Reprojection Errors as Prompts for Efficient Scene Coordinate Regression

Ting-Ru Liu[1], Hsuan-Kung Yang[2], Jou-Min Liu[1*], Chun-Wei Huang[1*], Tsung-Chih Chiang[1*], Quan Kong[2], Norimasa Kobori[2], and Chun-Yi Lee[1]

[1]National Tsing Hua University [2]Woven by Toyota, Inc.
{tingru, diasimui, alan, george, cymaxwelllee}@elsa.cs.nthu.edu.tw
{hsuan-kung.yang, quan.kong, norimasa.kobori}@woven.toyota
https://tingru0203.github.io/egfs

Abstract. Scene coordinate regression (SCR) methods have emerged as a promising area of research due to their potential for accurate visual localization. However, many existing SCR approaches train on samples from all image regions, including dynamic objects and texture-less areas. Utilizing these areas for optimization during training can potentially hamper the overall performance and efficiency of the model. In this study, we first perform an in-depth analysis to validate the adverse impacts of these areas. Drawing inspiration from our analysis, we then introduce an error-guided feature selection (EGFS) mechanism, in tandem with the use of the Segment Anything Model (SAM) [1]. This mechanism seeds low reprojection areas as prompts and expands them into error-guided masks, and then utilizes these masks to sample points and filter out problematic areas in an iterative manner. The experiments demonstrate that our method outperforms existing SCR approaches that do not rely on 3D information on the Cambridge Landmarks and Indoor6 datasets.

1 Introduction

The objective of visual localization is to estimate a 6-DoF camera pose from images, a key component in fields such as Augmented Reality, Virtual Reality, and autonomous driving. Contemporary leading methods in visual localization typically involve establishing 2D-3D correspondences and then utilizing Perspective-n-Point (PnP) [2] with RANSAC [3] for camera pose estimation. These methods can be broadly classified into two main directions: feature-matching [4] and scene coordinate regression (SCR) [5]. Feature-matching approaches reconstruct a 3D scene using Structure from Motion (SfM), identify and describe key points in 2D images [6,7], and link these to 3D coordinates [8,9]. Nevertheless, they may encounter challenges such as high computational demands, significant storage requirements, and potential privacy concerns [10]. On the other hand, SCR methods [11–17] employ deep neural networks (DNNs) to predict the 3D coordinates of pixels and then utilize PnP with RANSAC for camera pose estimation, which

* indicates equal contribution.

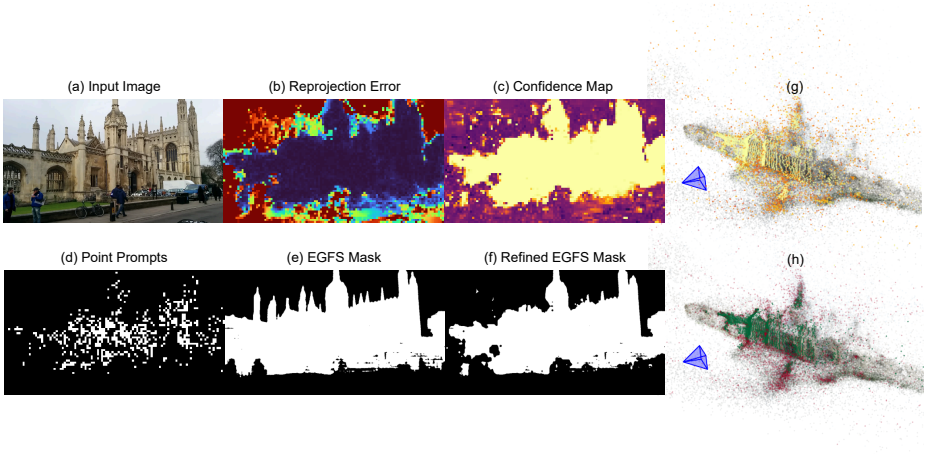

Fig. 1: Visualization of the primary components (i.e., (d)-(h)) introduced in the proposed visual localization scheme. (d) illustrates the point prompts selected from (b) with low reprojection errors, while (e) presents an error-guided mask expanded from the prompted points in (d) using SAM. (f) displays the proposed error-guided feature selection (EGFS), which refines the mask from (e) with the predicted confidence map (c) to ensure high-quality scene coordinates are sampled for estimating the final camera pose. The point cloud constructed from the predicted scene coordinates is shown on the right-hand side (i.e., (g)-(h)), with the confidence (yellow parts) and the refined EGFS mask (green for selected areas; red for rejected areas).

provide benefits such as accuracy in smaller scenes, reduced training times, as well as minimized storage requirements. Given these advantages, SCR is thus the primary focus of this study and presents potential for further enhancements.

Albeit effective, previous SCR methods face two primary challenges to be addressed: the presence of dynamic objects and texture-less regions. Dynamic objects, such as pedestrians and cars, pose difficulties for these techniques due to their changing nature and the unsuitable features extracted from these regions. As illustrated in Fig. 1, these regions could lead to high reprojection errors during training. The second challenge is that current methods struggle with flat, texture-less surfaces. Despite their apparent simplicity, these areas often result in inaccurate scene coordinates due to difficulties in feature extraction. While previous methods [14, 15, 17] employ techniques like RANSAC to filter out outliers and differentiable end-to-end optimization approaches to disregard these outliers, they demand significant computational time for end-to-end processing and do not account for the semantic meaning of the selected areas, which may still result in the selection of outlier areas. Thus, they might not effectively prevent the models from unstable weight updates and can lead to training instability.

In light of the aforementioned issues, this study aims to explore the possibility of guiding the sampling process to favor regions with low reprojection errors, and seeks to leverage visual information in an image to expand masks

that bear semantic regional meanings. The core philosophy of our methodology focuses on selecting robust regions for training without the need to manually define explicit areas or categories. Such robust regions might differ across scenes, which make it inappropriate for manual definition, as elaborated in Section 4. More specifically, we propose a strategy, named *Error-Guided Feature Selection (EGFS)*, for deriving low reprojection error samples as point prompts, and expanding them to encompass a complete mask with similar semantic meanings. The masks are iteratively updated after being frozen for a specified period, which enable the proposed model to dynamically renew the focused regions. To achieve this prompt-based semantic regions expansion, we employ the Segment Anything Model (SAM) [1], a vision foundation model capable of providing a general understanding of the scene. This error-guided scheme ensures that our training process concentrates on regions with low reprojection errors and adaptively reduces the variability introduced by dynamic or texture-less objects. Moreover, EGPS adopts a predicted confidence map to refine the expanded error-guided mask for updates, which ensures the regions marked by the masks are sufficiently reliable. An illustrations of each component is provided in Fig. 1.

To validate the effectiveness of our proposed method, we conduct evaluations on both outdoor and indoor datasets, including the Cambridge Landmarks dataset [18] and the Indoor6 [19] dataset, for evaluating the performance of our proposed SCR methodology. The experimental results suggest that our method can be broadly applied to previous methods and provides benefits, and leads to state-of-the-art (SOTA) performance when compared with SCR approaches without leveraging any 3D information in terms of estimated translational and rotational errors, while requiring less training time and smaller model sizes. We further perform a series of detailed analyses and ablation studies to confirm the effectiveness of our error-guided sampled regions and the proposed methodology.

2 Related Work

Scene Coordinate Regression. Conventional SCR approaches typically begin by establishing 2D-3D correspondences and then utilize RANSAC-based optimization to estimate the camera pose. To establish 2D-3D correspondences, previous endeavors have explored the use of random forests [5, 20–22] or convolutional neural networks (CNNs) [11–15, 17, 23] to regress scene coordinates from images. Since SCR typically requires scene-specific training, training time has emerged as a critical concern in practical applications. The ACE [17] method, which can be trained within five minutes, stands out as a promising solution. Moreover, unlike methods that map every 2D pixel to a 3D coordinate, some works [19, 24, 25] have aimed at learning to detect 3D landmarks to establish 2D-3D correspondences. However, they often require the reconstruction of a 3D model, which incurs additional computational time and costs. Our method seeks to utilize SAM to derive error-guided masks from 2D images, while maintaining the training efficiency.

Emphasis on Robust Features for Localization. Several prior studies [24, 26–31] in visual localization have discovered that not all regions within an image contribute

equally to localization accuracy. As a result, these approaches have attempted to distinguish robust and invariant features, found that prioritizing the recognition and comprehension of those regions can enhance visual localization tasks more efficiently and effectively. However, a common limitation of these methods is their reliance on pre-defined semantics or the necessity of a 3D model, which could hinder their generalizability and practicality in real-world applications.

3 Preliminary of Scene Coordinate Regression (SCR)

SCR-based visual localization determines the camera pose of an RGB image $I \in \mathbb{R}^{H \times W \times 3}$ by predicting 3D scene coordinates $\mathcal{Y}$ for a set of pixels and establishing the 2D-3D correspondences between pixel coordinates in the image and those predicted 3D scene coordinates. The primary objective of the training process is to establish a mapping function $f(\cdot)$ such that $\mathcal{Y} = f(I)$. The training data comprise RGB images paired with their respective camera poses, but do not necessarily include depth information. To generate accurate $\mathcal{Y}$, SCR methods often involve optimizing the reprojection error $\mathcal{R}$, which is measured by projecting a 3D predicted scene coordinate $\mathrm{y_i} \in \mathcal{Y}$ back onto its 2D pixel coordinate and calculating the discrepancy from the actual coordinate of the image patch, where i denotes the patch index. The reprojection error for each patch $\mathrm{r_i} \in \mathcal{R}$ can be formulated as $r(\mathrm{p_i}, \mathrm{y_i}, \mathrm{h}^*) = \|\mathrm{p_i} - K\mathrm{h}^{*-1}\mathrm{y_i}\|$, where K denotes the camera intrinsic matrix and h^* denotes the ground truth camera pose. After obtaining $\mathcal{Y}$, the camera pose h can be estimated according to the following equation:

$$\mathrm{h} = g(\mathcal{C}) = g^{PnP}(\mathcal{C}_{\mathcal{I}}), \text{with } \mathcal{C} = \{(\mathrm{p_i}, \mathrm{y_i}) | \mathrm{p_i} \in I, \mathrm{y_i} \in \mathcal{Y}\}, \tag{1}$$

where $g(\cdot)$ involves PnP with RANSAC, followed by Levenberg–Marquardt-based refinement, $\mathcal{C}$ represents the set of all 2D-3D correspondences and $\mathcal{C}_{\mathcal{I}}$ signifies inlier correspondences after refinement, which is a subset of $\mathcal{C}$. Inlier correspondences denote those that are better consistent with the estimated camera pose, while correspondences that do not align well with the estimated camera pose are termed outliers. The final camera pose estimation is determined by the inlier correspondences $\mathcal{C}_{\mathcal{I}}$. In the following paragraphs, we introduce two scene coordinate regression methods that have demonstrated superior performance.

DSAC Variants. DSAC [13] introduces a differentiable variant of RANSAC, and enables end-to-end training of the entire pipeline. This design combines the benefits of RANSAC for handling outliers with the power of deep neural networks to learn complex patterns. In addition, DSAC++ [14] removes the requirement of depth information for SCR training and adopts robust loss terms to downweight outliers. DSAC* [15] further simplifies and refines DSAC++, and encourages the network to focus on reliable scene structures while ignoring outlier predictions.

ACE. ACE [17] is a fast SCR method that can achieve promising localization accuracy within a short training time of only five minutes. ACE splits its regression network into a pre-trained scene-agnostic backbone f_B, and a scene-specific

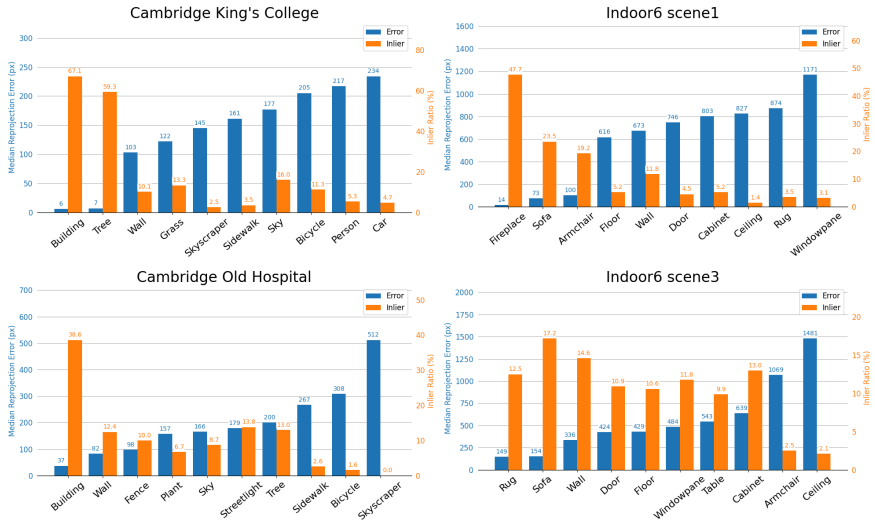

Fig. 2: Analysis between reprojection error and semantic meaning. The analysis result indicates the regions with low reprojection errors tend to have higher inlier ratios, while the errors do not always align with specific semantic categories, e.g., "tree" and "rug".

multi-layer perceptron (MLP) f_H. The backbone f_B extracts feature vectors f_i from image patches p_i and the MLP f_H predicts the scene coordinates y_i based on the extracted f_i. Between f_B and f_H, the features are shuffled to decorrelate gradients within a batch, which enables ACE to enhance its training efficiency.

4 In-Depth Evaluation of Scene Coodinate Regression

4.1 Challenges in Scene Coordinate Regression

Several challenging factors have been identified that may compromise the accuracy of SCR. Two key challenges among them are dynamic objects and textureless surfaces [27, 32–34]. Dynamic objects exhibit characteristics of motion or disappearance across different frames, while textureless surfaces lack distinctive feature points or salient characteristics. These challenges highlight the complexity of achieving precise SCR in complex visual scenarios. To address them, previous SCR approaches [13–15, 17] have employed RANSAC to filter out outliers, and some researches [14, 15, 17] have attempted to utilize end-to-end learning to concentrate on reliable scene structures while down-weighting outlier predictions during training. Unfortunately, these prior endeavors failed to fully exploit the rich semantic information in the visual content, which can be crucial for enhancing the robustness and accuracy of camera pose estimation in complex scenes. Furthermore, RANSAC-based approachs for camera pose estimation could potentially fail in scenarios with insufficient number of inlier correspondences [35].

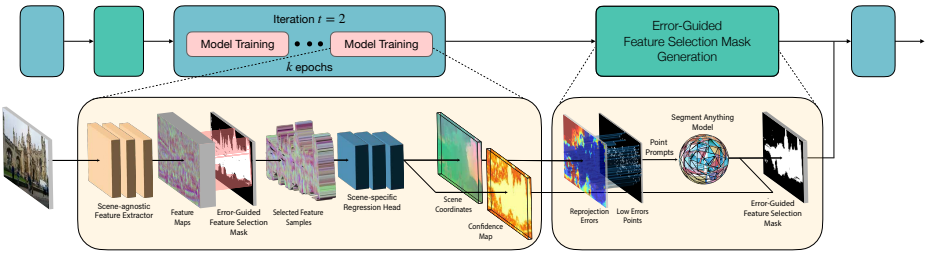

Fig. 3: An overview of the training framework.

4.2 Analysis between Reprojection Error and Semantic Meaning

For a more comprehensive understanding of the correlation and relationship between reprojection errors and semantic meaning, we investigate the performance of ACE on the reprojection errors and the inlier correspondences selected by RANSAC for estimating the final camera pose during the inference process. First, image pixels are labeled by a semantic segmentation model ViT-Adapter [36]. Subsequently, the reprojection errors and inlier correspondences of ACE for each semantic class are calculated. Fig. 2 presents the median reprojection errors and the ratios of inlier correspondences for the top ten most frequent classes in each scene. The left two figures correspond to the outdoor dataset Cambridge Landmarks [18], while the right two figures pertain to the indoor dataset Indoor6 [19]. It is observed that regions with low reprojection errors tend to have relatively higher inlier ratios, indicating a greater impact on the estimated camera poses. As a result, improving the accuracy of low-error areas could lead to more precise camera pose predictions. Nevertheless, it is also observed that the reprojection errors for the same class can vary significantly across different scenes. For example, in the Cambridge Landmarks dataset, the error for "tree" is low in King's College but high in Old Hospital. Similarly, in the Indoor6 dataset, the error for "rug" is high in scene1 but low in scene3. This indicates that low reprojection errors do not always align with specific semantic categories, which presents a challenge in directly using certain specific semantic segmentation labels for area selection. Motivated by these insights, the primary objective of this study is to develop a methodology that does not rely on pre-defined semantic categories to identify low reprojection error areas for enhancing localization accuracy in SCR.

5 Methodology

5.1 Problem Definition and Framework Overview

Fig. 3 provides an overview of the proposed training framework for SCR, which iteratively samples features to train a scene-specific MLP. The scene-specific MLP consists of a scene coordinate head and a confidence head. In each iteration, the model is trained for k epochs. During the initial iteration, features are randomly sampled from all parts of images in order to derive the first set

of reprojection errors. In subsequent iterations, features are selected based on error-guided feature selection (EGFS) masks generated according to reprojection errors and a confidence map. This iterative mechanism enables the model to dynamically update its focus areas throughout the training process. During the inference phase, the network estimates scene coordinates $\mathcal{Y}$ and a confidence map for the refinement, while the masks generated during training are not required for the inference phase. The inference procedure is depicted in Fig. 4.

5.2 Error-Guided Feature Selection (EGFS) with SAM

In alignment with the discussion in Section 4 on identifying potential areas contributing to the final camera pose estimation during the training stage, we employ the Segmentation Anything Model (SAM) [1], which is capable of leveraging diverse visual cues to generate high-quality object masks. Specifically, in this study, SAM is utilized to extend point prompts with low reprojection errors to form a complete mask sharing a similar semantic context. The rationale behind this approach is the high likelihood that these comprehensive semantic regions correspond to low error areas and can contribute to final pose estimation. Such a concept harnesses the capabilities of the foundational SAM model to uncover more viable areas, and hence, enabling a thorough understanding of the image context. In our experiments, we utilize EfficientViT-SAM-L0 [37], a lightweight and efficient variant of SAM, to identify salient regions by selecting the τ percentage of points with the lowest reprojection errors as point prompts, where τ is adjustable and set to ten in our case. After obtaining the error-guided mask from the SAM model, the predicted confidence map c is utilized to refine the error-guided mask to ensure the points marked by the mask is sufficiently reliable. Specifically, in each image, the points with confidence scores below a threshold σ are filtered out. The design for confidence refinement and its optimization are described in Section 5.3. In each iteration, the refined error-guided masks serve as guidance for training the scene-specific MLP. Only the features corresponding to the regions selected by the masks are sampled into the training buffer. This selective inclusion is facilitated by the nature of f_H [17], which employs 1×1 convolutions to treat and process each selected sample independently using the same set of shared weights. The proposed feature selection mechanism ensures that the training process focuses on the crucial regions identified in the current iteration, and therefore enables the model to enhance its overall performance.

5.3 Scene Coordinate and EGFS Refinement with Confidence

In order to further refine the error-guided mask and ensure the selected points are of sufficiently high quality for predicting accurate camera pose, we incorporate the design of a confidence map [38, 39]. Specifically, we predict a confidence score pixel-wise from the confidence head and optimize it jointly with the reprojection error. This confidence head replaces the last three 1×1 convolution layers of the original scene-specific MLP with two 1×1 convolution layers for

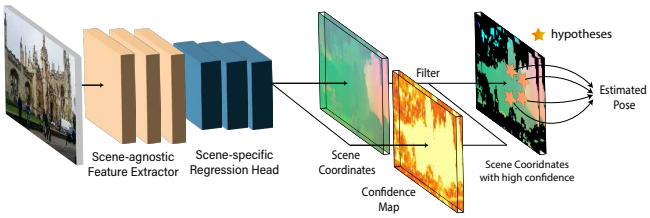

Fig. 4: An overview of the inference procedure.

confidence score prediction, thus enabling adaptive weights for the reprojection loss at each position. The training loss used by our method is formulated as:

$$\ell(\mathrm{p_i}, \mathrm{y_i}, \mathrm{h_i^*}) = \begin{cases} \mathrm{c_i} \cdot \hat{r}(\mathrm{p_i}, \mathrm{y_i}, \mathrm{h_i^*}) - \alpha \log \mathrm{c_i} & \text{if } \mathrm{y_i} \in \mathcal{V} \\ \|\mathrm{y_i} - \bar{\mathrm{y}}_\mathrm{i}\|_0 - \alpha \log(1 - \mathrm{c_i}) & \text{otherwise} \end{cases}, \tag{2}$$

where $\hat{r}(\mathrm{p_i}, \mathrm{y_i}, \mathrm{h_i^*})$ represents a `tanh` clamping of the reprojection error that changes over time, $\mathcal{V}$ denotes the set of 2D pixels with valid scene coordinate predictions, as explained in [15, 17], $\mathrm{c_i}$ is the confidence score at pixel i, α is a hyperparameter that balances the confidence regularization term, and $\bar{\mathrm{y}}_\mathrm{i}$ is a dummy scene coordinate derived from the ground truth camera pose, assuming a constant image depth of 10m. When the scene coordinate prediction is valid, the model down-weights the reprojection errors in challenging regions using confidence scores and focuses more on regions with reliable predictions. On the other hand, when the scene coordinate prediction is invalid, the model is incentivized to lower the confidence scores for inaccurate predictions. The confidence score is also used in the inference phase to select reliable scene coordinates. For each query image, the median of the confidence scores is calculated, and only scene coordinates with confidence greater than this median are selected. This thresholding ensures that only the most reliable 2D-3D correspondences are used to generate the camera pose by the PnP with RANSAC pose solver. By focusing on high-confidence correspondences, a more accurate camera pose can be estimated.

6 Experimental Results

6.1 Experimental Setups

Datasets. Our evaluation of the proposed method was conducted on two representative datasets: the outdoor dataset Cambridge Landmarks [18] and the indoor dataset Indoor6 [19]. The details of the datasets are described as follows:

Cambridge Landmarks. The Cambridge Landmarks Dataset [18] is a renowned outdoor dataset extensively utilized for visual localization tasks, which includes five distinct scenes. Every scene in this dataset is composed of RGB images along with their respective ground truth camera poses reconstructed using the SfM technique. It includes a broad spectrum of conditions and various viewpoints, and is ideal for evaluating the resilience and effectiveness of various SCR methods.

Table 1: A comparison of model sizes, training times, and median translation and rotation errors (in cm/°) evaluated on the Cambridge Landmarks dataset. '*dual*' and '*quad.*' denote the ensemble versions of two and four models, respectively.

	Method	Size	Mapping Time	King's College	Great Court	Old Hospital	Shop Facade	St Mary's Church	Average
FM	hLoc (SP+SG) [4, 6, 8]	∼800MB	∼35 min	12/0.2	16/0.1	15/0.3	4/0.2	7/0.2	11/0.2
	pixLoc [26]	∼600MB		30/0.1	14/0.2	16/0.3	5/0.2	10/0.3	15/0.2
SCR (w/ 3D model)	DSAC* [15]	28MB	15 hr	15/0.3	49/0.3	21/0.4	5/0.3	13/0.4	21/0.3
	FocusTune [31]	4MB	6 min	19/0.3	38/0.1	18/0.4	6/0.3	15/0.5	19/0.3
	FocusTune (*quad.*) [31]	16MB	24 min	15/0.3	29/0.1	17/0.4	5/0.2	9/0.3	15/0.3
	NeuMap [40]	-	-	14/0.2	6/0.1	19/0.3	6/0.2	17/0.5	12/0.3
	SACReg-L [41]	-	-	11/0.2	13/0.1	13/0.2	6/0.2	5/0.3	10/0.2
SCR	DSAC* [15]	28MB	15 hr	18/0.3	34/0.2	21/0.4	5/0.3	15/0.6	19/0.4
	ACE [17]	4MB	5 min	28/0.4	43/0.2	31/0.6	5/0.3	18/0.6	25/0.4
	ACE (*quad.*) [17]	16MB	20 min	18/0.3	**28/0.1**	25/0.5	5/0.3	**9/0.3**	17/0.3
	EGFS	4.5MB	12 min	**14/0.3**	31/0.1	21/0.4	5/0.3	15/0.5	17/0.3
	EGFS (*dual*)	9MB	21 min	**14/0.3**	**28/0.1**	**19/0.4**	**5/0.2**	10/0.3	**15/0.3**

Table 2: A comparison of model sizes, training times, and the proportions of translation and rotation errors that are below 5cm/5° evaluated on the Indoor6 dataset.

	Method	Size	Mapping Time	scene1	scene2a	scene3	scene4a	scene5	scene6	Average
FM	hLoc [4]	∼1.5GB	∼3.3 hr	70.5%	52.1%	86.0%	75.3%	58.0%	86.7%	71.4%
SCR (w/ 3D model)	DSAC* [15]	28MB	15 hr	18.7%	28.0%	19.7%	60.8%	10.6%	44.3%	30.4%
	SLD* (300 landmarks) [25]	15MB	∼5.5 hr	47.2%	48.2%	56.2%	67.7%	33.7%	52.0%	50.8%
	SLD* (1000 landmarks) [25]	120MB	∼44 hr	68.5%	62.6%	76.2%	77.2%	57.8%	78.0%	70.1%
SCR	DSAC* [15]	28MB	15 hr	23.0%	33.9%	26.0%	67.1%	10.6%	50.2%	35.1%
	ACE [17]	4MB	5 min	26.0%	32.3%	31.4%	62.0%	14.2%	47.4%	35.6%
	ACE (*quad.*) [17]	16MB	20 min	52.9%	52.5%	62.9%	69.6%	**31.1%**	**82.4%**	58.6%
	EGFS	4.5MB	21 min	46.4%	**60.6%**	56.4%	**78.7%**	22.8%	71.6%	56.1%
	EGFS (*dual*)	9MB	30 min	**58.5%**	59.1%	**67.0%**	76.1%	30.6%	75.9%	**61.2%**

Indoor6. The Indoor6 [19] dataset comprises six distinct indoor scenes captured over several days, with ground truth camera poses computed using COLMAP [42]. Each scene includes multiple rooms and contains illumination variations, making it challenging for different types of visual localization tasks.

Implementation Details. The training process of our methodology includes 20 epochs, with the generation of masks every five epochs. We set the confidence regularization parameter α to ten and the reprojection error threshold τ to ten. For mask refinement, the parameter σ is set to the median confidence score of each image. Our method builds upon the ACE [17] architecture as the backbone, and thus, the remaining hyperparameters are kept the same as those used in [17].

6.2 Results on the Cambridge Landmarks and Indoor6 Dataset

Cambridge Landmarks Dataset. Table 1 presents the experimental results evaluated on the Cambridge Landmarks dataset, which affirm the effectiveness and efficiency of our proposed methodology. It can be observed that EGFS maintains a model size similar to ACE, with only a 0.5MB increase for the confidence head, while reducing average translational and rotational errors. The results further reveal that EGFS not only surpasses DSAC* in performance and performs

Table 3: Comparison of the effectiveness between (a) directly sampling solely based on pixel points with low reprojection errors, and (b) sampling through the EGFS masks.

Dataset	Scene	$r_i < Q_{0.3}(R)$	$r_i < Q_{0.4}(R)$	$r_i < Q_{0.5}(R)$	$r_i < Q_{0.6}(R)$	$r_i < Q_{0.7}(R)$	*EGFS Masks*
Cambridge	King's College	20/0.3	19/0.3	20/0.3	20/0.3	20/0.3	**14/0.3**
	Great Court	39/0.2	37/0.1	36/0.2	38/0.2	39/0.2	**31/0.1**
	Old Hospital	26/0.5	25/0.5	23/0.5	25/0.5	24/0.5	**21/0.4**
	Shop Facade	6/0.3	6/0.3	6/0.3	6/0.3	**5/0.3**	**5/0.3**
	St Mary's Church	17/0.5	17/0.5	16/0.5	17/0.5	16/0.5	**15/0.5**
	Average	22/0.4	21/0.4	20/0.4	21/0.4	21/0.4	**17/0.3**
Indoor6	scene1	34.4%	34.9%	34.9%	32.0%	30.2%	**46.4%**
	scene2a	39.8%	40.6%	46.1%	41.3%	42.5%	**60.6%**
	scene3	45.2%	43.9%	44.9%	41.3%	42.3%	**56.4%**
	scene4a	63.2%	63.9%	62.6%	63.2%	67.1%	**78.7%**
	scene5	19.5%	18.1%	17.1%	18.3%	16.9%	**22.8%**
	scene6	59.7%	60.3%	60.0%	57.2%	58.1%	**71.6%**
	Average	43.6%	43.6%	44.3%	42.2%	42.9%	**56.1%**

on par with the ensemble version of ACE (i.e., ACE (*quad.*)), but also features a smaller model size and reduced training time. Furthermore, our model's ensemble version (i.e., EGFS (*dual.*)) is able to exceed the performance of DSAC* and ACE (*quad.*) while requiring a smaller model size and less training time.

Indoor6 Dataset. Table 2 presents the evaluation results on Indoor6. EGFS significantly outperforms DSAC* and ACE across all scenes, while also achieving comparable performance to ACE (*quad.*) with a much smaller model size (4.5MB compared to 16MB). Moreover, EGFS (*dual.*) demonstrates superior performance relative to ACE (*quad.*), which further substantiates the effectiveness of EGFS in optimization through error-guided masks and confidence refinement.

6.3 Effectiveness of Error-Guided Feature Selection

To further substantiate the advantages of employing the EGFS approach, this experiment compares several baseline schemes that sample points with different reprojection error thresholds against EGFS. The results are summarized in Table 3. Specifically, these baseline schemes select sample points based on various quantiles of the reprojection error map, ranging from 30% to 70%, denoted as $Q_{0.3}(R)$ through $Q_{0.7}(R)$, respectively. In each baseline scheme, a point is chosen if its reprojection error r falls below the threshold specified by $Q_{quantile}(R)$. The results suggest that relying solely on reprojection errors does not guarantee optimal learning of scene coordinates. This is due to the fact that areas of low reprojection errors can be scattered and may not encompass entire semantic objects. In contrast, the proposed EGFS expands the points into complete semantic masks followed by confidence refinement, which enables our model to determine scene coordinates with minimal transitional and rotational errors. This experiment thus confirms the effectiveness of leveraging SAM's ability to interpret image context alongside confidence maps for enhanced sample point selection.

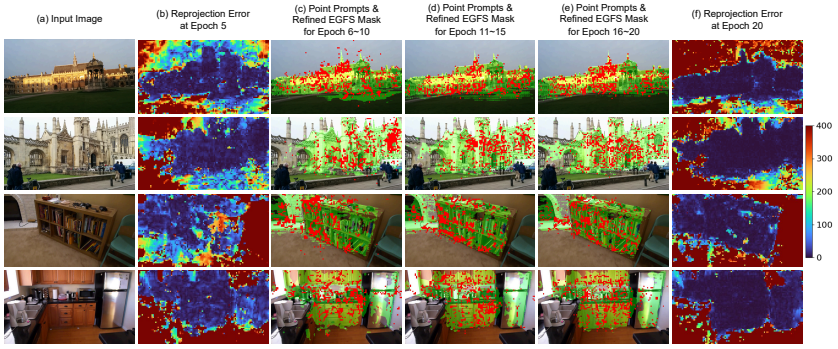

Fig. 5: Visualization of the EGFS mask refinement process at every five epochs, which depicts the reprojection errors at the beginning (epoch 5) and the end (epoch 20), as well as the refined error-guided masks used throughout training. The red dots represent low reprojection errors that serve as prompts, while the light green overlay denotes the refined EGFS masks. It can be observed that the EGFS masks enhances over epochs.

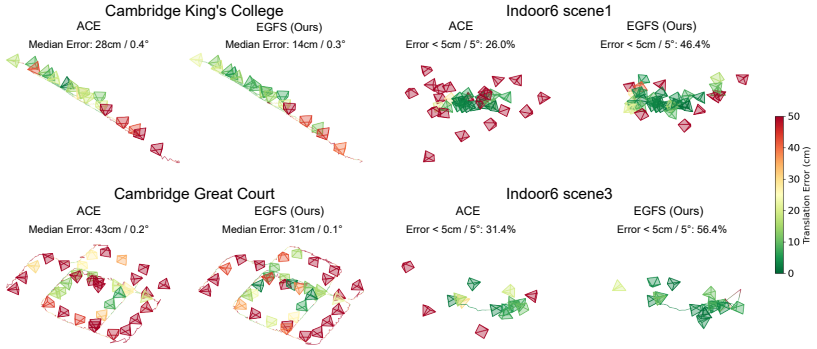

Fig. 6: Visualization of the estimated camera pose trajectories from testing sequences with the camera frustums colored based on translational errors. Pose errors denoted.

6.4 Qualitative Results

Iterative Error-guided Feature Selection Mask. Fig. 5 illustrates the refinement of EGFS masks every five epochs and compares the reprojection errors at the beginning and the end. It can be observed that EGFS is not only capable of expanding points with low reprojection errors (i.e., $r < Q_{0.1}(R)$) into regions that bear similar semantic meanings, but also refines the masks with a confidence map to exclude uncertain areas. The masks are observed to be iteratively enhanced from coarse to fine. In the beginning (i.e., at epoch five), the reprojection errors appear noisy, whereas by the end (i.e., at epoch 20), the EGFS masks become more refined and concentrated on key regions beneficial for SCR.

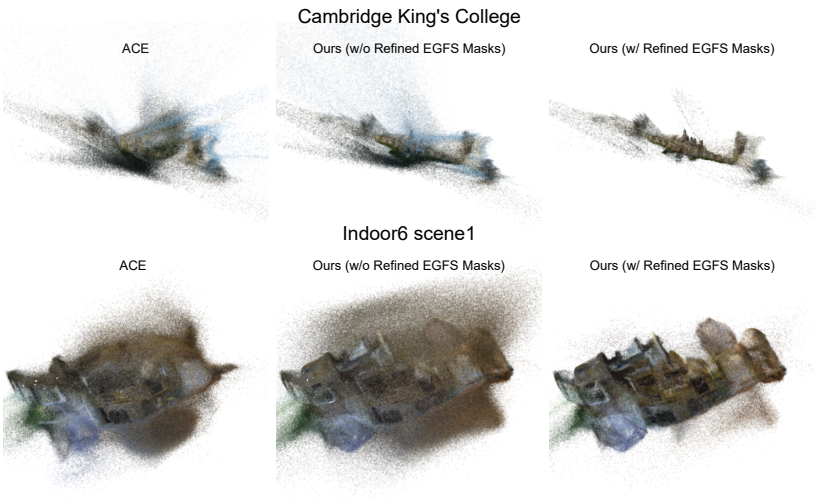

Fig. 7: Visualization of point clouds reconstructed from estimated scene coordinates of the training sequence. The point clouds derived from the scene coordinates estimated by the model trained with the proposed EGFS are clearer compared to those from the ACE model, even without the application of refined EGFS masks at inference time.

Visualization of the Estimated Camera Poses. In this section, we present the estimated camera poses from the testing sequences, and compare the proposed EGFS with the ACE baseline. The results are illustrated in Fig. 6. To visualize camera trajectories, we connect consecutive camera positions with lines and indicate camera orientations with frustums. Each sampled camera pose is associated with a color to represent the translation error. It is observed that our proposed EGFS can estimate more accurate camera poses as compared to ACE.

Scene Coordinates and 3D Point Clouds w/ and w/o EFGS. To validate the effectiveness of EGFS qualitatively and its impact on the quality of estimated scene coordinates that learned from training sequence, we depict the 3D point cloud generated from the scene coordinates from the training sequence, and apply colors based on the corresponding queried pixels in Fig. 7. To compare the 3D point clouds reconstructed from the scene coordinates, we compare the baseline ACE approach with the proposed EGFS approach. For our approach, we visualize the point clouds both before and after the application of EGFS masks and confidence maps when collecting the scene coordinates. Please note that both before and after versions which shown in the figure are trained with EGFS refinement. The 3D point cloud constructed by the proposed approach appears clearer, in contrast to the ACE-generated point cloud, which exhibits blurriness due to inaccurate scene coordinate estimation. In addition, the point cloud reconstructed from raw scene coordinates predicted without EGFS from our approach contains noise and floaters due to the inclusion of scene coordinates

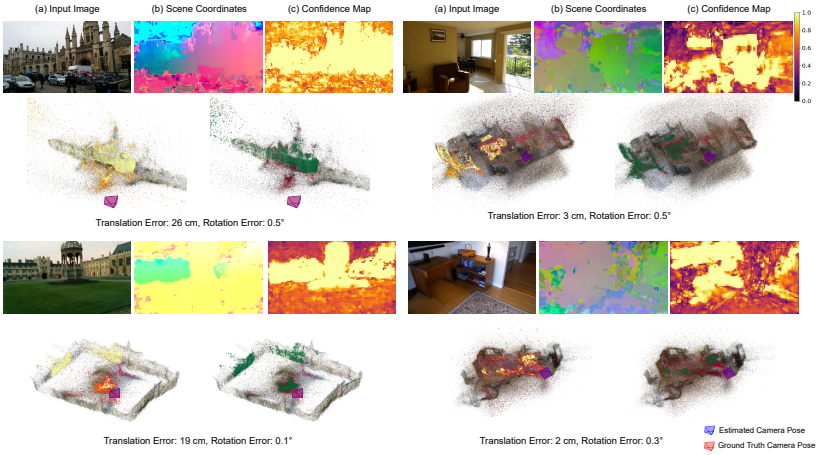

Fig. 8: Visualization of the estimated scene coordinates and the confidence maps on unseen samples. The RGB color in the scene coordinates represents the XYZ coordinates. The point clouds are colored green for selected areas and red for rejected areas.

with low confidence scores. On the other hand, the point cloud reconstructed from the scene coordiantes with EGFS refinement exhibits clearer contours and less noise. This justifies EGFS's ability to eliminate areas of low confidence values and produce high-quality dense scene coordinates.

Visualization of Estimated Confidence Map on Unseen Samples. In this section, we qualitatively evaluate the confidence maps estimated from the testing sequence by visualizing them in both the 2D image plane and as 3D projections onto the point cloud. The visualized point cloud is reconstructed from the training sequence, whereas the plotted samples are selected from the testing sequence and were not included in the training procedure. These visualizations are provided in Fig. 8. It can be observed that the estimated confidence map assists in rejecting scene coordinates located in areas where accurate prediction is challenging, such as in the air or on texture-less surfaces.

6.5 Ablation Study

Error-Guided Feature Selection (EGFS) and Confidence Refinement. We first present an ablation analysis to validate the effectiveness of the EGFS masks and confidence refinement in enhancing our method's performance. Table 4 reports the average results on the Cambridge Landmarks and Indoor6 datasets. The results demonstrate that the model trained with the proposed EGFS mechanism outperforms those trained without it, which indicates that focusing on training with robust features leads to improved performance. Furthermore, the incorporation of the confidence map also enhances the overall

Table 4: Ablation study on the impact of each proposed components.

Error-Guided Masks	*Confidence Refinement*	Cambridge (cm/°)	Indoor6 (%)
✗	✗	25/0.4	35.6
✓	✗	19/0.3	41.6
✗	✓	19/0.4	50.9
✓	✓	**17/0.3**	**56.1**

Table 5: Impact of varying proportions of point prompts (τ).

Proportions of point prompts (τ)	Cambridge (cm/°)	Indoor6 (%)
5	**17/0.3**	54.3
10	**17/0.3**	**56.1**
15	**17/0.3**	55.4
20	**17/0.3**	55.3

performance. The final row of Table 4 demonstrates that combining both techniques (i.e., our EGFS) achieves the best results, which further confirms the effectiveness of EGFS in improving localization accuracy.

Analysis of Proportions of Point Prompts. We evaluate the performance of different proportions of point prompts used for expanding into EGFS masks, and the results are reported in Table 5. It can be observed that the proportions of point prompts do not significantly impact the performance on the Cambridge Landmark dataset. This may be attributed to the fact that the scenes in the dataset typically feature a primary architectural structure with fewer complicated details in the surroundings. On the other hand, for the Indoor6 dataset, the proportions of point prompts significantly affect the performance. The rationale behind these observations is that the indoor scenes are more complicated, and feature various small items that necessitate a larger number of low-error point prompts to comprehensively capture the entire scene structure. Please note that we employ $\tau = 10$ as our setting for all the experiments presented.

7 Conclusion

This paper addressed key challenges in SCR for visual localization, and specifically focused on the impacts of dynamic objects and texture-less regions. Our approach introduced an innovative EGFS mechanism through the use of SAM and confidence maps to enhance the performance of SCR. This technique effectively filtered out problematic areas by concentrating on regions with low reprojection errors. Moreover, we used confidence maps to further refine the pixels selected by EGFS and perform this process iteratively to enable dynamic updates of the focused regions. The experimental results on the Cambridge Landmarks and Indoor6 datasets suggest that our method can provide improvements in terms of training efficiency, model size, and accuracy. Our findings highlighted the importance of carefully selecting low-reprojection error pixels by taking into account semantic information and confidence scores. Furthermore, we quantitatively and qualitatively demonstrated that dynamically updating the masks enables more robust selection of error points, thus enabling better training efficiency and effectiveness. Our ablation studies validated the effectiveness of the techniques adopted in our method, and solidified their roles in enhancing the performance.

Acknowledgements

The authors gratefully acknowledge the support from the National Science and Technology Council (NSTC) in Taiwan under grant numbers MOST 111-2223-E-007-004-MY3, 113-2221-E-007-122-MY3, 113-2640-E-002-003, 113-2221-E-007-104-MY3, as well as the financial support from Woven by Toyota, Inc., Japan. The authors would also like to express their appreciation for the donation of the GPUs from NVIDIA Corporation and NVIDIA AI Technology Center (NVAITC) used in this work. Furthermore, the authors extend their gratitude to the National Center for High-Performance Computing (NCHC) for providing the necessary computational and storage resources.

References

1. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.
2. 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. 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*, 1981.
4. Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.
5. 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.
6. Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 224–236, 2018.
7. Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8922–8931, 2021.
8. Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.
9. Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. *arXiv preprint arXiv:2306.13643*, 2023.
10. Pablo Speciale, Johannes L Schonberger, Sing Bing Kang, Sudipta N Sinha, and Marc Pollefeys. Privacy preserving image-based localization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5493–5503, 2019.
11. Xiaotian Li, Juha Ylioinas, Jakob Verbeek, and Juho Kannala. Scene coordinate regression with angle-based reprojection loss for camera relocalization. In *Proc. European Conf. on Computer Vision Workshop (ECCVW)*, 2018.
12. Xiaotian Li, Juha Ylioinas, and Juho Kannala. Full-frame scene coordinate regression for image-based localization. *RSS*, 2018.

13. Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. DSAC-Differentiable RANSAC for camera localization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.
14. Eric Brachmann and Carsten Rother. Learning less is more - 6D camera localization via 3D surface regression. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.
15. Eric Brachmann and Carsten Rother. Visual camera re-localization from RGB and RGB-D images using DSAC. *TPAMI*, 2021.
16. Eric Brachmann and Carsten Rother. Expert sample consensus applied to camera re-localization. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2019.
17. Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using RGB and poses. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.
18. Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, December 2015.
19. Tien Do, Ondrej Miksik, Joseph DeGol, Hyun Soo Park, and Sudipta N. Sinha. Learning to detect scene landmarks for camera localization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.
20. Abner Guzman-Rivera, Pushmeet Kohli, Ben Glocker, Jamie Shotton, Toby Sharp, Andrew Fitzgibbon, and Shahram Izadi. Multi-output learning for camera relocalization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.
21. Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and carsten Rother. Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.
22. Julien Valentin, Matthias Niessner, Jamie Shotton, Andrew Fitzgibbon, Shahram Izadi, and Philip H. S. Torr. Exploiting uncertainty in regression forests for accurate camera relocalization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.
23. Tao Xie, Kun Dai, Siyi Lu, Ke Wang, Zhiqiang Jiang, Jinghan Gao, Dedong Liu, Jie Xu, Lijun Zhao, and Ruifeng Li. OFVL-MS: Once for visual localization across multiple indoor scenes. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2023.
24. Zhaoyang Huang, Han Zhou, Yijin Li, Bangbang Yang, Yan Xu, Xiaowei Zhou, Hujun Bao, Guofeng Zhang, and Hongsheng Li. VS-Net: Voting with segmentation for visual localization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.
25. Tien Do and Sudipta N. Sinha. Improved scene landmark detection for camera localization. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2024.
26. Paul-Edouard Sarlin, Ajaykumar Unagar, Måns Larsson, Hugo Germain, Carl Toft, Victor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the Feature: Learning Robust Camera Localization from Pixels to Pose. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.
27. Mohammad Altillawi. PixSelect: Less but reliable pixels for accurate and efficient localization. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 4156–4162. IEEE, 2022.

28. Fei Xue, Ignas Budvytis, Daniel Olmeda Reino, and Roberto Cipolla. Efficient large-scale localization by global instance recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 17348–17357, 2022.
29. Fei Xue, Ignas Budvytis, and Roberto Cipolla. SFD2: Semantic-guided feature detection and description. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5206–5216, 2023.
30. Bach-Thuan Bui, Dinh-Tuan Tran, and Joo-Ho Lee. D2S: Representing local descriptors and global scene coordinates for camera relocalization. *arXiv preprint arXiv:2307.15250*, 2023.
31. Son Tung Nguyen, Alejandro Fontan, Michael Milford, and Tobias Fischer. FocusTune: Tuning visual localization through focus-guided sampling. In *Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 3606–3615, January 2024.
32. Johanna Wald, Torsten Sattler, Stuart Golodetz, Tommaso Cavallari, and Federico Tombari. Beyond controlled environments: 3d camera re-localization in changing indoor scenes. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 467–487, 2020.
33. Siyan Dong, Qingnan Fan, He Wang, Ji Shi, Li Yi, Thomas Funkhouser, Baoquan Chen, and Leonidas J Guibas. Robust neural routing through space partitions for camera relocalization in dynamic indoor environments. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.
34. Donghwan Lee, Soohyun Ryu, Suyong Yeon, Yonghan Lee, Deokhwa Kim, Cheolho Han, Yohann Cabon, Philippe Weinzaepfel, Nicolas Guerin, Gabriela Csurka, and Martin Humenberger. Large-scale localization datasets in crowded indoor spaces. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3227–3236, June 2021.
35. Hongyi Fan, Joe Kileel, and Benjamin Kimia. On the instability of relative pose estimation and RANSAC's role. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8935–8943, June 2022.
36. Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.
37. Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. EfficientViT: Lightweight multi-scale attention for high-resolution dense prediction. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 17302–17313, 2023.
38. Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.
39. Sheng Wan, Tung-Yu Wu, Wing H Wong, and Chen-Yi Lee. Confnet: predict with confidence. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2921–2925. IEEE, 2018.
40. 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 (CVPR)*, pages 929–939, June 2023.
41. Revaud, Jérôme and Cabon, Yohann and Brégier, Romain and Lee, JongMin and Weinzaepfel, Philippe. SACReg: Scene-Agnostic Coordinate Regression for Visual Localization, 2023.
42. Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2016.