GSLoc: Efficient Camera Pose Refinement via 3D Gaussian Splatting

Changkun Liu[1*] Shuai Chen[2†] Yash Bhalgat[2‡] Siyan Hu[1] Zirui Wang[2†]
Ming Cheng[3] Victor Adrian Prisacariu[2†] Tristan Braud[1]

[1]HKUST [2]University of Oxford [3]Dartmouth College

Abstract

We leverage 3D Gaussian Splatting (3DGS) as a scene representation and propose a novel test-time camera pose refinement framework, GSLoc. This framework enhances the localization accuracy of state-of-the-art absolute pose regression and scene coordinate regression methods. The 3DGS model renders high-quality synthetic images and depth maps to facilitate the establishment of 2D-3D correspondences. GSLoc obviates the need for training feature extractors or descriptors by operating directly on RGB images, utilizing the 3D vision foundation model, MASt3R, for precise 2D matching. To improve the robustness of our model in challenging outdoor environments, we incorporate an exposure-adaptive module within the 3DGS framework. Consequently, GSLoc enables efficient pose refinement given a single RGB query and a coarse initial pose estimation. Our proposed approach surpasses leading NeRF-based optimization methods in both accuracy and runtime across indoor and outdoor visual localization benchmarks, achieving state-of-the-art accuracy on two indoor datasets. The project page is available at https://gsloc.active.vision.

1. Introduction

Camera localization, the task of determining the 6-DoF pose of a camera within a given environment, is crucial for numerous applications including robotics, autonomous vehicles, augmented reality (AR), and virtual reality (VR). Current methods for camera pose estimation primarily fall into the categories of structure-based approaches and absolute pose regression (APR) techniques. Classic structure-based pipelines [13, 25, 31, 35–37, 44] rely on 2D-3D correspondences between a point cloud and the reference image. Another class of structure-based methods - Scene Coordinate Regression (SCR) [3, 4, 6, 51] - uses neural networks for direct regression of 2D-3D correspondences. These 2D-3D correspondences are fed into Perspective-n-Point (PnP) [15] for pose estimation. APR methods [8, 22, 40, 49] employ neural networks to directly infer camera poses from query images. While APR approaches offer fast inference times, they often struggle with accuracy and generalization [38]. SCR methods generally achieve higher accuracy but at the cost of increased computational complexity.

Given the above limitations, there has been a growing interest in pose refinement methods to enhance accuracy of the *initial* pose estimates of an underlying pose-estimation method. Recent approaches have leveraged Neural Radiance Fields (NeRF) for this purpose. For instance, LENS [29] uses NeRF as a novel view synthesizer to augment APR training sets, while NeFeS [10] proposes a test-time refinement pipeline. However, these methods offer limited improvements in accuracy and suffer from slow convergence due to the computational demands of NeRF rendering and the requirement for backpropagation through the pose estimation model. Furthermore, a recent NeRF-based localization method - CrossFire [30] - establishes explicit 2D-3D matches using features rendered from NeRF. However, it requires training a customized scene model together with the matching model, and exhibits a lower accuracy compared to structure-based methods.

To address these challenges of slow convergence, limited accuracy, and the requirement of customized feature matching training, we propose a novel test-time pose refinement framework, which we call GSLoc. GSLoc utilizes 3D Gaussian Splatting (3DGS) [23] as the scene representation and leverages the high-quality and fast rendering capabilities of 3DGS to generate synthetic images and depth maps, facilitating efficient establishment of 2D-3D correspondences with the query image and the *initial* pose estimate from the underlying pose estimator (e.g., APR, SCR). We incorporate an exposure-adaptive module into the 3DGS model to improve its robustness to the domain shift between the query image and the rendered images. Secondly, our method operates directly on RGB images, utilizing the

*This research was conducted during Changkun Liu's visit at Active Vision Lab, University of Oxford.

†Active Vision Lab, University of Oxford

‡Visual Geometry Group, University of Oxford

3D vision foundation model MASt3R [24] for precise 2D matching, eliminating the need for training scene-specific feature extractors or descriptors [10, 30]. This significantly accelerates our method compared to iterative NeRF-based refinement methods [10], and makes our framework easier to deploy than CrossFire [30] and it's variants [17, 27, 54].

Lastly, we conduct comprehensive quantitative evaluations and ablation studies on the 7Scenes [18, 41], 12Scenes [48], and Cambridge Landmarks [22] benchmarks. GSLoc significantly enhances the pose estimation accuracy of both APR and SCR methods across these benchmarks, achieving state-of-the-art accuracy on the two indoor datasets. Unlike previous NeRF-based methods [10], which fail to improve SCR methods, such as ACE [6], our method offers substantial improvements and outperforms other leading NeRF-based methods [17, 27, 30, 54].

2. Related Work

Pose Estimation without 3D Representation. A straightforward approach for a coarse pose estimation is using image retrieval [1, 16, 19] to average poses from top-retrieved images, but this lacks precision. Absolute Pose Regression (APR) methods [8, 9, 11, 20–22, 40, 49] directly regress a pose from a query image using trained models, bypassing 3D representations and geometric relationships. Despite being fast, APR methods suffer in accuracy and generalization [26, 38] compared to geometry-based techniques. LENS [29] enhances APR by augmenting views with NeRF, but matching the accuracy of 3D structure-based methods remains challenging. To improve APR methods' accuracy, we used 3DGS as a 3D representation and utilized its geometry information to optimize the initial prediction.

Structure-based Pose Estimation. Classical 3D structure-based methods, like the hierarchical localization pipeline (HLoc) [13, 25, 31, 35–37, 44], predict camera poses using a point cloud and a database of reference images, requiring descriptor storage and 2D-3D correspondence through image retrieval. In contrast, Scene Coordinate Regression (SCR) methods [3, 4, 6, 51] directly regress 2D-3D correspondences using neural networks and apply PnP [15] and RANSAC [14] for pose estimation. Our GSLoc eliminates the need for reference images and descriptor databases by using a 3DGS model for scene representation, further optimizing SCR outputs like ACE [6].

NeRF-based Pose Estimation. NeRF-based pose estimation methods [10, 53] rely on iterative rendering and pose updates, leading to slow convergence and limited accuracy. While NeFeS [10] improves APR pose estimation, it faces difficulties in enhancing SCR results and and suffers from long refinement runtime. HR-APR [26] speeds up optimization by 30%, but the average runtime of each query still takes several seconds on a high-performance GPU. Other NeRF-based methods like FQN [17], CrossFire [30], NeRFLoc [27], and NeRFMatch [54] improve positioning by establishing 2D-3D matches but require specialized feature extractors and suffer from slow rendering and quality issues.

3DGS-based Pose Estimation. With the novel view synthesis (NVS) field transitioning from NeRF to 3DGS, iComMa [43], like iNeRF [53], uses an inefficient iterative refinement process for camera pose estimation by inverting 3DGS. In contrast, 6DGS [2] achieves a one-shot estimate by projecting rays from an ellipsoid surface, avoiding iteration. While both methods use 3DGS for visual localization, neither has been tested on large benchmarks [22, 48] or compared with mainstream methods like SCR and APR. We propose an approach using 3DGS for 2D-3D correspondences, similar to CrossFire [30], but without requiring a trained feature extractor or feature matchers. Our method generates high-quality synthetic images and employs direct 2D-2D matching, making it faster and easier to deploy than previous NeRF-based methods such as NeFeS, CrossFire, and other variants [17, 26, 27, 54].

3. Proposed Method

Our proposed method, GSLoc, is a test-time camera pose refinement framework. We assume availability of a pretrained pose estimator and a 3DGS model of the scene. For a query image, we first obtain an initial estimated pose from the pose estimator. Our goal is to output a refined pose.

Specifically, given a query image $I_q \in \mathbb{R}^{H\times W\times 3}$ with camera intrinsics $K \in \mathbb{R}^{3\times 3}$, a pose estimator $\mathcal{F}$ (typically an APR or SCR model) predicts an *initial* 6-DOF pose $\hat{p} = [\hat{\mathbf{t}}|\hat{\mathbf{R}}]$, where $\hat{\mathbf{t}} \in \mathbb{R}^3$ and $\hat{\mathbf{R}} \in \mathbb{R}^{3\times 3}$ represent the estimated translation and rotation respectively. Subsequently, for the viewpoint $\hat{p}$, a pretrained 3DGS model $\mathcal{H}$ renders an image $\hat{I}_r \in \mathbb{R}^{H\times W\times 3}$ and a depth map $\hat{I}_d \in \mathbb{R}^{H\times W\times 1}$. We use an exposure-adaptive affine color transformation (ACT) module $\mathcal{E}$ during this rendering process to enhance the robustness of our model to challenging outdoor environments (see Section 3.1). A matcher $\mathcal{M}$ then establishes dense 2D-2D correspondences between I_q and $\hat{I}_r$. Then we can establish the 2D-3D matches based on $\hat{I}_q$ and $\hat{I}_d$ (see Section 3.2). Finally, we obtain the refined pose $\hat{p}'$ from these 2D-3D matches (see Section 3.2). An overview of our framework is depicted in Figure 1. We also explore a faster pose refinement framework without 2D-3D matches depicted in Figure 2 (see Section 3.3).

3.1. 3DGS Test-time Exposure Adaptation

Existing literature [23, 28] shows that 3DGS achieves high-quality novel view renderings but assumes training and testing without significant photometric distortions. In visual relocalization, mapping and query sequences often differ in lighting due to varying times, weather, and exposure. This

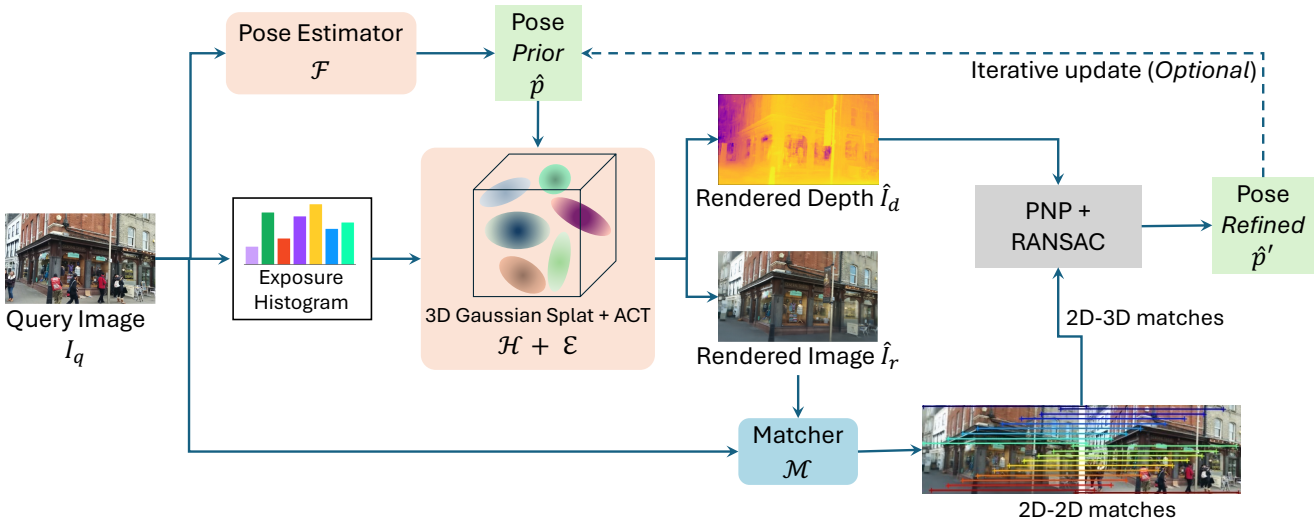


Figure 1. Overview of GSLoc. We assume the availability of a pre-trained pose estimator $\mathcal{F}$ and a pre-trained 3DGS model $\mathcal{H}$ of the scene. For a query image I_q, we first obtain an initial estimated pose $\hat{p}$ from the pose estimator $\mathcal{F}$. Our goal is to output a refined pose $\hat{p}'$. For less precise initial poses $\hat{p}$, we can repeat the same operation with the optimized pose $\hat{p}'$ as pose prior again.

creates a significant appearance gap between 3DGS renderings and query images, negatively impacting 2D-2D matching performance.

To address this issue, we apply an exposure-adaptive affine color transformation module $\mathcal{E}$ [9, 10] to 3DGS, allowing the 3DGS to adaptively render appearances during testing and accurately reflecting the exposure of I_q. Specifically, we use a 4-layer MLP that takes the luminance histogram of the query image as input and produces a 3x3 matrix $\mathbf{Q}$ along with a 3-dimensional bias vector $\mathbf{b}$. These outputs are then directly applied to the rendered pixels of the 3DGS as shown in Equation 1, ensuring a closer match to the exposure of the query image.

$$\hat{\mathbf{C}}(\mathbf{r}) = \mathbf{Q}\hat{\mathbf{C}}_{\text{rend}}(\mathbf{r}) + \mathbf{b} \tag{1}$$

where $\hat{\mathbf{C}}(\mathbf{r})$ is the final per-pixel color and $\hat{\mathbf{C}}_{\text{rend}}(\mathbf{r})$ is the rendered per-pixel color obtained from the 3DGS model $\mathcal{H}$.

3.2. Pose Refinement with 2D-3D Correspondences

GSLoc estimates the camera pose by establishing 2D-3D correspondences between the query image I_q and the scene representation. This process involves the following steps:

2D-2D Matching. First, an image $\hat{I}_r$ is rendered from the initial estimated viewpoint $\hat{p}$. A Matcher $\mathcal{M}$ is then used to establish 2D-2D pixel correspondences $C_{q,r}$ between the query image I_q and the rendered image $\hat{I}_r$. In our implementation, the matcher $\mathcal{M}$ is a recently released 3D vision foundation model, MASt3R [24]. Given that MASt3R is trained on both synthetic and real data, it demonstrates strong robustness for 2D-2D matching across images pair with the sim-to-real domain gap.

3D Coordinate Map Generation. Simultaneously, we use our trained 3DGS model $\mathcal{H}$ to render a depth map $\hat{I}_d$ from the viewpoint $\hat{p}$. We modify the rasterization engine of 3DGS to render the depth map as follows:

$$\hat{I}_d = \sum_{i \in N} d_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{2}$$

where d_i is the z-depth of each Gaussian in the viewspace and α_i is the learned opacity multiplied by the projected 2D covariance of the i^{th} Gaussian. In our framework, ground truth depth maps are not required for supervision during the training of the 3DGS model $\mathcal{H}$. Using the rendered depth map $\hat{I}_d$, camera intrinsics K, and pose $\hat{p}$, we obtain the 3D coordinate map $X_r^d \in \mathbb{R}^{H \times W \times 3}$ for the rendered image $\hat{I}_r$.

Establishing 2D-3D Correspondences. By combining the 2D-2D correspondences $C_{q,r}$ with the 3D coordinate map X_r^d, we establish 2D-3D correspondences between I_q and the scene. For each matched pixel in I_q, we obtain its corresponding 3D coordinate from X_r^d.

Pose Refinement. Finally, we obtain the refined pose $\hat{p}'$ by feeding these 2D-3D correspondences into a PnP [15] solver with RANSAC [14] loop. This process does not require backpropagation through the pose estimator $\mathcal{F}$ or the 3DGS model $\mathcal{H}$, ensuring efficient computation and enabling its usage with any black-box pose estimator model.

The use of 2D-3D correspondences, coupled with PnP + RANSAC, provides a robust pose estimation that is more reliable than methods relying solely on 2D-2D matching. Furthermore, our method eliminates the requirement of training specialized feature descriptors that previous approaches [9, 10, 30] rely on for robustness.

Optional Iterative Refinement. For less precise initial poses $\hat{p}$, we can optionally repeat the above process with the optimized pose $\hat{p}'$ as the initial pose, further improving the accuracy of the estimation.

3.3. Faster Alternative with Relative Post Estimation

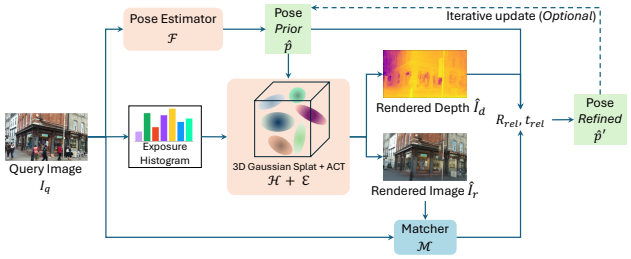


Figure 2. Overview of GSLoc$_{\text{rel}}$. Different from GSLoc in Figure 1, we use $\hat{I}_d$ to get the scale s of $\mathbf{t}_{\text{rel}}$. Then we calculate the refined pose $\hat{p}'$ based on $\mathbf{R}_{\text{rel}}$ and $s\mathbf{t}_{\text{rel}}$ without PnP + RANSAC.

While GSLoc provides high accuracy through 2D-3D correspondences, we also explore an alternative approach that prioritizes computational efficiency. This variant, which we call GSLoc$_{\text{rel}}$, utilizes MASt3R's point map registration capabilities to estimate relative pose. Figure 2 shows an overview of the GSLoc$_{\text{rel}}$ approach.

Specifically, MASt3R generates point maps $\mathbf{P}_q$ and $\mathbf{P}_r$ for both the query image I_q and the rendered image $\hat{I}_r$ and predicts the relative rotation $\mathbf{R}_{\text{rel}}$ and translation $\mathbf{t}_{\text{rel}}$ between the two images. However, this relative pose predicted by MASt3R needs to be aligned to the scene's scale. We recover the scale s by aligning the pointmap $\mathbf{P}_r$ with the depth map $\hat{I}_d$ rendered from the 3DGS model $\mathcal{H}$. The final refined pose $\hat{p}'$ is computed as:

$$\hat{p}' = [\hat{\mathbf{R}}'|\hat{\mathbf{t}}'] = [\mathbf{R}_{\text{rel}}\hat{\mathbf{R}}|\hat{\mathbf{t}} + s\mathbf{R}_{\text{rel}}\mathbf{t}_{\text{rel}}] \tag{3}$$

where $\hat{\mathbf{R}}$, $\hat{\mathbf{t}}$ are the initial rotation and translation estimates. As shown in Tables 5 to 7, GSLoc$_{\text{rel}}$ offers a trade-off between speed and accuracy, making it suitable for scenarios where faster processing is desired, particularly when refining initial estimates from APR methods, such as DFNet [9].

4. Experiments

4.1. Evaluation Setup

Datasets. We evaluate the performance of GSLoc across three widely-used public visual localization datasets. The 7Scenes dataset [18, 41] comprises seven indoor scenes with volumes ranging from 1m^3 to 18m^3. The 12Scenes dataset [48] features 12 larger indoor scenes, with volumes spanning from 14m^3 to 79m^3. The Cambridge Landmarks dataset [22] represents a large-scale outdoor scenario, characterized by challenges such as moving objects and varying lighting conditions between query and training images.

Evaluation Metrics. We report two types of metrics to compare the performance of different methods. The first metric is the median translation and rotation error. The second metric is the recall rate, which measures the percentage of test images localized within a cm and b°.

Baselines. In our experiment, to demonstrate the improvement capabilities of our framework, we use the initial estimates of APR and SCR methods as our baseline. The term *APR/SCR + GSLoc* denotes the one-shot refinement, while *APR/SCR + GSLoc$_n$* indicates the refinement with n steps using the method shown in Figure 1. Similar naming convention applies to *APR/SCR + GSLoc$_{rel}$* and *APR/SCR + GSLoc$_{rel_n}$*. We employ our method on top of the prevailing APR methods, DFNet [9] and Marepo [11], as well as a well-known SCR method, ACE [6], as the pose estimator $\mathcal{F}$. We follow the default settings of these pose estimators to obtain the initial pose prior for each query image[1].

Implementation Details. GT Poses: For both the 7Scenes and 12Scenes datasets, we adopt the SfM ground truth (GT) provided by [5]. As demonstrated in NeFeS [10], SfM GT can render superior geometric details compared to dSLAM GT for the 7Scenes dataset. Gaussian Splatting: For the training of the 3DGS model of each scene, we utilize the sparse point cloud of training frames generated by COLMAP [39] as the initial input. We select Scaffold-GS [28] as our 3DGS representation, incorporating modifications detailed in Sections 3.1 and 3.2 to adapt exposure and enable depth rendering. Scaffold-GS reduces redundant Gaussians while delivering high-quality rendering compared to the vanilla 3DGS [23]. For the exposure-adaptive ACT module, we follow the default setting in [10], computing the query image's histogram in the YUV color space and binning the luminance channel into 10 bins. In addition, we apply temporal object filtering to filter out moving objects in the dynamic scene using an off-the-shelf method [12], leading to better accurate scene reconstruction quality and pixel-matching performance. Training Details: We employ the official pre-trained MASt3R [24] model without fine-tuning for 2D-2D matching and resize all images to 512 pixels on their largest dimension. The modified Scaffold-GS model is trained for each scene for 30,000 iterations on an NVIDIA A6000 GPU. We implement our framework with PyTorch [32]. Additional details can be found in the supplementary materials.

[1]Note that the original paper of Marepo reports results on 7Scenes using dSLAM GT; we retrained the ACE head of Marepo using SfM GT.

Table 1. Comparisons on 7Scenes dataset. The median translation and rotation errors (cm/°) of different methods. The best results are in bold (lower is better). NRP denotes neural render pose estimation. **GSLoc $_2$ (ours)** indicates the refinement of the pose over two iterations. **GSLoc** refers to the refinement of the pose over a single iteration.

	Methods	Chess	Fire	Heads	Office	Pumpkin	Redkitchen	Stairs	Avg. ↓ [cm/°]
APR	PoseNet [22]	10/4.02	27/10.0	18/13.0	17/5.97	19/4.67	22/5.91	35/10.5	21/7.74
	MS-Transformer [40]	11/6.38	23/11.5	13/13.0	18/8.14	17/8.42	16/8.92	29/10.3	18/9.51
	DFNet [9]	3/1.12	6/2.30	4/2.29	6/1.54	7/1.92	7/1.74	12/2.63	6/1.93
	Marepo [11]	1.9/0.83	2.3/0.92	2.1/1.24	2.9/0.93	2.5/0.88	2.9/0.98	5.9/1.48	2.9/1.04
SCR	DSAC* [3]	**0.5**/0.17	**0.8**/0.28	**0.5**/0.34	1.2/0.34	1.2/0.28	**0.7**/0.21	2.7/0.78	1.1/0.34
	ACE [6]	**0.5**/0.18	**0.8**/0.33	**0.5**/0.33	**1**/0.29	1/0.22	0.8/0.2	2.9/0.81	1.1/0.34
	GLACE [51]	0.6/0.18	0.9/0.34	0.6/0.34	1.1/0.29	**0.9**/0.23	0.8/0.20	3.2/0.93	1.2/0.36
NRP	FQN-MN [17]	4.1/1.31	10.5/2.97	9.2/2.45	3.6/2.36	4.6/1.76	16.1/4.42	139.5/34.67	28/7.3
	CrossFire [30]	1/0.4	5/1.9	3/2.3	5/1.6	3/0.8	2/0.8	12/1.9	4.4/1.38
	DFNet + NeFeS$_{50}$ [10]	2/0.57	2/0.74	2/1.28	2/0.56	2/0.55	2/0.57	5/1.28	2.4/0.79
	HR-APR [26]	2/0.55	2/0.75	2/1.45	2/0.64	2/0.62	2/0.67	5/1.30	2.4/0.85
	NeRFMatch [54]	0.9/0.3	1.3/0.4	1.6/1.0	3.3/0.7	3.2/0.6	1.3/0.3	7.2/1.3	2.7/0.7
	DFNet + **GSLoc $_2$ (ours)**	1.3/0.35	2/0.71	1.1/0.71	2.2/0.5	2.5/0.62	2.2/0.47	3.6/1.05	2.1/0.63
	Marepo + **GSLoc (ours)**	1.3/0.4	1.6/0.5	1.4/0.68	2.2/0.5	2/0.47	2.2/0.48	3.8/0.86	2.1/0.56
	ACE + **GSLoc (ours)**	**0.5/0.15**	0.9/**0.27**	**0.5/0.28**	**1/0.25**	1.1/**0.21**	0.8/**0.17**	**2.3/0.58**	**1.0/0.27**

Table 2. We report the average percentage (%) of frames below a 5cm, 5° pose error across 7Scenes. The best results are in bold (higher is better). IR denotes image retrieval, and NRP denotes-neural render pose estimation. **GSLoc $_2$ (ours)** indicates the refinement of the pose over two iterations. **GSLoc** refers to the refinement of the pose over a single iteration.

	Methods	Avg. ↑ [5cm, 5°]
APR	DFNet [9]	43.1
	Marepo [11]	84.0
IR+SfM points	HLoc [35, 36]	95.7
	DVLAD+R2D2 [34, 45]	95.7
SCR	DSAC* [3]	97.8
	ACE [6]	97.1
	GLACE [51]	95.6
NRP	DFNet + NeFeS$_{50}$ [10]	78.3
	HR-APR [26]	76.4
	NeRFMatch [54]	75.4
	NeRFLoc [27]	89.5
	DFNet + **GSLoc $_2$ (ours)**	82.9
	Marepo + **GSLoc (ours)**	91.1
	ACE + **GSLoc (ours)**	**98.9**

4.2. Localization Accuracy

We conduct quantitative experiments on three datasets to evaluate the improved localization accuracy of our framework compared to the APR and SCR methods.

7Scenes Dataset. Using the 7Scenes dataset, we evaluate the performance of DFNet, Marepo and ACE with GSLoc. Table 1 demonstrates that GSLoc significantly reduces pose estimation errors for Marepo and ACE with one-show refinement. Since DFNet's estimations are less accurate than Marepo and ACE, we experimentally find 2 rounds refinement can get a more accurate result than one-shot refinement, our DFNet + GSLoc $_2$ outperforms DFNet + NeFeS$_{50}$, which required 50 rounds of optimization. Table 2 shows that GSLoc significantly improves the proportion of query images below 5cm and 5° pose error. It is worth noting that ACE + GSLoc outperforms HLoc, indicating that 3DGS has the potential to replace traditional point cloud based visual localization pipelines. Figure 3 (a) shows that after refinement using our GSLoc, the rendered image of the estimated pose better matches the real image.

12Scenes Dataset. Using the 12Scenes dataset, we conduct the quatitative evaluation using Marepo and ACE with GSLoc. The former works [6, 51] report the percentage of frames below a 5cm, 5° pose error. Since SCR methods have already achieved good results with this metric, in this paper we use a more stringent standard (2cm, 2°). Table 3 shows that GSLoc significantly improves the percentage of query images below 2cm, 2° pose error for Marepo and ACE. Even though the initial pose prior provided by ACE is accurate in most scenes, our approach can still improve accuracy in challenging scenes, such as office1/lounge, office2/5a and office2/5b. Figure 3 (b) shows that after refinement using our GSLoc, the rendered image with our pose estimation aligns better with the real image.

Cambridge Landmarks Dataset. We conduct a quantitative evaluation by deploying DFNet and ACE with GSLoc on the Cambridge Landmarks Dataset. Marepo is not included in this comparison due to the absence of an official

Table 3. We report the average percentage (%) of frames below a 2cm, 2° pose error across 12Scenes. The best results are in bold (higher is better).

Scene	apt1		apt2				office1				office2		Avg. ↑
Methods	kitchen	living	bed	kitchen	living	luke	gates362	gates381	lounge	manolis	5a	5b	[2cm, 2°]
Marepo	52.9	63.1	50.4	47.4	57.1	41.2	61.7	39.8	57.8	55.1	50.1	28.6	50.4
DSAC*	99.7	98.2	**100**	**99.6**	95.3	95.5	99.7	95	89.9	91.3	94	**95.1**	96.7
ACE	99.4	99.8	97.1	**99.6**	99.4	98.4	99.5	96.7	93.3	96.7	93.8	93.1	97.2
GLACE	99.4	99.6	**100**	**99.6**	**100**	96.2	99.7	96.9	97.6	**97.6**	91.1	92.6	97.5
Marepo + **GSLoc (ours)**	71.7	86.6	66.4	76.1	77.7	51.4	76.9	59.7	70.3	74.7	76.1	55.8	70.3
ACE + **GSLoc (ours)**	**100**	**100**	98	99.1	99.4	**98.9**	**100**	**98.5**	**99.4**	**97.6**	**95.8**	94.8	**98.5**

model for this dataset. Table 4 demonstrates that GSLoc significantly reduces pose estimation errors for both DFNet and ACE. Specifically, the accuracy of DFNet + GSLoc with one-shot optimization significantly surpasses that of CrossFire and DFNet + NeFeS with 30 and even 50 steps of optimization (see Table 4). This result fully demonstrates the efficiency of our GSLoc. On the Kings College scene, DFNet + GSLoc outperforms ACE after our refinement. ACE provides accurate initial estimation, ACE + GSLoc consistently improves ACE accuracy across all four scenes, as illustrated in Figure 3 (c). Refining the pose using our method results in a rendered image that aligns more accurately with the ground truth image.

Table 4. Comparisons on Cambridge Landmarks dataset. We report the median translation and rotation errors (cm/°) of different methods. NRP denotes neural render pose estimation. Best results are in bold (lower is better) among the NRP-based approaches.

	Methods	Kings	Hospital	Shop	Church	Avg. ↓ [cm/°]
APR	PoseNet	93/2.73	224/7.88	147/6.62	237/5.94	175/5.79
	MS-Transformer	85/1.45	175/2.43	88/3.20	166/4.12	129/2.80
	LENS [29]	33/0.5	44/0.9	27/1.6	53/1.6	39/1.15
	DFNet	73/2.37	200/2.98	67/2.21	137/4.02	119/2.90
SCR	ACE	29/0.38	31/0.61	5/0.3	19/0.6	21/0.47
	GLACE	19/0.32	18/0.42	5/0.22	9/0.3	13/0.32
NRP	FQN-MN [17]	28/0.4	54/0.8	13/0.6	58/2	38/1
	CrossFire	47/0.7	43/0.7	20/1.2	39/1.4	37/1
	DFNet + NeFeS$_{30}$[1]	37/0.64	98/1.61	17/0.60	42/1.38	49/1.06
	DFNet + NeFeS$_{50}$	37/0.54	52/0.88	15/0.53	37/1.14	35/0.77
	HR-APR	36/0.58	53/0.89	13/0.51	38/1.16	35/0.78
	DFNet + **GSLoc (ours)**	26/0.34	48/0.72	10/0.36	27/0.62	28/0.51
	ACE + **GSLoc (ours)**	**25/0.29**	**26/0.38**	**5/0.23**	**13/0.41**	**17/0.33**

[1] Results of DFNet + NeFeS$_{30}$ taken from [26].

GSLoc vs. GSLoc$_{rel}$. We compare GSLoc, a pose refinement framework that use 2D-3D correspondence, with GSLoc$_{rel}$, a faster pose refinement framework that use relative pose from MASt3R. Both frameworks are evaluated on 7Scenes and Cambridge Landmarks datasets using DFNet and ACE predictions. Tables 5 and 6 show that GSLoc$_{rel}$ achieves notable accuracy improvement with DFNet on both indoor and outdoor datasets, though it is less effective than GSLoc. However, GSLoc$_{rel}$ is significantly faster than GSLoc and other NeRF-based methods, as discussed in Section 4.3. While GSLoc$_{rel}$ improves coarse pose estimates from APR methods like DFNet, it struggles with accurate pose estimates from SCR methods. For ACE, GSLoc$_{rel}$ results in performance degradation because our pose refinement relies on the relative pose estimator MASt3R, which struggles to provide more accurate relative pose estimates when the ACE-predicted pose is already sufficiently close to the GT pose. Higher median rotation and translation errors in Tables 5 and 6 compared to GSLoc indicate that scale recovery is not the only challenge for GSLoc$_{rel}$, as rotation is scale-independent.

Table 5. Results of GSLoc$_{rel}$ in Figure 2 and GSLoc in Figure 1 on Cambridge Landmarks dataset. We report the median translation and rotation errors (cm/°).

Methods	Kings	Hospital	Shop	Church	Avg. ↓
DFNet	73/2.37	200/2.98	67/2.21	137/4.02	119/2.9
DFNet + GSLoc$_{rel}$	54/0.54	95/0.73	26/0.43	43/**0.59**	55/0.57
DFNet + GSLoc	**28/0.34**	**48/0.72**	**10/0.36**	**26**/0.66	**28/0.53**
ACE	29/0.38	31/0.61	**5**/0.3	19/0.6	21/0.47
ACE + GSLoc$_{rel}$	56/0.56	70/0.60	24/0.46	38/0.55	47/0.54
ACE + GSLoc	**26/0.34**	**26/0.38**	**5/0.23**	**13/0.41**	**18/0.36**

Table 6. Results of GSLoc$_{rel}$ in Figure 2 and GSLoc in Figure 1 on 7Scenes dataset. We report the average accuracy (%), the percentage of frames meeting a 5cm, 5° pose error threshold, and the median translation and rotation errors (cm/°).

Methods	Avg Acc ↑ [5cm, 5°]	Avg Err ↓ [cm/°]
DFNet [9]	43.1	6/1.93
DFNet + **GSLoc$_{rel}$ (ours)**	80.5	2.7/**0.38**
DFNet + **GSLoc $_2$ (ours)**	**82.9**	**2.1**/0.63
ACE [6]	97.1	1.1/0.34
ACE + **GSLoc$_{rel}$ (ours)**	79.9	2.8/0.43
ACE + **GSLoc (ours)**	**98.9**	**1/0.27**

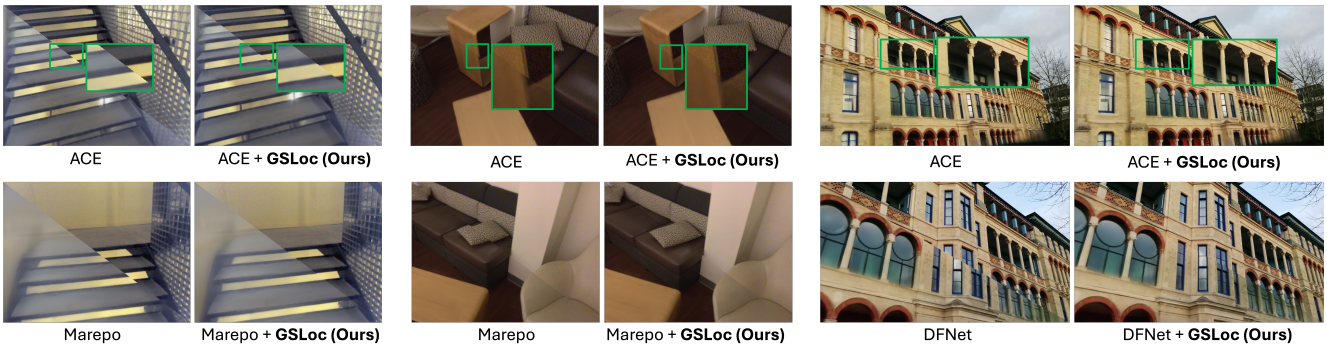


Figure 3. Our GSLoc enhances pose predictions for Marepo, DFNet, and ACE. The examples presented are from three benchmarks. Each subfigure is divided by a diagonal line, with the **bottom left** part rendered using the estimated pose and the **top right** part displaying the ground truth image. For the initial estimates provided by Marepo and DFNet, we can clearly see the inaccuracies. For ACE estimates, patches highlighting visual differences are emphasized with green insets for enhanced visibility.

4.3. Runtime Analysis

We evaluate the processing time of the proposed framework using an NVIDIA GeForce GTX 4090 GPU. On average, 3DGS rendering takes 3.7 ms on 7Scenes dataset and 12 ms on Cambridge Landmarks dataset (due to higher scene complexity and image resolution). MASt3R relative pose estimation takes 71 ms. MASt3R 2D-2D matching takes additional 42 ms, and PnP+RANSAC takes another 52 ms. As a result, our $\text{GSLoc}_{\text{rel}}$ only adds 71 ms to the inference time of the pose estimator $\mathcal{F}$ and our GSLoc adds less than 180 ms overhead. All time measurements are averaged over 1,000 runs. We compare the runtime and accuracy with other methods in Table 7. Although $\text{GSLoc}_{\text{rel}}$ is less accurate than GSLoc, it is more efficient. $\text{GSLoc}_{\text{rel}}$ provides a feasible solution to APR pose refinement when time budget is important.

Table 7. Runtime Analysis (test on Cambridge Landmarks). Numbers in bold represent the best performance (lower is better).

Methods	Avg. ↓ [cm/°]	Avg. ↓ time (s)
CrossFire	37/1.0	0.3
DFNet + NeFeS_{50}	35/0.8	≈ 10
HR-APR	35/0.8	≈ 8.5
DFNet + **$\text{GSLoc}_{\text{rel}}$ (ours)**	55/0.6	**0.07**
DFNet +**GSLoc (ours)**	**28/0.5**	0.2

4.4. Ablation study

In this section, we first demonstrate the rationale behind selecting MASt3R as the matcher $\mathcal{M}$ in GSLoc. Subsequently, we show that ACT effectively reduces the domain gap between the query image and the rendered image, thereby enhancing the refinement accuracy.

Different Matchers. We compare three matching methods: LoFTR [42], DUSt3R [52], and MASt3R [24] – within GSLoc on the 7Scenes dataset. For DUSt3R and MASt3R, we resize all images to 512 pixels on their largest dimension. For LoFTR, we use the pre-trained model for indoor scenes and maintain the frames in the 7Scenes dataset at 640×480. As shown in Table 8, Marepo + GSLoc and ACE + GSLoc using MASt3R as $\mathcal{M}$ achieve the highest improvement. Conversely, Marepo/ACE + GSLoc using DUSt3R does not yield any improvement, and Marepo/ACE + GSLoc using LoFTR shows lower improvement compared to MASt3R. These results validate our design choice of using MASt3R as the 2D-2D matcher.

Table 8. Results of different matchers (LoFTR, DUSt3R, and MASt3R) on the 7Scenes dataset. GSLoc $^{\text{L}}$ denotes using LoFTR as the matcher $\mathcal{M}$, GSLoc $^{\text{D}}$ denotes using DUSt3R as $\mathcal{M}$, and GSLoc $^{\text{M}}$ denotes using MASt3R as $\mathcal{M}$. The table presents median translation and rotation errors (cm/°) of the different methods.

Methods	Avg. ↓ [cm/°]
Marepo	2.9/1.04
Marepo + GSLoc $^{\text{L}}$	2.4/0.66
Marepo + GSLoc $^{\text{D}}$	2.7/0.81
Marepo + GSLoc $^{\text{M}}$	**2.1/0.56**
ACE	1.1/0.34
ACE + GSLoc $^{\text{L}}$	1.1/0.31
ACE + GSLoc $^{\text{D}}$	1.7/0.49
ACE + GSLoc $^{\text{M}}$	**1.0/0.27**

Affine Color Transformation. To enhance the robustness of the 3DGS model in image rendering and to reduce the domain gap between the rendered image and the query image, we incorporated an ACT module into the Scaffold-GS model, as described in Section 3.1. Figure 4 illustrates

the improvement in image rendering quality with the ACT module applied. The performance enhancement on GSLoc from ACT module is demonstrated in Table 9. On Cambridge Landmarks dataset, employing the ACT module in DFNet + GSLoc setup reduces average median translation and rotation error by 17.6% and by 13.6%, respectively.

(a) Ground Truth (b) Scaffold-GS/PSNR:16.5 dB (c) Ours/PSNR:18.4 dB

Figure 4. Benefit of the ACT module. A regular 3DGS model tends to render images based on the lighting conditions and the appearance of its training frames, as demonstrated by the synthetic view of Scaffold-GS in (b). However, in challenging visual localization datasets, such as ShopFacade in the Cambridge Landmarks, some query frames may have different exposures compared to the training frames. (c) Our proposed Scaffold-GS + ACT can adaptively adjust the exposure based on the query's histogram.

Table 9. Ablation study for ACT module on Cambridge Landmarks dataset. We report the median translation and rotation errors (cm/°).

Methods	Kings	Hospital	Shop	Church	Avg. ↓
DFNet + GSLoc (w/o. ACT)	34/0.46	55/0.84	12/**0.34**	34/0.72	34/0.59
DFNet + GSLoc (w. ACT)	**26/0.34**	**48/0.72**	**10**/0.36	**27/0.62**	**28/0.51**

4.5. Discussion

In this section, we provides additional insights and discussion of our design choices.

Replace Feature Descriptors. Given that 3DGS can render high-quality synthetic images $\hat{I}_r$ in real-time, we show that using a pre-trained 3D vision fundation model, MASt3R, can directly establish accurate 2D-2D correspondences $C_{q,r}$ between I_q and $\hat{I}_r$ with sim-to-real domain gap. As demonstrated in Section 4.2, GSLoc achieves significantly higher accuracy than NeRF-based refinement pipelines that rely on feature rendering. Direct RGB matching makes our framework more compact, reduces runtime, eliminates the need for training additional neural radiance features, and simplifies both deployment and usage.

Efficient and Effective Pose Refinement. In the experiments detailed in Section 4.2, we demonstrate that the accuracy of the refined pose $\hat{p}'$ is proportional to the accuracy of the pose prior $\hat{p}$ provided by the pose estimator $\mathcal{F}$. DFNet provides less accurate predictions than Marepo and ACE, but NeFeS report the best results over DFNet. To ensure a fair comparison with NeFeS, we present examples in Figure 5 illustrating that our GSLoc outperforms NeFeS in both efficiency and effectiveness. With only one or two rounds of optimization, our GSLoc achieves higher accuracy than NeFeS with 50 optimization iterations when combined with DFNet on both the indoor 7Scenes and outdoor Cambridge Landmarks datasets. This superior performance is due to our method's leverage of 3D geometry (depth rendering) of the representation, unlike previous NeRF-based refinement methods [10, 53] that use only 2D feature/photometric information in an iterative process, rendering candidate poses and comparing them with the target image.

(a) 7Scenes

(b) Cambridge

Figure 5. A comparison between DFNet + GSLoc and DFNet + NeFeS$_{50}$. Each subfigure is divided by a diagonal line, with the **bottom left** part rendered using the estimated pose and the **top right** part displaying the ground truth image. (a) Our DFNet + GSLoc $_2$ achieves higher accuracy with two steps refinement from DFNet's predictions on 7Scenes dataset. (b) Our DFNet + GSLoc achieves higher accuracy with one-shot refinement from DFNet's predictions on Cambridge Landmarks dataset. Patches highlighting visual differences are emphasized with green insets for enhanced visibility.

5. Conclusion

To improve the localization accuracy of state-of-the-art absolute pose regression (APR) and scene coordinate regression (SCR) methods, we present GSLoc, a novel test-time camera pose refinement framework leveraging 3DGS for scene representation. By rendering high-quality images and depth maps, the 3DGS model facilitates the establishment of 2D-3D correspondences without the need to train feature extractors or descriptors. This is achieved by directly operating on RGB images and utilizing the 3D vision foundation model, MASt3R, for precise 2D matching. To improve robustness in challenging outdoor environments, we incorporated a test-time exposure-adaptive module within the 3DGS model. GSLoc enables efficient pose refinement

using only a single RGB query and a coarse initial pose estimate from APR and SCR methods. Our approach outperforms existing NeRF-based optimization methods in both accuracy and runtime across various indoor and outdoor visual localization benchmarks, achieving state-of-the-art accuracy on two indoor datasets. These results demonstrate the effectiveness and efficiency of our proposed framework in diverse visual localization scenarios.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2

[2] Matteo Bortolon, Theodore Tsesmelis, Stuart James, Fabio Poiesi, and Alessio Del Bue. 6dgs: 6d pose estimation from a single image and a 3d gaussian splatting model. *arXiv preprint arXiv:2407.15484*, 2024. 2

[3] Eric Brachmann and Carsten Rother. Visual camera relocalization from rgb and rgb-d images using dsac. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5847–5865, 2021. 1, 2, 5

[4] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6684–6692, 2017. 1, 2

[5] Eric Brachmann, Martin Humenberger, Carsten Rother, and Torsten Sattler. On the limits of pseudo ground truth in visual camera re-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6218–6228, 2021. 4

[6] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5044–5053, 2023. 1, 2, 4, 5, 6

[7] 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*, pages 2616–2625, 2018. 2

[8] Shuai Chen, Zirui Wang, and Victor Prisacariu. Direct-posenet: absolute pose regression with photometric consistency. In *2021 International Conference on 3D Vision (3DV)*, pages 1175–1185. IEEE, 2021. 1, 2

[9] Shuai Chen, Xinghui Li, Zirui Wang, and Victor A Prisacariu. Dfnet: Enhance absolute pose regression with direct feature matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 1–17. Springer, 2022. 2, 3, 4, 5, 6, 1

[10] Shuai Chen, Yash Bhalgat, Xinghui Li, Jia-Wang Bian, Kejie Li, Zirui Wang, and Victor Adrian Prisacariu. Neural refinement for absolute pose regression with feature synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20987–20996, 2024. 1, 2, 3, 4, 5, 8

[11] 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*, pages 20665–20674, 2024. 2, 4, 5, 1

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 4

[13] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8092–8101, 2019. 1, 2

[14] 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. 2, 3

[15] 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. 1, 2, 3

[16] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *European Conference on Computer Vision*, 2020. 2

[17] Hugo Germain, Daniel DeTone, Geoffrey Pascoe, Tanner Schmidt, David Novotny, Richard Newcombe, Chris Sweeney, Richard Szeliski, and Vasileios Balntas. Feature query networks: Neural surface description for camera pose refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5071–5081, 2022. 2, 5, 6

[18] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time rgb-d camera relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 173–179. IEEE, 2013. 2, 4

[19] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. *IJCV*, 2017. 2

[20] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE international conference on Robotics and Automation (ICRA)*, pages 4762–4769. IEEE, 2016. 2

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

[22] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international*

conference on computer vision, pages 2938–2946, 2015. 1, 2, 4, 5

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2, 4

[24] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 2, 3, 4, 7

[25] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 1, 2

[26] Changkun Liu, Shuai Chen, Yukun Zhao, Huajian Huang, Victor Prisacariu, and Tristan Braud. Hr-apr: Apr-agnostic framework with uncertainty estimation and hierarchical refinement for camera relocalisation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8544–8550, 2024. 2, 5, 6

[27] Jianlin Liu, Qiang Nie, Yong Liu, and Chengjie Wang. Nerf-loc: Visual localization with conditional neural radiance field. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9385–9392. IEEE, 2023. 2, 5

[28] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024. 2, 4

[29] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*, pages 1347–1356. PMLR, 2022. 1, 2, 6

[30] Arthur Moreau, Nathan Piasco, Moussab Bennehar, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Crossfire: Camera relocalization on self-supervised features from an implicit representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 252–262, 2023. 1, 2, 3, 5

[31] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017. 1, 2

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 4

[33] Chengyu Qiao, Zhiyu Xiang, Yuangang Fan, Tingming Bai, Xijun Zhao, and Jingyun Fu. Transapr: Absolute camera pose regression with spatial and temporal attention. *IEEE Robotics and Automation Letters*, 8(8):4633–4640, 2023. 2

[34] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 5

[35] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 1, 2, 5

[36] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 5

[37] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016. 1, 2

[38] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3302–3312, 2019. 1, 2

[39] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 4

[40] 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*, pages 2733–2742, 2021. 1, 2, 5

[41] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 2, 4

[42] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 7

[43] Yuan Sun, Xuan Wang, Yunfan Zhang, Jie Zhang, Caigui Jiang, Yu Guo, and Fei Wang. icomma: Inverting 3d gaussians splatting for camera pose estimation via comparing and matching. *arXiv preprint arXiv:2312.09031*, 2023. 2

[44] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 1, 2

[45] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1808–1817, 2015. 5

[46] Gabriele Trivigno, Carlo Masone, Barbara Caputo, and Torsten Sattler. The unreasonable effectiveness of pre-trained features for camera pose refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12798, 2024. 1, 2

[47] Abhinav Valada, Noha Radwan, and Wolfram Burgard. Deep auxiliary learning for visual localization and odometry. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 6939–6946. IEEE, 2018. 2

[48] Julien Valentin, Angela Dai, Matthias Nießner, Pushmeet Kohli, Philip Torr, Shahram Izadi, and Cem Keskin. Learning to navigate the energy landscape. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 323–332. IEEE, 2016. 2, 4

[49] Bing Wang, Changhao Chen, Chris Xiaoxuan Lu, Peijun Zhao, Niki Trigoni, and Andrew Markham. Atloc: Attention guided camera localization. *arXiv preprint arXiv:1909.03557*, 2019. 1, 2

[50] Bing Wang, Changhao Chen, Chris Xiaoxuan Lu, Peijun Zhao, Niki Trigoni, and Andrew Markham. Atloc: Attention guided camera localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10393–10401, 2020. 2

[51] Fangjinhua Wang, Xudong Jiang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Glace: Global local accelerated coordinate encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21562–21571, 2024. 1, 2, 5

[52] 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*, pages 20697–20709, 2024. 7

[53] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021. 2, 8

[54] Qunjie Zhou, Maxim Maximov, Or Litany, and Laura Leal-Taixé. The nerfect match: Exploring nerf features for visual localization. *arXiv preprint arXiv:2403.09577*, 2024. 2, 5, 1

GSLoc: Efficient Camera Pose Refinement via 3D Gaussian Splatting

Supplementary Material

6. GT Poses Details

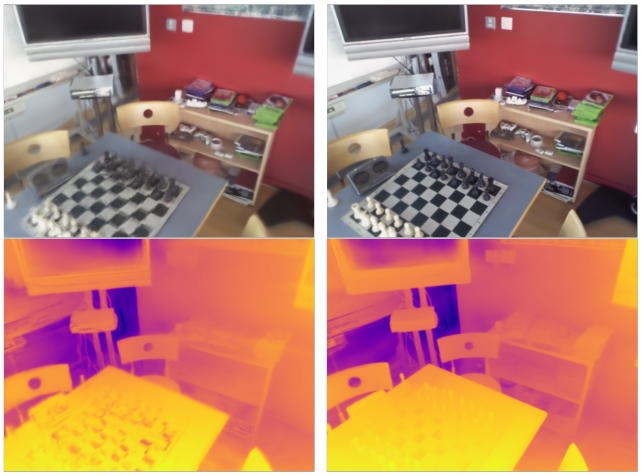
In Section 4.2, we report evaluation results based on the SfM ground truth (GT) poses for 7Scenes dataset, as these poses can render higher quality images [10]. Since NeFeS [10] demonstrates the superior accuracy of SfM poses using NeRF as the scene representation, we provide a quantitative comparison in Table 10 and illustrative rendering examples in Figure 6. These results affirm that SfM poses are more accurate, leading to higher quality rendered images and depth maps when using 3DGS. In the main paper, we present only the results for SfM GT. We attach the results of dSLAM GTs on 7Scenes in Table 11. It shows that our approach and other methods [6, 11, 46] are more competitive on more accurate SfM GTs.

Table 10. Quatitative comparison between the 3DGS models implemented in Section 4.1 trained by dSLAM GT poses and SfM GT poses. We report the average PSNR (dB) for the test frames in each scene. The best results are in bold (higher is better).

	dSLAM GT	SfM GT
Scenes	avg. PSNR ↑	avg. PSNR ↑
chess	19.6	**23.1**
fire	19.8	**21.2**
heads	18.4	**19.7**
office	19.4	**21.7**
pumpkin	20.3	**23.2**
redkitchen	18.5	**21.4**
stairs	19.7	**20.1**
avg.	19.4	**21.5**

7. Comparison with Other Camera Relocalization Approaches

While our paper primarily focuses on evaluating improvements over APR and SCR methods and comparing them with NeRF-based neural render pose estimation methods [10, 30, 54], we also include a comparison here with the pose refinement pipeline, MCLoc [46], which is agnostic to scene representation. MCLoc reports visual localization results using meshes or 3DGS models across various datasets. We compare our state-of-the-art ACE + GSLoc with MCLoc on the 7Scenes and Cambridge Landmarks datasets, as MCLoc provides results using 3DGS models as scene representations for these datasets. Our ACE + GSLoc



(a) dSLAM GT (b) SfM GT

Figure 6. Render performance example (dSLAM GT vs. SfM GT). The 3DGS model trained with SfM GT poses (b) renders superior geometric details compared to the dSLAM 3DGS (a) for the same query image, particularly in the chessboard and pieces area.

demonstrates higher accuracy than MCLoc, as shown in Table 11 and Table 12. On the Cambridge Landmarks dataset, MCLoc requires an average of 2.4s per query with 80 iterations [46]. In contrast, our ACE + GSLoc with one-shot optimization only takes 0.2s per query. Therefore, in terms of efficiency and improvement, our GSLoc is better than MCLoc when using 3DGS as scene representation.

We also compare our results with sequential-based APR methods. Tables 11 and 12 show that when compared with sequential-based APR methods, our single-frame GSLoc achieves more accurate results on both Cambridge Landmarks dataset and 7Scenes dataset.

8. Accuracy Bound

Here we show the effect of iteration rounds on the accuracy improvement of different methods.

GSLoc: We demonstrate that for DFNet [9], which provides less accurate initial predictions than Marepo [11] and ACE [6], at least two rounds of iterative optimization are required to achieve accuracy exceeding DFNet + NeFeS_{50} [10] in Table 13. For ACE, precise initial predictions enable a single round of GSLoc optimization to achieve convergent accuracy, with further iterations offering no improvement in median translation and rotation errors, as shown in Table 13.

Table 11. Comparisons on 7Scenes dataset. The median translation and rotation errors (cm/°) of different methods. The best results are in bold (lower is better). NRP denotes neural render pose estimation. **GSLoc** refers to the refinement of the pose over a single iteration.

	Methods	Chess	Fire	Heads	Office	Pumpkin	Redkitchen	Stairs	Avg. ↓ [cm/°]
1-frame APR	Marepo (dSLAM GT)	2.6/1.35	2.5/1.42	2.3/2.21	3.6/1.44	4.2/1.55	5.1/1.99	6.7/1.83	3.9/1.68
	Marepo (SfM GT)	1.9/0.83	2.3/0.92	2.1/1.24	2.9/0.93	2.5/0.88	2.9/0.98	5.9/1.48	2.9/1.04
SCR	ACE (dSLAM GT)	1.9/0.7	1.9/0.9	0.9/0.6	2.7/0.8	4.2/1.1	4.2/1.3	3.9/1.1	2.8/ 0.93
	ACE (SfM GT)	**0.5**/0.18	**0.8**/0.33	**0.5**/0.33	**1**/0.29	**1**/0.22	0.8/0.2	2.9/0.81	1.1/0.34
Seq. APR	VLocNet [47]	3.6/1.71	3.9/5.34	4.6/6.64	3.9/1.95	3.7/2.28	3.9/2.2	9.7/6.48	4.8/3.8
	MapNet + PGO [7]	9/3.24	20/9.29	12/8.45	19/5.42	19/3.96	20/4.94	27/10.57	18/6.55
	AtLoc+ [50]	10/3.18	26/10.8	14/11.4	17/5.16	20/3.94	16/4.9	29/10.2	19/7.1
	TransAPR+ [33]	8/3.4	21/8.4	14/9.5	17/5.5	18/4.1	19/4.7	23/8.5	17/6.3
NRP (dSLAM GT)	MCLoc [46]	5/1.8	4/2.0	4/1.9	10/3.6	10/3.7	8/3.1	10/2.5	7.3/2.66
	Marepo + **GSLoc (ours)**	2.7/0.85	2.3/0.92	1.8/0.94	3.3/0.89	4.9/1.09	5/1.43	4.6/1.21	3.5/1.0
	ACE + **GSLoc (ours)**	2/0.7	2/0.9	1/0.6	2.8/0.8	5/1.1	4.4/1.3	3.5/0.9	3/0.9
NRP (SfM GT)	MCLoc [46]	2/0.8	3/1.4	3/1.3	4/1.3	5/1.6	6/1.6	6/2.0	4.1/1.43
	Marepo + **GSLoc (ours)**	1.3/0.4	1.6/0.5	1.4/0.68	2.2/0.5	2/0.47	2.2/0.48	3.8/0.86	2.1/0.56
	ACE + **GSLoc (ours)**	**0.5/0.15**	0.9/**0.27**	**0.5/0.28**	**1/0.25**	1.1/**0.21**	**0.8/0.17**	**2.3/0.58**	**1.0/0.27**

Table 12. Comparisons on Cambridge Landmarks dataset. We report the median translation and rotation errors (cm/°) of different methods. NRP denotes neural render pose estimation. Best results are in bold (lower is better) among the NRP-based approaches.

	Methods	Kings	Hospital	Shop	Church	Avg. ↓ [cm/°]
Seq. APR	TransAPR [33]	59/0.86	142/2.29	54/2.18	121/3.16	94/2.12
	VLocNet [47]	84/1.42	108/2.41	59/3.53	63/3.91	78/2.82
NRP	MCLoc [46]	31/0.42	39/0.73	12/0.45	26/0.8	27/0.6
	DFNet + **GSLoc (ours)**	26/0.34	48/0.72	10/0.36	27/0.62	28/0.51
	ACE + **GSLoc (ours)**	**25/0.29**	**26/0.38**	**5/0.23**	**13/0.41**	**17/0.33**

GSLoc$_{\text{rel}}$: We demonstrate that one-shot GSLoc$_{\text{rel}}$ refinement over DFNet already achieves the accuracy upper bound as shown in Table 13. DFNet + GSLoc$_{\text{rel}\ 2}$ results in performance degradation. DFNet + GSLoc$_{\text{rel}}$ achieves higher rotation accuracy than DFNet + GSLoc $_2$. Although ACE + GSLoc indicates that the accuracy bound of GSLoc$_{\text{rel}}$ remains a bit lower, but the fast refinement offered by this method maintains its competitiveness.

9. Supplementary Video

To complement our quantitative analysis, we provide a supplementary video offering a qualitative perspective, focusing on pixel-wise alignment using novel feature synthesis based on 3DGS across three datasets. The video primarily demonstrates the camera pose accuracy improvements of our GSLoc over DFNet [9], Marepo [11], and ACE [6] as well as by GSLoc$_{\text{rel}}$ over DFNet. We also compare GSLoc with NeFeS [10] using DFNet as the pose estimator.

Table 13. Results of GSLoc$_{\text{rel}}$ in Figure 2 and GSLoc in Figure 1 on 7Scenes dataset. We report the average median translation and rotation errors (cm/°) across 7 scenes.

Methods	Avg Err ↓ [cm/°]
DFNet [9]	6/1.93
DFNet + NeFeS$_{50}$ [10]	2.4/0.79
DFNet + **GSLoc$_{\text{rel}}$ (ours)**	2.7/**0.38**
DFNet + **GSLoc$_{\text{rel}\ 2}$ (ours)**	2.9/0.4
DFNet + **GSLoc (ours)**	4.0/1.0
DFNet + **GSLoc $_2$ (ours)**	**2.1**/0.63
ACE [6]	1.1/0.34
ACE + **GSLoc$_{\text{rel}}$ (ours)**	2.8/0.43
ACE + **GSLoc (ours)**	**1.0/0.27**
ACE + **GSLoc $_2$ (ours)**	**1.0/0.27**