SCALE-CONSISTENT FEATURE FIELDS FOR RETRIEVAL-FREE VISUAL RELOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature-based scene representation plays a critical role in visual relocalization, enabling robots and augmented reality systems to estimate their positions and orientations without global positioning. However, this task is often undermined by significant variations in viewpoint-induced scale and scene appearance. Recently, Gaussian splatting has emerged as a promising scene model due to its explicit geometry and real-time rasterization. In this paper, we propose Scale-Place, a retrieval-free framework that learns Scale-Consistent Gaussian Feature Fields (SCGFF) for retrieval-free centimeter-level visual relocalization. Specifically, our framework performs pose estimation through an integrated sparse-to-dense pipeline built upon a rasterizable Gaussian field, comprising three key aspects: First, Descriptor-Attached Gaussian Representation (DAGR) yields a view-dependent feature field that can be both sparsely sampled for initialization and densely rendered for refinement; Second, Multi-Resolution Feature Supervision (MRFS) facilitates cross-scale consistency through Gaussian rendering at multiple resolutions with aligned photometric and SuperPoint objectives; Third, Scale-Aware Pose Estimation (SAPE) empowers the sparse matcher and dense alignment with resolution-consistent primitives to support both initialization and iterative refinement without retrieval or pose priors. Extensive evaluations across public datasets demonstrate that the proposed approach substantially outperforms existing baselines and ablation variants in diverse indoor and outdoor environments.

1 INTRODUCTION

Visual localization estimates the six-degree-of-freedom camera pose of a query image with respect to a known scene Li et al. (2025). It is a core capability for robotics, augmented reality, and indoor navigation, where GPS is unavailable and centimeter-level accuracy is often required. The problem remains challenging because query viewpoints, illumination, and apparent object scale can differ substantially from the images used to build the map Li & Shang (2026). Existing methods address this task with different scene representations and pose estimators.

Scene-based pipelines (SbP) recover pose from image-to-scene correspondences Sarlin et al. (2021). They are accurate in well-textured environments, yet typically depend on retrieval, sparse SfM points, or hand-crafted matching heuristics. Consequently, they degrade when the query viewpoint is poorly covered or when local descriptors become unstable under scale change. Scene-coordinate regression (SCR) methods directly predict 3D coordinates for image pixels and then solve PnP Brachmann & Rother (2021). This avoids retrieval, but the mapping is implicit and often scene-specific. Accuracy can drop in weakly textured regions, as illustrated by the large error of D2S on Stairs in 7-Scenes. Absolute pose regression (APR) networks directly regress camera pose from a single image Kendall et al. (2015). They are compact and fast, but generally remain far from centimeter-level ac-

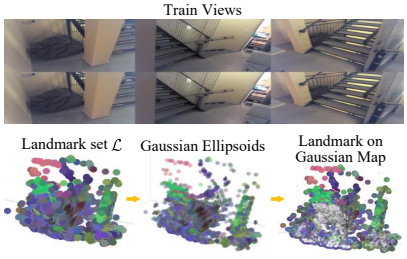


Figure 1: Construction of the localization map. Training views are used to learn a Gaussian field. Matching-oriented landmarks $\mathcal{L}$ and Gaussian ellipsoids are fused into the final map.

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.

curacy, especially on large outdoor scenes such as Cambridge Landmarks. Relative pose regression (RPR) methods estimate the transformation between a query and one or more reference views. Pairwise models such as RelFormer and Map-free Reloc struggle with large baselines, while multi-view methods such as ReLoc3r Dong et al. (2025) and DeMT Wang et al. (2026) improve robustness by aggregating multiple references Wang et al. (2026). Even the strongest RPR baseline, DeMT, still yields coarser rotation estimates than explicit neural scene representations. Neural rendering methods based on NeRF or 3D Gaussian Splatting (GS) provide a denser scene model and have recently become competitive for localization. DFNet+NeFeS Chen et al. (2024), CROSSFIRE Moreau et al. (2023), PNeRFLoc Zhao et al. (2024), and NeRFMatch Zhou et al. (2024) render appearance or features for pose alignment, but implicit NeRFs are expensive to query and difficult to sample sparsely. Explicit Gaussian fields alleviate this limitation: STDLoc attaches localization descriptors to 3D Gaussians and performs sparse-to-dense matching, achieving strong indoor and outdoor results. Nevertheless, a single-resolution feature field remains sensitive to viewpoint-induced scale change, and sparse landmarks selected without matching supervision may be suboptimal for 2D-3D correspondence.

We present ScalePlace, a two-stage localization framework that builds a scale-aware Gaussian scene representation and estimates pose by hierarchical matching. In the mapping stage, Descriptor-Attached Gaussian Representation (DAGR) equips each Gaussian with a learnable localization descriptor, while Multi-Resolution Feature Supervision (MRFS) trains the field at multiple rasterization scales, as shown in Figure 1. Matching-Oriented Sampling (MOS) then selects a compact landmark set from Gaussians that are both spatially distributed and easy to match. In the localization stage, Scale-Aware Pose Estimation (SAPE) first matches query keypoints to landmarks for PnP initialization and then refines the pose by coarse-to-fine dense 2D–3D matching. The pipeline requires neither image retrieval nor a pose prior.

2 Related Work

2.1 Correspondence-Based Localization

Structure-based pipelines build an SfM map and recover pose by 2D-3D matching and PnP Larsson et al. (2019). Active Search Sattler et al. (2016) predicts visible landmarks to accelerate matching, while hierarchical methods combine image retrieval with learned matchers. PixLoc Sarlin et al. (2021) further refines pose by dense feature-metric alignment, and later systems including DeViLoc Giang et al. (2024) improve matching robustness. These scene-based pipelines are strong in textured environments, but they usually rely on retrieval or sparse SfM points and can fail when the query viewpoint is poorly covered or when descriptors vary with scale.

Scene-coordinate regression (SCR) replaces explicit 2D-3D matching with a learned mapping from pixels to 3D coordinates, followed by RANSAC PnP Donadi & Pretto (2024). DSAC* makes RANSAC differentiable Brachmann & Rother (2021), ACE trains a compact scene-specific coordinate network, and subsequent methods such as NeuMap Tang et al. (2023) introduce hierarchical or uncertainty-aware prediction. SCR avoids retrieval and can be accurate indoors, yet the scene model remains implicit and is often brittle in weakly textured or repetitive regions. ScalePlace instead uses an explicit Gaussian landmark set, so the same primitives support both sparse matching and dense feature rendering, without a retrieval database.

2.2 Absolute and Relative Pose Regression

Absolute pose regression (APR) directly predicts camera pose from a single image. PoseNet established this direction; later models such as PAE Shavit & Keller (2022) add geometric constraints or neural scene priors. APR is compact and fast, but its accuracy is typically far from centimeter level, especially on large outdoor landmarks.

Relative pose regression (RPR) estimates the transform between a query and one or more reference views. Pairwise methods, including RelocNet Balntas et al. (2018), AnchorNet Saha et al. (2018), NC-EssNet Zhou et al. (2020), Map-free Reloc Arnold et al. (2022), and RelFormer Idan et al. (2026), are sensitive to large baselines. Multi-view methods such as CamNet, RelPoseGNN, ReLoc3r, and DeMT aggregate multiple references and are more robust, yet still lag behind explicit neural maps in rotation accuracy. These approaches are attractive in map-free settings, but our

goal is centimeter-level absolute localization against a scene-specific map rather than relative pose prediction.

2.3 Neural Scene Representations for Localization

Neural rendering provides a denser alternative to sparse SfM maps. NeRF-based localization treats pose estimation as inverse rendering: iNeRF Yen-Chen et al. (2021) optimizes pose by photometric alignment, while LENS Moreau et al. (2022) and DFNet Chen et al. (2022) improve initialization or joint structure-pose learning. CROSSFIRE, PNeRFLoc, and NeRFMatch render features for 2D-3D matching. Implicit volumes capture dense geometry, but they are expensive to query and difficult to sample sparsely.

3D Gaussian Splatting offers an explicit and real-time representation. Feature Gaussians attach learned descriptors for segmentation or language grounding. For localization, RelLoc3r Dong et al. (2025) combines Gaussians with foundation matchers, while STDLoc Huang et al. (2025) learns scene-specific descriptors and performs sparse-to-dense matching. STDLoc is the closest prior work and already outperforms most NeRF/GS baselines on 7-Scenes and Cambridge Landmarks. Its feature field, however, is trained at a single rendering scale, and landmarks are not selected with matching as the explicit objective.

Scale-space pyramids and multi-resolution detectors, from SIFT to SuperPoint DeTone et al. (2018), have long been used for scale-invariant sparse features. We instead enforce scale consistency at the scene-representation level: the same 3D Gaussians are rendered and supervised at multiple resolutions. Relative to STDLoc, ScalePlace adds multi-resolution feature supervision, matching-oriented landmark sampling, and coarse-to-fine pose refinement, yielding a retrieval-free localizer that remains robust under viewpoint-induced scale change.

3 Method

ScalePlace builds a Gaussian map from calibrated training views to localize queries with known intrinsics. As shown in Figure 2, DAGR provides a shared field for landmark sampling and dense feature rendering, MRFS supervises it at multiple resolutions, and SAPE initializes pose from selected landmarks before refining it with rendered features. We define $\mathbf{P} = (\mathbf{R}, \mathbf{t}) \in \mathrm{SE}(3)$ as the world-to-camera transformation, mapping $\mathbf{x}$ to $\mathbf{R}\mathbf{x} + \mathbf{t}$. Pixel coordinates denote pixel centers, with intrinsics $\mathbf{K}$ defined for the corresponding image grid.

3.1 Descriptor-Attached Gaussian Representation

3.1.1 Field Parameterization

ScalePlace uses one anisotropic Gaussian field for sparse initialization and dense refinement. Each primitive stores geometry, opacity, appearance, and a learnable localization descriptor. Sampled primitives provide explicit landmarks, while tile-based splatting renders images and feature maps from the full field.

The scene contains N Gaussians,

$$\mathcal{G} = \left\{ (\boldsymbol{\mu}_i, \mathbf{s}_i, \mathbf{q}_i, \alpha_i, \mathbf{c}_i, \boldsymbol{d}_i) \right\}_{i=1}^{N}, \tag{1}$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^3$ is the world-space mean, $\mathbf{s}_i \in \mathbb{R}^3_+$ gives scales along the local principal axes, $\mathbf{q}_i$ is a unit rotation quaternion, and $\alpha_i \in [0, 1]$ is opacity. Spherical-harmonic coefficients $\mathbf{c}_i$ parameterize view-dependent color. Each Gaussian also stores a descriptor $\boldsymbol{d}_i \in \mathbb{R}^D$ with $D = 256$. Its covariance is

$$\boldsymbol{\Sigma}_i = \mathbf{R}(\mathbf{q}_i)\,\mathrm{diag}(\mathbf{s}_i)^2\mathbf{R}(\mathbf{q}_i)^\top, \tag{2}$$

where $\mathbf{R}(\mathbf{q}_i)$ converts the quaternion into a rotation matrix. Descriptors are randomly initialized and optimized jointly with the field. SuperPoint supervision aligns them with the query embedding used during localization.

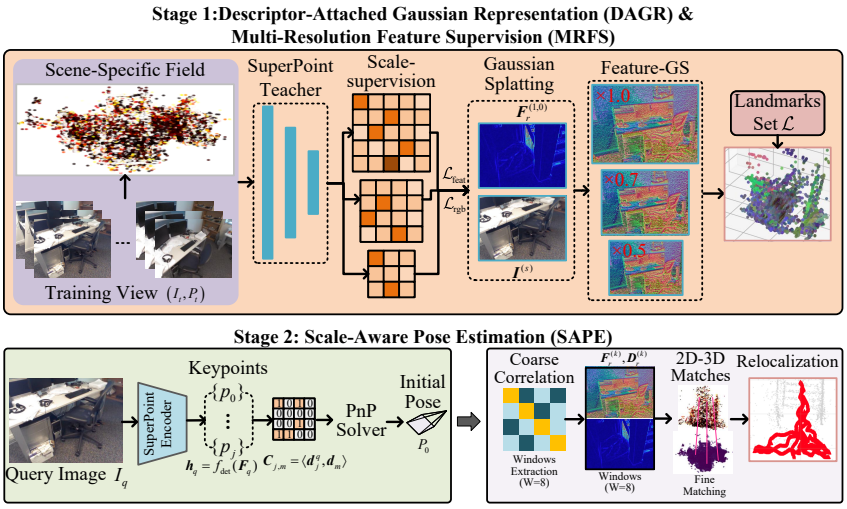


Figure 2: ScalePlace pipeline. Top: the descriptor-attached Gaussian field is trained with multi-resolution photometric and feature supervision; landmarks are subsequently selected from the frozen field. Bottom: query-to-landmark matching initializes the pose, which is refined through coarse-to-fine matching against rendered feature and depth maps.

3.1.2 Differentiable Rasterization

For a calibrated camera, projected Gaussians are composited in front-to-back order. Let $a_i(\mathbf{p}) \in [0,1]$ denote the effective opacity of primitive i at pixel $\mathbf{p}$, including its projected footprint and opacity. The transmittance preceding that primitive and the rendered feature are

$$T_i(\mathbf{p}) = \prod_{j<i} \big(1 - a_j(\mathbf{p})\big), \tag{3}$$

$$\mathbf{F}_r(\mathbf{p}) = \sum_i T_i(\mathbf{p}) a_i(\mathbf{p}) \widetilde{\boldsymbol{d}}_i. \tag{4}$$

The ordering is local to the pixel. The descriptor supplied to the rasterizer is $\widetilde{\boldsymbol{d}}_i = \boldsymbol{d}_i$, or its normalized form when pre-render normalization is enabled. This convention is fixed across scales within a training and localization configuration. Color is composited analogously, with the configured background contribution. Antialiased rasterization is used throughout.

Individual descriptors are view-independent; the rendered image $\mathbf{F}_r \in \mathbb{R}^{D\times H\times W}$ is pose-dependent because visibility, projected footprints, and blending weights change with the camera. Photometric and feature losses both influence the shared geometry, while feature supervision additionally optimizes the descriptors. The rasterizer's variable output resolution enables the multi-resolution objective below.

3.1.3 Matching-Oriented Landmark Sampling

After mapping, Matching-Oriented Sampling (MOS) selects a compact landmark subset. Let $\mathcal{J} \subseteq \{1,\dots,N\}$ be the selected indices and $\mathcal{L} = \{(\boldsymbol{\mu}_i, \boldsymbol{d}_i) : i \in \mathcal{J}\}$ the sparse map. Define descriptor normalization as $\nu(\mathbf{x}) = \mathbf{x}/\max(\|\mathbf{x}\|_2, \epsilon)$, with a small $\epsilon > 0$.

For primitive i, let $\mathcal{V}_i$ contain training views in which its projection is valid and the rasterizer marks it visible. Its matching score is

$$u_i = \frac{1}{\max(1, |\mathcal{V}_i|)} \sum_{t\in\mathcal{V}_i} \Big\langle \nu(\boldsymbol{d}_i), \widehat{\mathbf{F}}_t^{\mathrm{SP}}(\mathbf{p}_{it}) \Big\rangle, \tag{5}$$

where $\widehat{\mathbf{F}}_t^{\mathrm{SP}}$ is the normalized training-image feature map and $\mathbf{p}_{it}$ is the projected pixel of $\boldsymbol{\mu}_i$ in view t. The implementation samples features at projected pixel indices. Unobserved primitives receive score zero.

MOS draws random seed primitives with target budget $M = 16{,}384$. For each seed, its $k_{\mathrm{nn}} = 32$ nearest spatial neighbors are ranked by u_i, and the highest-scoring unused candidate is selected. Neighborhoods without unused candidates are skipped, so the unique landmark count can be below M. This combines local descriptor agreement with spatially distributed seeds; it does not explicitly maximize descriptor separation. Sparse localization reads these selected means and descriptors, while dense refinement renders the same underlying field.

3.2 Multi-Resolution Feature Supervision

3.2.1 Multi-Resolution Rasterization

For each training view, SuperPoint is evaluated once on the image $\mathbf{I}$. Its output is bilinearly resized to the native mapping grid $H \times W$ and normalized along the descriptor dimension to obtain $\mathbf{F}^{\mathrm{SP}}$. MRFS uses rendering scales $\mathcal{S} = \{1.0, 0.7, 0.5\}$, where 1.0 denotes this native resolution. At scale s, the field is rendered under the same pose to

$$H_s = \max(1, \lfloor sH \rfloor), \qquad W_s = \max(1, \lfloor sW \rfloor). \tag{6}$$

The corresponding targets are

$$\mathbf{I}^{(s)} = \mathrm{Resize}(\mathbf{I}; H_s, W_s), \tag{7}$$

$$\mathbf{F}^{\mathrm{SP},(s)} = \nu\big(\mathrm{Resize}(\mathbf{F}^{\mathrm{SP}}; H_s, W_s)\big), \tag{8}$$

where ν acts independently at each pixel. The renderer produces aligned outputs $\mathbf{I}_r^{(s)}$ and $\mathbf{F}_r^{(s)}$. With pixel-center coordinates, intrinsics are adjusted as

$$\mathbf{K}^{(s)} = \mathrm{diag}(W_s/W, H_s/H, 1)\mathbf{K}. \tag{9}$$

This preserves the field of view while changing the sampling density. Every scale updates the same field parameters.

3.2.2 Mapping Objective

Let $\ell_1(\mathbf{A}, \mathbf{B})$ denote the mean absolute difference over channels and pixels. Using a mean rather than an unnormalized sum prevents the loss magnitude from growing with resolution. The native-resolution objective retains photometric and feature terms,

$$\begin{aligned}\mathcal{L}_{\mathrm{rgb}} = (1 - \lambda_{\mathrm{dssim}})\ell_1(\mathbf{I}_r^{(1)}, \mathbf{I}^{(1)}) \\ + \lambda_{\mathrm{dssim}}\big(1 - \mathrm{SSIM}(\mathbf{I}_r^{(1)}, \mathbf{I}^{(1)})\big),\end{aligned} \tag{10}$$

$$\mathcal{L}_{\mathrm{feat}} = \ell_1(\mathbf{F}_r^{(1)}, \mathbf{F}^{\mathrm{SP},(1)}), \tag{11}$$

where SSIM denotes structural similarity and $\lambda_{\mathrm{dssim}} = 0.2$. Each coarser scale contributes

$$\mathcal{L}^{(s)} = \ell_1(\mathbf{I}_r^{(s)}, \mathbf{I}^{(s)}) + \ell_1(\mathbf{F}_r^{(s)}, \mathbf{F}^{\mathrm{SP},(s)}). \tag{12}$$

The complete mapping objective is

$$\mathcal{L}_{\mathrm{map}} = \mathcal{L}_{\mathrm{rgb}} + \mathcal{L}_{\mathrm{feat}} + \lambda_{\mathrm{MS}} \sum_{s \in \mathcal{S} \setminus \{1\}} \mathcal{L}^{(s)}, \tag{13}$$

MRFS constrains the shared field to reconstruct aligned teacher responses at several sampling densities. It introduces neither a second descriptor space nor a separate pairwise consistency loss. The teacher targets are resampled features, not independent SuperPoint encodings of resized images. Scale consistency here refers to this shared reconstruction constraint; changes in rendering resolution do not reproduce every geometric or appearance change caused by a new viewpoint.

3.3 Scale-Aware Pose Estimation

SAPE uses the landmark set for sparse initialization and pose-conditioned rendering for refinement. The query is encoded by the same frozen SuperPoint teacher, resized to the localization grid, and channel-normalized to give $\mathbf{F}_q \in \mathbb{R}^{D \times H_q \times W_q}$. No database image or external pose prior is required.

For a world point $\mathbf{x}$ with positive camera depth, perspective projection is

$$\pi(\mathbf{K}, \mathbf{P}, \mathbf{x}) = \mathrm{dehom}\big(\mathbf{K}(\mathbf{R}\mathbf{x} + \mathbf{t})\big), \tag{14}$$

where $\mathrm{dehom}([a, b, c]^\top) = [a/c, b/c]^\top$. Given correspondences $\mathcal{M} = \{(\mathbf{p}, \mathbf{x})\}$, RANSAC identifies a consensus set $\mathcal{I} \subseteq \mathcal{M}$. Pose refinement on that set is expressed as

$$\widehat{\mathbf{P}}(\mathcal{M}) = \arg \min_{\mathbf{P} \in \mathrm{SE}(3)} \sum_{(\mathbf{p}, \mathbf{x}) \in \mathcal{I}} \|\mathbf{p} - \pi(\mathbf{K}, \mathbf{P}, \mathbf{x})\|_2^2. \tag{15}$$

The inliers are estimated during robust pose estimation, rather than assumed known beforehand. Sparse and dense stages differ in how they construct the correspondences supplied to this solver.

3.3.1 Sparse Initialization

After mapping and landmark selection, a lightweight scene-specific detector f_{det} is trained with the field and teacher frozen. Its input is the SuperPoint feature map of a training image. Visible landmarks projected into that view provide a binary target heatmap: projected landmark pixels are positive and other pixels are negative. The detector is optimized with pixelwise binary cross-entropy. Thus, training relates query-image features to the selected map landmarks; it does not use Gaussian-rendered features as detector inputs.

At inference, the detector produces $\mathbf{h}_q = f_{\mathrm{det}}(\mathbf{F}_q)$. Non-maximum suppression with radius four retains $n_q \le N_{\max} = 2048$ positive-score pixels. Each keypoint $\mathbf{p}_j$ supplies descriptor $\boldsymbol{d}_j^q = \mathbf{F}_q(\mathbf{p}_j)$. Landmark descriptors are normalized, giving similarity

$$C_{jm} = \langle \boldsymbol{d}_j^q, \nu(\boldsymbol{d}_m) \rangle, \qquad j = 1, \dots, n_q, \quad m \in \mathcal{J}. \tag{16}$$

The default top-one matcher selects

$$m_j = \arg \max_{m \in \mathcal{J}} C_{jm}, \tag{17}$$

$$\mathcal{M}_0 = \{(\mathbf{p}_j, \boldsymbol{\mu}_{m_j}) : C_{jm_j} > \theta_s\}, \tag{18}$$

with acceptance threshold $\theta_s = 0$. The initial pose is $\mathbf{P}_0 = \widehat{\mathbf{P}}(\mathcal{M}_0)$, using a 12-pixel reprojection threshold and RANSAC confidence 0.99999. Empty or degenerate correspondences do not yield a valid initialization. An identity fallback is an implementation convention; subsequent rendering alone does not guarantee recovery from such a failure.

MRFS affects sparse localization through the trained landmark descriptors and the landmark-derived detector targets. The query encoder, top-one matching rule, and PnP solver themselves are unchanged.

3.3.2 Dense Iterative Alignment

Starting from $\mathbf{P}_0$, refinement alternates feature matching and pose estimation for iterations $\ell = 0, \dots, L-1$. At each iteration, the field is rendered at $\mathbf{P}_\ell$ to produce $\mathbf{F}_r^{(\ell)}$ and an expected camera-axis depth map $\mathbf{D}_r^{(\ell)}$. A vanishing rendered feature image leaves the current estimate unchanged.

Coarse matching The query and rendered feature maps are bilinearly downsampled by a factor $b = 8$ and channel-normalized. The fine grid dimensions are chosen to be divisible by b, giving $H_c = H_q/b$ and $W_c = W_q/b$. Flatten the coarse maps into $\mathbf{Q}, \mathbf{A} \in \mathbb{R}^{n_c \times D}$ with $n_c = H_c W_c$. Their similarity matrix is $S_{uv} = \langle \mathbf{Q}_u, \mathbf{A}_v \rangle$. Dual-softmax reweighting is

$$C_{uv}^c = \frac{\exp(S_{uv}/\tau)}{\sum_{v'} \exp(S_{uv'}/\tau)} \, \frac{\exp(S_{uv}/\tau)}{\sum_{u'} \exp(S_{u'v}/\tau)}, \qquad \tau = 0.1. \tag{19}$$

The two factors normalize over rendered and query locations, respectively; their product is a matching score, not a doubly stochastic assignment. Mutual nearest neighbors above $\theta_c = 0$ form coarse pairs. Fewer than three coarse pairs abort the iteration.

Fine matching Each coarse pair identifies corresponding $b \times b$ blocks in the fine grids. Dual-softmax matching with the same temperature is applied within each block pair. Mutual nearest

Table 1: Comparison with state-of-the-art visual localization methods on the 7-Scenes dataset (indoor scenes). We report the median translation and rotation errors in centimeters and degrees, respectively.

Method		7Scenes Median Error (cm/ °) ↓							
		Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Average
SbP	AS	4/2.00	3/1.50	2/1.50	9/3.60	8/3.10	7/3.40	3/2.20	5.1/2.47
	PixLoc	2/0.80	2/0.73	1/0.82	3/0.82	4/1.21	3/1.20	5/1.30	2.9/0.98
	DeViLoc	2/0.78	2/0.74	1/0.65	3/0.82	4/1.02	3/1.19	4/1.12	2.7/0.90
	FaVoR	1/0.20	1/0.40	1/0.60	2/0.40	1/0.30	1/0.30	6/1.60	1.9/0.50
SCR	DSAC*	2/1.10	2/1.24	1/1.82	3/1.15	4/1.34	4/1.68	3/1.16	2.7/1.36
	ACE	2/1.10	2/1.80	2/1.10	3/1.40	3/1.30	3/1.30	3/1.20	2.7/1.31
	NeuMap	2/0.81	3/1.11	2/1.17	3/0.98	4/1.11	4/1.33	4/1.12	3.1/1.09
	HSCNet++	2/0.70	2/0.72	1/0.80	2/0.69	4/1.00	4/1.15	3/1.02	2.6/1.36
	D2S	2/0.57	2/0.74	1/0.75	2/0.62	3/0.83	3/1.04	13/2.02	3.7/0.94
	SACNet†	2/0.53	2/0.71	1/0.59	3/0.65	2/0.65	3/0.93	2/0.40	2.1/0.64
APR	PoseNet	32/8.12	47/14.4	29/12.0	48/7.68	47/8.42	59/8.64	47/13.8	44.1/10.4
	PAE	12/4.95	24/9.31	14/12.5	19/5.79	18/4.89	18/6.19	25/8.74	18.6/7.48
	DFNet	5/1.88	17/6.45	6/3.63	8/2.48	10/2.78	22/5.45	16/3.29	12.0/3.71
	LENS†	3/1.30	10/3.70	7/5.80	7/1.90	8/2.20	9/2.20	14/3.60	8.3/2.96
	PMNet†	4/1.70	10/4.51	7/4.23	7/1.96	14/3.33	14/3.36	16/3.62	10.3/3.24
	Marepo	2/1.24	2/1.39	2/2.03	3/1.26	4/1.48	4/1.71	6/1.67	3.3/1.54
RPR (pw)	NN-Ne	13/6.50	26/12.7	14/12.3	21/7.40	24/6.40	24/8.00	27/11.8	21.3/9.30
	ReLocNet	12/4.10	26/10.4	14/10.5	18/5.30	26/4.20	23/5.10	28/7.50	21.0/6.73
	AnchorNet	8/4.12	16/11.1	9/11.2	11/5.38	14/3.55	13/5.29	21/11.9	13.1/7.51
	NC-EssNet	12/5.60	26/9.60	14/10.7	20/6.70	22/5.70	22/6.30	31/7.90	21.0/7.50
	Map-free	9/2.66	13/4.54	11/4.81	11/2.77	16/3.11	14/3.48	18/4.70	13.1/3.72
	RelFormer	11/4.01	23/8.57	17/10.9	16/4.92	15/4.15	19/4.89	24/6.46	17.8/6.27
RPR (mv)	CamNet†	4/1.73	3/1.74	5/1.98	4/1.62	4/1.64	4/1.63	4/1.51	4.0/1.69
	RelPoseGNN	8/2.70	21/7.50	13/8.70	15/4.10	15/3.50	19/3.70	22/6.50	16.1/5.24
	ReLoc3r	3/0.99	4/1.13	2/1.23	5/0.88	7/1.14	5/1.23	12/2.25	5.4/1.26
	DeMT	2/0.74	3/0.95	2/1.18	4/0.72	5/0.97	4/1.16	4/1.80	3.6/1.12
NeRF/GS	DFNet+NeFeS50	2/0.57	2/0.74	2/1.28	2/0.56	2/0.55	2/0.57	5/1.28	2.43/0.79
	CROSSFIRE	1/0.4	5/1.9	3/2.3	5/1.6	3/0.8	2/0.8	12/1.9	4.43/1.38
	PNeRFLoc	2/0.80	2/0.88	1/0.83	3/1.05	6/1.51	5/1.54	32/5.73	7.29/1.76
	NeRFMatch	0.95/0.30	1.11/0.41	1.34/0.92	3.09/0.87	2.21/0.60	1.03/0.28	9.26/1.74	2.71/0.73
	STDLoc	0.46/0.15	0.57/0.24	0.45/0.26	0.86/0.24	0.93/0.21	0.63/0.19	1.42/0.41	0.76/0.24
	ScalePlace	**0.40/0.13**	**0.45/0.18**	**0.40/0.22**	**0.71/0.21**	**0.85/0.19**	**0.56/0.12**	**1.10/0.31**	**0.63/0.19**

neighbors above $\theta_f = 0$ yield pixel pairs $(\mathbf{p}_q, \mathbf{p}_r)$. This restores pixel-level correspondence within the matched blocks without an additional Gaussian rendering pass.

Depth lifting and pose update Let $z = \mathbf{D}_r^{(\ell)}(\mathbf{p}_r)$ denote rendered camera-axis depth and $\mathbf{K}_q$ the intrinsic matrix for the fine grid. A rendered pixel is mapped into world coordinates as

$$\mathbf{x}_w = \mathbf{R}_\ell^\top \left(z\mathbf{K}_q^{-1} \begin{bmatrix} \mathbf{p}_r \\ 1 \end{bmatrix} - \mathbf{t}_\ell \right) \in \mathbb{R}^3. \tag{20}$$

This relation assumes finite positive depth. Integer pixel indices are shifted by $(0.5, 0.5)$ before backprojection and PnP to follow the pixel-center convention. Because depth is blended over contributing Gaussians, the lifted point need not coincide with a Gaussian mean.

The resulting correspondences $\mathcal{M}_{\ell+1} = \{(\mathbf{p}_q, \mathbf{x}_w)\}$ give the update $\mathbf{P}_{\ell+1} = \widehat{\mathbf{P}}(\mathcal{M}_{\ell+1})$ with an 8-pixel reprojection threshold. Insufficient correspondences leave the pose unchanged. With the stated setting $L = 4$, the final estimate is $\mathbf{P}_L$.

Training and inference use different notions of resolution: MRFS supervises rasterizations at $\{1.0, 0.7, 0.5\}$ during mapping, while SAPE downsamples a native-resolution rendering for coarse matching. The added mapping scales therefore introduce no extra inference renders. Their role is to regularize the field consumed by the existing sparse-to-dense matching procedure.

4 Experiments

4.1 Experimental Details

We evaluate ScalePlace on 7-Scenes Shotton et al. (2013) and Cambridge Landmarks Kendall et al. (2015). A frozen SuperPoint teacher DeTone et al. (2018) supervises 256-dimensional Gaussian

Table 2: Comparison with state-of-the-art visual localization methods on the Cambridge Landmarks dataset (outdoor scenes).

Method		Cambridge Landmarks Median Error(cm/°) ↓					
		King's College	Hospital	Shop	St Mary's Church	Court	Average
SbP	AS	42/0.60	44/1.00	12/0.40	19/0.50	-	29/0.63
	PixLoc	14/0.24	16/0.32	5/0.23	10/0.34	30/0.14	15/0.23
	DeViLoc	12/0.21	13/0.28	4/0.18	7/0.23	18/0.11	110.18
	FaVoR	18/0.30	27/0.50	5/0.30	11/0.40	29/0.20	18/0.34
SCR	DSAC*	18/0.30	21/0.40	5/0.30	15/0.60	49/0.30	220.38
	ACE	28/0.40	31/0.60	5/0.30	18/0.60	43/0.20	25/0.42
	NeuMap	19/0.14	36/0.19	25/0.06	53/0.17	10/0.06	28.6/0.12
	HSCNet++	19/0.34	20/0.31	6/0.24	9/0.30	39/0.23	190.28
	D2S	**7/0.12**	15/0.29	3/0.17	8/0.25	23/0.11	11/0.19
	SACNet†	17/0.30	18/0.30	6/0.30	12/0.30	47/0.32	20/0.30
APR	PoseNet	192/5.40	231/5.40	146/8.10	266/8.50	-	209/6.85
	PAE	90/1.49	207/2.58	99/3.88	164/4.16	-	140/3.03
	DFNet	73/2.37	200/2.98	67/2.21	137/4.03	-	119/2.90
	LENS†	33/0.50	44/0.90	27/1.60	53/1.60	-	39/1.15
RPR (pw)	AnchorNet	79/0.95	211/3.05	77/3.25	122/3.02	589/3.53	216/2.76
	NC-EssNet	61/1.60	95/2.70	71/3.40	112/3.60	-	85/2.83
	Map-free	244/2.54	373/5.23	97/3.17	291/5.10	840/4.56	369/4.12
	RelFormer	83/2.90	184/3.80	86/3.70	117/4.10	367/3.80	167/3.66
RPR (mv)	RelPoseGNN	48/1.00	114/2.50	48/2.50	152/3.20	320/2.20	136/2.28
	ReLoc3r	47/0.41	87/0.66	18/0.53	41/0.73	171/0.94	73/0.65
	DeMT	9/0.54	8/0.56	5/0.37	6/0.44	11/0.39	8/0.46
NeRF/GS	DFNet+NeFeS50	37/0.54	52/0.88	15/0.53	37/1.14	-	35/0.77
	CROSSFIRE	47/0.7	43/0.70	20/1.2	39/1.4	-	37/1.00
	PNeRFLoc	24/0.29	28/0.37	6/0.27	40/0.55	81/0.25	36/0.35
	NeRFMatch	13/0.23	21/0.38	8/0.40	11/0.35	20/0.09	15/0.29
	STDLoc	15/0.17	12/0.21	**3**/0.13	**5**/0.14	16/**0.06**	10/0.14
	ScalePlace	8/**0.12**	**11/0.19**	**3/0.10**	**5/0.12**	**15/0.06**	**8/0.12**

Table 3: Ablation study of the proposed modules on 7-Scenes and Cambridge Landmarks.

Variant	DAGR	MRFS	MOS	SAPE	7-Scenes	Cambridge
w/o DAGR		✓	✓	✓	1.38/0.41	16/0.26
w/o MRFS	✓		✓	✓	0.74/0.23	10/0.14
w/o MOS	✓	✓		✓	0.71/0.22	9/0.14
w/o SAPE	✓	✓	✓		0.97/0.28	13/0.20
ScalePlace	✓	✓	✓	✓	**0.63/0.19**	**8/0.12**

descriptors, optimized with geometry and appearance for $30{,}000$ Adam iterations. Images have a longest edge of 640 pixels; MRFS uses $1.0, 0.7, 0.5$. MOS targets $16{,}384$ landmarks with 32 neighbors. A scene-specific detector learns from SuperPoint features and projected landmarks, extracting up to $2{,}048$ keypoints with NMS radius of 4. PoseLib initializes pose through robust PnP, followed by four SAPE refinement iterations using rendered features and depth. Localization takes approximately 200 ms per query without retrieval or pose priors. MRFS adds no rendering scales during inference.

4.2 Comparison with state-of-the-art methods

Tables 1 and 2 compare ScalePlace with existing localization methods. On 7-Scenes, ScalePlace achieves the lowest errors across all seven scenes, averaging 0.63 cm/$0.19°$, reducing translation and rotation errors by 17.1% and 20.8% relative to STDLoc. On Cambridge Landmarks, it achieves the lowest reported average errors among NeRF/GS methods, improving from 10 cm/$0.14°$ to 8 cm/$0.12°$. King's College improves from 15 cm/$0.17°$ to 8 cm/$0.12°$. ScalePlace also matches DeMT's reported average translation accuracy while lowering rotation error from $0.46°$ to $0.12°$. These results support accurate localization in both indoor and outdoor environments using scene-specific Gaussian maps, although some outdoor baselines remain stronger on individual scenes and metrics.

Table 4: Multi-resolution feature supervision ablation.

Training scales	Chess		King's College	
	T (cm) ↓	R (°) ↓	T (cm) ↓	R (°) ↓
1.0	0.46	0.15	9.2	0.14
0.7	0.44	0.14	8.8	0.13
0.5	0.48	0.16	9.8	0.15
{1.0, 0.7}	0.42	0.14	8.5	0.13
{1.0, 0.5}	0.43	0.14	8.6	0.13
{0.7, 0.5}	0.44	0.14	8.9	0.13
{1.0, 0.7, 0.5}	**0.40**	**0.13**	**8.0**	**0.12**

Table 5: Ablation on Matching-Oriented Sampling and Detector.

Stage	Err.↓ [cm/°]	5m 10°↑
Random sampling	12.5/0.18	92.1%
Farthest point sampling	10.8/0.16	94.5%
MOS (k=16)	9.2/0.14	97.8%
MOS (k=32)	**8.0/0.12**	**99.4%**
w/o detector	11.5/0.16	93.2%
Generic detector	9.5/0.14	96.8%
Scene-specific detector	**8.0/0.12**	**99.4%**

4.3 Ablation Study

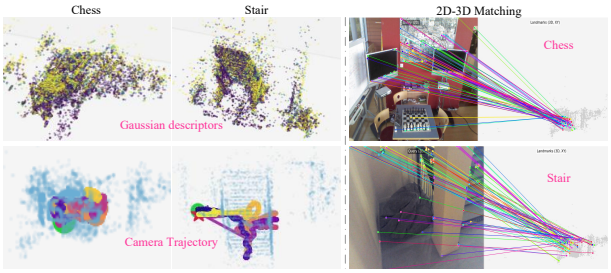


Figure 3: Visualized results on Chess and Stairs.

Table 3 shows that removing DAGR causes the largest degradation, reaching 1.38 cm/0.41° on 7-Scenes and 16 cm/0.26° on Cambridge. Omitting SAPE also substantially increases errors, while disabling MRFS or replacing MOS with random sampling causes smaller declines. The complete configuration performs best, supporting the combined representation, supervision, sampling, and refinement design.

Table 4 shows that MRFS with $1.0, 0.7, 0.5$ achieves the lowest errors on Chess and King's College: 0.40 cm/$0.13°$ and 8.0 cm/$0.12°$, respectively. Translation errors decrease by approximately 13% relative to single-scale training on both scenes. The results favor three-scale supervision over the evaluated alternatives without requiring additional rendering scales during test-time pose estimation.

Table 5 shows that MOS with $k = 32$ improves King's College localization from random sampling's 12.5 cm error and 92.1% recall to 8.0 cm/$0.12°$ and 99.4%. Scene-specific detection reduces translation error from 11.5 cm without the detector and 9.5 cm with a generic detector to 8.0 cm, supporting both sampling and detection choices within the Gaussian feature space.

4.4 Visualization

Figure 3 visualizes Gaussian descriptors, camera trajectories, and 2D-3D correspondences on Chess and Stairs. PCA colors reveal spatially coherent feature patterns, while trajectories show viewpoint coverage relative to sampled landmarks. Correspondence lines link query keypoints to matched landmarks around textured regions, illustrating feature associations within the shared Gaussian representation across viewpoints.

5 Conclusion

We presented ScalePlace, a retrieval-free framework that unifies sparse matching and dense refinement within a descriptor-attached Gaussian field. MRFS supervises shared primitives at multiple resolutions, while MOS selects landmarks using descriptor agreement and spatial neighborhoods. SAPE initializes pose from 2D-3D correspondences and refines it through coarse-to-fine matching against rendered features and depth. The pipeline requires no external pose prior, and additional training scales introduce no extra inference renders. On 7-Scenes and Cambridge Landmarks, reported average errors improve over SOTA methods. Ablations favor the complete configuration, supporting multi-resolution supervision, matching-oriented sampling, and iterative refinement. These results demonstrate the effectiveness of the shared Gaussian representation for accurate localization in mapped indoor and outdoor scenes.

REFERENCES

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

Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *European conference on computer vision*, pp. 782–799. Springer, 2018.

Eric Brachmann and Carsten Rother. Visual camera re-localization from rgb and rgb-d images using dsac. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5847–5865, 2021.

Shuai Chen, Xinghui Li, Zirui Wang, and Victor A Prisacariu. Dfnet: Enhance absolute pose regression with direct feature matching. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2022.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20987–20996. IEEE, 2024.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 337–33712. IEEE, 2018.

Ivano Donadi and Alberto Pretto. Kvn: keypoints voting network with differentiable ransac for stereo pose estimation. *IEEE Robotics and Automation Letters*, 9(4):3498–3505, 2024.

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

Khang Truong Giang, Soohwan Song, and Sungho Jo. Learning to produce semi-dense correspondences for visual localization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19468–19478. IEEE, 2024.

Zhiwei Huang, Hailin Yu, Yichun Shentu, Jin Yuan, and Guofeng Zhang. From sparse to dense: Camera relocalization with scene-specific detector from feature gaussian splatting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27059–27069. IEEE, 2025.

Ofer Idan, Yoli Shavit, and Yosi Keller. Beyond familiar landscapes: Exploring the limits of relative pose regressors in new environments. *Computer Vision and Image Understanding*, pp. 104629, 2026.

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

Viktor Larsson, Torsten Sattler, Zuzana Kukelova, and Marc Pollefeys. Revisiting radial distortion absolute pose. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1062–1071, 2019.

Zhenyu Li and Tianyi Shang. A2gc: Asymmetric aggregation with geometric constraints for locally aggregated descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19423–19431, 2026.

Zhenyu Li, Tianyi Shang, Pengjie Xu, and Zhaojun Deng. Place recognition meet multiple modalities: a comprehensive review, current challenges and future development. *Artificial Intelligence Review*, 58(11):363, 2025.

Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *Conference on robot learning*, pp. 1347–1356. PMLR, 2022.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 252–262. IEEE, 2023.

Soham Saha, Girish Varma, and CV Jawahar. Improved visual relocalization by discovering anchor points. *arXiv preprint arXiv:1811.04370*, 2018.

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

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

Yoli Shavit and Yosi Keller. Camera pose auto-encoders for improving pose regression. In *European Conference on Computer Vision*, pp. 140–157. Springer, 2022.

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*, pp. 2930–2937, 2013.

Shitao Tang, Sicong Tang, Andrea Tagliasacchi, Ping Tan, and Yasutaka Furukawa. Neumap: Neural coordinate mapping by auto-transdecoder for camera localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 929–939. IEEE, 2023.

Wenshuai Wang, Hong Liu, Shengquan Li, Peifeng Jiang, and Runwei Ding. Debiased multiplex tokenizer for efficient map-free visual relocalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10145–10153, 2026.

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)*, pp. 1323–1330. IEEE, 2021.

Boming Zhao, Luwei Yang, Mao Mao, Hujun Bao, and Zhaopeng Cui. Pnerfloc: Visual localization with point-based neural radiance fields. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7450–7459, 2024.

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

Qunjie Zhou, Maxim Maximov, Or Litany, and Laura Leal-Taixé. The nerfect match: Exploring nerf features for visual localization. In *European Conference on Computer Vision*, pp. 108–127. Springer, 2024.