens, r and image patch tokens, f, to form the input to our transformer block as follows: $t = [g, r, f]$, where [.] is a concatenation operation.

Finally, our transformer model ingests all the tokens and first perform self (or frame) attention followed by cross (or global) attention across all the tokens t similar to Wang et al. (2025a). This allows each token, especially query image tokens, to be grounded with scene's 3D spatial context which pair-wise peers lag. The transformer block is repeated L times where after every frame attention we add the ground truth encoded pose tokens, y, for long-term context as $g^n = g^o + y$ where g^o is the output from previous layer and $g^o = g$ for the first transformer block.

Camera Pose Estimation. The learnable camera tokens are sampled from the $L - th$ transformer block's output and input into a simple four-layer self-attention-based camera head followed by a linear layer. The camera head predicts both the relative camera extrinsic and intrinsic parameters as:

$$C(.) : g^n \mapsto o \in [q, c, f_x, f_y]$$

where q_i is the quaternion representation $\in \mathbb{R}^4$, $c_i \in \mathbb{R}^3$ is the translation component at a normalized scale and finally f_x and f_y are the focal lengths.

3.2 Visual Re-Localization

We retrieve the top-10 reference images using NetVLAD Arandjelovic et al. (2016) and process them with the query as described in section 3.1. The query camera token predicts $o_{1,q}$ in the first reference frame at normalized scale. Rescaling its translation by s and composing with the first reference pose yields the global metric pose without geometric optimization. We compare this direct recovery with Umeyama alignment and motion averaging in the appendix Table 5.

3.3 Training Details

We initialize MultiLoc's weight with VGGT Wang et al. (2025a) and finetune on 4 H200s with a total of 1M iterations. Our model comprises of 1.1 billion parameters with $L = 24$ total transformer blocks and is trained using AdamW with learning rate of 1×10^{-5} decaying to 1×10^{-8}. The aspect ratio of input images is randomized to be between 0.33 and 1, where the width is resized to be of maximum length as 518 pixels. Data augmentations and input image sampling is applied similar to VGGT. For training we utilize Co3D Reizenstein et al. (2021), DL3DV Ling et al. (2024), VKITTI2 Cabon et al. (2020), BlendedMVS Yao et al. (2020), Re10K Zhou et al. (2018), MegaDepth Li & Snavely (2018), ARKitScenes Baruch et al. (2021) and ScanNet++ Yeshwanth et al. (2023).

Loss function. For training we extend the homoscedastic uncertainty-based geometric loss function Kendall et al. (2018) to regress on both the camera extrinsic and intrinsics parameters as follows:

$$\text{loss}_x = |x^{\text{pred}} - x^{\text{gt}}| \qquad (1) \qquad\qquad \text{loss}(.) = \sum_x \text{loss}_{\text{x}} \cdot e^{-s_x} + s_x \qquad (2)$$

where x is a variable being translation, rotation or focal length component and $s_x \in \{s_c, s_q, s_f\}$ are the learnable homoscedastic uncertainty scalar values for each of the x task variables respectively.

4 Experiments

Setup. Given that MultiLoc is a large-scale trained model, we evaluate its zero-shot generalization across diverse environments, including indoor, outdoor, natural and in the-wild domains. Our benchmarking focuses on two primary tasks: relative camera pose estimation and visual re-localization. We report AUC@θ for relative pose estimation following Dong et al. (2025), and median translation error (m or cm) and median rotation error (°) for visual re-localization.

Table 1: Relative camera pose estimation results on ACID, MegaDepth-1500, and ScanNet-1500. **Non-PR** denotes non-pose-regression methods, and **PR** relative pose regression. Shaded rows use multiple views; unshaded rows are pairwise. Red indicates the best and **bold** the second-best result.

		Method	ACID			MegaDepth-1500			ScanNet-1500		
			5°	10°	20°	5°	10°	20°	5°	10°	20°
Non-PR	Pair	Efficient LoFTR	–	–	–	52.80	69.19	81.18	19.20	37.00	53.60
		RoMa	57.98	68.53	76.62	60.92	75.63	85.69	34.13	55.28	71.69
		DUSt3R	21.50	35.95	49.70	27.90	46.00	63.30	23.81	45.91	65.57
		MASt3R (Ess.)	54.42	66.23	75.26	42.79	62.55	77.61	35.25	57.64	74.48
		NoPoSplat	48.60	61.70	72.80	–	–	–	31.80	53.80	71.70
	MV	MASt3R (Ess.)	49.51	64.68	76.08	48.68	68.16	82.20	33.40	55.27	73.42
		RoMa	58.54	70.10	78.79	68.05	**81.23**	89.76	28.22	49.91	68.88
PR	Pair	ReLoc3r	38.18	56.39	70.34	49.59	67.84	81.22	34.78	58.31	75.52
		VGGT	22.59	40.83	59.04	40.93	60.81	76.61	37.10	58.93	75.51
		MapAnything	26.31	45.95	63.46	27.17	47.83	67.04	24.65	47.69	67.64
		MultiLoc (Ours)	23.83	40.70	58.08	46.64	63.16	75.82	32.62	55.03	72.65
	MV	VGGT	22.55	40.85	58.75	46.25	66.46	80.70	38.82	60.31	76.48
		MapAnything	37.24	57.50	73.34	34.45	54.82	72.08	34.97	56.77	73.62
		MultiLoc (Ours)	76.56	87.15	93.31	69.49	81.98	**89.61**	62.34	76.32	86.57

For both evaluation tasks, we benchmark MultiLoc against representative methods from two broad categories: pose-regression and non-pose-regression approaches, including feature matching and scene-coordinate regression for comparative analysis on visual re-localization. However, it should be noted that MultiLoc is a relative pose regression method.

4.1 Relative Camera Pose Estimation

Relative camera pose estimation recovers relative pose between images. We evaluate pairwise relative pose estimation on ScanNet-1500 Dai et al. (2017); Sarlin et al. (2020), MegaDepth-1500 Sun et al. (2021), and the Aerial Coastline Imagery Dataset (ACID) Liu et al. (2021a). The datasets cover complementary settings, including indoor scenes, crowd-sourced outdoor landmark photos with diverse viewpoints, and sequential drone footage of natural coastlines. MegaDepth-1500 is drawn from the same source as our training MegaDepth. ScanNet-1500 scenes are sampled from Scannet and are completely disjoint our training on ScanNet++ while ACID's domain is entirely unseen.

Baselines. We categorize baselines into: *Pose Regression* (PR) including state-of-the-art ReLoc3r Dong et al. (2025) along with feed-forward multi-view models VGGT Wang et al. (2025a) and MapAnything Keetha et al. (2026). And *Non-Pose-Regression* (Non-PR) with Eff. LoFTR Wang et al. (2024d), RoMa Edstedt et al. (2024) as feature matchers along with DUSt3R Wang et al. (2024c), MASt3R Leroy et al. (2024), and the pose-free NoPoSplat Ye et al. (2024).

Since, MultiLoc is designed to be guided by additional views, we evaluate every method in two regimes. In the **Pairwise (Pair)** regime the model sees only the intended pair (I_q, I_t). In the **Multi-view (MV)** regime we use k=10 support views, randomly sampled per pair, together with their camera poses. The AUC is always measured on the intended pair (I_q, I_t) alone, and every MV method — pose and non-pose regression alike — is given the *same* support set. We additionally report MultiLoc and its multi-view counterparts in the Pairwise regime for completeness. Among the pose regressors, MapAnything, like MultiLoc, ingests both the support images and their poses, whereas VGGT uses the images alone. To extend the same advantage to the Non-PR methods and to verify that any gain is not merely due to the smaller viewpoint changes afforded by support views, we also supply them the support set. For the matchers (RoMa, MASt3R) we run matching on the intended pair (I_q, I_t) and on each query–support pair (I_q, I_{s_i}), select the pair with the highest inlier count, and report the performance using query and the selected view. Although all pairs are candidates, the oracle selection can underperform Pairwise because higher inlier count does not guarantee better pose accuracy, see appendix Figure 7 and Section B.2.

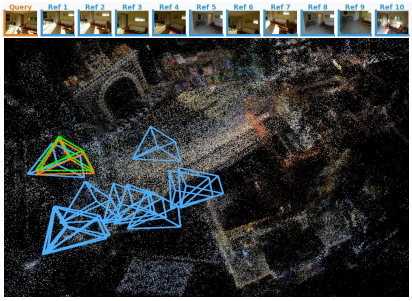
Conclusion. As shown in Table 1, MultiLoc is competitive with dedicated regressors in Pairwise mode and achieves the best MV AUC on ACID and ScanNet-1500, and at $5°/10°$ on MegaDepth-1500. It substantially outperforms MapAnything with identical posed support, while VGGT gains little from the same multi-view images. MultiLoc also surpasses the feature-matching baselines that

utilize the same additional support set. These results highlight that grounding query in a posed 3D sub-scene induces geometrically and sub-scene–consistent features, enabling better pose accuracy.

4.2 VISUAL RE-LOCALIZATION

We evaluate MultiLoc for visual re-localization against methods drawn from three families for comparative benchmarking. First is our own *relative pose regression* (RPR) family. Second is *matching-based* methods (MC) which localize through feature matching, in either a 2D–3D or a 2D–2D fashion, and, finally, *scene coordinate regression* (SCR) based methods. Alongside performance metrics, we report per-query latency on a single H100 GPU and mapping time. For SCR methods, latency is dominated by the CPU-side RANSAC/PnP solver, so it varies with scene difficulty, implementation, and hardware. We therefore, for SCR, treat absolute latency as indicative, while relative comparisons remain meaningful. Furthermore, mapping time depends on factors such as scene size. Therefore, we report bounds using all database images, without subsampling.

Baselines. Within the matching-based family we include two complementary approaches. MASt3R+Kapture Leroy et al. (2024) establishes 2D–3D correspondences between the query and a pre-built 3D map and recovers the pose with a PnP solver. E5+1 Panek et al. (2026) instead relies purely on 2D–2D correspondences between the query and a set of posed reference images, obtained with the dense matcher RoMa Edstedt et al. (2024). For scene coordinate regression we benchmark state-of-the-art ACE Brachmann et al. (2023), and its large-scale extension GLACE Wang et al. (2024a). Finally, within our own RPR family, we compare against state-of-the-art ReLoc3r Dong et al. (2025) and GeLoc3r Li et al. (2025), wherever possible since it's code is not public yet. Please note all the methods relying on image retrieval are provided with the same references. Implementation details for all baselines are mentioned in appendix.



(a)

Categ.	Method	Avg (cm/°)	Lat.
MC	MASt3R + Kapture	3.1 / **0.30**	0.85 s
	E5+1 (RoMa)	**2.3** / 0.45	7.30 s
SCR	ACE	11.0 / 1.80	∼0.10 s
	GLACE	**3.5** / **0.60**	∼0.10 s
RPR	ReLoc3r	10.7 / 0.80	0.23 s
	MultiLoc (ours)	**2.9** / **0.43**	0.15 s

(b)

Figure 3: **Visual re-localization on the Indoor-6 benchmark. (a)** Top row shows the query and retrieved reference images. Bottom row shows the Indoor6 reconstruction, where MultiLoc's predicted pose closely matches the ground truth in position and orientation. Blue frustums indicate reference poses. **(b)** Median translation / rotation error averaged over six scenes (cm / °; lower is better) and per-query latency; **bold** is best within each category. Per-scene results are in appendix.

4.2.1 INDOOR VISUAL RE-LOCALIZATION.

To benchmark cross-dataset transferability of our method, we conduct zero-shot evaluations on Indoor6 Do et al. (2022) dataset. Unlike static benchmarks, Indoor6 introduces dynamic environmental factors, including day-night transitions and illumination changes across six diverse scenes. Table 6 from appendix and Figure 3b highlights the efficacy of MultiLoc, where it achieves competitive state-of-the-art (SOTA) performance, validating its ability to handle the geometric and photometric complexities of indoor scenes without prior fine-tuning. Figure 3a demonstrates the qualitative result of visual re-localization performance of MultiLoc on Indoor6 benchmark. MultiLoc outperforms existing relative pose regressors, reducing median translation and rotation errors by 72% and 50%, respectively. Remarkably, it surpasses both of the scene coordinate regression (SCR) methods and a map-based matching approach, MASt3R+Kapture, without relying on explicit 3D reconstructions.

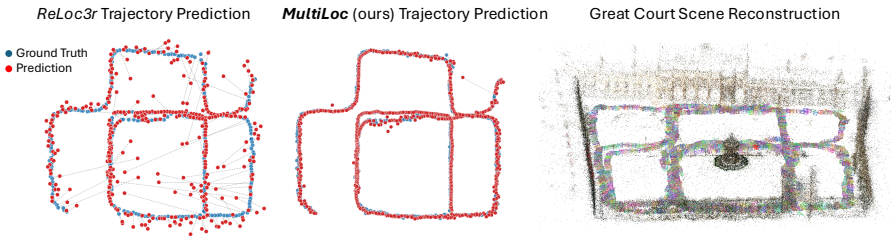


Figure 4: **MultiLoc v/s ReLoc3r trajectory error** depiction on GreatCourt sequence of Cambridge dataset. Gray dashed lines indicate the translational error between ground truth and predicted poses. MultiLoc achieves significantly lower error by leveraging 3D sub-scene context from multiple images and their pose, whereas ReLoc3r's reliance on pairwise images lead to higher drift.

4.2.2 OUTDOOR AND IN-THE-WILD VISUAL RE-LOCALIZATION.

We evaluate MultiLoc on outdoor Cambridge Landmarks Kendall et al. (2015) and in-the-wild WaySpots Brachmann et al. (2023) benchmarks. Cambridge Landmarks dataset provides diverse scenes characterized by significant environmental clutter and temporal variations in lighting and weather. On the other hand, WaySpots dataset, incorporates both small places of interests like lawn or sculptures and challenging environments like wintersign. Again, both of these datasets are entirely not seen by our model. Our method significantly enhances re-localization performance, yielding improvements of around 43% and 38% on the Cambridge , see Table 2, and 80% and 72% on WaySpots benchmark for translation and rotation, respectively in the RPR family, Table 3. The robustness of our method is also qualitatively validated in Figure 4 by comparing the trajectory prediction.

Across both benchmarks, MultiLoc is competitive with *MC* and *SCR* methods. On Cambridge, its translation error is 0.29 m, comparable to E5+1 (0.21 m) and ACE (0.26 m). Rotation errors among state-of-the-art methods differ by at most 0.21°, making fine-grained comparisons less meaningful. On WaySpots, MultiLoc achieves the lowest translation error across method families, while its rotation error is within 0.1° of E5+1. These results suggest that aggregating multi-view images and camera poses into a shared scene representation helps the model learn geometry-aware, spatially consistent features and resolve pose ambiguity.

Table 2: Visual re-localization benchmark on Cambridge Landmarks with median errors as (m / °). Green, blue, and red cells indicate the best, middle, and worst comparative results, respectively.

Category	Method	GreatCourt	Kings	OldHosp	ShopFac	StMary	Avg	Map. Time	Latency
MC	MASt3R + Kapture	**0.10 / 0.10**	**0.08 / 0.12**	**0.13 / 0.24**	**0.03 / 0.13**	**0.04 / 0.12**	**0.08 / 0.14**	$> 10min$	0.85 s
	E5+1 (RoMa)	0.37 / 0.16	0.21 / 0.35	0.29 / 0.64	0.07 / 0.27	0.11 / 0.37	0.21 / 0.36	$< 1min$	7.30 s
SCR	ACE	0.44 / 0.20	0.31 / 0.40	0.30 / 0.60	0.06 / 0.30	0.20 / 0.60	0.26 / 0.42	$> 2min$	∼0.10 s
	GLACE	**0.19 / 0.10**	**0.19 / 0.30**	**0.18 / 0.40**	**0.05 / 0.20**	**0.10 / 0.30**	**0.14 / 0.26**	$> 9min$	∼0.10 s
RPR	ReLoc3r	1.22 / 0.73	0.42 / 0.36	0.62 / 0.55	0.13 / 0.58	0.34 / 0.58	0.55 / 0.56	$< 1min$	0.23 s
	GeLoc3r	1.07 / 0.52	0.41 / 0.34	0.64 / 0.54	0.15 / 0.64	0.27 / 0.58	0.51 / 0.52	$< 1min$	–
	MultiLoc (ours)	**0.45 / 0.26**	**0.32 / 0.32**	**0.39 / 0.53**	**0.08 / 0.29**	**0.23 / 0.37**	**0.29 / 0.35**	$< 1min$	0.15 s

Table 3: Visual re-localization benchmark on the WaySpots dataset in median translation / rotation error as (cm / °). MASt3R+Kapture is omitted because it's superset, Map-Free, was used to train it.

Categ.	Method	Bears	Cubes	Insc	Lawn	Map	SqB	Statue	Tendrils	Rock	WinterSign	Avg
MC	E5+1 (RoMa)	**2.2 / 0.72**	**8.0 / 0.85**	**7.9 / 1.00**	**11.6 / 3.29**	**7.7 / 0.92**	**12.3 / 0.74**	**59.3 / 1.48**	**22.2 / 3.59**	**3.1 / 0.73**	**170.0 / 1.55**	**30.4 / 1.49**
SCR	ACE	3.6 / 1.15	**5.3 / 0.76**	**10.0** / 1.50	120.0 / 37.00	7.5 / 1.10	**5.7 / 0.70**	595.0 / 15.70	166.0 / 45.70	**2.9 / 0.74**	**423.0 / 1.19**	134.0 / 10.55
	GLACE	**3.0 / 0.90**	5.9 / 0.80	**10.0 / 1.30**	**115.0 / 26.30**	**7.2 / 1.00**	**5.7 / 0.70**	**517.0 / 10.30**	**155.0 / 32.20**	3.2 / 0.80	465.0 / 1.30	**128.7 / 7.56**
RPR	ReLoc3r	13.1 / 3.63	69.4 / 2.14	33.0 / 3.28	41.5 / 11.03	16.4 / 2.00	27.7 / 1.73	331.2 / 11.00	82.9 / 15.31	8.3 / 1.34	607.0 / 5.15	123.1 / 5.66
	MultiLoc (ours)	**4.0 / 0.77**	**8.3 / 0.93**	**9.2 / 1.09**	**16.0 / 2.84**	**6.9 / 1.12**	**9.9 / 0.75**	**89.0 / 1.46**	**34.0 / 3.55**	**4.9 / 0.72**	**65.0 / 2.61**	**24.7 / 1.58**

4.3 ABLATION STUDY

Impact of *k* for localization. In this section, we analyze the impact of parameter k—representing number of reference images—on both localization accuracy and computational latency. As illus-

Table 4: Median pose error [cm / °] for visual re-localization (all 5 Cambridge, 10 WaySpots sequences) and AUC@5/10 for camera pose estimation (ACID, ScanNet-1500). ❄: frozen (no pose intrusion); 🔥: trained MultiLoc. Red denotes best, blue second best result.

	Visual Reloc.		Camera Pose Est.	
State	Cambridge	WaySpots	ACID	ScanNet-1500
Frozen (no pose) ❄	41.0 / 0.374	66.6 / 1.80	22.55 / 40.85	38.82 / 60.31
MultiLoc (with pose) 🔥	29.5 / 0.35	24.7 / 1.39	76.56 / 87.15	62.34 / 76.32

trated in Figure 5, both translation and rotation error decreases as k increases; however, this gain comes at the cost of higher latency. We observe that $k = 10$ provides an optimal balance for MultiLoc, delivering high-precision results while maintaining efficient real-time performance. It is also worth noting that with just two or four retrieved images, our method achieves state-of-the-art RPR localization at 60 and 77 ms, respectively, making it the fastest baseline.

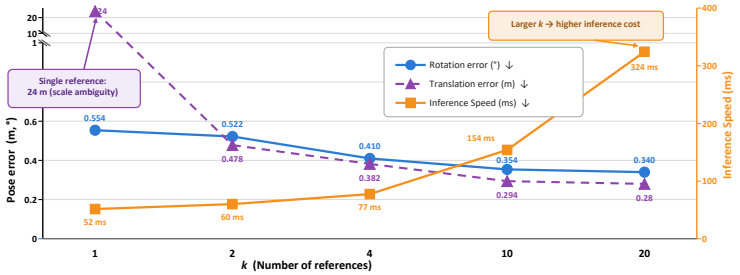


Figure 5: Visual Localization performance (average across all scenes) and inference speed comparison across varying numbers of input views *(k)* on the Cambridge Landmarks.

Effect of training. We understand the effectiveness of our task-specific training, or developing 3D sub-scene context, by comparing MultiLoc with the frozen image-only variant that is pose free as shown Table 4. Training consistently improves both re-localization benchmarks. Median translation error falls from 41.0 to 29.5 cm on Cambridge (∼28%) and from 66.6 to 24.7 cm on WaySpots (∼63%). Rotation error improves more modestly, by approximately 6% and 22%, respectively. For camera pose estimation, AUC@5/10 increases from 22.55/40.85% to 76.56/87.15% on ACID and from 38.82/60.31% to 62.34/76.32% on ScanNet-1500.

For pose-free visual re-localization, we use VGGT's aggregator and camera weights as the pose regressor, then recover scale with motion averaging, as described in the appendix. These results suggest that the frozen backbone already encodes strong rotational and geometric priors, while task-specific training substantially improves metric translation by calibrating predictions to each scene's scale and layout. MultiLoc thus learns a more accurate, scene-grounded notion of pose, which off-the-shelf geometric features leave underdetermined.

5 Conclusion and Limitations

We present MultiLoc, a large-scale, multi-view pose regressor that recasts visual localization as a 3D sub-scene representation-learning problem for pose regression. It's regressor learns to encode a set of reference views together with their camera poses into a single, pose-conditioned scene representation from which the query pose is directly decoded. Such a formulation provides effective geometric grounding, generalization to diverse environments and robust performance competitive to non-pose-regression methods. Nevertheless, performance depends on retrieval quality as with any relative pose regression work. For example, we observe references lacking co-visibility with the query contributes high pose errors, see appendix. Secondly, attending to all image tokens also incurs high latency and memory costs, which token merging or pruning could reduce. Retrieval memory and mapping costs scale with database size; as prior work shows, subsampling redundant frames can reduce them with minimal accuracy loss.

AI USE STATEMENT

In this work, we used generative AI tools for paraphrasing and polishing our writing. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for modifying figures. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We have no concerns regarding this.

REPRODUCIBILITY STATEMENT

Our work and the results reported are reproducible. The code for reproducibility will be made public once accepted along with the trained AI model's checkpoint.

REFERENCES

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*, pp. 5297–5307, 2016.

Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pp. 690–708. Springer, 2022.

Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 751–767, 2018.

Axel Barroso-Laguna, Tommaso Cavallari, Victor Prisacariu, and Eric Brachmann. A scene is worth a thousand features: Feed-forward camera localization from a collection of image features. In *International Conference on Learning Representations*, volume 2026, pp. 45134–45156, 2026.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Gabriele Berton and Carlo Masone. Megaloc: One retrieval to place them all. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2861–2867, 2025.

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.

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*, pp. 5044–5053, 2023.

Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2616–2625, 2018.

Leonard Bruns, Axel Barroso-Laguna, Tommaso Cavallari, Aron Monszpart, Sowmya Munukutla, Victor Adrian Prisacariu, and Eric Brachmann. Ace-g: Improving generalization of scene coordinate regression through query pre-training. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26751–26761. IEEE, 2025.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Robert Castle, Georg Klein, and David W Murray. Video-rate localization in multiple maps for wearable augmented reality. In *2008 12th IEEE International Symposium on Wearable Computers*, pp. 15–22. IEEE, 2008.

Kefan Chen, Noah Snavely, and Ameesh Makadia. Wide-baseline relative camera pose estimation with directional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3258–3268, 2021a.

Shuai Chen, Zirui Wang, and Victor Prisacariu. Direct-posenet: Absolute pose regression with photometric consistency. In *2021 International Conference on 3D Vision (3DV)*, pp. 1175–1185. IEEE, 2021b.

Shuai Chen, Tommaso Cavallari, Victor Adrian Prisacariu, and Eric Brachmann. Map-relative pose regression for visual re-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20665–20674, 2024.

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*, pp. 5828–5839, 2017.

Tien Do, Ondrej Miksik, Joseph DeGol, Hyun Soo Park, and Sudipta N Sinha. 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.

Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16739–16752, 2025.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19790–19800, 2024.

Martin A Fischler and Robert C Bolles. 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.

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.

Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1071–1081. IEEE, 2025.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction; map-anything. github. io. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509. IEEE, 2026.

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*, pp. 2938–2946, 2015.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Ramil Khafizov, Artem Komarichev, Ruslan Rakhimov, Peter Wonka, and Evgeny Burnaev. G-cut3r: Guided 3d reconstruction with camera and depth prior integration. *arXiv preprint arXiv:2508.11379*, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pp. 71–91. Springer, 2024.

Jingxing Li, Yongjae Lee, and Deliang Fan. Geloc3r: Enhancing relative camera pose regression with geometric consistency regularization. *arXiv preprint arXiv:2509.23038*, 2025.

Xiaotian Li, Shuzhe Wang, Yi Zhao, Jakob Verbeek, and Juho Kannala. 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.

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*, pp. 2041–2050, 2018.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14458–14467, 2021a.

Dongfang Liu, Yiming Cui, Xiaolei Guo, Wei Ding, Baijian Yang, and Yingjie Chen. Visual localization for autonomous driving: Mapping the accurate location in the city maze. In *2020 25th international conference on pattern recognition (ICPR)*, pp. 3170–3177. IEEE, 2021b.

Son Tung Nguyen, Alejandro Fontan, Michael Milford, and Tobias Fischer. Focustune: Tuning visual localization through focus-guided sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3606–3615, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Vojtech Panek, Qunjie Zhou, Yaqing Ding, Sérgio Agostinho, Zuzana Kukelova, Torsten Sattler, and Laura Leal-Taixé. A guide to structureless visual localization: V. panek et al. *International Journal of Computer Vision*, 134(6):263, 2026.

Haosong Peng, Hao Li, Yalun Dai, Yushi Lan, Yihang Luo, Tianyu Qi, Zhengshen Zhang, Yufeng Zhan, Junfei Zhang, Wenchao Xu, et al. Omnivggt: Omni-modality driven visual geometry grounded transformer. *arXiv preprint arXiv:2511.10560*, 2025.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10901–10911, 2021.

Jerome Revaud, Yohann Cabon, Romain Brégier, JongMin Lee, and Philippe Weinzaepfel. Sacreg: Scene-agnostic coordinate regression for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 688–698, 2024.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4937–4946. IEEE, 2020.

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*, pp. 3247–3257, 2021.

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.

Yoli Shavit, Ron Ferens, and Yosi Keller. Learning multi-scene absolute pose regression with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2733–2742, 2021.

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*, pp. 8922–8931, 2021.

Lauri Suomela, Jussi Kalliola, Atakan Dag, Harry Edelman, and Joni-Kristian Kämäräinen. Benchmarking visual localization for autonomous navigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2945–2955, 2023.

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*, pp. 7199–7209, 2018.

Shitao Tang, Chengzhou Tang, Rui Huang, Siyu Zhu, and Ping Tan. Learning camera localization via dense scene matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1831–1841, 2021.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 2002.

Fangjinhua Wang, Xudong Jiang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Glace: Global local accelerated coordinate encoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5819–5828. IEEE, 2024a.

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

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

Shuzhe Wang, Zakaria Laskar, Iaroslav Melekhov, Xiaotian Li, Yi Zhao, Giorgos Tolias, and Juho Kannala. Hscnet++: Hierarchical scene coordinate classification and regression for visual localization with transformer. *International Journal of Computer Vision*, 132(7):2530–2550, 2024b.

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

Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. 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, 2024d.

Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17969–17980, 2023.

Dominik Winkelbauer, Maximilian Denninger, and Rudolph Triebel. Learning to localize in new environments from synthetic training data. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5840–5846. IEEE, 2021.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1790–1799, 2020.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Enliang Zheng and Changchang Wu. Structure from motion using structure-less resection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2075–2083, 2015.

Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To learn or not to learn: Visual localization from essential matrices. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3319–3326. IEEE, 2020.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

APPENDIX

A ADDITIONAL TRAINING CONFIGURATION

A.1 ARCHITECTURE ABLATION

MultiLoc's pose regressor employs learnable camera tokens for both reference and query images. While providing learnable tokens for images with known ground truth poses may initially appear redundant, empirical evidence suggests that utilizing a collective set of learnable tokens, l, significantly enhances training stability. As illustrated in fig. 6, restricting learnable tokens to the query image alone (i.e., last learnable token) induces volatile gradient spikes, whereas our 'all-learnable' configuration facilitates smoother convergence. A similar holistic token learning design is also employed in concurrent work Peng et al. (2025), which also reports reduced error when learnable tokens and ground-truth poses are used together. Due to the high computational overhead of full-scale training and availability of limited compute, these architectural ablations were initially conducted on the ScanNet++ Yeshwanth et al. (2023) dataset as a representative for our broader multi-dataset training.

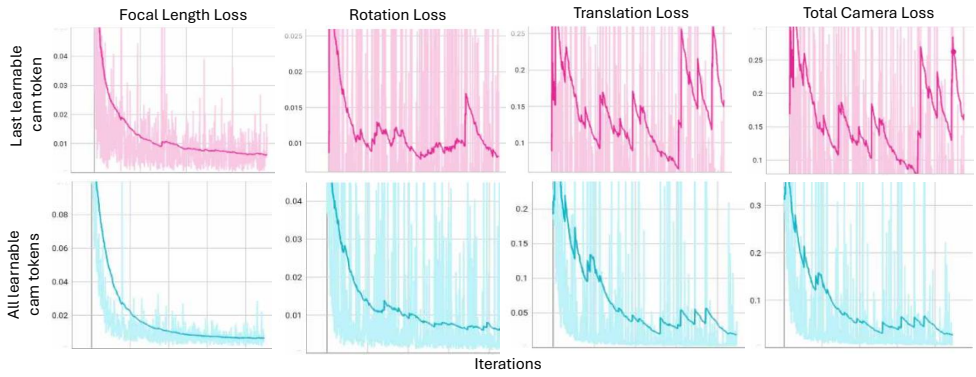


Figure 6: **Ablation study for MultiLoc's architecture**. Using learnable camera token only for the last image (top) leads to unstable training and sudden gradient jumps. Whereas, using learnable tokens for all the images in a sequence (bottom) and regressing upon them reduces this instability.

B MORE EXPERIMENTS

B.1 SCALE RECOVERY

We highlight the scale recovery of our relative camera poses using three methodologies. The first is a solution to the Procrustes Analysis; Umeyama Alignment Umeyama (2002), and second one is a motion averaging module that was suggested in Dong et al. (2025). Finally, third utilizes the scene scale s that is known from normalizing the camera pose centers of reference images as discussed in the Section 3.1. For the former, to resolve the scale and gauge ambiguity between our model's relative predictions and the global coordinate system, we perform a 7-DoF similarity alignment using the k reference frames. For third, the output pose prediction from our regressor is just multiplied with s to derive the metric pose. We utilize the Umeyama algorithm to compute a transformation $S \in \mathrm{Sim}(3)$ that maps the local relative coordinate frame to the metric world space. This transformation is then applied to the query pose to obtain its final global representation. For each scene, we report the median translation and rotation errors. As Table 5 shows, recovering the metric pose with the scale factor s—derived directly from the reference images—outperforms both Umeyama alignment and motion averaging. Although motion averaging comes closest, it relies on post-hoc geometric optimization, whereas scale recovery from s requires no such step and is effectively compute-free.

B.2 MULTI-VIEW FOR 2D-2D FEATURE MATCHING

The premise of using multiple views for correspondence based methods, as suggested in section 4.1, is that a support view with a smaller baseline or higher overlap yields more correspondences and,

Table 5: Scale Recovery benchmark of MultiLoc (ours) on Cambridge dataset for visual re-localization. Median translation error is in meters (m) and median rotation error in degrees (°).

Methods	GreatCrt	KingsClg	OldHosp	ShopFac	StMarys	Avg Rot (°↓)	Avg Trans. (m↓)
Umeyama Align.	0.44 / 0.37	0.33 / 0.41	0.39 / 0.70	0.09 / 0.33	0.23 / 0.51	0.46	0.30
Motion Avg.	0.45 / 0.25	0.32 / 0.33	0.40 / 0.52	0.08 / 0.28	0.24 / 0.36	**0.35**	0.30
Scale (s)	0.45 / 0.26	0.32 / 0.32	0.39 / 0.53	0.08 / 0.29	0.23 / 0.37	**0.35**	**0.29**

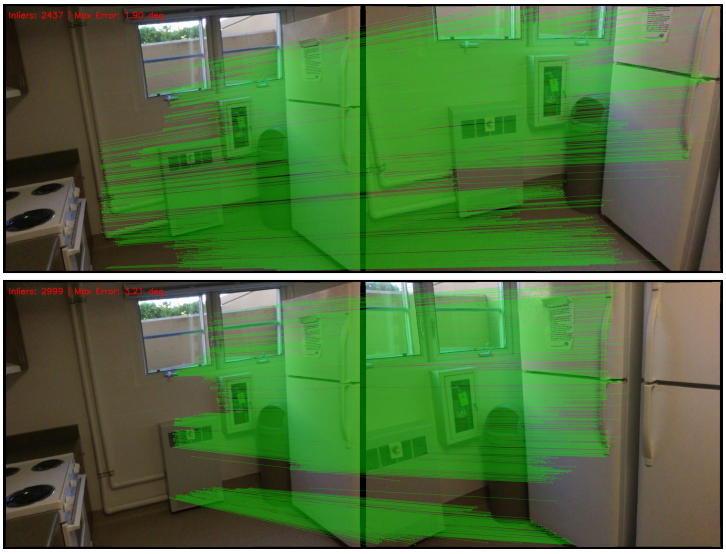


Figure 7: **RoMA comparison.** Pairwise (top) vs. best-reference (bottom) matching. While additional views increased inliers from 2437 to 2999 (top left in red), pose error, max(trans,rot), rose from 1.90 ∘ to 3.21 ∘ showing failure to exploit multi-view cues.

in turn, a better-conditioned pose. This does not hold in practice: the inlier count reflects the quantity and spatial support of correspondences, not the geometric conditioning of the resulting pose. Figure 7 shows a representative case, with error measured as $\max(e_R, e_t)$ (rotation and translation-direction error, in degrees). Matching the query directly to the intended target yields 2437 inliers and an error of $1.90°$, whereas the highest-inlier support view yields 2999 inliers yet a *larger* error of $3.21°$. The selector therefore discards the more accurate direct pair in favor of a higher-inlier but geometrically inferior one.

B.3 INDOOR6 RESULTS

In this section we provide the full table for the indoor6 benchmark as shown in Table 6. We also provide results for FastForward as discussed in Section 1.

B.4 CAMBRIDGE RESULTS

Full visual re-localization results, including FastForward and HLoc, are provided in Table 7. Note that, FastForward uses top-20 retreived images for localization on Cambridge benchmark, wheras MultiLoc utilize only 10 while achieving similar accuracy.

B.5 NON-CO-VISIBILITY IN RETRIEVAL

MultiLoc uses NetVLAD as its primary method for image retrieval and reference view selection. We also evaluate it against other retrieval methods, including the co-visibility-driven features introduced in MegaLoc Berton & Masone (2025). In this section, we present several failure cases in which the retrieved images—whether selected by MegaLoc or NetVLAD—are not co-visible with the query, as shown in Figure 8. One of the interesting scenarios from Figure 8 highlights the street viewing

Table 6: Visual re-localization benchmark on the Indoor-6 dataset. We report median translation / rotation error (cm / °); lower is better. The best result *within each category* is in **bold**. "–" denotes results not yet available.

Category	Method	Scene1	Scene2a	Scene3	Scene4a	Scene5	Scene6	Avg
MC	MASt3R + Kapture	3.4 / **0.31**	3.4 / **0.15**	2.9 / **0.28**	**2.5 / 0.39**	4.2 / **0.45**	2.3 / **0.23**	3.1 / **0.30**
	E5+1 (RoMa)	**2.2** / 0.43	**2.1** / 0.23	**1.5** / 0.37	3.0 / 0.68	**3.8** / 0.65	**1.3** / 0.31	**2.3** / 0.45
SCR	ACE	17.7 / 2.40	8.0 / 0.80	9.0 / 1.60	4.6 / 0.90	21.6 / 4.00	5.2 / 0.90	11.0 / 1.80
	GLACE	**3.8 / 0.70**	**3.9 / 0.40**	**2.9 / 0.60**	**2.3 / 0.50**	**6.1 / 0.90**	**2.1 / 0.40**	**3.5 / 0.60**
	FastForward	–	–	–	–	–	–	4.0 / **0.60**
RPR	ReLoc3r	8.0 / 0.60	11.0 / 0.78	4.0 / 0.44	9.0 / 0.87	29.0 / 1.60	3.0 / 0.52	10.7 / 0.80
	MultiLoc (ours)	**2.8 / 0.41**	**2.9 / 0.23**	**1.9 / 0.40**	**3.4 / 0.62**	**4.8 / 0.49**	**1.6 / 0.40**	**2.9 / 0.43**

Table 7: Visual re-localization benchmark on the Cambridge Landmarks dataset in median translation / rotation error as (m / °). Green, blue, and red cells indicate the best, middle, and worst comparative results, respectively.

Category	Method	GreatCourt	Kings	OldHosp	ShopFac	StMary	Avg
MC	HLoc (SP+SG)	0.16 / **0.10**	0.12 / 0.20	0.15 / 0.30	0.04 / 0.20	0.07 / 0.20	0.11 / 0.20
	MASt3R + Kapture	**0.10 / 0.10**	**0.08 / 0.12**	**0.13 / 0.24**	**0.03 / 0.13**	**0.04 / 0.12**	**0.08 / 0.14**
	E5+1 (RoMa)	0.37 / 0.16	0.21 / 0.35	0.29 / 0.64	0.07 / 0.27	0.11 / 0.37	0.21 / 0.36
SCR	ACE	0.44 / 0.20	0.31 / 0.40	0.30 / 0.60	0.06 / 0.30	0.20 / 0.60	0.26 / 0.42
	GLACE	**0.19 / 0.10**	**0.19 / 0.30**	**0.18 / 0.40**	**0.05 / 0.20**	**0.10 / 0.30**	**0.14 / 0.26**
	FastForward	0.62 / 0.40	0.24 / 0.40	0.26 / 0.50	0.08 / 0.40	0.14 / 0.50	0.27 / 0.40
RPR	Map-free (regress)	8.40 / 4.56	2.44 / 2.54	3.73 / 5.23	0.97 / 3.17	2.91 / 5.10	3.69 / 4.12
	ReLoc3r	1.22 / 0.73	0.42 / 0.36	0.62 / 0.55	0.13 / 0.58	0.34 / 0.58	0.55 / 0.56
	GeLoc3r	1.07 / 0.52	0.41 / 0.34	0.64 / 0.54	0.15 / 0.64	0.27 / 0.58	0.51 / 0.52
	MultiLoc (ours)	**0.45 / 0.26**	**0.32 / 0.32**	**0.39 / 0.53**	**0.08 / 0.29**	**0.23 / 0.37**	**0.29 / 0.35**

direction that fails to successfully retrieve reference images. Furthermore, the exact scenarios where MegaLoc performance is suboptimal is mentioned in Berton & Masone (2025). Given the rapid advancements in feed-forward 3D geometry models, optimizing reference view selection with covisibility criterion remains a critical challenge, which we designate for future works.

C Details about baselines

For every baseline we follow the authors' original public implementations and default settings, and we take care to place all methods on a comparable footing. For MASt3R Leroy et al. (2024), input images are resized so that the longer side is 512 and then center-cropped to a multiple of 16 (model's patch constraint), with the intrinsics adjusted for both operations. For RoMa Edstedt et al. (2024), different from ReLoc3r Dong et al. (2025) and NoPoSplat's Ye et al. (2024) implementation, we adopt RoMa's original MegaDepth benchmarking pipeline, which accepts an arbitrary input size, resizes to 560×560 resolution, and finally upsamples 864×864 resolution. Sparse correspondences are then mapped back to a per-dataset pixel frame in which the RANSAC threshold is interpreted: the longer side is set to 1200 px for MegaDepth-1500, the shorter side to 480 px for ScanNet-1500 following DKM/RoMa convention, and native resolution is used for ACID. We use the outdoor RoMa weights for MegaDepth-1500, ACID, Cambridge, and WaySpots, and the indoor weights for ScanNet-1500 and Indoor6 benchmarks.

For the two-view relative-pose benchmarks, both RoMa and MASt3R estimate the pose from 2D–2D correspondences via an essential matrix, so that MASt3R's metric point maps are not used and the comparison is purely a matching comparison. From RoMa's dense field we sample 5000 correspondences per pair. MASt3R contributes up to 12000 reciprocal-nearest-neighbour matches. In both

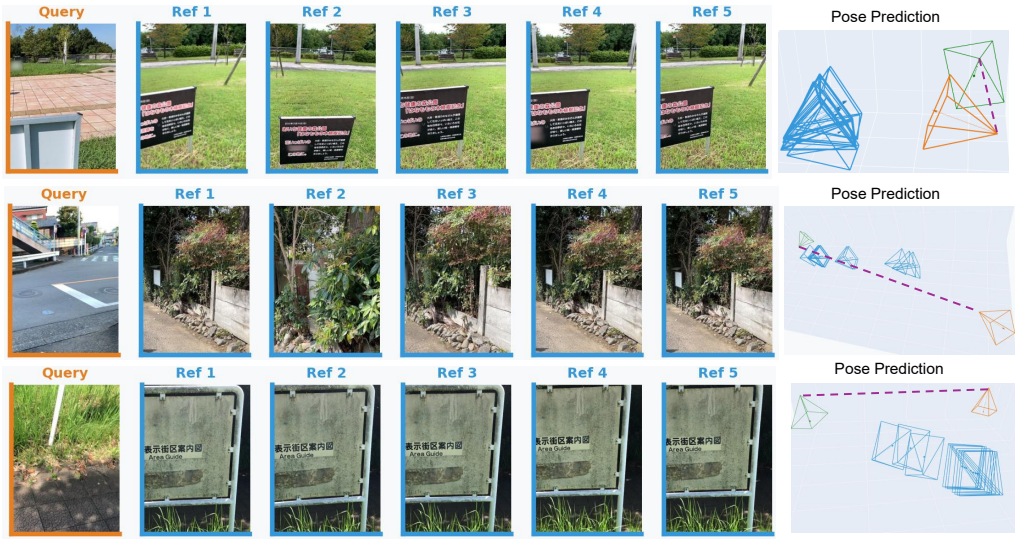


Figure 8: **Failure cases when retrieved images are not co-visible.** The following cases illustrate scenarios where co-visibility-based retrieval features fail to identify appropriate reference images. We demonstrate using top-5 of 10 such reference images per query. Such instances result in significant translation (purple dashed line) and rotation errors between the ground truth and predicted camera poses.

cases the RANSAC inlier threshold is a 0.5 px error normalized by the mean focal length. This full-image essential-matrix setting reflects each matcher's strongest two-view configuration, since the feature matcher observes the entire image rather than a cropped region.

We further evaluate MASt3R in a structure-based localization setting as denoted in Section 4 as MASt3r+Kapture. Given a 3D map of posed database images, the query is matched to the map, and its absolute pose is recovered by PnP using PoseLib with a reprojection threshold of 0.008× the image diagonal.

For ReLoc3r, we follow its resize-and-center-crop protocol where the longer side is 512 and shorter side often being 384 pixels. For VGGT Wang et al. (2025a) and MultiLoc we resize the longest dimension to be 518 pixels based on VGGT's implementation. Finally, for the scene-coordinate-regression methods ACE Brachmann et al. (2023) and GLACE Wang et al. (2024a) we use their original implementations and recommended settings.