GeLoc3r: Enhancing Relative Camera Pose Regression with Geometric Consistency Regularization

Anonymous authors
Paper under double-blind review

Abstract

ReLoc3R achieves competitive performance with fast 33ms inference and state-of-the-art regression accuracy, yet our analysis reveals subtle geometric inconsistencies in its internal representations that prevent reaching the precision ceiling of correspondence-based methods like MASt3R (which requires 294ms per pair). In this work, we present GeLoc3r, a novel approach to relative camera pose estimation that enhances pose regression methods through Geometric Consistency Regularization (GCR). GeLoc3r overcomes the speed-accuracy dilemma by training regression networks to produce geometrically consistent poses without inference-time geometric computation. During training, GeLoc3r leverages ground-truth depth to generate dense 3D-2D correspondences, weights them using a FusionTransformer that learns correspondence importance, and computes geometrically-consistent poses via weighted RANSAC. This creates a consistency loss that transfers geometric knowledge into the regression network. Unlike the FAR model, which requires both regression and geometric solving at inference, GeLoc3r only uses the enhanced regression head at test time, maintaining ReLoc3R's fast speed and approaching MASt3R's high accuracy. Against a vanilla ReLoc3R trained on the same data, GeLoc3r improves AUC@5° from 34.20% to 41.42% on CO3Dv2 (21% relative), from 63.54% to 68.66% on RealEstate10K, and from 39.86% to 50.45% on MegaDepth1500. By teaching geometric consistency during training rather than enforcing it at inference, GeLoc3r offers a new perspective on how neural networks learn camera geometry, achieving both the speed of regression and the geometric understanding of correspondence methods.

1 Introduction

Estimating relative camera poses between image pairs is fundamental in 3D vision applications such as structure-from-motion (Schonberger & Frahm, 2016), visual SLAM (Mur-Artal et al., 2015), and augmented reality. The field has traditionally faced a speed-accuracy dilemma where correspondence-based methods that explicitly match features and solve geometric constraints achieve high accuracy but require 300-400ms per pair (Leroy et al., 2025; Wang et al., 2024a), while early regression-based methods offered fast inference but with significantly lower accuracy. Recent ReLoc3R (Dong et al., 2025) achieves strong results by leveraging Vision Transformers with cross-attention mechanisms and large-scale training to deliver both fast inference (33ms) and competitive accuracy, substantially closing the historical performance gap. However, despite this impressive advance, a persistent accuracy ceiling remains—ReLoc3R and other regression methods still cannot fully match the precision of correspondence-based approaches. FAR (Rockwell et al., 2024) partially addressed this by demonstrating complementary strengths and fusing both paradigms at inference, but this requires running both methods, negating the speed advantage.

We trace this ceiling to a specific gap: ReLoc3R is trained only to minimize pose error, never to keep its predicted pose consistent with the scene's 3D structure. As a result, projecting pixels across views with its predicted pose leaves descriptors at corresponding locations poorly matched (Figure 1). Correspondence-based methods avoid this by explicitly enforcing geometric constraints, but regression methods cannot adopt them directly: such constraints require depth, available at training but not at inference.

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.

In this work, we propose GeLoc3r, which introduces Geometric Consistency Regularization (GCR)—a novel training paradigm that teaches regression networks to understand 3D geometry without requiring geometric computation at inference. By leveraging ground-truth depth available in training datasets to provide geometric supervision, our method maintains the 33ms inference speed of pure regression while approaching the accuracy of correspondence-based methods.

GeLoc3r achieves state-of-the-art regression results across seven benchmarks. Compared with a vanilla ReLoc3R trained on the same data, it improves AUC@5° from 34.20% to 41.42% on CO3Dv2 and from 63.54% to 68.66% on RealEstate10K (Zhou et al., 2018). On MegaDepth1500, it raises AUC@20° from 76.00% to 82.55% under extreme viewpoint changes. Furthermore, GeLoc3r demonstrates strong generalization on entirely unseen visual localization tasks, achieving 0.04m/1.03° on indoor 7-Scenes and 0.51m/0.52° on outdoor Cambridge Landmarks. These consistent improvements validate that incorporating geometric consistency during training enhances regression networks without compromising inference efficiency.

In summary, our main contributions are:

- **Geometric Consistency Regularization (GCR)**: A training paradigm using ground-truth depth to generate 3D-2D correspondences for RANSAC-based geometric supervision, transferring geometric knowledge to regression networks.
- **FusionTransformer-Weighted RANSAC**: An attention module that learns correspondence importance weights for weighted RANSAC sampling with prior-guided hypothesis scoring.
- **State-of-the-art Performance**: GeLoc3r achieves best regression results across seven benchmarks while maintaining 33ms inference speed.

2 Related Work

Pose Regression (PR) methods directly predict camera poses through end-to-end neural networks without explicit geometric solving.

Learning-based Pose Estimation. Recent works have explored end-to-end learning for relative pose estimation. RelPoseNet (Laskar et al., 2017) uses CNNs for direct regression, while ReLoc3R (Dong et al., 2025) leverages Vision Transformers with cross-attention for improved geometric reasoning. Chen et al. (2021) propose DirectionNet, which decomposes relative pose into 3D direction vectors on the sphere for wide-baseline scenarios. Xu et al. (2023) integrate rotation-consistent global features with rotation-invariant local features. Despite these advances, regression methods remain less accurate than structure-based approaches and can behave more like image retrieval (Sattler et al., 2019). Balancing rotation and translation terms with fundamentally different units and scales is also difficult (Kendall & Cipolla, 2017). We address these limitations from a different angle: training-time geometric consistency rather than another backbone redesign.

Dense Correspondence Methods. DUSt3R (Wang et al., 2024a) regresses dense 3D pointmaps for an image pair, and MASt3R (Leroy et al., 2025) adds a head that predicts dense local descriptors, enabling robust matching and geometric pose estimation. These methods excel when sufficient correspondences exist but require matching and solver steps at inference. Naive dense reciprocal matching requires $O(W^2H^2)$ comparisons, which MASt3R reduces to $O(kWH)$ with fast reciprocal matching (Leroy et al., 2025). Nevertheless, both remain substantially slower than direct regression. GeLoc3r seeks to transfer some of their geometric accuracy into a solver-free inference path.

Geometric Supervision in Pose Learning. Geometric constraints have been incorporated through differentiable RANSAC (Brachmann & Rother, 2022), pose-map consistency (Brahmbhatt et al., 2018), and epipolar supervision for descriptor learning (Wang et al., 2020). These works demonstrate that geometry supplies information absent from pose-only objectives. Our setting differs in using privileged depth and a non-differentiable weighted solver exclusively during training, with the resulting pose supervising the regression branch.

Hybrid Approaches. FAR (Rockwell et al., 2024) combines regressed and correspondence-solved poses through learned fusion at inference, improving robustness at the cost of running both branches.

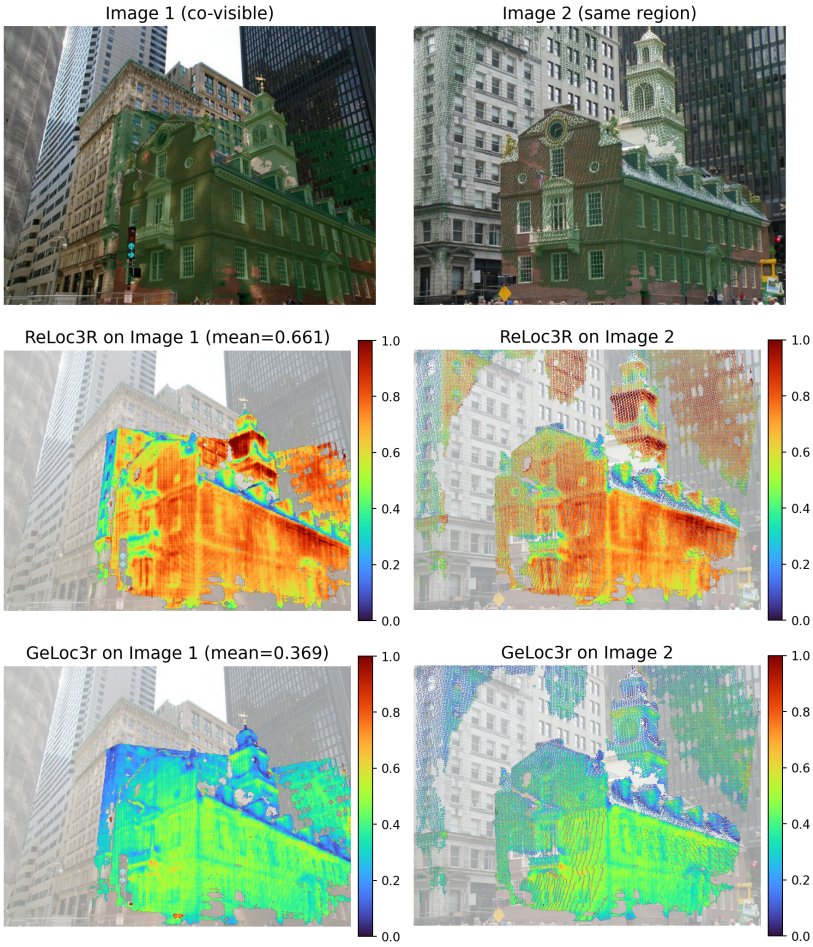

Figure 1: **Geometric consistency on a representative MegaDepth1500 pair.** The top row shows the co-visible region obtained from GT depth, relative pose, and occlusion checks. For each co-visible source pixel, we unproject its depth, transform the 3D point with the predicted pose, and compare its frozen MASt3R descriptor with the descriptor sampled at the target projection. The remaining panels show this cosine distance for each method (rows: ReLoc3R, GeLoc3r) in the source and target frames (columns; lower is better). GeLoc3r reduces the mean error from 0.661 to 0.369.

In contrast, GeLoc3r confines the FusionTransformer, depth, and RANSAC solver to training. Their geometric supervision is transferred into the regression network, preserving ReLoc3R's test-time architecture and runtime.

3 METHOD

Problem setup. Given an image pair (I_1, I_2), our goal is to predict the relative camera pose $\mathbf{P} = [\mathbf{R} \mid \mathbf{t}]$ from I_1 to I_2, where $\mathbf{R} \in SO(3)$ and $\mathbf{t}$ is a unit-norm translation direction with a separate scale. At training time we additionally have access to ground-truth depth and camera intrinsics, which we treat as privileged information used only to construct geometric supervision. At inference, the model receives only the image pair and predicts $\mathbf{P}$ directly, with no depth, intrinsics, or geometric solving.

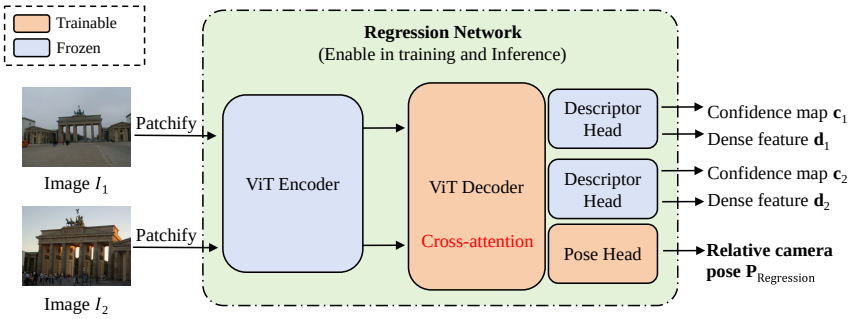

Figure 2: **GeLoc3r inference architecture.** A shared ViT encoder and bidirectional cross-attention decoder process the image pair. A trainable pose head directly predicts $\mathbf{P}_{regression}$. The frozen MASt3R descriptor heads provide dense features and confidence maps for training supervision, but their outputs are unused at test time. Inference therefore follows the ReLoc3R regression path without depth, correspondence matching, or geometric solving.

3.1 Motivation: Addressing Descriptor Inconsistency in Pose Regression

Pose-only training does not require a regressed pose to agree with the scene's 3D structure. Consequently, source points transformed by a plausible ReLoc3R pose can still land on target locations with dissimilar descriptors (Figure 1). We address this gap with a descriptor reprojection loss that requires projected source points to retain high similarity to the frozen MASt3R descriptors sampled in the target image.

This constraint complements pose-space supervision: it tests whether the predicted transformation preserves correspondence relationships in feature space rather than only minimizing a global rotation and translation error. Because descriptors are sampled densely, it supplies many local geometric signals for each training pair. The improvement in Figure 1 illustrates that GCR learns poses whose projections agree more closely with scene appearance.

We further use a FusionTransformer to prioritize reliable correspondence candidates for weighted RANSAC, suppressing ambiguous regions caused by occlusion or weak texture. Together, these signals form GCR: privileged depth supplies geometric supervision during training, while inference remains pure pose regression with no depth or solver.

3.2 Architecture Overview: A Novel Training Framework

At inference (Figure 2), GeLoc3r operates identically to ReLoc3R: the image pair passes through the shared encoder, cross-attention decoder, and pose head. During training (Figure 3), GCR additionally extracts frozen MASt3R descriptors, constructs depth-derived 3D–2D correspondences, predicts correspondence weights, and applies weighted RANSAC consistency. All geometric components are training-only, so the enhanced geometric supervision does not alter the test-time architecture or runtime.

3.3 Regression Network Backbone with Encoder-Decoder

We employ a Siamese encoder with shared weights to extract features from both images. Each image I_i is divided into 16×16 patches and projected to dimension 1024 with 2D Rotary Positional Embeddings (RoPE; Su et al. (2024)). The decoder uses bidirectional cross-attention to establish implicit correspondences:

$$\mathbf{f}_i^{enc} = \text{Encoder}(I_i + \text{RoPE}(p_i)), \quad i \in \{1, 2\} \tag{1}$$

$$\mathbf{f}_1^{dec}, \mathbf{f}_2^{dec} = \text{CrossDecoder}(\mathbf{f}_1^{enc}, \mathbf{f}_2^{enc}) \tag{2}$$

This cross-attention mechanism enables features from one image to query relevant features in the other, creating representations that capture both appearance and geometric relationships crucial for pose estimation. At inference, these features are passed directly to the pose head (Figure 2), while during training they additionally feed the GCR module.

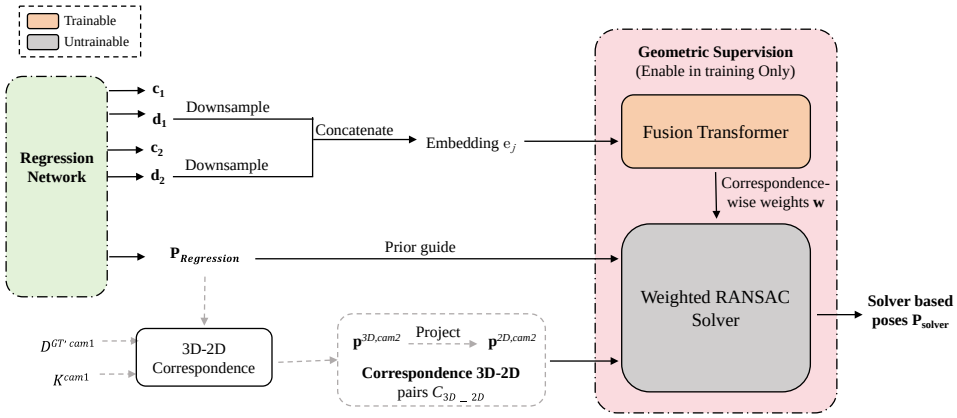

Figure 3: **GeLoc3r Training with Geometric Consistency Regularization (GCR).** During training, the model leverages ground-truth depth maps and camera intrinsics as privileged information. Dense features from frozen descriptor heads are downsampled and concatenated to form correspondence embeddings $\mathbf{e}_j$, which are processed by the FusionTransformer to produce per-correspondence weights $\mathbf{w}$. Simultaneously, GT depth is unprojected to 3D points, transformed using the predicted pose $\mathbf{P}_{regression}$, and projected to form 3D-2D correspondence pairs $\mathcal{C}_{3D-2D}$. The weighted RANSAC solver uses these correspondences with FusionTransformer weights and the regression pose as a prior to compute $\mathbf{P}_{solver}$. The consistency loss between $\mathbf{P}_{regression}$ and $\mathbf{P}_{solver}$ provides geometric supervision that teaches the regression network to produce geometrically consistent poses. The pink background highlights the training-only GCR module that is not used during inference.

3.4 Pose Head and Descriptor Head

The prediction heads operate on concatenated encoder and decoder features to leverage both local and global information.

Pose Head. A shared pose head processes decoder features to predict relative transformations. For rotation, it predicts a 9D representation (flattened 3×3 matrix) converted to SO(3) via SVD (Zhou et al., 2019). For translation, it predicts a normalized direction $\hat{\mathbf{t}}$ and scale $s \in [0, s_{max}]$:

$$\mathbf{P}_{\text{regression}} = \text{PoseHead}(\mathbf{f}^{dec}), \quad \mathbf{t} = s \cdot \hat{\mathbf{t}} \tag{3}$$

Descriptor Head. Frozen descriptor heads (pre-trained from MASt3R) extract 24-dimensional descriptors and confidence maps:

$$\mathbf{d}_i, \mathbf{c}_i = \text{DescHead}([\mathbf{f}_i^{enc}; \mathbf{f}_i^{dec}]), \quad i \in \{1, 2\} \tag{4}$$

These provide geometric anchors for training-time supervision.

3.5 Correspondence-Wise Weight

Our FusionTransformer is a specialized attention module that learns to assess correspondence quality for geometric supervision. It consists of two standard transformer blocks with hidden dimension 768 and 4 attention heads, processing concatenated descriptor pairs through self-attention layers to produce per-correspondence weights that guide RANSAC sampling. The key innovation is using attention mechanisms to model inter-correspondence relationships—correspondences are not evaluated independently but in context of all other correspondences, enabling the network to identify consistent geometric patterns and suppress outliers.

The FusionTransformer operates solely on descriptor correspondences without directly using pose predictions as input. Instead, it learns to assess correspondence quality through self-attention, exploiting the geometric principle that valid correspondences exhibit consistent transformation patterns while outliers deviate from this consensus. It first computes putative correspondences by subsampling descriptors with an appropriate stride s and concatenating them to form correspondence

embeddings:

$$\mathbf{e}_j = [\mathbf{d}_{1,j}; \mathbf{d}_{2,j}] \in \mathbb{R}^{48} \tag{5}$$

The FusionTransformer processes these through self-attention to produce normalized weights:

$$\mathbf{w} = \text{Softmax}(\text{MLP}(\text{FusionTransformer}(\mathbf{e}))) \tag{6}$$

where $\sum_i w_i = 1$. These weights guide RANSAC sampling during training to prioritize geometrically consistent correspondences. Unlike traditional matching that requires explicit feature matching, it leverages the cross-attention decoder's implicit spatial alignment: descriptors at the same spatial location in both views after subsampling are treated as potential correspondences. Specifically, for each spatial position i in the subsampled grid, it pairs $\mathbf{d}_1^{\text{sub}}[i]$ with $\mathbf{d}_2^{\text{sub}}[i]$. This approach provides dense correspondence candidates while maintaining computational tractability.

3.6 Geometric Consistency Regularization (GCR) During Training

GCR regularizes the regression network by transferring geometric supervision into the pose head—a training signal that, unlike prior DUSt3R/MASt3R-style methods, updates the regression branch rather than the descriptor representation. It operates through a chain of implications: **correct pose** $\rightarrow$ **accurate projections** $\rightarrow$ **matching descriptors**. We leverage this chain to create multiple regularizations for training dynamics.

The Consistency Chain: Our framework follows a logical progression where if $\mathbf{P}_{\text{regression}}$ is geometrically correct, it should transform 3D points accurately between views. If transformations are accurate, projected points should land at correct pixel locations. If projections are correct, descriptors at corresponding locations should match. This motivates our dual supervision strategy: Direct and Indirect Supervision.

Direct Supervision - Pose Consistency: Our framework directly verifies pose quality by comparing with a geometrically-solved pose. To generate $\mathbf{P}_{solver}$, it employs a weighted RANSAC algorithm that leverages both GT depth-derived 3D-2D correspondences and FusionTransformer weights. We extend standard RANSAC (Fischler & Bolles, 1981) with two key enhancements: learned importance sampling and prior-guided hypothesis scoring. The weighted RANSAC operates as follows: (1) samples 6-point sets with probability $p_i = w_i / \sum_j w_j$, where weights w_i are learned by the FusionTransformer from descriptor correspondences, biasing selection toward reliable matches; (2) solves the PnP problem (Lepetit et al., 2009) to obtain candidate pose $\mathbf{P}_{candidate}$; and (3) evaluates each hypothesis using a hybrid scoring function $\text{Score} = N_{inliers} + \beta \cdot \exp(-||\mathbf{P}_{candidate} - \mathbf{P}_{\text{regression}}||_F/\tau)$. This scoring combines classical geometric verification (inlier counting with a 2-pixel threshold) with a soft prior that favors poses near the regression prediction, using $\beta = 0.5$ to balance both terms and $\tau = 10$ for the exponential decay rate. After 50 iterations, the highest-scoring pose becomes $\mathbf{P}_{solver}$.

The consistency loss then measures the angular error between the regression and solver poses:

$$\mathcal{L}_{consistency} = \text{AngularError}(\mathbf{P}_{\text{regression}}, \mathbf{P}_{solver}) \tag{7}$$

Note that $\mathbf{P}_{solver}$ is detached from the computation graph: Equation (7) backpropagates only through $\mathbf{P}_{\text{regression}}$, so no gradients flow through the non-differentiable RANSAC procedure. The FusionTransformer weights w_i are instead supervised directly through $\mathcal{L}_{descriptor}$, where they appear explicitly.

Indirect Supervision - Projection Verification: The framework tests pose quality through its projections. A correct pose should project 3D points to locations with matching descriptors:

$$\mathcal{L}_{descriptor} = -\sum_i w_i \cdot \text{sim}(\mathbf{d}_i^{src}, \mathbf{d}_i^{proj}) \tag{8}$$

where $\mathbf{d}_i^{proj} = \text{Descriptor}(\Pi(\mathbf{P}_{\text{regression}} \cdot \mathbf{p}_i^{3D}))$ and w_i are the FusionTransformer weights from Equation (6), through which the FusionTransformer receives gradient and learns to focus the constraint on reliable correspondences. The benefit of this approach is **descriptor similarity serves as a verification mechanism for pose quality**. If the regression pose is wrong, it will project points to incorrect locations where descriptors don't match, creating a strong error signal. This teaches the network that correct poses must not only minimize pose error but also maintain correspondence

consistency across views. The pose loss $\mathcal{L}_{pose}$ measures the discrepancy between the predicted pose $\mathbf{P}_{\text{regression}}$ and the ground-truth pose $\mathbf{P}_{gt}$, combining a geodesic rotation error with a translation-direction error. The final training objective is

$$\begin{aligned}\mathcal{L}_{total} = \lambda_{pose}\mathcal{L}_{pose} + \lambda_{consistency}\mathcal{L}_{consistency} \\ + \lambda_{desc}\mathcal{L}_{descriptor}\end{aligned} \tag{9}$$

Note that $\mathcal{L}_{pose}$ is the standard supervision used in pose regression, while the latter two terms are our proposed GCR that exists only during training. This training paradigm transfers geometric knowledge into the network weights, enabling the model to produce geometrically consistent poses at inference without any additional computation.

4 Experiments

We conduct extensive experiments to evaluate GeLoc3r against state-of-the-art visual localization methods. Our goal is to verify that geometric consistency regularization (GCR) improves pose accuracy at inference while preserving the efficiency of regression-based approaches.

4.1 Setup

Training data. GeLoc3r is trained on 10 datasets: CO3Dv2 (Reizenstein et al., 2021), ScanNet++ (Yeshwanth et al., 2023), ARKitScenes (Dehghan et al., 2021), MegaDepth (Li & Snavely, 2018), Waymo Open (Sun et al., 2020), StaticThings3D (Schröppel et al., 2022), Habitat (Savva et al., 2019), BlendedMVS (Yao et al., 2020), RealEstate10K, and WildRGB-D (Xia et al., 2024). We sample 50,000 image pairs from each dataset per epoch, pad and resize images between $(512, 160)$ and $(512, 384)$, and apply color jitter.

Most datasets provide the depth required by GCR. RealEstate10K is the exception, so its samples use pose-regression supervision only. Sensor depth from ScanNet++ and ARKitScenes contains holes and noise, while SfM depth from MegaDepth and BlendedMVS introduces reconstruction artifacts. The consistent gains in Table 1 suggest that learned correspondence weighting and RANSAC tolerate these realistic imperfections. Depth remains privileged training information and is never required at inference.

Implementation details. GeLoc3r uses a 24-block encoder and 12-block decoder followed by a two-layer convolutional pose head. We initialize from MASt3R 512-DPT (Leroy et al., 2025). The encoder and descriptor heads are frozen; the decoder, pose head, and FusionTransformer are trainable. The FusionTransformer has two blocks, hidden dimension 768, and four attention heads. Weighted RANSAC runs 50 iterations with six-point sets and a 2-pixel reprojection threshold; consistency supervision is applied when more than 50 valid correspondences remain.

We set $\lambda_{pose} = 0.8$, $\lambda_{desc} = 0.1$, and $\lambda_{consistency} = 0.1$. Training uses eight NVIDIA L40S GPUs with batch size 10. AdamW (Loshchilov & Hutter, 2019) uses gradient clipping at 1.0 and cosine learning-rate decay from 10^{-5} to 10^{-7}. All evaluations and speed measurements conducted locally use one 24GB RTX 4090.

All quantitative results from our experiments are averaged over three independent runs.

4.2 Relative Camera Pose Estimation

We evaluate relative pose estimation on ScanNet1500 (Dai et al., 2017; Sarlin et al., 2020), RealEstate10K (Zhou et al., 2018), MegaDepth1500 (Li & Snavely, 2018), BlendedMVS (Yao et al., 2020), and CO3Dv2 (Reizenstein et al., 2021). We report AUC at angular thresholds of 5°, 10°, and 20°, where the error is the maximum of rotation and translation angular errors. Table 1 summarizes the results. GeLoc3r yields consistent improvements over the same-data vanilla ReLoc3R baseline. On CO3Dv2, GeLoc3r achieves 41.42% AUC@5° vs. 34.20% for vanilla ReLoc3R, validating the benefit of GCR. While non-PR methods (e.g., ROMA) can be more accurate on some datasets, they require substantially more computation (typically 300–2000 ms). In contrast, GeLoc3r runs in 33 ms.

Additional VGGT results on RealEstate10K and CO3Dv2 are provided in Appendix C, with differences in evaluation protocols noted.

Table 1: Relative pose AUC at $5°/10°/20°$ on five benchmarks. Bold marks the best PR result. Published baselines follow ReLoc3R (Dong et al., 2025), except VGGT (matching), whose accuracy is from Table 4 of Wang et al. (2025). The VGGT matching runtime is measured on an RTX 4090. ReLoc3R-512 is our vanilla model trained on the same data as GeLoc3r and evaluated locally. CO3Dv2 is evaluated on 93,690 pairs from held-out sequences of the official `fewview_dev` split, disjoint from all training sequences.

Methods	ScanNet1500			RealEstate10K			MegaDepth1500			BlendedMVS			CO3Dv2			Time
	AUC@5°	AUC@10°	AUC@20°	AUC@5°	AUC@10°	AUC@20°	AUC@5°	AUC@10°	AUC@20°	AUC@5°	AUC@10°	AUC@20°	AUC@5°	AUC@10°	AUC@20°	(ms)
Non-PR Methods																
VGGT (matching) (Wang et al., 2025)	33.9	55.2	73.4	-	-	-	-	-	-	-	-	-	-	-	-	115
Efficient LoFTR (Wang et al., 2024b)	19.20	37.00	53.60	-	-	-	56.4	72.2	83.5	-	-	-	-	-	-	40
ROMA (Edstedt et al., 2024)	28.90	50.40	68.30	54.60	69.80	79.70	62.6	76.7	86.3	-	-	-	-	-	-	300
DUSt3R (Wang et al., 2024a)	23.81	45.91	65.57	39.70	56.88	70.43	27.9	46.0	63.3	-	-	-	-	-	-	441
MASt3R (Leroy et al., 2025)	28.01	50.24	68.83	63.54	76.39	84.50	42.4	61.5	76.9	-	-	-	-	-	-	294
NoPoSplat (Ye et al., 2024)	31.80	53.80	71.70	69.10	80.60	87.70	-	-	-	-	-	-	-	-	-	> 2000
PR Methods																
Map-free (Regress-SN)	1.84	8.75	25.33	0.83	4.06	13.97	-	-	< 10	-	-	-	-	-	-	10
Map-free (Regress-MF)	0.50	3.48	13.15	1.61	6.74	18.38	-	-	< 10	-	-	-	-	-	-	10
ExReNet (SN)	2.30	10.71	26.13	2.17	7.94	20.43	-	-	< 10	-	-	-	-	-	-	17
ExReNet (SUNCG)	1.61	7.00	18.03	3.27	12.06	27.85	-	-	< 10	-	-	-	-	-	-	17
ReLoc3R-224 (Dong et al., 2025)	28.34	52.60	71.56	59.70	75.05	84.71	39.9	59.7	75.4	-	-	-	-	-	-	15
ReLoc3R-512	33.57	57.17	74.86	63.54	77.96	86.90	39.86	59.98	76.00	66.34	80.05	88.69	34.20	55.57	72.57	33
GeLoc3r-512 (Ours)	**34.40**	**58.12**	**75.56**	**68.66**	**81.49**	**89.18**	**50.45**	**69.29**	**82.55**	**72.17**	**83.51**	**90.67**	**41.42**	**61.66**	**76.63**	33
Delta (Δ) vs. ReLoc3R-512	**+2.47%**	**+1.66%**	**+0.94%**	**+8.06%**	**+4.53%**	**+2.62%**	**+26.57%**	**+15.52%**	**+8.62%**	**+8.79%**	**+4.32%**	**+2.23%**	**+21.11%**	**+10.96%**	**+5.59%**	**0%**

Table 2: Median localization error (m/°) on unseen 7-Scenes (Shotton et al., 2013). Bold marks the best result. Published baselines follow ReLoc3R (Dong et al., 2025); ReLoc3R-512 is our same-data vanilla baseline.

Methods	Chess	Fire	Heads	Office	Pumpkin	RedKitchen	Stairs	Avg.
EssNet (CL)	-	-	-	-	-	-	-	0.57/80.06
NC-EssNet (CL)	-	-	-	-	-	-	-	0.48/32.97
Relative PN (U)	0.31/15.05	0.40/19.00	0.24/22.15	0.38/14.14	0.44/18.24	0.41/16.51	0.35/23.55	0.36/18.38
RelocNet (SN)	0.21/10.90	0.32/11.80	0.15/13.40	0.31/10.30	0.40/10.90	0.33/10.30	0.33/11.40	0.29/11.30
ImageNet+NCM	-	-	-	-	-	-	-	0.19/4.30
Map-free (Match)	0.10/2.93	0.12/4.95	0.11/5.40	0.12/3.01	0.16/3.19	0.14/3.45	0.21/4.50	0.14/3.92
Map-free (Regress)	0.09/2.66	0.13/4.54	0.11/4.81	0.11/2.77	0.16/3.11	0.14/3.48	0.18/4.70	0.13/3.72
ExReNet (SN)	0.06/2.15	0.09/3.20	0.04/3.30	0.07/2.17	0.11/2.65	0.09/2.57	0.33/7.34	0.11/3.34
ExReNet (SUNCG)	0.05/1.63	0.07/2.54	0.03/2.71	0.06/1.75	0.07/2.04	0.07/2.10	0.19/4.87	0.08/2.52
ReLoc3R-224 (Dong et al., 2025)	**0.03**/0.99	0.04/1.13	0.02/1.23	0.05/0.88	0.07/1.14	0.05/1.23	0.12/2.25	0.05/1.26
ReLoc3R-512	**0.03**/0.90	0.03/0.94	**0.01**/0.89	**0.04/0.82**	0.06/1.11	**0.04**/1.28	0.08/1.93	**0.04**/1.12
GeLoc3r-512 (Ours)	**0.03/0.80**	**0.02/0.82**	**0.01/0.83**	**0.04/0.82**	**0.05/1.00**	**0.04/1.22**	**0.07/1.74**	**0.04/1.03**

Table 3: Median localization error (m/°) on unseen Cambridge Landmarks (Kendall et al., 2015). Bold marks the best result. Published baselines follow ReLoc3R (Dong et al., 2025); ReLoc3R-512 is our same-data vanilla baseline.

Methods	GreatCourt	KingsCollege	OldHospital	ShopFacade	StMarysChurch	Avg. (4)	Avg.
EssNet (7S)	-	-	-	-	-	10.36/85.75	-
NC-EssNet (7S)	-	-	-	-	-	7.98/24.35	-
Map-free (Match)	9.09/5.33	2.51/3.11	3.89/6.44	1.04/3.61	3.00/6.14	2.61/4.83	3.90/4.93
Map-free (Regress)	8.40/4.56	2.44/2.54	3.73/5.23	0.97/3.17	2.91/5.10	2.51/4.01	3.69/4.12
ExReNet (SN)	10.97/6.52	2.48/2.92	3.47/3.90	0.90/3.27	2.60/4.98	2.36/3.77	4.08/4.32
ExReNet (SUNCG)	9.79/4.46	2.33/2.48	3.54/3.49	0.72/2.41	2.30/3.72	2.22/3.03	3.74/3.31
ImageNet+NCM	-	-	-	-	-	0.83/1.36	-
ReLoc3R-224 (Dong et al., 2025)	1.71/0.94	0.47/0.41	0.87/0.66	0.18/**0.53**	0.41/0.73	0.48/0.58	0.73/0.65
ReLoc3R-512	1.31/0.76	0.45/0.39	0.69/0.60	0.22/0.63	0.35/0.68	0.43/0.58	0.60/0.61
GeLoc3r-512 (Ours)	**1.07/0.52**	**0.41/0.34**	**0.64/0.54**	**0.15**/0.64	**0.27/0.58**	**0.37/0.53**	**0.51/0.52**

4.3 Visual Localization

We evaluate GeLoc3r on visual localization with unseen test datasets: 7-Scenes (Shotton et al., 2013) (indoor) and Cambridge Landmarks (Kendall et al., 2015) (outdoor). Following prior work, we use NetVLAD (Arandjelovic et al., 2016) for retrieval and take the top-10 retrieved reference images for each query. We report median translation error (meters) and median rotation error (degrees) under the standard evaluation protocol. Note that both datasets are unseen during training, corresponding to the *unseen* Relative Pose Regression (RPR) setting. Complete results, including Absolute Pose Regression and seen-RPR methods, are given in Section B.

Indoor localization. Table 2 reports unseen RPR results on 7-Scenes. Early approaches (e.g., EssNet, NC-EssNet) fail with large angular errors, while recent methods substantially improve robustness. The same-data vanilla ReLoc3R obtains 0.04 m/1.12° on average, whereas GeLoc3r attains 0.04 m/1.03°, matching translation accuracy while improving rotation accuracy.

Outdoor localization. Table 3 shows unseen RPR results on Cambridge Landmarks. The same-data vanilla ReLoc3R obtains 0.60 m/0.61° on average, and GeLoc3r improves to 0.51 m/0.52°, indicating that GCR improves localization accuracy in challenging outdoor scenes.

To analyze behavior on extended trajectories, we evaluate the GreatCourt test sequence (612 poses, 38.52 m). After Sim3 alignment, GeLoc3r reduces trajectory RMSE from 10.862 m to 6.411 m (41%), and lowers the maximum and mean errors by 58% and 45%, respectively. A qualitative trajectory comparison against ReLoc3R is shown in Figure 4.

4.4 Ablation Studies

We conduct ablation studies on ScanNet to analyze the contribution of each component in our GCR framework. Table 4 reports the progressive improvements from the baseline to the full model. All models are initialized from the MASt3R 512-DPT checkpoint (Leroy et al., 2025). To keep the ablation tractable, every variant is trained and evaluated on ScanNet only, using 10,000 training pairs (one fifth of the ScanNet training data), rather than on the 10-dataset mixture used for Table 1. Absolute AUC values are therefore lower than the ScanNet1500 results in Table 1; the ablation is intended to isolate the relative contribution of each component under a controlled setting.

Table 4: Ablation on ScanNet. Each GCR component improves relative pose estimation accuracy. All models start from MASt3R 512-DPT initialization (Leroy et al., 2025) and are trained on 10,000 ScanNet pairs only, so absolute values are not comparable to Table 1.

Method	AUC@5°	AUC@10°	AUC@20°
Baseline (pose regression only)	19.4	44.9	66.6
+ Descriptor loss	20.3	46.3	67.5
+ Weighted RANSAC consistency (Full GeLoc3r)	**22.4**	**48.3**	**68.8**

The results show consistent gains from each component of Geometric Consistency Regularization (GCR). Starting from the MASt3R-initialized baseline (19.4% AUC@5°), adding the descriptor loss improves performance to 20.3% AUC@5°, and incorporating weighted RANSAC pose consistency further improves to 22.4% AUC@5°. Overall, the full model achieves a +3.0 point gain in AUC@5° over the baseline, supporting our hypothesis that training-time geometric supervision improves regression-based pose estimation.

5 Conclusion

We introduce GeLoc3r, a framework for relative camera pose estimation that addresses limitations of regression-based methods via Geometric Consistency Regularization (GCR). By leveraging geometric supervision during training (when depth is available), GeLoc3r improves pose accuracy without adding inference-time geometric computation. The FusionTransformer learns correspondence weights that enable effective RANSAC-based consistency regularization, transferring geometric priors into the regression model. Across benchmarks, GeLoc3r achieves state-of-the-art performance among regression methods while maintaining fast inference. Future work includes extending GCR to multi-view settings and exploring alternative forms of geometric supervision for visual localization.

Reproducibility Statement

Section 3 describes the architecture and training objective; Section A.1 gives the weighted RANSAC procedure (Algorithm 1). Section 4.1 lists the training datasets, sampling, resolutions, optimizer, schedule, loss weights and hardware, and Table 1 together with Section B describes the evaluation benchmarks and protocols, including the held-out CO3Dv2 split. We will release our code, the processed data and data splits, and the complete evaluation protocol upon acceptance.

AI USE STATEMENT

We used generative AI tools to polish the writing of this paper, including grammar, wording and readability. All AI-assisted edits were reviewed by the authors, and we take full responsibility for the content of this work.

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 (CVPR)*, June 2016. 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, 2022. doi: 10.1109/TPAMI.2021.3070754. 2

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 (CVPR)*, June 2018. 2

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 (CVPR)*, pp. 3258–3268, June 2021. 2

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443, 2017. doi: 10.1109/CVPR.2017.261. 7

Afshin Dehghan, Gilad Baruch, Zhuoyuan Chen, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, and Elad Shulman. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-D data. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/66f041e16a60928b05a7e228a89c3799-Abstract-round1.html. 7

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16739–16752, June 2025. 1, 2, 8, 14

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 (CVPR)*, pp. 19790–19800, June 2024. 8

Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. ISSN 0001-0782. doi: 10.1145/358669.358692. URL https://doi.org/10.1145/358669.358692. 6

Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6555–6564. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.694. URL https://doi.org/10.1109/CVPR.2017.694. 2

Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2938–2946, 2015. doi: 10.1109/ICCV.2015.336. 8, 14

Zakaria Laskar, Iaroslav Melekhov, Surya Kalia, and Juho Kannala. Camera relocalization by computing pairwise relative poses using convolutional neural network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017. 2

Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. *Int. J. Comput. Vision*, 81(2):155–166, February 2009. ISSN 0920-5691. doi: 10.1007/s11263-008-0152-6. URL https://doi.org/10.1007/s11263-008-0152-6. 6

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 71–91, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73220-1. 1, 2, 7, 8, 9

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 2041–2050. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00218. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Li_MegaDepth_Learning_Single-View_CVPR_2018_paper.html. 7

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7. 7

Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robotics*, 31(5):1147–1163, 2015. doi: 10.1109/TRO.2015.2463671. URL https://doi.org/10.1109/TRO.2015.2463671. 1

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotný. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10881–10891, 2021. URL https://api.semanticscholar.org/CorpusID:237371959. 7

Chris Rockwell, Nilesh Kulkarni, Linyi Jin, Jeong Joon Park, Justin Johnson, and David F. Fouhey. Far: Flexible accurate and robust 6dof relative camera pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19854–19864, June 2024. 1, 2

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 (CVPR)*, June 2020. 7

Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixé. Understanding the limitations of cnn-based absolute camera pose regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3302–3312. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00342. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Sattler_Understanding_the_Limitations_of_CNN-Based_Absolute_Camera_Pose_Regression_CVPR_2019_paper.html. 2

Manolis Savva, Jitendra Malik, Devi Parikh, Dhruv Batra, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, and Vladlen Koltun. Habitat: A platform for embodied AI research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 9338–9346. IEEE, 2019. doi: 10.1109/ICCV.2019.00943. URL https://doi.org/10.1109/ICCV.2019.00943. 7

Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

Philipp Schröppel, Jan Bechtold, Artemij Amiranashvili, and Thomas Brox. A benchmark and a baseline for robust multi-view depth estimation. In *2022 International Conference on 3D Vision (3DV)*, pp. 637–645, 2022. doi: 10.1109/3DV57658.2022.00074. 7

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 *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2930–2937, 2013. doi: 10.1109/CVPR.2013.377. 8, 14

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C), February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063. 4

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurélien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for

autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451, 2020. doi: 10.1109/CVPR42600.2020.00252. 7

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025. doi: 10.1109/CVPR52734.2025.00499. 8, 15

Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 757–774, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58452-8. 2

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 (CVPR)*, pp. 20697–20709, June 2024a. 1, 2, 8

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 (CVPR)*, pp. 21666–21675, June 2024b. 8

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22378–22389, 2024. doi: 10.1109/CVPR52733.2024.02112. 7

Meng Xu, Zhihuang Zhang, Yuanhao Gong, and Stefan Poslad. Regression-based camera pose estimation through multi-level local features and global features. *Sensors*, 23(8), 2023. ISSN 1424-8220. doi: 10.3390/s23084063. URL https://www.mdpi.com/1424-8220/23/8/4063. 2

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 1787–1796. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00186. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yao_BlendedMVS_A_Large-Scale_Dataset_for_Generalized_Multi-View_Stereo_Networks_CVPR_2020_paper.html. 7

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, 2024. URL https://arxiv.org/abs/2410.24207. 8

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Niesner, and Angela Dai. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22, Los Alamitos, CA, USA, October 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.00008. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00008. 7

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images, 2018. URL https://arxiv.org/abs/1805.09817. 2, 7

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5

A Implementation Details

A.1 Weighted RANSAC Algorithm - Full Details

The weighted RANSAC algorithm used to generate $\mathbf{P}_{solver}$ (described in Section 3.6) is formalized in Algorithm 1.

Algorithm 1 Weighted RANSAC with Prior Guidance

Require: 3D-2D correspondences $\mathcal{C}_{3D-2D}$, weights $\mathbf{w}$, prior $\mathbf{P}_{\text{regression}}$
Ensure: Refined pose $\mathbf{P}_{solver}$
1: **for** $iter = 1$ to 50 **do**
2: Sample 6 correspondences using probability $p_i = w_i / \sum_j w_j$
3: Solve PnP to get $\mathbf{P}_{candidate}$
4: Compute inliers with reprojection error < 2 pixels
5: Score = $N_{inliers} + \beta \cdot \exp(-||\mathbf{P}_{candidate} - \mathbf{P}_{\text{regression}}||_F / \tau)$
6: Update best if score improves
7: **end for**
8: **return** Best scoring pose

The algorithm enhances standard RANSAC through: (1) importance-weighted sampling based on FusionTransformer weights, and (2) prior-guided scoring that balances geometric evidence with the regression pose prior. The values $\beta = 0.5$ and $\tau = 10$ (Section 3.6) were determined through ablation studies.

B Complete Visual Localization Results

This section provides complete evaluation results, including Absolute Pose Regression (APR) and Relative Pose Regression (RPR) methods under both seen and unseen settings. Table 5 and Table 6 report the full per-scene results on 7-Scenes and Cambridge Landmarks, expanding the averaged numbers given in the main paper. For each benchmark, methods are grouped into APR, RPR (Seen) trained on the target dataset, and RPR (Unseen) evaluated without dataset-specific training. GeLoc3r is evaluated strictly under the RPR (Unseen) setting and attains the best average accuracy among unseen RPR methods on both benchmarks.

B.1 Qualitative Trajectory Comparison

Figure 4 shows the qualitative trajectory comparison referenced in Section 4.3. We visualize the estimated camera positions on the GreatCourt sequence of Cambridge Landmarks, the largest scene in the benchmark, comprising 612 query poses over a 38.52 m trajectory. Each predicted position is drawn as a point and connected to its ground-truth position by a line whose length is proportional to the localization error. Compared with ReLoc3R (Fig. 4a), GeLoc3r (Fig. 4b) exhibits markedly fewer large-error outliers, especially at the trajectory corners where wide-baseline relative poses are hardest to estimate. After Sim(3) alignment, GeLoc3r reduces the trajectory RMSE from 10.862 m to 6.411 m (a 41% reduction), confirming that the gains of GCR yield spatially coherent trajectories rather than isolated per-pair improvements.

C VGGT Results under Different Evaluation Protocols

Table 7 places our results alongside the published VGGT results on two shared datasets. VGGT uses ten random input views and reports AUC@30°, whereas our evaluation uses image pairs and reports AUC@5°/10°/20°. The sampling protocols and training data also differ: VGGT does not train on RealEstate10K, while GeLoc3r includes its training split. These results provide context rather than a controlled comparison; their numerical differences do not establish relative accuracy or improvement.

Table 5: Complete visual localization results on 7-Scenes dataset Shotton et al. (2013). We report median pose errors in meters and degrees. Methods are categorized into APR, RPR (Seen) trained on 7-Scenes, and RPR (Unseen) evaluated without dataset-specific training. Published baseline results are from Dong et al. (2025); ReLoc3R-512 is our vanilla model trained on the same data as GeLoc3r, as in Table 2.

Methods	Chess	Fire	Heads	Office	Pumpkin	RedKitchen	Stairs	Average
APR								
LENS	0.03/1.30	0.10/3.70	0.07/5.80	0.07/1.90	0.08/2.20	0.09/2.20	0.14/3.60	0.08/3.00
PMNet	0.03/1.26	0.04/1.76	**0.02**/1.68	0.06/1.69	0.07/1.96	0.08/2.23	0.11/2.97	0.06/1.93
DFNet+NeFeS	**0.02/0.57**	**0.02/0.74**	**0.02/1.28**	**0.02/0.56**	**0.02/0.55**	**0.02/0.57**	**0.05/1.28**	**0.02/0.79**
Marepo	**0.02**/1.24	**0.02**/1.39	**0.02**/2.03	0.03/1.26	0.04/1.48	0.04/1.71	0.06/1.67	0.03/1.54
RPR (Seen)								
EssNet (7S)	-	-	-	-	-	-	-	0.22/8.03
Relative PN (7S)	0.13/6.46	0.26/12.72	0.14/12.34	0.21/7.35	0.24/6.35	0.24/8.03	0.27/11.82	0.21/9.30
NC-EssNet (7S)	-	-	-	-	-	-	-	0.21/7.50
RelocNet (7S)	0.12/4.14	0.26/10.4	0.14/10.5	0.18/5.32	0.26/4.17	0.23/5.08	0.28/7.53	0.21/6.73
Relpose-GNN	0.08/**2.70**	0.21/7.50	0.13/8.70	0.15/4.10	0.15/3.50	0.19/3.70	0.22/6.50	0.16/5.20
AnchorNet	0.06/3.89	0.15/10.3	0.08/10.9	0.09/5.15	0.10/2.97	0.08/4.68	0.10/9.26	0.09/6.74
CamNet	**0.04**/1.73	**0.03/1.74**	**0.05/1.98**	**0.04/1.62**	**0.04/1.64**	**0.04/1.63**	**0.04/1.51**	**0.04/1.69**
RPR (Unseen)								
EssNet (CL)	-	-	-	-	-	-	-	0.57/80.06
NC-EssNet (CL)	-	-	-	-	-	-	-	0.48/32.97
Relative PN (U)	0.31/15.05	0.40/19.00	0.24/22.15	0.38/14.14	0.44/18.24	0.41/16.51	0.35/23.55	0.36/18.38
RelocNet (SN)	0.21/10.9	0.32/11.8	0.15/13.4	0.31/10.3	0.40/10.9	0.33/10.3	0.33/11.4	0.29/11.3
ImageNet+NCM	-	-	-	-	-	-	-	0.19/4.30
Map-free (Match)	0.10/2.93	0.12/4.95	0.11/5.40	0.12/3.01	0.16/3.19	0.14/3.45	0.21/4.50	0.14/3.92
Map-free (Regress)	0.09/2.66	0.13/4.54	0.11/4.81	0.11/2.77	0.16/3.11	0.14/3.48	0.18/4.70	0.13/3.72
ExReNet (SN)	0.06/2.15	0.09/3.20	0.04/3.30	0.07/2.17	0.11/2.65	0.09/2.57	0.33/7.34	0.11/3.34
ExReNet (SUNCG)	0.05/1.63	0.07/2.54	0.03/2.71	0.06/1.75	0.07/2.04	0.07/2.10	0.19/4.87	0.08/2.52
ReLoc3R-224	**0.03**/0.99	0.04/1.13	0.02/1.23	0.05/0.88	0.07/1.14	0.05/1.23	0.12/2.25	0.05/1.26
ReLoc3R-512	**0.03**/0.90	0.03/0.94	**0.01**/0.89	**0.04/0.82**	0.06/1.11	**0.04**/1.28	0.08/1.93	**0.04**/1.12
GeLoc3r-512 (Ours)	**0.03/0.80**	**0.02/0.82**	**0.01/0.83**	**0.04/0.82**	**0.05/1.00**	**0.04/1.22**	**0.07/1.74**	**0.04/1.03**

Table 6: Complete visual localization results on Cambridge Landmarks Kendall et al. (2015). We report median pose errors in meters and degrees. Methods are categorized into APR, RPR (Seen) trained on Cambridge, and RPR (Unseen) evaluated without dataset-specific training. Published baseline results are from Dong et al. (2025); ReLoc3R-512 is our vanilla model trained on the same data as GeLoc3r, as in Table 3.

Methods	GreatCourt	KingsCollege	OldHospital	ShopFacade	StMarysChurch	Average (4)	Average
APR							
LENS	-	0.33/**0.50**	**0.44**/0.90	0.27/1.60	0.53/1.60	0.39/1.20	-
PMNet	-	**0.31**/0.55	**0.44/0.79**	0.17/0.86	**0.31/0.96**	**0.31**/0.79	-
DFNet+NeFeS	-	0.37/0.54	0.52/0.88	**0.15/0.53**	0.37/1.14	0.35/**0.77**	-
RPR (Seen)							
EssNet (CL)	-	-	-	-	-	1.08/3.41	-
Relpose-GNN	3.20/2.20	**0.48**/1.00	**1.14/2.50**	**0.48**/2.50	1.52/3.20	0.91/2.30	1.37/2.30
NC-EssNet (CL)	-	-	-	-	-	0.85/2.82	-
AnchorNet	-	0.57/**0.88**	1.21/2.55	0.52/**2.27**	**1.04/2.69**	**0.84/2.10**	-
RPR (Unseen)							
EssNet (7S)	-	-	-	-	-	10.36/85.75	-
NC-EssNet (7S)	-	-	-	-	-	7.98/24.35	-
Map-free (Match)	9.09/5.33	2.51/3.11	3.89/6.44	1.04/3.61	3.00/6.14	2.61/4.83	3.90/4.93
Map-free (Regress)	8.40/4.56	2.44/2.54	3.73/5.23	0.97/3.17	2.91/5.10	2.51/4.01	3.69/4.12
ExReNet (SN)	10.97/6.52	2.48/2.92	3.47/3.90	0.90/3.27	2.60/4.98	2.36/3.77	4.08/4.32
ExReNet (SUNCG)	9.79/4.46	2.33/2.48	3.54/3.49	0.72/2.41	2.30/3.72	2.22/3.03	3.74/3.31
ImageNet+NCM	-	-	-	-	-	0.83/1.36	-
ReLoc3R-224	1.71/0.94	0.47/0.41	0.87/0.66	0.18/**0.53**	0.41/0.73	0.48/0.58	0.73/0.65
ReLoc3R-512	1.31/0.76	0.45/0.39	0.69/0.60	0.22/0.63	0.35/0.68	0.43/0.58	0.60/0.61
GeLoc3r-512 (Ours)	**1.07/0.52**	**0.41/0.34**	**0.64/0.54**	**0.15**/0.64	**0.27/0.58**	**0.37/0.53**	**0.51/0.52**



(a) ReLoc3R trajectory visualization (b) GeLoc3r trajectory visualization

Figure 4: **Trajectory comparison on Cambridge Landmarks GreatCourt.** We visualize localization on 612 test poses spanning a 38.52 m trajectory. Points denote predicted camera positions and lines connect predictions to ground truth, with length proportional to error. GeLoc3r exhibits fewer large-error outliers than ReLoc3R, reducing RMSE from 10.862 m to 6.411 m after Sim3 alignment.

Table 7: Results under different protocols, shown for reference without cross-protocol ranking. VGGT values are from Table 1 of Wang et al. (2025); GeLoc3r values are from Table 1. AUC values are percentages. A dash indicates an unreported metric. BA denotes bundle adjustment.

Dataset	Method	Input views	AUC@5°	AUC@10°	AUC@20°	AUC@30°
RealEstate10K	GeLoc3r	2	68.66	81.49	89.18	–
	VGGT (feed-forward)	10	–	–	–	85.3
	VGGT + BA	10	–	–	–	93.5
CO3Dv2	GeLoc3r	2	41.42	61.66	76.63	–
	VGGT (feed-forward)	10	–	–	–	88.2
	VGGT + BA	10	–	–	–	91.8