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

Supplementary Material

In this supplementary material, we first detail our experimental setup for both the Two-Stage Training Framework and the Consistent Retrieval Feature. Following this, we present several additional results to demonstrate the effectiveness of our proposed method. Finally, we discuss the efficiency of our work.

A. Implementation Details

A.1. Two-Stage Training Framework

Image data augmentation We build upon the data augmentation strategy of our baseline [21], applying the same random rotations uniformly from [-15°, 15°], random scaling uniformly from [0.67, 1.5], and brightness jitter uniformly from [0.9, 1.1]. However, to better address the challenge of severe illumination variations, we enhance this pipeline by incorporating two additional augmentations. First, we introduce contrast jitter, using the same [0.9, 1.1] factor range. Second, we apply Random Gamma Correction with a 60% probability, sampling the gamma value uniformly from [0.8, 3.0]. This allows us to simulate a much broader and more realistic range of lighting appearances, thereby improving the model's generalization capabilities.

Number of training pixels. We focus on the actual quantity of points involved in the training process. In this section, we report the pixel counts used by the baseline [21] alongside the specific numbers of points used in the respective stages of CoLoR.

For the baseline, the standard protocol involves multi-GPU training where each GPU employs a different random seed to sample a fixed number of pixels per image. Consequently, for a fair comparison, we calculate the total number of unique pixels actually sampled by the baseline across the entire training process for each scene.

In contrast, CoLoR dynamically adjusts the sample size based on the output of our efficient partitioning. Specifically, in the first stage, we utilize all identified multi-view points; as a result, each image contributes a varying number of points depending on its geometric connectivity. In the second stage, although the generated pseudo-depth maps theoretically allow us to supervise all points in the scene, we adopt a balanced sampling strategy: for each image, we sample additional single-view points equivalent to 0.5 times its multi-view point count.

Tab. 4 presents a comparison of the actual training point counts per scene between the baseline and our method.

Scenes	R-SCoRe [21]	CoLoR-stage1	CoLoR-stage2
aachen	25437778	13809562	20701301
Dept. 1F	73913789	34774488	49742681
Dept. 4F	43878927	15378739	22147642
Dept. B1	82703472	56675275	84900359

Table 4. **Comparison of the number of training pixels per scene.** We report the total count of unique pixels sampled by the baseline [21] and provide a breakdown of the two stages in CoLoR.

Network architecture Given that CoLoR applies strong supervision to all points in the scene, we are able to slightly increase the network depth to enhance its capacity. However, to ensure a fair comparison, we simultaneously reduce the network width as we increase the depth, thereby maintaining a map size comparable to that of the baseline.

Specifically, we modified the two ResNet blocks within the baseline architecture. We increased the number of blocks from 3 and 2 to 5 and 4, respectively, while reducing the MLP ratio from 2.0 to 1.0.

A.2. Consistent Retrieval Feature

Evaluation for PR Curves. In this section, we detail the evaluation protocol used to generate the Precision-Recall curves in Fig. 3b of the main paper. For the evaluation of global features in isolation, we directly adopt the results and experimental protocol from [21], visualizing the capability of the features to classify image-level co-visibility. In contrast, for the scenarios involving the concatenation of global features with either traditional local features or our proposed pixel-level retrieval features, we employed a fixed random seed to sample an identical subset of 3D points from the SfM point cloud of the Aachen dataset. Since each 3D point is associated with multiple 2D observations, this setup allows us to evaluate the features' ability to determine whether two distinct pixels correspond to the same underlying 3D point. For all feature comparisons presented in this figure, the L2 distance is used as the metric.

B. Additional Results

B.1. Analysis of Multi/Single-view Point Distribution.

Following the definitions of multi-view and single-view points established in the main text, we analyzed their distribution across three datasets, as summarized in Tab. 6. These datasets—7Scenes, Aachen Day-Night, and Hyundai Department Store—are ordered from smallest to largest based

	Dept. 1F Validation	Dept. 4F Validation	Dept. B1 Validation
HLoc+D2-Net [13, 34]	(83.2 / 89.2 /94.5) / 398GB	(72.1 / 85.3 / 98.5) / 183GB	(70.2/ 78.0 / 86.1) / 505GB
HLoc+R2D2 [33, 34]	(85.8 / 89.9 / 94.4) / 166GB	(72.6/ 84.6 / 98.3) / 76GB	(71.6/ 78.0 / 86.0) / 210GB
Neumap [41]	(75.5 / 88.2 / 95.8) / 726MB	(70.4 / 85.4 / 99.0) / 431MB	(46.0 /66.5 / 79.8) / 857MB
ESAC (×50) [2]	(49.7 / 71.5 / 84.1) /1.4GB	(45.2 / 69.9 / 85.1) / 1.4GB	(5.4 / 9.1 / 14.2) / 1.4GB
ACE (×50) [5]	(14.2 / 49.9 / 77.8) / 205MB	(29.3 / 80.0 / 96.7) / 205MB	(2.6 / 14.0 / 28.2) / 205MB
GLACE [44]	(4.9 / 24.4 / 53.5) / 42MB	(24.5 / 57.5 / 85.4) / 42MB	(1.0 / 4.5 / 13.8) / 42MB
R-SCoRe [21]	(70.6 / 86.6 / 95.5) / 127MB	(63.9 / 84.2 / 98.3) / 50MB	(57.7 / 74.7 / 86.7) / 130MB
CoLoR	(81.2 / 90.6 / 96.6) / 127MB	(71.9 / 85.6 / 99.0) / 50MB	(62.6 / 76.2 / 87.9) / 130MB

Table 5. **Hyundai Department Store Validation Set evaluation.** The percentages of query images within three thresholds: (0.1m,1°), (0.25m, 2°), and (1m, 5°) and the map size are reported. CoLoR significantly outperforms the baseline R-SCoRe and other SCR methods, achieving accuracy comparable to Feature Matching (FM) approaches while maintaining a compact map size.

Datasets	**Multi (%)**	**Single (%)**
7Scenes	74.86	25.14
Aachen Day-Night	57.05	42.95
Hyundai Dept. Store	23.36	76.64

Table 6. **Analysis of Multi/Single-view Point Distribution.**

Scenes	**Dept. 1F Test**	**Dept. 4F Test**
R-SCoRe [21]	61.4 / 80.2 / 90.9	60.2 / 79.3 / 87.9
R-SCoRe w/ Our MLP config	59.6 / 79.1 / 88.9	60.7 / 80.6 / 88.2

Table 7. **Ablation study on network architecture.** We evaluate the baseline [20] equipped with our modified network architecture (narrower and deeper) on the Hyundai 4F and 1F datasets.

on their spatial coverage. The statistics reveal a clear trend: as the scene scale increases, the proportion of multi-view points steadily decreases, while the percentage of single-view points correspondingly rises.

B.2. Evaluation on Hyundai Validation Set

While the main paper presents results on the test set of the Hyundai Department Store dataset [22], in this section, we report the corresponding performance on the validation set. As shown in the Tab. 5, our baseline R-SCoRe [21] underperforms Neumap [41] on the 1F and 4F floors. In contrast, CoLoR consistently outperforms all SCR based methods. Notably, on the 1F and 4F floors, our method achieves an accuracy comparable to FM based approaches, while retaining the significant advantage of a compact map size.

B.3. Ablation on network architecture

In the baseline configuration [21], the network architecture adapts its capacity based on the scale of the dataset, resulting in two distinct settings. For datasets with smaller training sets, such as Aachen Day-Night (4,328 training images) and Hyundai 4F (7,428 training images), a base width of 768 is employed. Conversely, for larger environments like Hyundai 1F (16,222 training images) and B1 (20,579 training images), the base width is increased to 1,280. To comprehensively evaluate the effectiveness of our design, we conducted ablation studies on both Hyundai 4F and 1F. This allows us to assess the impact of our narrower and deeper architecture variant across these two distinct capacity configurations.

The results in Tab. 7 indicate that the architectural modification alone has a negligible impact on the baseline's performance. This serves as a strong validation for our proposed framework. It demonstrates that the state-of-the-art performance of CoLoR is not a byproduct of a deeper network capacity, but rather the result of our novel designs.

B.4. Training Dynamics

To further validate the effectiveness of CoLoR, we compare its training dynamics against the baseline R-SCoRe [21]. In Fig. 4, we visualize the evolution of three key metrics during the training process: the median reprojection error, the ratio of inlier training predictions (defined as predictions with reprojection errors below 10 pixels), and the mean projection error of these inliers.

As observed, CoLoR demonstrates superior convergence properties compared to the baseline. Specifically, our method reduces the median reprojection error more rapidly and stabilizes at a significantly lower value. Furthermore, CoLoR consistently achieves a higher ratio of inlier predictions throughout the training process. This evidence confirms that our proposed explicit partitioning with strong supervision, combined with the consistent retrieval feature, effectively guides the network to learn more accurate and robust scene coordinates compared to the baseline.

B.5. Additional Visualization Results

In this section, we provide additional qualitative comparisons of the local point clouds reconstructed by the baseline R-SCoRe [21] and our proposed CoLoR. To generate these

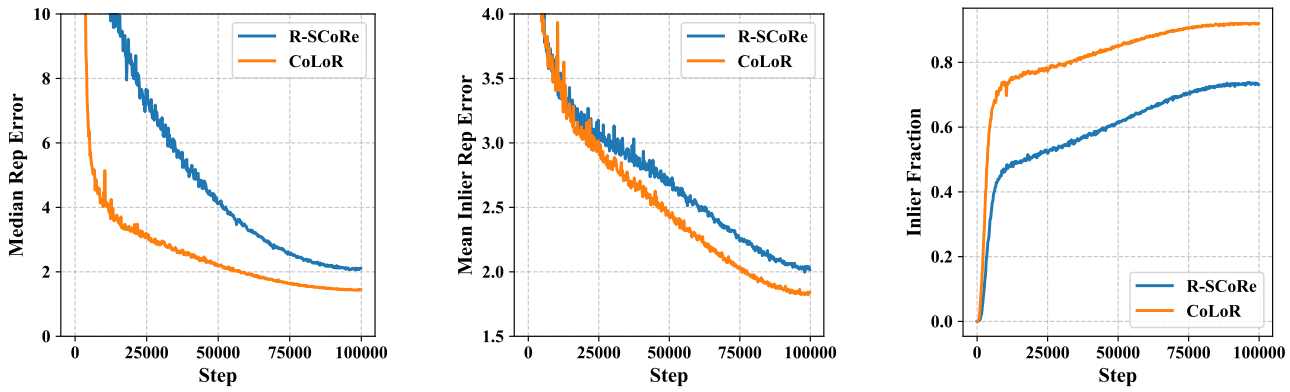


Figure 4. **Training dynamics comparison.** From left to right, we present the median reprojection error, the ratio of inlier training predictions with reprojection errors below 10 pixels, and the mean projection error of these inliers.

	Aachen Day-Night	Dept. 4F
Efficient Partitioning	2.6 min	3.8 min
Stage One	2.8 hour	2.8 hour
Pseudo Depth Generation	4.5 min	6.6 min
Stage Two	1.2 hour	1.2 hour

Table 8. **Efficiency Analysis.**

visualizations, we sampled images covering identical local regions and filtered out unreliable predictions by discarding pixels with a reprojection error exceeding 10 pixels.

As illustrated in Fig. 5, CoLoR consistently reconstructs significantly denser and more complete point clouds across both indoor and outdoor environments compared to the baseline. This visual evidence corroborates that our method effectively achieves comprehensive training for all scene points, successfully recovering scene geometry even in challenging single-view or textureless areas where the baseline often fails.

C. Efficiency Analysis

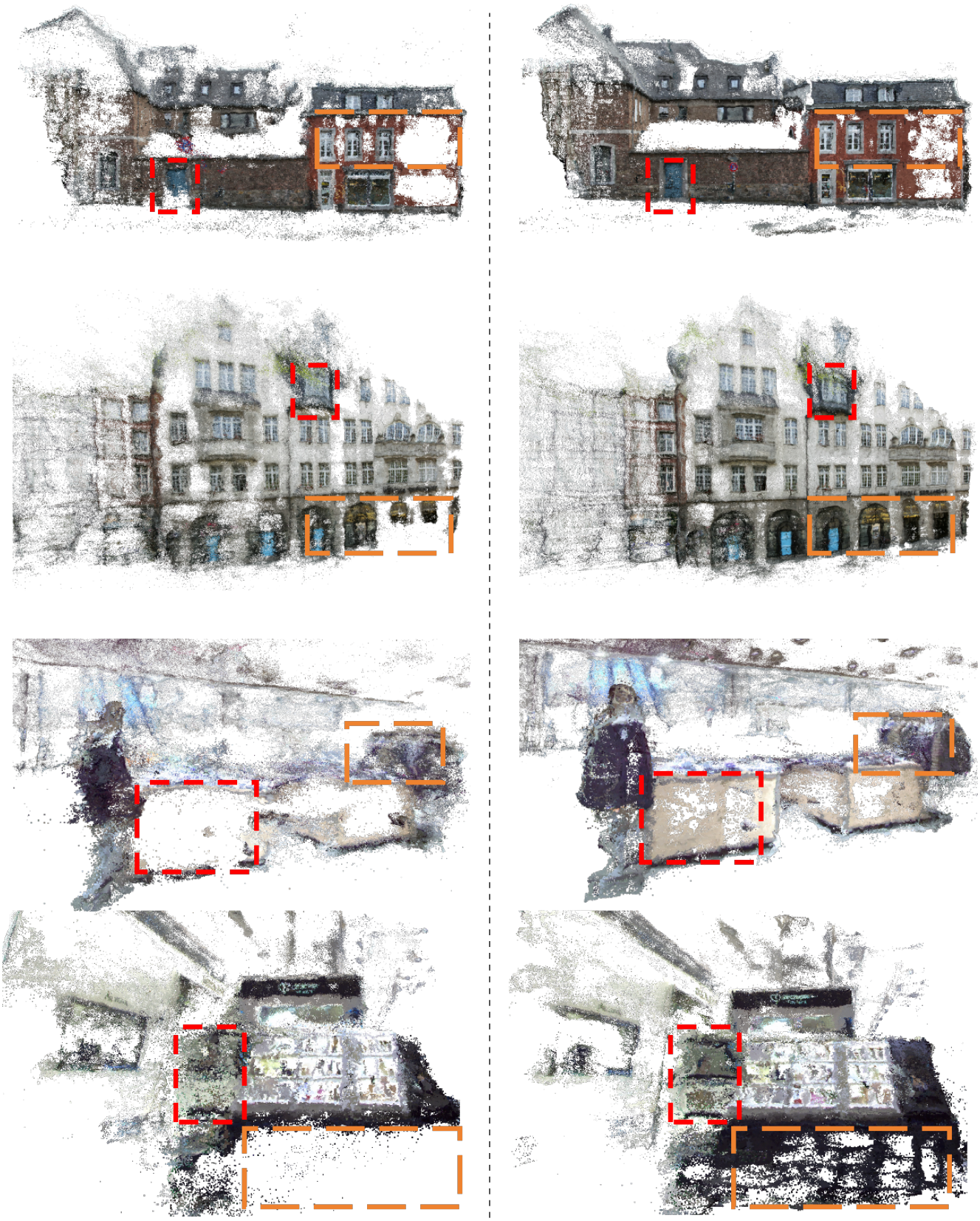
Inference efficiency. Since our method shares a similar network architecture and the PnP-RANSAC backend with the baseline [21], the inference latency remains unchanged. For reference, the baseline reported a total inference time of approximately 140 to 270 ms per query image (evaluated on an Intel i7-9700K CPU and NVIDIA RTX 2080 GPU). This confirms that CoLoR preserves the high runtime efficiency characteristic of SCR approaches.

Training and pre-processing efficiency. Our framework introduces two additional pre-processing steps during the training phase: Efficient Partitioning and Pseudo Depth Generation. We report the time costs for these steps in Tab. 8. All efficiency statistics were measured on a server equipped with 4 NVIDIA RTX 4090 GPUs.

For efficient partitioning, we maximize throughput by adopting a cached matching strategy: keypoints and descriptors for all training images are extracted once and stored in RAM. Feature matching is then performed exclusively against the top-k co-visible neighbors for each image. We impose a maximum cap of 5,000 multi-view points per image, triggering an early termination of the matching loop once this threshold is reached.

For pseudo depth generation, we utilize the ViT-B variant of [47] to generate monocular depth priors.

Overall, these two additional pre-processing steps introduce a negligible computational overhead, constituting only a marginal fraction of the overall training duration.



(a) R-SCoRe [21] (b) CoLoR

Figure 5. **Qualitative comparison of local point cloud reconstructions.** We visualize the point clouds generated by the baseline R-SCoRe [21] and CoLoR across varying indoor and outdoor environments.