This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CoLoR: The Devil is in Scene Coordinate Regression for Large-Scale Visual Localization

Xindong Mao[1] Hang Li[1,3] Yuchen Wu[1] Jiahe Li[1] Xiao Bai[1,2†] Jin Zheng[1,2,3†]
[1]School of Computer Science and Engineering, Beihang University
[2]State Key Laboratory of Complex & Critical
Software Environment, Jiangxi Research Institute, Beihang University
[3]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University



Figure 1. **Comparison with the state of the art.** *Left*: Qualitative Comparison of Point Cloud Reconstruction on Aachen Day-Night [36, 38] between R-SCoRe [21] and CoLoR. *Right*: Quantitative comparison of accuracy and map size on the Dept. 4F test set [22]. Our method CoLoR significantly outperforms Scene Coordinate Regression (SCR) based methods. Furthermore, it achieves an accuracy only marginally lower than Feature Matching (FM) based approaches, while requiring a map size several orders of magnitude smaller.

Abstract

Scene Coordinate Regression (SCR) has emerged as a memory-efficient paradigm for visual localization. While SCR has demonstrated performance comparable to classic feature matching based approaches in small-scale scenes, it has consistently underperformed in large-scale environments. Large-scale localization is hampered by two challenges: sparse co-visibility and local appearance ambiguity. In this work, we propose ***CoLoR****, a novel training framework tailored for large-scale SCR. First, we explicitly and efficiently partition scene points into multi-view and single-view sets and introduce a two-stage bootstrapping paradigm to provide complete and strong supervision for all points. Second, we propose a multi-granularity retrieval feature, which unifies the conventional global and local features as retrieval-oriented representations at the image and pixel levels, respectively, to enforce feature consistency. Our method achieves state-of-the-art performance on multiple challenging large-scale datasets and significantly narrows the accuracy gap with classical feature matching based approaches while retaining a compact map size.*

1. Introduction

Visual localization refers to determining the 6-DoF camera pose of a query image within a known environment. As a core problem in computer vision, it serves as a key component in a variety of applications, including self-driving vehicles, and Augmented and Virtual Reality systems.

Classical feature matching (FM) approaches [12, 34, 35, 37] have been extensively developed over the years and now achieve reliable and precise localization. Nevertheless, these methods generally rely on explicit 3D scene representations that store a large number of descriptors, leading to significant memory requirements. By contrast, scene coordinate regression (SCR) [3–5, 8, 9, 21, 26, 39, 43, 44] seeks to encode scene information implicitly within the weights of a neural network. SCR is a structure-based method

† Corresponding author.

that directly regresses 3D scene coordinates from 2D image pixels, whose key innovation is learning to encode the scene geometry within a neural network. Although SCR has demonstrated performance comparable to FM in small-scale environments, a significant accuracy gap still persists when scaling to large-scale scenes.

The primary challenges in large-scale scenes are sparse co-visibility and local appearance ambiguity [21, 44]. Sparse co-visibility can be understood by re-examining the scene's composition through the frequency of 3D point co-visibility. From this angle, every point in the scene can be classified as either a single-view point (seen once) or a multi-view point (seen more than once). In large-scale environments, the scene becomes overwhelmingly dominated by single-view points, and the co-observation frequency for multi-view points diminishes. This shift in scene composition reveals a critical flaw in prior methods, which both are grounded solely in multi-view geometry and utilize a form of weak supervision whose efficacy degrades sharply as co-observation becomes sparse. Local appearance ambiguity describes the challenge where visually similar, or even identical, local image patches originate from physically distinct 3D locations. This issue is particularly prevalent in large-scale environments, which naturally contain a high degree of repetitive textures and structural elements.

To address these challenges, this paper proposes CoLoR. CoLoR begins by efficiently partitioning the 3D points in the scene into multi-view and single-view categories, which are then trained separately. For multi-view points, the inherent correspondences from the partitioning process can directly generate strong multi-view reprojection supervision. For single-view points, however, multi-view geometry alone cannot provide supervision signals, and conventional monocular depth, lacking metric scale, fails to offer strong supervision. Therefore, we bootstrap the supervision for single-view points by introducing a rapidly maturing depth completion model [47] as an additional prior. This is achieved by leveraging the sparse depth obtained from the SCR model's training on multi-view points, which is then used by the completion model to generate a dense pseudo-depth map. This pseudo-depth in turn provides the supervisory signal for the single-view points. By applying this meticulously designed strong supervision to both point categories, CoLoR is capable of imposing complete supervision on all points within the scene.

Secondly, prior large-scale SCR methods [21, 44], in their effort to encode an entire scene with a single model and thus avoid partitioning it for ensembling, typically concatenate retrieval-oriented global features with matching-oriented local features. This approach is intended to remedy the local feature's inability to distinguish between similar patches in different regions. However, our analysis reveals a fundamental misalignment in the inherent properties of these two feature types, which compromises their final discriminative capability. We propose a new perspective, re-framing them not as a mix of retrieval and matching features, but as a unified set of retrieval features at different scales: image-level and pixel-level. Motivated by this, we retrained the local feature, resolving the misalignment and achieving a truly consistent representation across scales.

Our contributions are summarized as follows:

- We introduce a two-stage training paradigm that, for the first time, provides supervision for the entire scene. Our method achieves this by categorizing points as either multi-view or single-view and applying a tailored supervisory signal to each, thereby transcending the reliance on implicit triangulation in prior works.
- We identify and address an inherent inconsistency between the global and local features used in mainstream large-scale SCR methods. In response, we propose a novel pixel-level retrieval feature that effectively aligns these two feature for a consistent representation.
- Our proposed method achieves state-of-the-art performance on multiple challenging large-scale datasets, spanning both indoor and outdoor environments. Notably, our results significantly narrow the long-standing accuracy gap between SCR and classical Feature Matching (FM) approaches, with extensive experiments validating the overall effectiveness of our framework.

2. Related Work

Feature Matching. Feature matching plays a fundamental role in many vision tasks such as structure-from-motion, localization, and image registration. Traditional methods rely on hand-crafted descriptors [27] which detect keypoints and match them based on descriptor similarity. While these approaches are robust to moderate viewpoint and illumination changes, their performance degrades significantly under large appearance variations or low-texture regions.

To overcome these limitations, learning-based methods [15, 32] have been developed to learn discriminative feature representations and matching strategies. Early works [12] introduced end-to-end trainable frameworks for keypoint detection and description. More recent approaches [16, 40, 42] have begun to jointly learn dense or semi-dense local features with improved robustness and generalization. However, these methods are primarily trained on image pairs of the same region, without extensive exposure to pairs from different regions, which reduces the model's ability to distinguish between similar local patches in different scenes.

Feature Matching Based Localization. Feature matching-based methods [34, 35, 35] are widely regarded as the state-of-the-art for high-accuracy visual localization. These methods typically operate on an explicit 3D map of the scene, often reconstructed using Structure-from-Motion

(SfM), which stores 3D points along with their corresponding high-dimensional visual descriptors. At test time, localization is achieved by matching features from a query image to the stored 3D descriptors to establish 2D-3D correspondences, followed by a PnP-and-RANSAC [1, 11, 17, 18, 31] step to solve for the camera pose. To scale to large environments, a hierarchical strategy is employed: image retrieval first narrows down a small set of candidate database images, before detailed local feature matching is performed.

However, the primary drawback of this paradigm is its significant memory footprint, stemming from the need to store a vast number of descriptors for the entire 3D model. This storage bottleneck becomes a critical challenge in large-scale scenes. While solutions like descriptor compression [6, 7, 14] or recent descriptor-free methods [30, 45, 49] that match directly against geometry have been explored, the fundamental reliance on an explicit and large map representation remains. This limitation motivates the exploration of alternative, more compact scene representations, such as those used in Scene Coordinate Regression.

Scene Coordinate Regression. Scene Coordinate Regression (SCR) methods [3–5, 8, 9, 21, 26, 39, 43, 44] also follow a structure-based paradigm but offer a significant advantage in memory efficiency. Instead of relying on an explicit 3D map, SCR methods implicitly encode the scene within the weights of a neural network, which learns to directly regress the 3D coordinate for any given 2D pixel.

Broadly, SCR methods can be categorized by their supervision source. The most straightforward approach [5, 26, 29] is to train the network with ground truth 3D coordinates, often obtained from SfM reconstructions or depth sensors. To handle large-scale scenes, these methods [23, 41, 46] resort to a divide-and-conquer strategy, partitioning the scene into smaller segments and training an expert model for each.

Alternatively, to avoid the need for 3D models, another line of research [5, 21, 26, 44] trains SCR models using only images along with their intrinsic and extrinsic parameters. These methods rely on minimizing a reprojection loss, forcing the network to learn scene geometry through implicit triangulation [26, 44]. To scale up, these approaches typically augment a local feature encoder with a global feature vector [21, 44] to provide the necessary scene context, allowing a single model to represent the entire scene.

However, this dominant paradigm for large-scale, 3D model-free SCR faces two fundamental, unaddressed challenges. First, its reliance on implicit triangulation means it can only provide weak supervision for multi-view points and cannot provide supervision for single-view points, a critical bottleneck in sparsely-observed environments. Second, there is a fundamental inconsistency between the properties of local patch features and global image descriptors, compromising the model's discriminative power.

3. Method

3.1. Preliminaries

Problem Definition. Given a scene represented by a set of training images I_{train} with camera poses T_{train} and intrinsics K_{train}, our goal is to estimate the 6-DoF camera pose T_{test} for a test image I_{test}. Following the dominant paradigm for training without explicit 3D models, we do not assume access to any ground truth 3D scene coordinates. The core of our task is to train a neural network F parameterized by θ, that maps the feature x at a 2D pixel coordinate in an image to its corresponding 3D scene coordinate y:

$$y = f(F(x)) \tag{1}$$

The final pose is then recovered from a set of 2D-3D correspondences using a PnP-and-RANSAC solver.

ACE Training Framework in Large-scale Scenes. The ACE [5] framework introduced an efficient training paradigm for SCR, referred to as Gradient Decorrelation Training. The process begins with a pre-computation step: pixels are randomly sampled from the training images I_{train} and encoded using a scene-agnostic encoder. These encoded local features are then stored in a large training buffer. During training, batches are constructed by randomly sampling a massive number of features from this buffer, ensuring each batch covers a diverse set of observations from the entire scene. In the absence of ground truth 3D coordinates, the network is supervised by leveraging the known camera pose and intrinsics to minimize a reprojection error, a process that relies on implicit triangulation.

$$e_{repro} = \| \, x' - \pi(K_{train}(R_{train})y + t_{train}) \, \|_2 \tag{2}$$

where x' is the ground truth 2D pixel coordinate, y is the model-predicted 3D scene coordinate, and π is the projection function that converts homogeneous coordinates to 2D pixel coordinates. The rotation matrix R_{train} and translation vector t_{train} are extracted from the camera pose T_{train}.

3.2. Two-Stage Training Framework

The prevailing training paradigm for SCR, inherited from ACE [5], is predicated on random sampling and the principle of implicit triangulation [26, 44]. To dissect the core issues of this paradigm, we analyze it from a new perspective: by categorizing scene points as either multi-view or single-view. This analysis reveals two fundamental flaws: 1) The triangulation bottleneck: The framework's reliance on multi-view consistency means that only co-visible points are trainable. Single-view points, despite being sampled, receive no effective supervision, leading to a significant waste of computational resources and model capacity. This problem becomes critical in large-scale environments, which are inherently dominated by sparsely-observed regions; as detailed in the supplementary material, the proportion of single-view points rises rapidly as

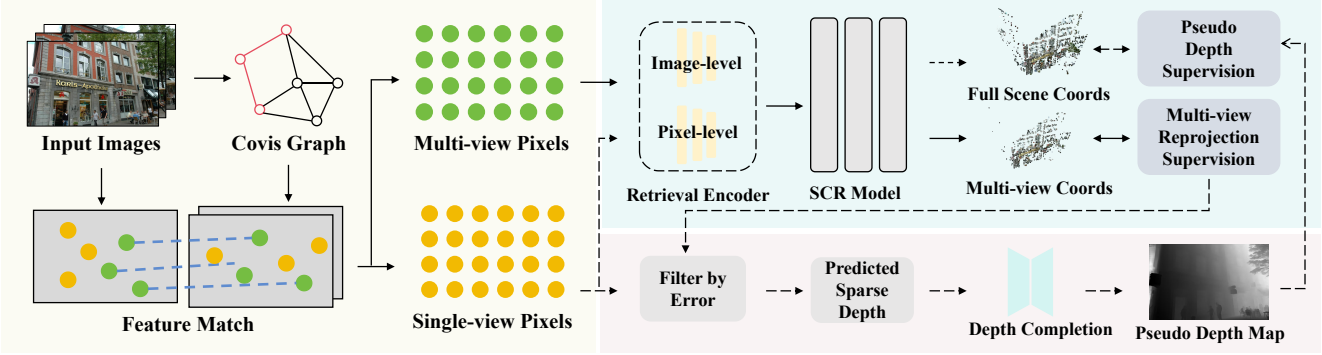


Figure 2. **Overview of CoLoR.** Our framework introduces a complete two-stage training paradigm and a consistent retrieval-based feature encoder. Solid lines denote stage one, while dashed lines represent stage two. We first partition pixels into multi-view and single-view sets, which are trained in stage one and stage two, respectively. Concurrently, our retrieval-oriented training for the local encoder ensures its consistency with the global encoder.

the scene scale expands. 2) The weakness of implicit constraints: The supervision for multi-view points is merely implicit; its strength is highly dependent on the frequency of co-observation. In large, sparsely co-visible scenes, this implicit signal becomes unreliable and weak. Motivated by this careful and detailed diagnosis, our approach abandons this flawed paradigm. As shown in Fig. 2, we explicitly partition the points and introduce one strong, explicit supervision sources for each category, thereby bypassing the reliance on implicit triangulation.

Efficient Partitioning and Multi-view Training. To distinguish between multi-view and single-view points, a feature matching pre-processing step is necessary. However, performing exhaustive pairwise matching across all images in the set $\mathcal{I} = \{I_1, \dots, I_N\}$ is computationally prohibitive. To circumvent this, we leverage the co-visibility score $S(I_i, I_j)$ [21], which is based on view frustum IoU. For each image I_i, we perform feature matching only against its set of top-K most co-visible neighbors, $\mathcal{N}_K(I_i)$:

$$\mathcal{N}_K(I_i) = \underset{I_j \in \mathcal{I} \setminus \{I_i\}}{\arg\operatorname{topk}} S(I_i, I_j). \tag{3}$$

Incorrect correspondences are filtered out by evaluating the epipolar error using the intrinsic and extrinsic parameters of the two views. Additionally, we extend the matching relations through transitive consistency to enrich the set of correspondences. Images producing an insufficient number of samples are skipped. This targeted matching strategy reduces the number of image pairs to be considered from $O(N^2)$ to $O(N)$, making the pre-computation efficient. Taking the Aachen Day-Night dataset [36, 38] as an example, this step completes in just three minutes on the same hardware as our baseline [21] (4x RTX 4090s). This constitutes about 1% of the four-hour training time, rendering the overhead entirely negligible.

The pre-computation step yields a set of valid 2D point correspondences, denoted as $\mathcal{C}$. This allows for a definitive partition of all 3D scene points, $\mathcal{P}$, into multi-view ($\mathcal{P}_{\text{mv}}$) and single-view ($\mathcal{P}_{\text{sv}}$) subsets. We define $\text{corr}(P_k) \subseteq \mathcal{C}$ as the subset of correspondences originating from a specific 3D point P_k. A point is classified as multi-view if this subset is non-empty; otherwise, it is a single-view point:

$$\mathcal{P}_{\text{mv}} = \{P_k \in \mathcal{P} \mid \text{corr}(P_k) \neq \emptyset\}, \tag{4}$$

$$\mathcal{P}_{\text{sv}} = \mathcal{P} \setminus \mathcal{P}_{\text{mv}}. \tag{5}$$

A consequence of our categorization process is that for every multi-view point $P_k \in \mathcal{P}_{\text{mv}}$, we inherently obtain the multi-view correspondences required for supervision. We formulate this as a multi-view reprojection loss, $\mathcal{L}_{\text{mv}}$:

$$\mathcal{L}_{\text{mv}}(P_k) = \sum_{p_j \in \mathcal{O}(P_k)} \|\pi_j(P_k) - p_j\|_2^2. \tag{6}$$

Here, $\mathcal{O}(P_k)$ is the set of all unique 2D observations for the 3D point P_k, derived from its correspondence set $\text{corr}(P_k)$. The term p_j is one such observed 2D location in image I_j, while $\pi_j(P_k)$ is the predicted 2D location obtained by projecting the estimated scene coordinates of P_k onto image I_j.

This formulation applies equal weighting to the reprojection losses from all visible views, thereby forcing the model to learn a 3D point that is geometrically consistent across multiple views. This cross-view consistency effectively creates a strong coupling between the lateral (xy) and depth (z) coordinates, as an error in depth (z) would inevitably lead to reprojection errors (xy) in at least one of the views—a crucial property that cannot be guaranteed by single-view reprojection supervision, and one that further lays the groundwork for our second training stage.

From a comparative standpoint, our stage-one training set is conceptually equivalent to the points effectively trained within the ACE paradigm. Our ablation studies

provide a key insight: simply isolating the training to this multi-view set (while using a standard single-view reprojection supervision) is sufficient to significantly outperform the baseline which adopted the ACE framework. By further substituting the weak implicit supervision with our strong explicit multi-view objective, we dismantle the "implicit" bottleneck and unlock an additional tier of performance.

Pseudo Depth Generation and Single-view Training. Having established strong supervision for multi-view points in stage one, our framework in stage two expands its focus to also incorporate the challenging set of single-view points. Triangulation-based supervision is fundamentally inapplicable to these points, necessitating the use of an external geometric prior. While monocular depth estimation is a common choice, it has a well-known intrinsic limitation: the lack of a true metric scale. This inherent scale ambiguity restricts its supervisory potential. For instance, one might resort to a relative objective like an ordinal depth loss [25], which constitutes another form of weak and implicit supervision, similar to the single-view reprojection losses in the original ACE paradigm. Fortunately, the strong xy-z coupling established in stage one provides a reliable indicator of depth accuracy. This allows us to leverage a depth completion prior, using the high-confidence depth predictions to bootstrap the training of all points in the scene.

Specifically, we first use the latest SCR model trained in stage one to perform inference on all training images. This produces a scene coordinate map M_i and, consequently, a depth map D_i and a reprojection error map E_i for each image I_i. We then generate a sparse depth map, $D_{\text{sparse},i}$, by filtering for high-confidence points. Concretely, we consider points with a reprojection error below a threshold τ (5 pixels in our experiments) as reliable, yielding a sparse depth map defined for each pixel u as:

$$D_{\text{sparse},i}(u) = \begin{cases} D_i(u) & \text{if } E_i(u) < \tau \\ \text{invalid} & \text{otherwise} \end{cases} \tag{7}$$

where $E_i(u)$ is the reprojection error at pixel u. Finally, we employ a depth completion model [47] to densify these sparse depth maps. This process yields a complete, metric, yet potentially noisy depth prior, which we hereafter refer to as the "pseudo depth map" for each image. This pseudo depth map $\tilde{d}$ can provide pseudo depth supervision for the depth component d of the predicted 3D coordinates.

$$l_{depth} = \| \tilde{d} - d \|_2 \tag{8}$$

Although our framework can now theoretically provide strong supervision for any point in the scene, in practice, not all points are beneficial for localization. We can leverage the error maps generated in our pipeline to effectively filter these nuisance points. To be precise, we compute a depth error map by comparing the depth map predicted by SCR model against the pseudo depth map. By jointly considering the reprojection error map and this depth error map, we can identify and discard low-quality points. For the second training stage, in addition to all multi-view points from stage one, we sample a subset of single-view points that exhibit the lowest combined reprojection and depth errors. The supervision in stage two is also expanded: alongside the multi-view reprojection supervision, we introduce a pseudo depth supervision. This new objective enforces that the model's predicted depth for a sampled point must be close to its corresponding pseudo depth value within a certain tolerance.

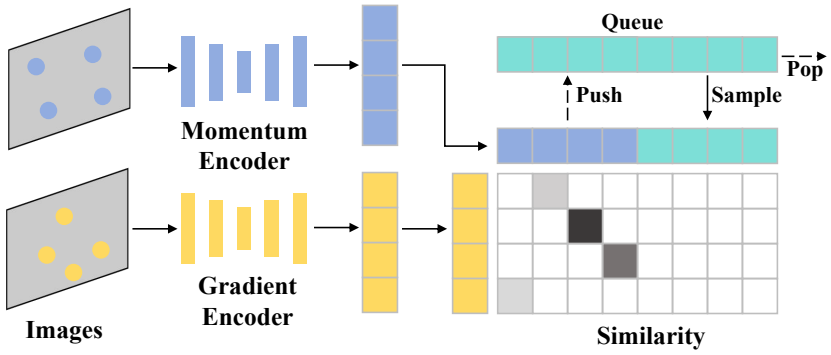
3.3. Consistent Retrieval Feature

The ideal feature for SCR must be able to distinguish a single 3D point from all others in the scene, which essentially defines it as a retrieval feature for 3D points within a scene.

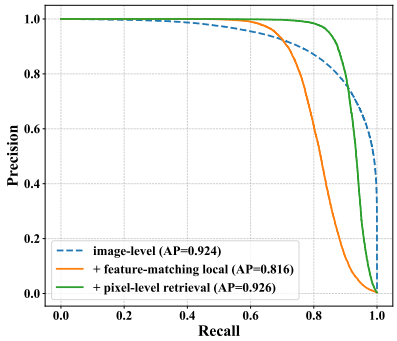
In previous works [21, 44], the input features are composed of two concatenated parts: a global feature obtained via image retrieval methods [19, 50] serving as an image-level descriptor, and a local feature derived from feature matching techniques [5, 15, 40] providing a pixel-level representation. As visualized in our Precision-Recall curves in Fig. 3b, this composition leads to a paradox. Global features alone are highly effective for their intended task of image co-visibility classification, achieving an Average Precision (AP) of 0.924. However, when these are concatenated with traditional local features, the granularity reaches the pixel level, but the ability to discriminate corresponding 3D points deteriorates significantly (AP = 0.816). We attribute this performance degradation to a fundamental inconsistency between the intrinsic properties of these two feature types: the global feature is optimized for large-scale retrieval, while the traditional local feature is optimized for local-region matching.

To resolve this dilemma, we propose a new perspective: instead of considering two disparate feature types, we reframe the input as a unified, multi-granularity retrieval representation. Under this view, the global feature acts as an image-level retrieval feature, and the local feature should be treated as a pixel-level retrieval feature. As illustrated in Fig. 3a, we resolve this inconsistency by re-training the local encoder with a retrieval-oriented[15] objective. As shown in Fig. 3b, this alignment of objectives restores and even surpasses the original performance, with the combined feature achieving an AP of 0.926.

Training Pixel-level Retrieval Encoders. We formulate the training of traditional feature matching descriptors as a form of contrastive learning. In each iteration, we sample a pair of images, I^A and I^B, that capture the same scene or region. Each image is associated with a set of keypoints, denoted as $K^A = \{K_i^A\}_{i=1}^N$ and $K^B = \{K_i^B\}_{i=1}^N$, respectively. For a given anchor keypoint $K_i^A \in K^A$, we



(a) Overview of pixel-level retrieval encoder training.



(b) Precision-Recall curves evaluation.

Figure 3. **Overview and Visualization of the Pixel-Level Retrieval Encoder.** (a) Our local features are trained for pixel-level retrieval using a MoCo-style contrastive learning strategy. The model learns to identify the correct positive feature for an anchor from a large queue of negative samples drawn from different scenes. (b) PR curves for different feature combinations on the Aachen Day-Night dataset.

extract its corresponding descriptor, which we refer to as the anchor feature F_i^A. The descriptor corresponding to the ground-truth match in the other image, $K_i^B \in K^B$, is considered the positive feature, F_i^B. Let $\mathcal{F}^B = \{F_j^B\}_{j=1}^N$ denote the set of all features within image I^B. We define a temperature-scaled cosine similarity function as $s(u, v) = u \cdot v/\tau$. The model's objective is to maximize the similarity between the anchor and its positive counterpart, while minimizing its similarity to all other features in $\mathcal{F}^B$. The standard contrastive loss is thus:

$$\mathcal{L}i = -\log \frac{\exp(s(F_i^A, F_i^B))}{\sum F_j \in \mathcal{F}^B \exp(s(F_i^A, F_j))} \tag{9}$$

To expose the model to more diverse negatives, we sample $\mathcal{M}$ images from other regions, each containing $\mathcal{N}$ keypoints. Let their corresponding features form an external negative set $\mathcal{F}_{\text{neg}} = \{F_k^{\text{neg}}\}_{k=1}^{\mathcal{M}\times\mathcal{N}}$. We then define an augmented set of candidates for the denominator:

$$\mathcal{F}\text{aug} = \mathcal{F}^B \cup \mathcal{F}\text{neg} \tag{10}$$

Our final loss is calculated using Equation 9, but with the summation in the denominator performed over this augmented set $\mathcal{F}_{\text{aug}}$.

Incorporation with MoCo. However, directly training with all negatives from different regions leads to a significant computational cost. Since negatives from different regions can be shared across batches, we adopt the MoCo [20] training strategy which decouples the number of negative samples from the batch size. This allows the model to see more negatives while reducing computational complexity.

Specifically, we maintain two encoders of the same architecture but different parameters: a standard encoder $encoder_g$ and a momentum encoder $encoder_m$. Both are initialized with the same parameters. During training, the standard encoder updates its parameters via backpropagation, while the momentum encoder updates its parameters using an exponential moving average (EMA) of the standard encoder's parameters. We denote the parameters of the standard encoder as θ_g and those of the momentum encoder as θ_m, which are updated as follows:

$$\begin{aligned} \theta_g &\leftarrow \theta_g - \eta \nabla L(\theta_g), \\ \theta_m &\leftarrow \lambda\theta_m + (1-\lambda)\theta_g \end{aligned} \tag{11}$$

We maintain a standard MoCo[20] queue that stores descriptors encoded by the momentum encoder $encoder_m$ along with their corresponding scene indices. Initially, the queue is populated by randomly sampling images and encoding their descriptors using the initial $encoder_m$. At the beginning, descriptors from scenes different from the current one are collected from the queue and used as additional negatives. After each iteration, the queue is updated with the newly encoded descriptors and their scene indices.

In practice, our training samples 2 co-visible and 64 negative images per iteration. Compared to a naive implementation that forwards all 66 images but only backpropagates gradients for 2, our approach only forwards the 2 co-visible images, reducing the computational cost to 1/33.

4. Experiments

4.1. Datasets

We validate our method on the Aachen Day-Night dataset [36, 38] and the Hyundai Department Store dataset [22].

Aachen Day-Night. The Aachen Day-Night dataset is a city-scale benchmark collected with mobile devices on the streets of Aachen, Germany. It covers an area of approximately 6 km² and contains 4,328 training images along with 922 test images captured under both daytime and nighttime conditions. Owing to its scale and diversity, it serves as an ideal large-scale outdoor visual localization dataset.

Hyundai Department Store. The Hyundai Department Store dataset was collected in a large shopping mall in Ko-

	Methods	w/o Depth	Map Size	Aachen Day			Aachen Night		
FM	HLoc+SPSG [12, 34, 35]	Yes	7.82GB	89.6	95.4	98.8	86.7	93.9	100
	AS [37]	Yes	750MB	85.3	92.2	97.9	39.8	49.0	64.3
	Cascaded [10]	Yes	140MB	76.7	88.6	95.8	33.7	48.0	62.2
	QP+R.Sift [28]	Yes	30MB	62.6	76.3	84.7	16.3	18.4	24.5
	Squeezer [48]	Yes	240MB	75.5	89.7	96.2	50.0	67.3	78.6
SCR	Neumap [41]	No	1.26GB	80.8	90.9	95.6	48.0	67.3	87.8
	HSCNet [23]	No	213MB	71.1	81.9	91.7	32.7	43.9	65.3
	HSCNet++ [46]	No	274MB	72.7	81.6	91.4	43.9	57.1	76.5
	ESAC (×50) [2]	No	1.31GB	42.6	59.6	75.5	6.1	10.2	18.4
	ACE (×50) [5]	No	205MB	6.9	17.2	50.0	0.0	1.0	5.1
	GLACE [44]	Yes	27MB	8.6	20.8	64.0	1.0	1.0	17.3
	R-SCoRe [21]	Yes	47MB	74.8	86.9	96.4	64.3	89.8	96.9
	CoLoR (ours)	Yes	47MB	78.6	89.4	96.8	67.3	88.8	96.9

Table 1. **Comparisons on Aachen Day-Night.** We compare feature matching (FM) methods and scene coordinate regression (SCR) methods in terms of map size and the percentages of test images within three thresholds: $(0.25\,\text{m}, 2^\circ)$, $(0.5\,\text{m}, 5^\circ)$, and $(5\,\text{m}, 10^\circ)$. Although FM approaches achieve the highest accuracy, they require substantially larger map sizes. Our CoLoR maintains the compact map size characteristic of SCR methods while further narrowing the accuracy gap with FM approaches.

Methods	Dept. 1F Test				Dept. 4F Test				Dept. B1 Test			
HLoc+D2-Net [13, 34]	398GB	78.0	82.8	88.0	183GB	84.2	89.8	92.0	505GB	73.7	79.3	87.2
HLoc+R2D2 [33, 34]	166GB	80.6	84.3	89.4	76GB	85.3	91.0	93.1	210GB	75.2	80.3	87.6
ESAC (×50) [2]	1.4GB	43.3	66.3	77.0	1.4GB	45.2	62.5	73.1	1.4GB	3.5	8.2	12.6
ACE (×50) [5]	205MB	14.1	54.4	75.5	205MB	27.3	70.9	84.1	205MB	2.7	14.4	29.3
GLACE [44]	42MB	5.6	21.3	48.6	42MB	8.4	29.8	51.6	42MB	0.9	4.4	11.9
R-SCoRe [21]	127MB	61.4	80.2	90.9	50MB	60.2	79.3	87.9	130MB	60.1	77.3	89.6
CoLoR (ours)	127MB	75.2	85.2	92.0	50MB	80.9	89.6	93.1	130MB	66.6	80.9	91.3

Table 2. **Comparison on the Hyundai Department Store Test Set.** We compare feature matching (FM) methods and scene coordinate regression (SCR) methods in terms of map size and the percentages of test images within three thresholds: $(0.1\,\text{m}, 1^\circ)$, $(0.25\,\text{m}, 2^\circ)$, and $(1\,\text{m}, 5^\circ)$. Our CoLoR significantly outperforms previous methods while maintaining the compact map size of SCR approaches, and achieves comparable performance to FM methods on the 1F floor.

rea, covering an area of approximately 10,000 m². It comprises three floors—B1, 1F, and 4F—which are treated as three sub-datasets, each with its own training and testing splits. In total, the dataset contains 44,283 training images and 5,927 test images. It features numerous dynamic objects and textureless regions, making visual localization particularly challenging. Notably, the B1 sub-dataset exhibits significant illumination variation: its training images were captured under dim lighting conditions, while its test images were captured under bright lighting conditions. These characteristics make it a challenging and representative large-scale indoor visual localization dataset.

4.2. Implementation Details

Two-Stage Training Framework. We follow R-SCoRe [21] to construct covisible pairs by selecting, for each image, its top-10 neighbors with the highest covisibility scores. Feature matching is then performed, and matches with an epipolar error greater than 1 are filtered out using the ground-truth poses. While we extend these matching relations through transitive consistency to modestly enrich the set of correspondences, this risks introducing undesirable cycles, which we mitigate by applying Non-Maximum Suppression (NMS) with a threshold of 1 to the covisibility scores. For a fair comparison, we follow the training configuration of our baseline, R-SCoRe. Specifically, the total number of training steps is set to 100,000, and the batch size is 81,920 per GPU. Our two-stage process is divided accordingly: the first stage runs for 70,000 steps, followed by a second stage of 30,000 steps.

We adapt the R-SCoRe architecture into a deeper, narrower variant while maintaining a comparable map size for

Model settings	Dept.	4F	Test
baseline	60.2	79.3	87.9
+pixel partition	70.2	83.9	88.8
+multi-view reprojection supervision	75.8	86.7	90.1
+pseudo depth supervision	76.6	86.7	90.3
+pixel-level retrieval feature (ours)	80.9	89.6	93.1

Table 3. **Ablation study of our key components.** We start from a baseline and progressively add our contributions: (1) Pixel Partitioning, (2) Multi-view Reprojection Supervision, (3) Pseudo-Depth Supervision for single-view points, and (4) Feature Consistency. The consistent performance improvement validates the effectiveness of each component in our final CoLoR model.

a fair comparison. Further details are in the Appendix.

Consistent Retrieval Feature. We adopt the same model architecture as DeDoDe [15], including its keypoint detector and descriptor, and train it on the same dataset MegaDepth [24]. Since our focus is solely on local feature learning, we freeze the detector and train the descriptor from scratch. During training, 5,000 keypoints and their corresponding descriptors are extracted from each image. The queue stores features and scene identifiers from the past 64 images, resulting in a total of 320,000 features. To balance GPU memory consumption and mitigate numerical issues of softmax when the number of classes is excessively large, we select only the top 2,000 negative features with the highest correlation to the anchor feature. We keep the batch size equivalent to DeDoDe by setting the batch size to 4 with gradient accumulation over 2 iterations, and use the same learning rate as DeDoDe.

4.3. Benchmark Results

Aachen Day-Night. The Aachen Day-Night benchmark evaluates the accuracy of the localization as the percentage of test images localized within three thresholds: (0.25 m, 2°), (0.5 m, 5°) and (5 m, 10°). As shown in Tab. 1, our approach significantly advances SCR-based methods by establishing a new state-of-the-art in efficiency, achieving highly competitive performance with a remarkably small memory footprint. During the day, our method achieves performance highly comparable to Neumap, while demanding only a fraction of the storage footprint—a mere 47MB compared to Neumap's 1.26GB. At night, our method demonstrates exceptional robustness, outperforming all rivals with the sole exception of the heavyweight HLOC+SPSG and its prohibitively large 7.82GB map.

Hyundai Department Store. The Hyundai Department Store test set evaluates the localization accuracy as the percentage of test images within three thresholds: (0.1 m, 1°), (0.25 m, 2°) and (1 m, 5°). As shown in Tab. 2, our method significantly improves the performance of SCR-based approaches under large-scale indoor scenarios. Crucially, it also substantially narrows the long-standing performance gap to some FM-based methods, all while maintaining a compact map size. Compared to baseline R-SCoRe, CoLoR improves the most challenging metric by 13.8% on 1F and 20.7% on 4F under the strict (0.1 m, 1°) threshold of the evaluation protocol.

4.4. Ablation Study Results

In our ablation studies, we investigate the impact of the different components in CoLoR. Our results systematically validate each of our contributions, as illustrated in Table 3.

Impact of Pixel Partitioning. First, we validate the benefit of partitioning the training samples. By explicitly identifying and training exclusively on the multi-view points (Row 2), we achieve a substantial performance improvement over the baseline. This result confirms our hypothesis that the original framework inefficiently expends model capacity on single-view points, which are fundamentally untrainable with its reprojection-based loss and act as noise.

Benefit of Multi-view Reprojection Supervision. With the multi-view correspondences established during pixel partitioning, we then substitute the baseline's single-view reprojection loss with our more robust multi-view supervision objective. As shown in Row 3, this transition from a weak, implicit signal to a strong, explicit geometric constraint yields another significant performance boost. This demonstrates the importance of providing a clear and consistent optimization signal to stabilize training.

Extending the Frontier with Pseudo-Depth Supervision. Row 4 demonstrates the value of our second stage, where we incorporate the previously discarded single-view points using our pseudo-depth supervision. This addition further improves the model's performance. While the numerical gain is moderate, its conceptual importance is paramount: it represents performance unlocked from a subset of the scene that is entirely beyond the reach of the baseline framework.

Effect of Feature Consistency. Finally, the complete CoLoR model, equipped with our pixel-level retrieval feature, achieves the best results in Row 5. This final improvement underscores the importance of enforcing consistency between global and local features for achieving a robust and discriminative scene representation.

5. Conclusion

We present CoLoR, a novel SCR framework for large-scale localization. By partitioning scene points into multi-view and single-view sets via a bootstrapped pseudo-depth prior, we overcome the limitations of implicit triangulation. Furthermore, we resolve global-local feature inconsistency through a contrastive pixel-level retrieval objective, yielding a unified representation. Extensive experiments on indoor and outdoor benchmarks confirm that CoLoR achieves new state-of-the-art performance.

Acknowledgements

This work is supported by the Key R&D Program of Jiangxi Province, China (20232BBE50019), and the National Natural Science Foundation of China (62276016, 62372029).

References

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

[2] Eric Brachmann and Carsten Rother. Expert sample consensus applied to camera re-localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7525–7534, 2019. 7, 2

[3] Eric Brachmann and Carsten Rother. Visual camera relocalization from RGB and RGB-D images using DSAC. *TPAMI*, 2021. 1, 3

[4] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. DSAC-Differentiable RANSAC for camera localization. In *CVPR*, 2017.

[5] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In *CVPR*, 2023. 1, 3, 5, 7, 2

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

[7] Song Cao and Noah Snavely. Minimal scene descriptions from structure from motion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 461–468, 2014. 3

[8] Tommaso Cavallari, Stuart Golodetz, Nicholas A Lord, Julien Valentin, Luigi Di Stefano, and Philip HS Torr. On-the-fly adaptation of regression forests for online camera relocalisation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4457–4466, 2017. 1, 3

[9] Tommaso Cavallari, Stuart Golodetz, Nicholas A Lord, Julien Valentin, Victor A Prisacariu, Luigi Di Stefano, and Philip HS Torr. Real-time rgb-d camera pose estimation in novel scenes using a relocalisation cascade. *IEEE transactions on pattern analysis and machine intelligence*, 42(10): 2465–2477, 2019. 1, 3

[10] Wentao Cheng, Weisi Lin, Kan Chen, and Xinfeng Zhang. Cascaded parallel filtering for memory-efficient image-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1032–1041, 2019. 7

[11] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized ransac. In *Joint pattern recognition symposium*, pages 236–243. Springer, 2003. 3

[12] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 1, 2, 7

[13] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*, 2019. 7, 2

[14] Marcin Dymczyk, Simon Lynen, Titus Cieslewski, Michael Bosse, Roland Siegwart, and Paul Furgale. The gist of maps-summarizing experience for lifelong localization. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 2767–2773. IEEE, 2015. 3

[15] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DeDoDe: Detect, Don't Describe — Describe, Don't Detect for Local Feature Matching. In *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024. 2, 5, 8

[16] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2

[17] 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. 3

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

[19] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016. 5

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 6

[21] Xudong Jiang, Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. R-score: Revisiting scene coordinate regression for robust large-scale visual localization. In *CVPR*, 2025. 1, 2, 3, 4, 5, 7

[22] Donghwan Lee, Soohyun Ryu, Suyong Yeon, Yonghan Lee, Deokhwa Kim, Cheolho Han, Yohann Cabon, Philippe Weinzaepfel, Nicolas Guérin, Gabriela Csurka, et al. Large-scale localization datasets in crowded indoor spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3227–3236, 2021. 1, 6, 2

[23] Xiaotian Li, Shuzhe Wang, Yi Zhao, Jakob Verbeek, and Juho Kannala. Hierarchical scene coordinate classification and regression for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11983–11992, 2020. 3, 7

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

[25] Qingming LIU, Yuan Liu, Jiepeng Wang, Xianqiang Lyu, Peng Wang, Wenping Wang, and Junhui Hou. MoDGS: Dynamic gaussian splatting from casually-captured monocular videos with depth priors. In *The Thirteenth International Conference on Learning Representations*, 2025. 5

[26] Ting-Ru Liu, Hsuan-Kung Yang, Jou-Min Liu, Chun-Wei Huang, Tsung-Chih Chiang, Quan Kong, Norimasa Kobori, and Chun-Yi Lee. Reprojection errors as prompts for efficient scene coordinate regression. In *ECCV*, 2024. 1, 3

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[28] Marcela Mera-Trujillo, Benjamin Smith, and Victor Fragoso. Efficient scene compression for visual-based localization. In *2020 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2020. 7

[29] Son Tung Nguyen, Alejandro Fontan, Michael Milford, and Tobias Fischer. Focustune: Tuning visual localization through focus-guided sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3606–3615, 2024. 3

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

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

[32] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2691, 2024. 2

[33] 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. 7, 2

[34] 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, 7

[35] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 2, 7

[36] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *BMVC*, page 4, 2012. 1, 4, 6

[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, 7

[38] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8601–8610, 2018. 1, 4, 6

[39] 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. 1, 3

[40] 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. 2, 5

[41] Shitao Tang, Sicong Tang, Andrea Tagliasacchi, Ping Tan, and Yasutaka Furukawa. Neumap: Neural coordinate mapping by auto-transdecoder for camera localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 929–939, 2023. 3, 7, 2

[42] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[43] Julien Valentin, Matthias Nießner, Jamie Shotton, Andrew Fitzgibbon, Shahram Izadi, and Philip HS Torr. Exploiting uncertainty in regression forests for accurate camera relocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4400–4408, 2015. 1, 3

[44] 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 (CVPR)*, 2024. 1, 2, 3, 5, 7

[45] Shuzhe Wang, Juho Kannala, and Daniel Barath. Dgc-gnn: Leveraging geometry and color cues for visual descriptor-free 2d-3d matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20881–20891, 2024. 3

[46] Shuzhe Wang, Zakaria Laskar, Iaroslav Melekhov, Xiaotian Li, Yi Zhao, Giorgos Tolias, and Juho Kannala. Hscnet++: Hierarchical scene coordinate classification and regression for visual localization with transformer. *International Journal of Computer Vision*, 132(7):2530–2550, 2024. 3, 7

[47] Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior, 2025. 2, 5, 3

[48] Luwei Yang, Rakesh Shrestha, Wenbo Li, Shuaicheng Liu, Guofeng Zhang, Zhaopeng Cui, and Ping Tan. Scenesqueezer: Learning to compress scene for camera relocalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8259–8268, 2022. 7

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

[50] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and

reranking transformer for place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19370–19380, 2023. 5