Exploring Matching Rates: From Keypoint Selection to Camera Relocalization

Hu Lin
Dalian University of Technology
Dalian, Liaoning, China
linhu@mail.dlut.edu.cn

Chengjiang Long*†
Meta Reality Labs
Burlingame, CA, USA
cjfykx@gmail.com

Yifeng Fei
Dalian University of Technology
Dalian, Liaoning, China
fyf0702@mail.dlut.edu.cn

Qianchen Xia*
Tsinghua University
Beijing, China
qianchenxia@tsinghua.edu.cn

Erwei Yin
Tianjin Artificial Intelligence Innovation Center
Tianjin, China
yinerwei1985@gmail.com

Baocai Yin
Beijing University of Technology
Beijing, China
Dalian University of Technology
Dalian, Liaoning, China
ybc@bjut.edu.cn

Xin Yang*
Key Laboratory of Social Computing and Cognitive Intelligence (Dalian University of Technology)
Ministry of Education
Dalian, Liaoning, China
xinyang@dlut.edu.cn

Abstract

Camera relocalization is a challenging task to estimate camera pose within a known scene, with wide applications in the fields of Virtual Reality (VR), Augmented Reality (AR), robotics, and etc. Most existing learning-based methods invariably utilize all the information within an image for pose estimation. Although these methods have demonstrated leading pose accuracy in some cases, they are still far from being sufficient to handle the robustness under challenging viewpoints with less impacts on the localization accuracy for viewpoints that are easier to localize. In this paper, we propose a novel two-branch camera pose estimation framework: one branch utilizes keypoint-guided partial scene coordinate regression, while the other employs full scene coordinate regression to assess the credibility of image poses, thereby enabling more accurate camera localization. In particular, we devise a keypoint selection method predicated on matching rates which is designed to measure the matching quality between a 3D keypoint and 2D keypoints across views. With these selected 3D keypoints, we can generate 2D supervision mask with the ground-truth camera pose to supervise the keypoint prediction from the keypoint selection network. Meanwhile, we further refine the 2D supervision mask through the optimization with reprojection errors on the scene coordinate network, which estimates the scene coordinates for points within the scene that truly warrant attention, also enhances the localization performance. We also introduce a gated camera pose estimation strategy on the two-branch pose estimation framework, employing an updated keypoint selection network for images with higher credibility and a more robust network for difficult viewpoints. By adopting an effective curriculum learning scheme, we achieve higher accuracy within a training span of just 20 minutes. Our method's superior performance is validated through rigorous experimentation. The code is released at https://github.com/DUT-ICCD/KP-Guided-Reloc.

CCS Concepts

• **Computing methodologies** → **Perception**; *Mixed / augmented reality*; *Virtual reality*; *Machine learning*.

Keywords

Camera relocalization, scene coordinates regression, keypoint selection, keypoint sets, keypoint guided

ACM Reference Format:
Hu Lin, Chengjiang Long, Yifeng Fei, Qianchen Xia, Erwei Yin, Baocai Yin, and Xin Yang. 2024. Exploring Matching Rates: From Keypoint Selection to Camera Relocalization. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3664647.3681628

*Corresponding Authors: Chengjiang Long (cjfykx@gmail.com), Qianchen Xia (qianchenxia@tsinghua.edu.cn), Xin Yang (xinyang@dlut.edu.cn).
†Dr. Long contributed to this work while employed at JD Finance America Corporation.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681628

1 Introduction

Certainty is a concept of profound importance in human endeavors, providing solace through the predictability of contemporary life. This pursuit of certainty is particularly evident in the field of camera relocalization, where there is a strong desire for pose estimation to achieve a similar degree of reliability. Camera relocalization involves recovering the six degrees of freedom (6-DOF) camera pose of a query image within a known scene. This functionality is crucial in various applications such as VR [45], AR [22], robotics [43, 48], and autonomous navigation, and therefore attracts a lot of researchers to constantly put efforts to make progress towards both camera relocalization accuracy and efficiency.

The evolution of camera relocalization began with image retrieval [35], where the primitive form, scene recognition, is relied on robust image retrieval for localization. The introduction of random regression forests [17, 38], capable of regressing scene coordinates, significantly improved camera relocalization accuracy. Later, deep learning frameworks [23–25] utilize deep networks for direct pose regression from images. Then, a new mainstream camera relocalization based on scene coordinate regression [3, 5–8] were developed with a network designed as the specific function of correlating input imagery with scene coordinates. With RANSAC plus PnP, the camera pose can be estimated indirectly from the selected 2D-3D keypoint correspondences. Although the above mentioned methods are able to estimate the camera poses in some cases, they are still far from sufficient to meet the high-accuracy requirements of real-world applications.

We argue that not full scene coordinates are required and selected partial scene coordinates might further improve the 2D-3D keypoints correspondences for pose estimation with RANSAC and PnP. As illustrated in Figure 1, we propose generating keypoints-guided scene coordinates, which requires integrating a learning-based keypoint selection process into the entire camera relocalization pipeline. This approach raises several questions, *i.e.*, *what are the criteria for keypoint selection? which keypoints should be chosen?* and *how should the network be trained to identify these keypoints?* It is worth mentioning that we also need to handle the special cases when only a smaller number of keypoints available to ensure the generalization performance.

In this paper, we design a novel camera pose estimation framework based on matching rates to select points genuinely suited for camera relocalization and introduced a keypoint selection network along with its corresponding training framework. As illustrated in Figure 2, we start selecting keypoints for camera relocalization and develop methods to assess the likelihood of obtaining accurate camera poses, thereby providing credibility support for practical applications. This approach not only allows for the evaluation of the reliability of pose estimations but also enhances the accuracy.

In particular, we explore the matching rates to evaluate each 3D point from the point clouds obtained via structure from motion (SfM) reconstruction to initially filter out those low-quality 3D keypoints. With the selected 3D keypoints, we can reproject them into a 2D space as supervision mask with the ground-truth camera pose as guidance to optimize the 2D keypoint selection network. Such a 2D supervision mask is further refined via reprojection errors via a learned scene coordinate network.

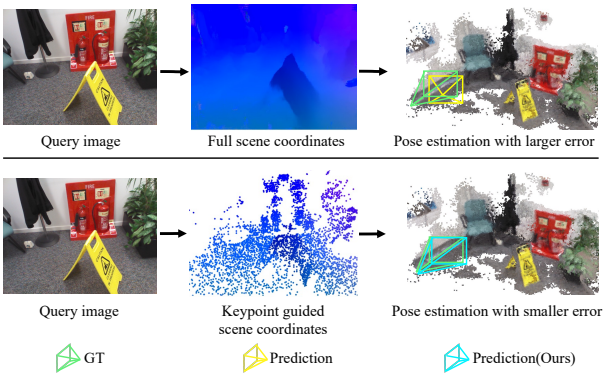

Figure 1: The camera relocalization task involves recovering the 6-DOF camera pose of a query image within a known scene. Unlike the existing scene coordinate regression based methods like ACE [9] predicting camera poses from the full scene coordinates (top), we propose to use partial scene coordinates instead (bottom) with 2D keypoints generated from a keypoint selection network where we can fully explore confidence measures and keypoint cues into the pose estimation process to enhance credibility in practical applications and improve the camera relocalization accuracy.

With the 2D keypoint mask predicted, we feed it together with the extracted feature from the feature backbone into the scene coordinate regression network to generate the partial scene coordinates, which can be used to predict the camera pose via RANSAC and PnP. Based on the predicted camera pose, we introduce the inlier ratio to assess the confidence on the estimated pose.

We shall emphasize that our proposed pose estimation framework is two-branch with the a well-designed gating mechanism upon the inlier ratio based confidences. Especially when the confidence of the inlier ratio on the estimated pose from the keypoint-guided branch is lower than the preset threshold, we resort to the generic full scene coordinates branch to avoid the final camera pose biasing towards low-quality keypoints heavily. The experiments conducted on both indoor and outdoor scenarios have clearly validated the precision and generalization ability of our proposed two-branch framework.

Our contributions can be summarized in four-fold as follows:

- We explore matching rates to select the high-quality 3D points to generate 2D supervision masks for the keypoint selection network.
- We propose to estimate camera pose from a partial scene coordinates with 2D keypoints as guidance, and output the confidence on the estimated pose based on the inlier ratio, which we introduce to enhance the reliability of pose estimations.
- We design a two-branch pose estimation framework with a gating mechanism to further improve the precision and evaluative capabilities of camera relocalization.
- We validate our proposed approach with superior performance to the competing state-of-the-art methods, *e.g.*, ACE [3], on both indoor and outdoor benchmark datasets.

2 Related Work

The related work involves *Camera Relocalization* and two relevant components: *Keypoint Selection* and *Pose Confidence*, which aim to determine better camera poses with increased confidence.

2.1 Camera Relocalization

The current mainstream camera relocalization approaches include image or feature retrieval-based methods, pose regression methods, and scene coordinate regression methods.

Image retrieval. The implementation of image retrieval methods involves searching for a query image [35] from a database of images with known poses to retrieve similar images, outputting the pose of the most similar retrieved image. Arandjelovic et al. proposed VLAD [41] for global image feature description. NetVLAD [1] obtains improvement upon VLAD features with features extracted by a CNN model. InLoc [39] optimizes pose estimation using synthesized virtual views. R2Former [50] handles both retrieval and re-ranking with a novel transformer model.

Absolute pose regression. Compared to the two-step process of extracting features from images and then performing retrieval, the PoseNet series [23–25] utilizes convolutional neural networks to extract features from images and subsequently to directly regress camera poses. Marepo [13] integrates scene coordinate regression and absolute pose regression methods by performing pose regression on the regressed scene coordinates. The main challenges of the above methods are the accuracy [34] of pose estimation.

Scene coordinate regression. Method [38] uses a random forest composed of regression decision trees to regress scene coordinates, which represent the coordinates of image pixels in the scene model. Then perform robust pose estimation using PnP within a RANSAC algorithm. Method [12] based on random forests enables real-time learning of new scenes. Method [17] enhances camera relocalization in dynamic scenes. However, using depth as input features in random forest methods complicates data acquisition.

Deep learning-based methods attempt to minimize reliance on depth as much as possible. In DSAC [5], a probabilistic model derived from reinforcement learning makes the optimal selection of RANSAC differentiable. DSAC++ [6] improves the network model, training methods, and data representation, further enhancing accuracy. ESAC [7] uses a hybrid expert model based on [5] to address the coverage of large datasets and ambiguity problems. SANet [44] proposes a scene-agnostic neural architecture that learns to construct hierarchical scene representations. DSAC* [8] improves [5] by using a better ResNet and an improved training loss function. AE-CRN [26] introduces RWEI to represent event data[46, 47], enabling its effective application to scene coordinate regression. SLD [15] attempts to implicitly encode the observations of scene landmarks into a CNN. SLD* [16] mitigated the issues of insufficient model capacity and noisy labels in SLD [15]. D2S [10] attempts to perform scene coordinate regression on hand-crafted features. CROSS-FIRE [28] utilizes dense local features obtained through Neural Radiance Fields (NeRF) rendering for matching. ACE [3] improves the DSAC series of methods [5–8] by separating feature extraction and scene coordinate regression. The ACE Zero [9] has designed an iterative loop for scene reconstruction [42] and pose estimation within its pipeline.

However, almost all of the aforementioned methods aim to utilize as much information from the image as possible. Specifically, among all scene coordinate regression approaches, there is a general tendency to estimate scene coordinates based on uniform sampling without distinguishing their importance, thereby delegating this challenging task to manually designed PnP and RANSAC algorithms. We believe that the selection of keypoints can also be achieved through learning-based methods, and these keypoints should be specially designed to select those that contribute to camera relocalization. The most significant difference between our approach and previous methods is that our method focuses more on **the truly important points in the scene for relocalization**, rather than solely on a global and extensive scene representation.

2.2 Keyoint Selection

Typical hand-crafted methods for keypoint description include SIFT [27], SURF [2], and ORB [30]. Currently, some learning-based methods have attempted to select keypoints in images. The work SuperPoint [14] learns to find corners by generating a set of synthetic shapes annotated with corners. R2D2 [29] proposes to jointly learn reliable and repeatable detectors and descriptors. SiLK [20] defines keypoints using high matching probability and models the cycle matching probability using a double softmax.

However, these methods address the more general problem of keypoint selection, description, and matching. For visual relocalization tasks, specifically designed approaches are more suitable [3]. Inspired by SiLK [20], we design a keypoint selection method based on the matching rates, which is capable of **selecting keypoints that are better suited for camera relocalization.**

2.3 Pose Confidence

Currently, there is limited work on confidence estimation for pose estimation, with most research focusing on pose verification based on multiple views. InLoc [39] performs pose verification using view synthesis. PV [40] focuses on pose verification, significantly enhancing it by combining different modalities, namely appearance, geometry, and semantics. MPV [21] re-estimates the pose iteratively by reorganizing local features and performing local feature matching in similar views. The paper [19] estimates pose confidence through the number of inliers and the spatial distribution of inliers.

Unlike methods that rely on the number of matches or inliers, or the distribution of inliers, our proposed evaluation method is based on a ratio inspired by Lowe's ratio test in SIFT [27]. We first use a network to learn from ground truth data to determine which points are likely to be suitable for pose estimation. Then, we use PnP and RANSAC to filter the actual inliers, allowing **our inlier ratio to be based on carefully selected data.**

3 Proposed Method

Camera relocalization involves learning the scene and estimating the 6-DOF pose of the camera based on the query image. In this paper, we propose to further enhance the accuracy of pose estimation in camera relocalization via splitting the scene coordinate regression into two parts, *i.e.*, a feature extraction network and the prediction of scene coordinates from the extracted features with

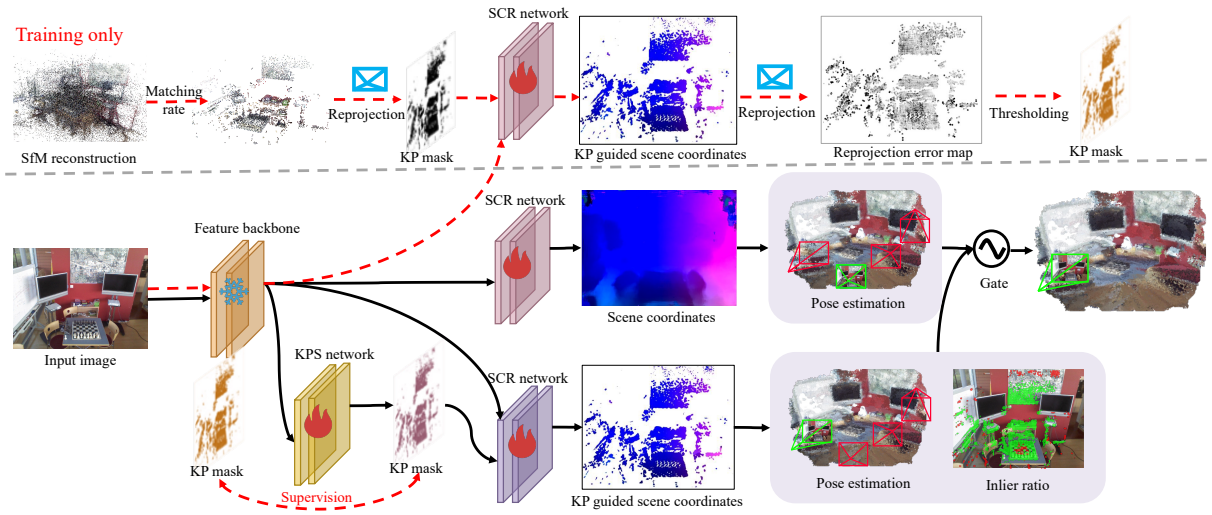

Figure 2: Our proposed pose estimation framework consists of two distinct branches with a well-designed gating mechanism for selecting the final pose. The upper generic branch is responsible for estimating the camera poses of images with a higher degree of generalizability but without guaranteed accuracy. The lower keypoint-guided branch is equipped with image confidence estimation, aimed at more accurately estimating the camera poses of images with higher confidence levels. Note that we explore the matching rates to select high-quality 3D keypoints to reproject into 2D space using the ground-truth camera pose as supervision to optimize the keypoint selection network. All query images undergo feature extraction via a feature backbone, and the extracted features are initially input into the keypoint selector network and the keypoint branch of the upper path. In the keypoint branch, only keypoints are utilized for pose estimation, generating a confidence metric, specifically the Inlier Ratio. Ultimately, in the Pose Selection phase, a gated mechanism based on the confidence values determine whether to (1) execute the generalist branch and adopt its output pose or (2) directly use the pose obtained from the keypoint-guided branch as the final pose output.

a well-designed keypoints selection module. As demonstrated in Figure 2, our proposed camera relocalization process comprises two branches. One branch is a keypoint-guided pose estimation process and the other one is a more generalized branch.

3.1 Explore Matching Rates to Generate 2D Supervision Mask for Keypoint Selection Network

To ensure the generated 2D keypoint mask from the keypoint selection network is consistent with what we expect, we first reconstruct the 3D point cloud from the multi-view images, and identify good-quality 3D keypoints by matching rates. The final 2D supervision mask is generated by reprojecting the selected 3D keypoints into the 2D space with the ground-truth camera poses.

3D keypoints selection is crucial for camera relocalization which relies on the establishment of matching relationships between cross-view images and the foundation of these relationships is the pairing of 2D keypoints. The selection and matching of keypoints also represent the primary intuition by which humans and all other visual animals perform localization. Inspired by SiLK [20], we posit that the fundamental criteria for keypoints should be their ease of matching and the likelihood of correct matches.

Matching rate. To evaluate whether a keypoint is easy to match and has a high likelihood of correct matches, we propose using the matching rates to measure these keypoints.

Let $P = \{p_1, p_2, \ldots, p_n\}$ be the set of all points in the multi-view images that successfully match with the same 3D point p, and $M(p)$ be a function that measures the matching rates of point p. The function $M(p)$ is designed to reflect the ease of matching and correctness of the matches, which can be formulated as:

$$M(p) = \sum_{i=1}^{n} w_i \cdot \mathrm{MQ}(p, p_i), \tag{1}$$

where w_i are weights assigned based on the relative importance of matching with point p_i, and $\mathrm{MQ}(p, p_i)$ is a function that returns a score representing the match quality between the 3D point p and the 2D point p_i.

$$\mathrm{MQ}(p, p_i) = \exp\left(-\frac{\|\phi(p) - \mathrm{desc}(p_i)\|^2}{2\sigma^2}\right), \tag{2}$$

where $desc(p_i)$ represents the feature descriptor of point p_i, $\phi(p) = \frac{1}{n}\sum_j^n \mathrm{desc}(p_j)$, and σ is a scaling parameter that adjusts the sensitivity of the match quality to differences in the descriptors. By default, the weights w_i and the scale σ are both set to 1.

Keypoint selection. Finally, we define the keypoints K with the matching rates as follows:

$$K = \{p \in P \mid M(p) \geq \tau\}, \tag{3}$$

where τ is a threshold value chosen based on the desired confidence level for the matches. The specific value of τ is related to the number

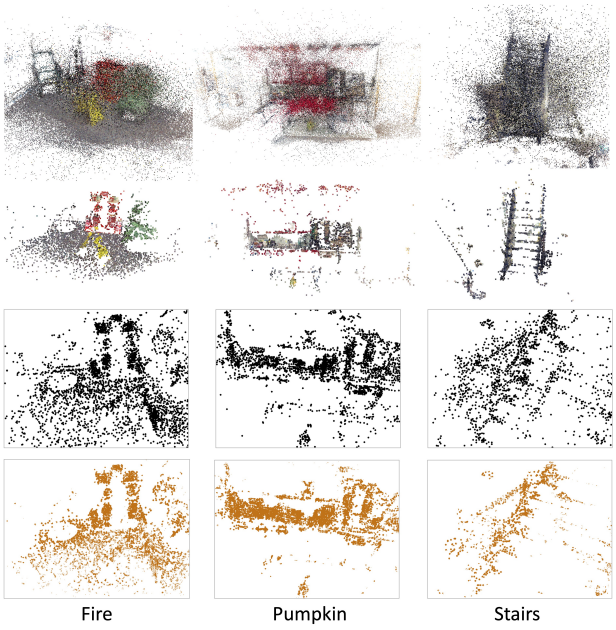

Figure 3: Visualization of the selected 3D keypoints with the matching rates and the corresponding 2D supervision masks on the 7Scenes dataset. From top to bottom are the point cloud obtained from SfM reconstruction, 3D keypoints selected via matching rates, 2D projected supervision mask, and the refined 2D supervision mask via reprojection errors, respectively.

of images used for reconstruction. Based on our observations, we suggest that the threshold should be greater than 1. We set its value as 1.5 in this paper.

We utilize COLMAP [36] to obtain 3D point collection because we realized that the poses obtained via SfM are more suitable for visual relocalization [4], as they use the same reprojection error as the loss function. Figure 3 visualizes the intermediate results on how we get the final supervision mask for the keypoint selection network. It is evident that our keypoint collection encompasses the primary and pivotal regions in the scene while disregarding areas that are deficient in texture or lack identifiable qualities.

3.2 Refine the 2D Supervision Mask with Reprojection Error from a 3D Coordinate Regression Network

Although we are able to filter out the points that should receive the most attention for the relocalization task based on the matching rate, in the keypoint branch of the entire camera relocalization framework, we need to identify these points in the query image and evaluate whether these points are still inliers from the perspective of RANSAC-based PnP procedure. Note that after feature extraction by the convolutional network, the performance of the scene coordinate regression network in regressing scene coordinates may exhibit randomness, which suggests that it does not always perform well at specific points. Therefore, we utilize the trained scene coordinate regression network to estimate the probability of these keypoints being inliers, which can further refine the keypoints selection from the keypoint selection network.

Keypoint probability. We propose a keypoint probability evaluation method based on reprojection error, which can be used to refine the generated 2D supervision mask to train the keypoint selection network. We use the following formula to convert the reprojection error into keypoint probability:

$$k_i = 1 - \frac{1}{1 + e^{\lambda - \epsilon_i}}, \tag{4}$$

where λ is the softness parameter that controls the tolerance level of the reprojection error. A smaller value indicates that a smaller reprojection error is required to achieve a higher keypoint probability. In this work, we set λ to 4. ϵ_i represents the reprojection error of the scene coordinate estimation, and k_i represents the keypoint probability of that point. This formula is inspired by the sigmoid function.

Regarding the calculation the keypoint probability, we first feed images into a generic scene coordinate estimation network to obtain the estimated scene coordinates. We then calculate the reprojection error of these scene coordinates using the ground-truth pose and apply the errors into the above Equation 4 to get the final keypoint probability. We keep out all the 2D keypoints with probability values larger than the threshold as the final 2D supervision mask.

3.3 Inlier Ratio Based Confidence Calculation on Estimated Poses

Currently, all vision-based camera relocalization methods aim to improve overall performance across the entire test set. However, it is unrealistic to expect camera relocalization methods to perform efficiently and consistently in challenging scenarios with limited sampled data or abundant repetitive textures. Therefore, we start from keypoints to evaluate the confidence of the input images and their output poses. When the confidence based on keypoint regions is higher, we should prioritize using the pose estimation results based on keypoints.

Inlier ratio. To address this, we propose an image pose confidence estimation based on the inlier ratio. The confidence estimation of an image is calculated using the following formula:

$$r_i = \frac{100 \cdot N^i}{N^a}, \tag{5}$$

where r_i represents the pose estimation confidence of image i, N^i is the number of inliers output by the RANSAC-based PnP algorithm, and N^a is the total number of 2D-3D correspondences provided as input. We consider confidence levels above 90% to be considerably confident, above 80% to be highly confident, above 60% to be moderately confident, and below 60% to be questionable.

We notice that existing methods perform poorly in terms of confidence in image pose estimation on the test set, particularly in the "stairs" scene, where the proportion of highly confident estimates is nearly zero. To address this, we propose a keypoint-guided inlier ratio estimation method. By introducing an additional network branch in the relocalization framework to evaluate the confidence

of scene coordinates. Consequently, our updated confidence estimation is given by:

$$r_i' = \frac{100 \cdot N^i}{N^{ar}}, \tag{6}$$

where N^{ar} denotes the subset of 2D-3D correspondences with reliable confidence, as estimated using the keypoint network.

3.4 Network Training

In this paper, we train the two-branch framework with a total of three networks: one to obtain keypoints with the extracted 2D supervision mask as guidance, and two scene coordinates regression networks. We adopt a simple network structure and accelerate the training of the keypoint selection network using curriculum learning. This approach enables us to achieve our goal of efficient training with minimal additional time.

We have ascertained the ground truth probabilities for keypoints at each point. We employ the Binary Cross Entropy (BCE) loss as the loss function for our keypoint selection network. This loss is applied to both the keypoint selection based on matching rates and reprojection error.:

$$L_k(k, \hat{k}) = -\frac{1}{N} \sum_{i=1}^{N} \Big(k_i \log(\hat{k}_i) + (1 - k_i) \log(1 - \hat{k}_i) \Big), \tag{7}$$

where N is the number of samples, k_i represents the ground truth keypoint probability for the i-th feature, and $\hat{k}_i$ represents the predicted keypoint probability for the i-th feature.

Consequently, during the supervision process, we are limited to methods such as reprojection error for guiding the learning of scene coordinates. Specifically, the loss used for supervising is:

$$L_\pi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}_i^*) = \begin{cases} e_\pi(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}_i^*) & \text{if } \mathbf{y}_i \in \mathcal{V} \\ ||\mathbf{y}_i - \bar{\mathbf{y}}_i||_0 & \text{otherwise.} \end{cases}, \tag{8}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the 2D-3D coordinate pairs, $\bar{\mathbf{y}}_i$ is the dummy 3D scene coordinate, $\mathbf{h}^*$ is the GT pose, and $\mathcal{V}$ is the group of 3D scene coordinate satisfied with the reprojection error e_π in [3].

We design a novel scene coordinate regression network. Enhancing the network depth of the scene coordinate regression network boosts its ability to solve scene coordinates in challenging scenarios. We follow the curriculum training technique adopted in [3] to train our scene coordinate regression network. This curriculum technique employs a dynamic inlier threshold based on reprojection error throughout the training process, which starts with a larger value and becomes progressively smaller as training progresses, making the training process increasingly strict. For more details on the network and training, please refer to the supplementary materials.

3.5 Implementation Details

Our method is implemented with PyTorch upon the publicly available code from ACE [3]. For the training of the keypoint selection network, a simple CNN network was adopted. The initial learning rate is set as 0.0001, and the AdamW optimizer is consistently employed across all network configurations. When initializing the typical scene coordinate regression network, we set up a training buffer with 8.8 million samples, with all features randomly sampled from random images and augmentations. For training the keypoint selection network, an 8 million sample training buffer was established, drawing features from randomly chosen images.

Note that when training the scene coordinate regression network for 2D supervision masks, we utilize all outputs from the intial keypoint masks projected from the selected 3D keypoints via matching rates with the ground-truth camera pose. During training the keypoint-guided scene coordinate regression network, we fine-tune the initialized head rather than starting the training from scratch, thus preserving the network's capability to regress scene coordinates for non-keypoint features. We follow the implementation of ACE [3] on the pose estimator but with some changes to accommodate our approach. Employing the keypoint selection network as a guide, we select points when their estimated probability value are greater than 0.8, which we observe the corresponding reprojection error is approximately 5.0.

4 Experiments

We conduct experiments on both the indoor 7Scenes [38] dataset and the outdoor Cambridge Landmarks [25][1], and report the performance with camera relocalization accuracy for errors within $(5^\circ, 5cm)$, $(2^\circ, 2cm)$ and $(1^\circ, 1cm)$.

4.1 Indoor Relocalization

The 7Scenes [38] dataset offers a variety of small-scale indoor scenes captured with handheld devices, along with depth information and camera poses. We utilize the camera poses obtained via structure from motion (SfM) provided in [4] as ground truth (GT). Besides the primary baseline ACE [3], we also compare against other scene coordinate regression approaches. As summarized in Table 2, we present the percentage performance of our method and others on the 7Scenes [38] dataset within $(5^\circ, 5cm)$ of error. We can clearly see that our proposed method is able to further enhance relocalization performance, especially in challenging scenes such as "stairs". We also report the performance of our method compared to others within $(2^\circ, 2cm)$ and $(1^\circ, 1cm)$ error margins in Table 3, respectively. As we can observe, our method not only excels in percentage performance within the $(5^\circ, 5cm)$ error margin but also significantly improves the precision of pose estimation. Such observations suggest that the precision of camera pose estimates from our method has been further elevated, offering a guarantee for the practical application of camera relocalization technology in real-world scenarios. As shown in Figure 4, we compare the differences in estimated translations between our method and the comparison methods. It is evident that our estimated trajectories are more accurate, as indicated by the less prominent green lines.

4.2 Outdoor Relocalization

The Cambridge Landmarks dataset [25] consists of images of various historical buildings in the old town area of Cambridge, with ground truth poses obtained using the SfM [4]. As we can observe

[1]The two datasets were received and exclusively accessed by the authors Hu Lin and Prof. Xin Yang for purely academic research only. Hu Lin produced the experimental results in this paper. Meta did not have access to the datasets as part of this research.

Table 1: Pose relocalization results compare with other methods on Cambridge Landmarks dataset.

Method	Publication	Year	Type	Depth	Cambridge Landmarks					Average
					Court	King's	Hospital	Shop	St.Mary's	(cm/°)
AS (SIFT) [33]	TPAMI	2016	FM	X	24/0.1	13/0.2	20/0.4	4/0.2	8/0.3	14/0.2
hLoc(SP+SG) [31]	CVPR	2019	FM	X	16/0.1	12/0.2	15/0.3	4/0.2	7/0.2	11/0.2
pixLoc [32]	CVPR	2021	FM	X	30/0.1	14/0.2	16/0.3	5/0.2	10/0.3	15/0.2
GoMatch [49]	ECCV	2022	FM	X	N/A	25/0.6	283/8.1	48/4.8	335/9.9	N/A
HybridSC [11]	CVPR	2019	FM	X	N/A	81/0.6	75/1.0	19/0.5	50/0.5	N/A
PoseNet17 [24]	CVPR	2017	APR	X	683/3.5	88/1.0	320/3.3	88/3.8	157/3.3	267/3.0
MS-Transformer [37]	ICCV	2021	APR	X	N/A	83/1.5	181/2.4	86/3.1	162/4.0	N/A
SANet [44]	ICCV	2019	SCR	✓	328/2.0	32/0.5	32/0.5	10/0.5	16/0.6	84/0.8
SRC [18]	3DV	2022	SCR	✓	81/0.5	39/0.7	38/0.5	19/1.0	31/1.0	42/0.7
DSAC* [8]	TPAMI	2021	SCR	X	98/0.5	27/0.4	33/0.6	11/0.5	56/1.8	45/0.8
ACE [3]	CVPR	2023	SCR	X	**43/0.2**	28/0.4	31/0.6	**5/0.3**	18/0.6	25/0.5
Ours	ACM MM	2024	SCR	X	46/0.5	**21/0.4**	**23/0.6**	5/0.4	**12/0.6**	**22/0.5**

Table 2: Pose relocalization results compared with other methods on 7Scenes dataset.

Scene	DSAC	DSAC++	Cas.	DSAC*	ACE	**Ours**
Chess	94.6%	93.8%	**100%**	96.7%	**100%**	**100%**
Fire	74.3%	75.6%	99.7%	92.9%	99.5%	**99.9%**
Heads	71.7%	18.4%	**100%**	98.2%	99.7%	**100%**
Office	71.2%	75.4%	99.5%	87.1%	**100%**	99.5%
Pumk.	53.6%	55.9%	90.9%	60.7%	99.9%	**99.9%**
Redki.	51.2%	50.7%	90.7%	65.3%	98.2%	**99.7%**
Stairs	4.5%	2.0%	**94.2%**	64.1%	81.9%	88.6%
Avg.	60.2%	60.4%	96.4%	80.7%	97.0%	**98.2%**

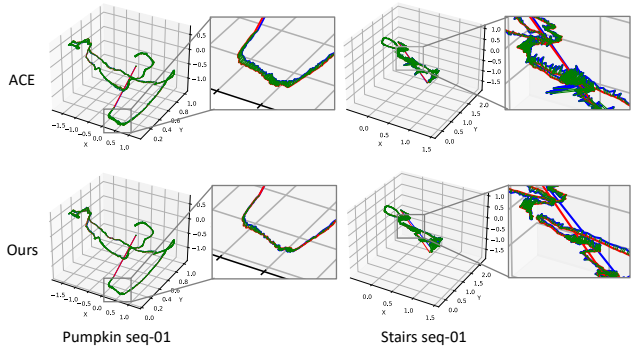

Figure 4: The difference between the pose trajectories estimated by our proposed method and the ACE compared to the GT pose trajectory. The blue represents the estimated translations, the red represents the GT translations, and the green indicates the discrepancies between the corresponding translations. The more prominent the green lines, the greater the difference.

from Table 1, our proposed method performs exceptionally well in some scenes, outperforming our main competing method ACE [3] in most scenarios. This is attributed to our network's enhanced

Table 3: Results under smaller thresholds. We further detail the camera relocalization accuracy for errors within (2°, 2cm) and (1°, 1cm).

	Within (2°, 2cm)			Within (1°, 1cm)		
	DSAC*	ACE	Ours	DSAC*	ACE	Ours
Chess	32.8%	99.0%	99.6%	0.5%	81.9%	91.8%
Fire	55.2%	87.1%	95.2%	14.8%	57.0%	63.9%
Heads	87.3%	98.2%	98.9%	40.0%	85.3%	87.1%
Office	32.1%	81.0%	91.2%	5.9%	28.0%	60.4%
Pumpk.	19.8%	84.5%	88.9%	4.7%	27.0%	60.7%
Redki.	14.9%	87.0%	94.9%	2.6%	45.5%	73.9%
Stairs	11.4%	24.1%	38.0%	1.1%	4.0%	8.1%
Average	36.2%	80.1%	86.7%	9.9%	47.0%	63.7%

capability for scene coordinate inference. During our experiments, we also observe that a significant reason for poorer performance on outdoor data was the sparsity of the dataset. Our method show commendable performance on the training set, yet its capacity to generalize to new viewpoints was somewhat limited.

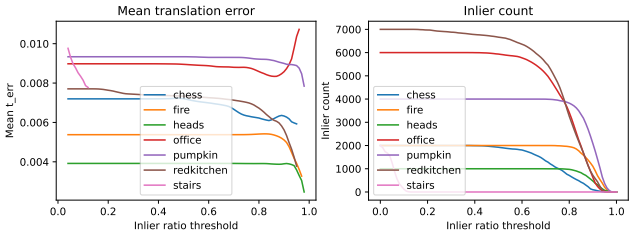

Figure 5: The mean translation error and inlier count beyond the inlier ratio.

4.3 Ablation Study

4.3.1 The relationship between inlier ratio and error. Our method introduces a confidence-based scene coordinate estimation approach. We ponder the potential outcomes if we solely rely on the poses

Table 4: Survival rate and camera relocalization accuracy. We measure the survival rate across three different stages, defined as the ratio of pose estimates that meet our set confidence criteria, in conjunction with the accuracy of the pose estimation. We set the confidence threshold to an inlier ratio exceeding 90%. "Init." denotes the network post-initialization, utilizing all coordinate estimates for pose estimation; "Init. With K.S." refers to the post-initialization network employing keypoint selection for pose estimation; "Keypoint fine-tune with K.S." indicates the network after keypoint fine-tuning, using keypoint selection for pose estimation. Note that almost all pose estimates deemed to meet the confidence criteria fall within (5°,5cm), and hence we only list the percentages for errors within (2°,2cm) and (1°,1cm).

	Init.			Init. with K.S.			Keypoint fine-tune with K.S.		
	Survival Rate	2°2cm	1°1cm	Survival Rate	2°2cm	1°1cm	Survival Rate	2°2cm	1°1cm
chess	97/2000	100%	91.8%	1714/2000	99.8%	93.6%	1718/2000	99.8%	93.9%
fire	92/2000	100%	100%	733/2000	93.7%	81.2%	736/2000	95.4%	83.4%
heads	0/1000	0.0%	0.0%	238/1000	100%	98.7%	236/1000	100%	99.2%
office	20/4000	100%	25.0%	3193/4000	86.1%	35.1%	3197/4000	87.4%	42.5%
pumpkin	48/2000	100%	22.9%	1393/2000	94.5%	39.8%	1388/2000	97.6%	54.7%
redkitchen	38/5000	100%	100%	3501/5000	94.6%	50.2%	3483/5000	96.4%	63.6%
stairs	0/1000	0.0%	0.0%	9/1000	100%	0.0%	29/1000	100%	24.1%
Average	42/2429	71.4%	48.5%	1540/2429	95.5%	56.9%	1541/2429	96.7%	65.9%

of images with high confidence. To validate the appropriateness of our confidence measure, we plotted a curve representing the relationship between the confidence utilized and the translational error, as illustrated in Figure 5. For clarity, we scaled and averaged the data. It is observable from the graph that an increase in error tends to follow a decrease in confidence. Although this correlation is not absolute, our observations suggest that there is generally an inverse relationship between inlier ratio and error, implying that a higher inlier ratio may be associated with increased error.

4.3.2 Confidence enhancement and survival rate in keypoints selection branch. We estimate scene coordinates on keypoint and employ the inlier ratio as a measure of confidence for the precision of image pose estimation. We preserve images with a confidence level of 90% and estimate their camera poses. Table 4 lists the survival rate and accuracy of our method at various stages for test images. It is evident that the survival rate of test images was quite low initially without keypoint guidance. The introduction of keypoint guidance resulted in a significant increase in the survival rate. We also observe that the survival rate of the network, further optimized using keypoint guidance, did not decrease significantly. Meanwhile, the localization accuracy for the test data with high survival rates was further improved within error margins of $(2^\circ, 2cm)$ and $(1^\circ, 1cm)$. Moreover, within the subset of high-confidence images, we nearly achieved 100% accuracy within an error range of $(5^\circ, 5cm)$, which validates the practical reliability of our method. Additionally, we note that in the Stairs scene, our confidence measurement method was not able to maintain a high survival rate. This could be attributed to the scene's higher difficulty and the presence of more repetitive textures, a phenomenon also observed in the Heads scene.

4.3.3 Usage of time and computational. Among all training processes, the network trained to identify keypoints based on the matching rates is the most time-consuming. This is attributed to the fact that the majority of points in an image are not keypoints, precluding targeted training. Conversely, all other training activities in the keypoint selection branch are confined to the regions of keypoints identified by the matching rate. Our approach does not significantly increase training duration or computational resource usage under standard conditions. The total training time is approximately 20 minutes using two Nvidia GeForce 2080 Ti GPUs, which is still acceptable given the desired higher accuracy. During testing, when the confidence is over the threshold, *i.e.*, 0.9, our computational load is lower than that of ACE. This is because we use fewer but more accurate 2D-3D matching points, reducing the required iterations. However, if the survival rate is below the threshold, we may incur the cost of performing pose estimation twice, *i.e.*, a keypoint-based pose estimation and then a more generalized pose estimation.

4.3.4 Limitations. We found several shortcomings that require further exploration. Firstly, our method faces the challenge of a low number of keypoints in difficult scenes with sparse textures. Consequently, fine-tuning the process becomes nearly infeasible, and we heavily rely on the initialized generalized network. Secondly, our approach is limited by the confidence threshold. Currently, we use a fixed confidence threshold and its performance varies across different scenes.

5 Conclusion

Measuring the confidence of camera pose estimation is critical in practical applications. We start selecting keypoints favorable for camera relocalization and design a keypoint selection scheme based on matching rate. We then design a keypoint evaluation method based on reprojection error. Finally, our gated camera pose estimation strategy based on confidence thresholding is introduced to combine keypoint-guided networks with more generalized networks to further enhance camera relocalization accuracy. Notably, our approach does not significantly increase training duration or volume compared to state-of-the-art methods, achieving greater accuracy within a training period of just 20 minutes. Through extensive experimental comparisons, we have demonstrated the effectiveness of our proposed method, surpassing state-of-the-art results.

Acknowledgments

This work was supported in part by the grants from the National Natural Science Foundation of China (No. 62332019, No. 62076250), the National Key Research and Development Program of China (No. 2022ZD0210500, No. 2023YFF1203900, No. 2023YFF1203903), the Distinguished Young Scholars Funding of Dalian (No. 2022RJ01), and the Ningbo Major Research and Development Plan Project of China (No. 2023Z225). The authors would like to thank the reviewers for their valuable comments and suggestions, which have significantly improved the quality of this manuscript.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*. 5297–5307.
[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *ECCV*. Springer, 404–417.
[3] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. 2023. Accelerated Coordinate Encoding: Learning to Relocalize in Minutes Using RGB and Poses. In *CVPR*. 5044–5053.
[4] Eric Brachmann, Martin Humenberger, Carsten Rother, and Torsten Sattler. 2021. On the limits of pseudo ground truth in visual camera re-localisation. In *ICCV*. 6218–6228.
[5] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. 2017. Dsac-differentiable ransac for camera localization. In *CVPR*. 6684–6692.
[6] Eric Brachmann and Carsten Rother. 2018. Learning less is more-6d camera localization via 3d surface regression. In *CVPR*. 4654–4662.
[7] Eric Brachmann and Carsten Rother. 2019. Expert sample consensus applied to camera re-localization. In *ICCV*. 7525–7534.
[8] Eric Brachmann and Carsten Rother. 2021. Visual camera re-localization from RGB and RGB-D images using DSAC. *TPAMI* 44, 9 (2021), 5847–5865.
[9] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. 2024. Scene Coordinate Reconstruction: Posing of Image Collections via Incremental Learning of a Relocalizer. In *ECCV*.
[10] Bach-Thuan Bui, Dinh-Tuan Tran, and Joo-Ho Lee. 2023. D2s: Representing local descriptors and global scene coordinates for camera relocalization. *arXiv preprint arXiv:2307.15250* (2023).
[11] Federico Camposeco, Andrea Cohen, Marc Pollefeys, and Torsten Sattler. 2019. Hybrid scene compression for visual localization. In *CVPR*. 7653–7662.
[12] Tommaso Cavallari, Stuart Golodetz, Nicholas A Lord, Julien Valentin, Victor A Prisacariu, Luigi Di Stefano, and Philip HS Torr. 2019. Real-time RGB-D camera pose estimation in novel scenes using a relocalisation cascade. *TPAMI* 42, 10 (2019), 2465–2477.
[13] Shuai Chen, Tommaso Cavallari, Victor Adrian Prisacariu, and Eric Brachmann. 2024. Map-Relative Pose Regression for Visual Re-Localization. In *CVPR*. 20665–20674.
[14] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2018. Superpoint: Self-supervised interest point detection and description. In *CVPRW*. 224–236.
[15] Tien Do, Ondrej Miksik, Joseph DeGol, Hyun Soo Park, and Sudipta N Sinha. 2022. Learning to detect scene landmarks for camera localization. In *CVPR*. 11132–11142.
[16] Tien Do and Sudipta N Sinha. 2024. Improved Scene Landmark Detection for Camera Localization. In *3DV*. IEEE, 975–984.
[17] Siyan Dong, Qingnan Fan, He Wang, Ji Shi, Li Yi, Thomas Funkhouser, Baoquan Chen, and Leonidas J Guibas. 2021. Robust neural routing through space partitions for camera relocalization in dynamic indoor environments. In *CVPR*. 8544–8554.
[18] Siyan Dong, Shuzhe Wang, Yixin Zhuang, Juho Kannala, Marc Pollefeys, and Baoquan Chen. 2022. Visual localization via few-shot scene region classification. In *3DV*. IEEE, 393–402.
[19] Luca Ferranti, Xiaotian Li, Jani Boutellier, and Juho Kannala. 2021. Can you trust your pose? confidence estimation in visual localization. In *ICPR*. IEEE, 5004–5011.
[20] Pierre Gleize, Weiyao Wang, and Matt Feiszli. 2023. Silk: Simple learned keypoints. In *ICCV*. 22499–22508.
[21] Janghun Hyeon, Joohyung Kim, and Nakju Doh. 2021. Pose correction for highly accurate visual localization in large-scale indoor spaces. In *ICCV*. 15974–15983.
[22] Reint Jansen, Frida Ruiz Mendoza, and William Hurst. 2023. Augmented reality for supporting geo-spatial planning: An open access review. *VI* 7, 4 (2023), 1–12.
[23] Alex Kendall and Roberto Cipolla. 2016. Modelling uncertainty in deep learning for camera relocalization. In *ICRA*. 4762–4769.
[24] Alex Kendall and Roberto Cipolla. 2017. Geometric loss functions for camera pose regression with deep learning. In *CVPR*. 5974–5983.
[25] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. Posenet: A convolutional network for real-time 6-DOF camera relocalization. In *ICCV*. 2938–2946.
[26] Hu Lin, Meng Li, Qianchen Xia, Yifeng Fei, Baocai Yin, and Xin Yang. 2022. 6-dof pose relocalization for event cameras with entropy frame and attention networks. In *ACM SIGGRAPH VRCAI*. 1–8.
[27] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *IJCV* 60, 2 (2004), 91–110.
[28] Arthur Moreau, Nathan Piasco, Moussab Bennehar, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. 2023. Crossfire: Camera relocalization on self-supervised features from an implicit representation. In *ICCV*. 252–262.
[29] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. 2019. R2d2: Reliable and repeatable detector and descriptor. *NIPS* 32 (2019).
[30] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *ICCV*. Ieee, 2564–2571.
[31] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. 2019. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*. 12716–12725.
[32] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. 2021. Back to the feature: Learning robust camera localization from pixels to pose. In *CVPR*. 3247–3257.
[33] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. 2016. Efficient & effective prioritized matching for large-scale image-based localization. *TPAMI* 39, 9 (2016), 1744–1756.
[34] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. 2019. Understanding the limitations of cnn-based absolute camera pose regression. In *CVPR*. 3302–3312.
[35] Grant Schindler, Matthew Brown, and Richard Szeliski. 2007. City-scale location recognition. In *CVPR*. 1–7.
[36] Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *CVPR*.
[37] Yoli Shavit, Ron Ferens, and Yosi Keller. 2021. Learning multi-scene absolute pose regression with transformers. In *ICCV*. 2733–2742.
[38] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. 2013. Scene coordinate regression forests for camera relocalization in RGB-D images. In *CVPR*. 2930–2937.
[39] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. 2018. InLoc: Indoor visual localization with dense matching and view synthesis. In *CVPR*. 7199–7209.
[40] Hajime Taira, Ignacio Rocco, Jiri Sedlar, Masatoshi Okutomi, Josef Sivic, Tomas Pajdla, Torsten Sattler, and Akihiko Torii. 2019. Is this the right place? geometric-semantic pose verification for indoor visual localization. In *ICCV*. 4373–4383.
[41] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 2015. 24/7 place recognition by view synthesis. In *CVPR*. 1808–1817.
[42] Huiqun Wang, Di Huang, and Yunhong Wang. 2022. GridNet: efficiently learning deep hierarchical representation for 3D point cloud understanding. *FCS* 16, 1 (2022), 161301.
[43] Boyan Wei, Xianfeng Ye, Chengjiang Long, Zhenjun Du, Bangyu Li, Baocai Yin, and Xin Yang. 2023. Discriminative active learning for robotic grasping in cluttered scene. *IEEE RA-L* 8, 3 (2023), 1858–1865.
[44] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. 2019. Sanet: Scene agnostic network for camera localization. In *ICCV*. 42–51.
[45] Ziyue Yuan, Shuqi He, Yu Liu, and Lingyun Yu. 2023. MEinVR: Multimodal interaction techniques in immersive exploration. *VI* 7, 3 (2023), 37–48.
[46] Jiqing Zhang, Bo Dong, Yingkai Fu, Yuanchen Wang, Xiaopeng Wei, Baocai Yin, and Xin Yang. 2024. A Universal Event-Based Plug-In Module for Visual Object Tracking in Degraded Conditions. *IJCV* 132, 5 (2024), 1857–1879.
[47] Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. 2022. Spiking transformers for event-based single object tracking. In *CVPR*. 8801–8810.
[48] Peiyao Zhao, Fei Zhu, Quan Liu, and Xinghong Ling. 2023. A stable actor-critic algorithm for solving robotic tasks with multiple constraints. *FCS* 17, 4 (2023), 174328.
[49] Qunjie Zhou, Sérgio Agostinho, Aljoša Ošep, and Laura Leal-Taixé. 2022. Is Geometry Enough for Matching in Visual Localization?. In *ECCV*. Springer, 407–425.
[50] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. 2023. R2former: Unified retrieval and reranking transformer for place recognition. In *CVPR*. 19370–19380.