IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 June 2024, accepted 25 July 2024, date of publication 31 July 2024, date of current version 8 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435851

RESEARCH ARTICLE

ST-PixLoc: A Scene-Agnostic Network for Enhanced Camera Localization

JING WANG, YIBO WANG, YUCHU JIN, CHENG GUO, AND XUHUI FAN
College of Communication and Information Engineering, Xi'an University of Science and Technology, Xi'an 710054, China
Corresponding author: Yibo Wang (wamgshuM@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 62301414.

ABSTRACT Visual localization is a significant problem in computer vision and robotics, involving estimating the six degrees of freedom pose of a camera relative to a known environment based on captured images. Most DL-based visual localization methods exhibit poor generalization capabilities. While some scene-independent visual localization methods demonstrate satisfactory generalization, they often suffer from low localization accuracy. To address the issue of low accuracy in scene-independent methods and the indiscriminate fusion of channel and spatial information when using neural networks for feature extraction, we propose a visual localization method ST-PixLoc, which effectively leverages image edge gradient information using the PixLoc framework. Firstly, we optimize the gradients of input images to enhance the weighting of gradient values within the image. Secondly, we employ ResNet50 as the feature extraction network's downsampling layers to enhance feature extraction capability, while introducing channel attention mechanisms in the upsampling layers of the feature extraction network. Notably, this mechanism focuses on relevant information and resolves the aforementioned indiscriminate fusion problem. Lastly, based on the feature maps of the considered images, we compute feature residuals and optimize the initial pose using optimization algorithms. Additionally, we optimize the loss function to improve the model's accuracy in complex scenes. Experimental results demonstrate that the proposed method achieves high-precision localization. The average rotation and translation errors on the indoor 7-Scenes dataset increased by 6.9% and 9.7%, respectively, while those on the outdoor Cambridge Landmarks dataset increased by 16.7% and 28.2%, validating the effectiveness of the proposed approach.

INDEX TERMS Visual localization, deep learning, residual network, channel attention mechanism, camera localization.

I. INTRODUCTION

The goal of visual localization is to estimate the six-degrees-of-freedom pose of a camera relative to a known environment. This involves determining the camera's position coordinates and angular deviations around the three coordinate axes. It is a critical problem in the fields of computer vision and robotics; solving this problem is essential for achieving truly autonomous robots, such as mobile robots and self-driving cars. This is crucial for applications in the field of SLAM [1], and it is also a prerequisite for augmented reality and virtual reality systems.

Existing visual localization methods can be categorized into two main types: traditional geometric and deep learning (DL) methods. Traditional geometric methods rely on matching local features in images [2], [3], [4], [5]; they match feature points in a query image with a sparse three-dimensional (3D) point [6], [7], [8] cloud in the scene and then use obtained correspondences to recover the precise camera pose. However, they are in-adequate in cases of changing scene illumination, blurry image pixels, and weak image textures, resulting in low localization accuracy or even failure. Thus, DL has been employed for the automatic and efficient extraction of high-quality image features [9], [10] for visual localization.

DL-based visual localization methods can be categorized based on the processing pipeline from input to output

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

into end-to-end and non-end-to-end methods. End-to-end methods involve convolutional neural networks (CNNs) that directly output predictions for camera poses, including absolute and relative pose regressions. Absolute pose regression, exemplified by PoseNet [11], uses neural networks (NNs) to extract features and then directly regresses the camera pose based on the feature vectors. Existing absolute pose regression methods vary in terms of their NN architectures [12], [13] and loss functions [14], [15]. Meanwhile, relative pose regression employs NNs to predict the pose of a query image relative to one or more reference images [16], where the reference images are obtained through implicit image retrieval using NNs. Compared with absolute pose regression methods, relative pose regression methods exhibit higher adaptability, higher generalization, and improved localization accuracy.

Although end-to-end methods are simple and efficient, they often underperform traditional geometric methods in terms localization accuracy. Consequently, non-end-to-end methods, such as scene coordinate regression, have been proposed. They do not directly regress camera poses using CNNs; instead, they regress 3D scene coordinates. Brachmann et al. [17] employed CNNs for feature extraction (FE) and feature matching to obtain 3D scene coordinates. They subsequently calculated the camera pose using the traditional perspective-n-point method—a local learning approach that combines the advantages of DL and traditional methods—and achieved localization accuracy comparable to that of traditional methods. However, non-end-to-end methods may encounter localization failures in outdoor large-scale scenes due to limited model capacity. Further, their models require training or fi-ne-tuning for each scene, indicating poor generalization ability to unseen scenes.

Inspired by direct image alignment [18], Sarlin et al. [19] introduced a scene-agnostic visual localization model called PixLoc. PixLoc uses NNs solely to extract image features and output multiscale feature maps. Subsequently, it employs optimization algorithms to minimize the feature residuals between query and reference im-ages to obtain an optimal pose. PixLoc's localization accuracy is comparable to those of scene coordinate regression methods in most scenes; however, it performs poorly in complex scenes with repetitive structures and low-texture regions.

In the field of camera localization algorithms, end-to-end models should focus on issues related to feature fusion and geometric information loss. Zhang et al. [20]. addressed the problem of alignment in motion change features using a divide-and-conquer strategy and proposed a video super-resolution method named LGDFNet. This method decomposes the overall features into multiple local features, with each local feature processed by a dedicated subnetwork, gradually merging from local to global features. The model introduces a self-calibrating dynamic filtering module to align features and utilizes a cross-attention feature fusion module to merge features before and after dynamic filtering. To address the loss of geometric information in voxel models during downsampling, Kang et al. [21]. proposed a new framework named PVB-SSD. This framework improves accuracy and computational efficiency through innovative use of Fourier embedding features, a global pre-module, and dynamic fusion of spatial-semantic features, providing a new technical pathway for practical applications such as autonomous driving.

In summary, the current state of visual localization based on DL shows that scene coordinate regression methods have superior performance but suffer from poor generalization ability, whereas models with better generalization ability exhibit lower localization accuracy. To address these issues, we propose an improved visual localization method. The major contributions of this study are as follows.

(1) We designed and implemented a scene-independent network model that performs localization by directly aligning multi-scale image features. The network does not need to regress the camera's pose; instead, it focuses on extracting appropriate image features to ensure the model can accurately generalize to other scenes. We use the UNet network as the encoder structure for extracting multi-scale features, which are then used to compute the optimal pose. To enhance the robustness of multi-scale image features, we replaced the FE components in the UNet network with ResNet50 and combined it with ECA-Net to form a new ResUNet-E network. This approach produces multi-scale feature maps of the image. By calculating the feature residuals between the query image and the reference image and using the LM algorithm, the optimal pose is iteratively derived from the initial pose.

(2) Design an image edge extraction network to enhance image feature gradients, To prevent the Levenberg-Marquardt(LM) [22], [23] algorithm from falling into local optima while iteratively minimizing the camera pose error.

(3) We employed ResNet50 in the downsampled part of the FE network to enhance FE capabilities while improving context aggregation and added channel attention mechanisms in the upsampled part to amplify important information and reduce the impact of less relevant information.

(4) We employed a joint loss function that combines pose estimation and geometric reprojection errors for model training. This approach improved model accuracy in complex scenes.

II. RELATED WORKS

In the following section, we discuss the main stream of research for solving visual camera localization. We also discuss PixLoc.

A. SCENE COORDINATE REGRESSION

In terms of accuracy, scene coordinate regression methods have shown significant improvement compared with end- to-end methods. The concept of scene coordinate regression can be traced back to 2013, when it was first applied to RGB-D image localization. Shotton et al. [24] proposed the use of random forests to predict two-dimensional–3D matches and learn how pixels in an image patch map to scene coordinates

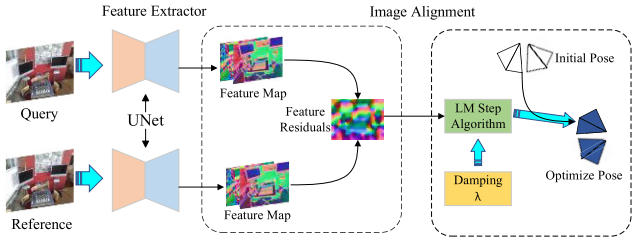


FIGURE 1. Overall framework of PixLoc.

in the scene model. In 2017, Brachmann et al. [17] extended this idea to predict scene coordinates using a visual geometry group (VGG)-style architecture and proposed differentiable RANSAC (DSAC) to learn a matching function that optimizes pose quality. The model comprised two CNNs: one directly regressed scene coordinates for all pixels in an image and the other scored pose hypotheses to select the best pose and further optimize it to obtain the final pose. Although DSAC achieves high-precision camera localization, it is resource-intensive because it predicts scene coordinates using image patches. Brachmann and Rother [25]. proposed DSAC++ that used entropy-controlled soft inlier counting during the scoring phase to mitigate overfitting and Shannon entropy and local linearization during the pose optimization phase to tackle the issues of high gradient variance. Subsequently, Brachmann and Rother [6]. designed a flexible localization system, DSAC*, which allowed the choice of using depth information and scene models. In DSAC*, the scene coordinate regression network is replaced with ResNet, thereby reducing memory consumption by 75% and achieving a certain degree of accuracy improvement. Cai et al. [26]. argued that multiview constraints benefited the learning process and final performance and proposed the incorporation of multiview geometric constraints in the training of scene coordinate regression, building upon DSAC++. These constraints not only accelerated network convergence but also improved accuracy. Compared with end-to-end camera-pose estimation methods, scene coordinate regression methods offer higher accuracy but require longer training times, have slower runtime, and exhibit poorer generalization. Moreover, scene coordinate.

B. CROSS-SCENE CAMERA-POSE ESTIMATION

Cross-scene camera-pose estimation refers to a model's ability to directly generalize to unseen scenes without retraining, which is also known as scene-agnostic camera-pose estimation. Models with strong generalization ability in this aspect are high-lighted in this section. Yang et al. [27]. proposed a scene-agnostic network called SANet, wherein scene and model parameters are independent of each other. The model begins by performing image retrieval to select a subset of scenes, which is then fed into the DRN38 network to generate feature maps at different resolutions. The feature maps are combined with 3D point cloud coordinates of the scenes and their corresponding pixel feature vectors to construct a scene

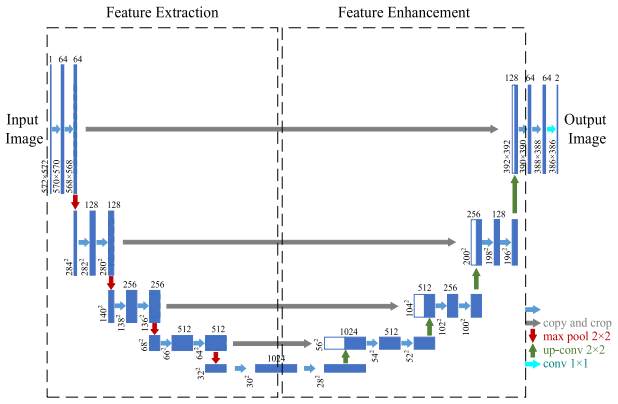


FIGURE 2. UNet architecture.

pyramid. Query images are processed using DRN38 to obtain a query image feature pyramid, and scene coordinates are predicted by matching the scene pyramid with the query image feature pyramid. SANet exhibits better generalization ability than other scene coordinate regression methods, although its accuracy is lower. Tang et al. [28]. introduced a new scene-agnostic model that leverages dense scene matching to create a cost volume between the query image and scene. Subsequently, a CNN processes this cost volume and its corresponding coordinates to predict dense scene coordinates. In terms of accuracy, this model significantly outperforms SANet and is comparable to DSAC++. Although the above methods are scene-agnostic and show improved accuracy, they face challenges in open scenes. Therefore, Sarlin et al. proposed a feature-map-based scene-agnostic model called PixLoc, which shows enhanced generalization due to the separation of model parameters from scene geometry. The model uses NNs to extract multiscale feature maps from reference and query images and computes feature residuals between these feature maps. An optimization algorithm is then used to minimize the feature residuals, yielding an optimized pose.

C. PIXLOC

The overall framework of PixLoc, which comprises a deep NN—UNet—and an image alignment module, is shown in Figure 1. UNet is used to extract features from the query and reference images, and the image alignment module [29] is employed to refine the pose and obtain the optimal pose.

UNet is a fully convolutional network introduced by Ronneberger et al. [30]. in 2015; its structure, which comprises FE and feature enhancement components, is shown in Figure 2. The FE component employs a VGG-style architecture to extract feature information from an image, performing downsampling and including the original image scale. Overall, the FE component produces feature maps at five different scales; larger-scale features contain more detailed information, whereas smaller-scale features contain higher-level semantic information. The feature enhancement component performs image upsampling and then combines it with the corresponding scale's feature map. Feature enhancement

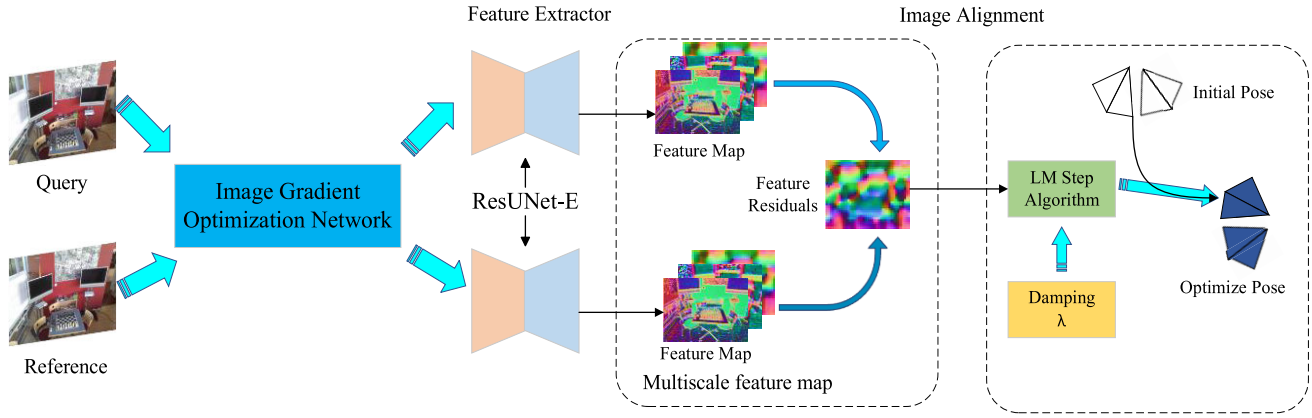


FIGURE 3. ST-PixLoc model in this article.

integrates both detail and semantic information, allowing for better capture of the contextual features of image pixels.

The deep features learned by UNet are ultimately used for pose estimation, which is achieved through direct image alignment. The image alignment module is designed to find the pose that minimizes the differences between the query and reference images. It uses the LM algorithm to determine the optimal pose. The LM algorithm takes as input the feature maps extracted by UNet from the reference and query images and outputs a predicted relative pose. LM is a commonly used optimization algorithm that involves computations of various factors, such as a robust cost function ρ and damping factors λ. In previous DL optimization algorithms, NNs were used to predict optimization parameters $\hat{\rho}'$, damping factors λ, and pose updates δ. Sarlin et al. argued that this approach severely limits a model's ability to generalize to new datasets, as it couples the optimizer with visual semantic information from the training dataset, and proposed decoupling the optimizer from semantic information to achieve scene-agnostic results. They treated λ as a fixed model parameter and learned it alongside the CNN through gradient descent.

Compared with the end-to-end and scene coordinate regression methods, PixLoc exhibits superior generalization ability. Particularly, a single model training can be used for camera-pose estimation across multiple scenes, enabling the model to generalize to unseen scenes without the need for specific training on each scene.

III. MATH

This article is inspired by direct image alignment methods, using multi-scale image features for camera localization. It measures the similarity of features between reference and query images as a metric, aiming to minimize feature differences. The idea of image feature alignment is equivalent to the direct method in visual odometry, which does not require extracting feature points from images but directly solves camera poses using pixel intensities, relying on the photometric invariance assumption. In the direct method, the alignment is based on the intensity information of the same pixels in two images. According to the photometric invariance assumption, the intensity of corresponding pixels is the same. However, in practical scenarios, inaccuracies in pose estimation lead to intensity differences. Therefore, it is necessary to construct a nonlinear optimization problem based on intensity differences. This optimization problem minimizes the total error of all pixels as the objective function by optimizing camera poses.

Image feature alignment consists of two steps: multi-scale FE and pose optimization, gradually reducing the error between the reference and query images. The model first utilizes the ResUNet-E network to extract multi-scale feature information from the reference and query images. Then, based on the pose of the reference image, it aligns the corresponding pixels in the feature maps. Finally, it computes the difference between two pixels (i.e., feature residuals) and treats the feature residuals as the objective function for the optimization algorithm. The LM algorithm is used to optimize the pose and obtain the optimal pose. The LM algorithm is heavily influenced by the gradients of image features during the iteration process. Due to the significant role of image gradients, this paper designs an image edge extraction network before the ResUNet-E network. By processing images in the dataset, it identifies points with large gradients in the images, thereby increasing the weight of points with large gradients in the images. The network model ST-PixLoc in this article is shown in Figure 3.

By using deep learning methods and scene coordinate regression methods to regress the camera pose from images, the model identifies specific visual features of the scene. However, its effectiveness is not significant when applied to unknown scenes. Therefore, we focus on learning robust and generalizable multi-scale features to enhance the performance of image alignment methods and improve the model's generalization capability. In the two steps of the image alignment method, we improve the multi-scale FE network used to calculate feature residuals and enhance the LM algorithm's capability to optimize camera pose.

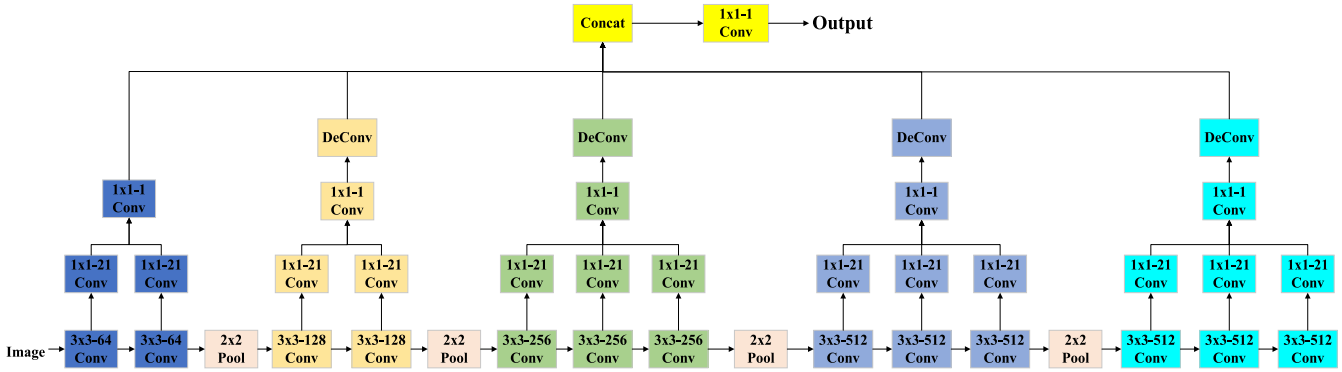
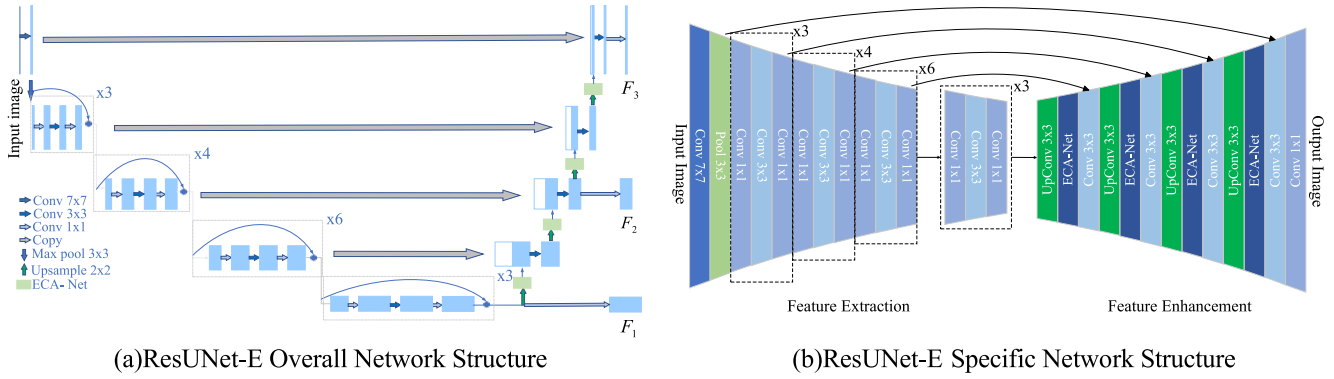


FIGURE 4. Image gradient optimization network.



(a)ResUNet-E Overall Network Structure (b)ResUNet-E Specific Network Structure

FIGURE 5. ResUNet-E network model (a) Overall network structure (b) Specific network structure.

A. IMAGE GRADIENT OPTIMIZATION NETWORK

In the ST-PixLoc model, the LM algorithm is a core component. It is a nonlinear least squares optimization method that does not rely on the geometry or features of a specific scene. ST-PixLoc uses the feature confidence predicted by the FE network to weight the feature residuals in the LM algorithm. This approach enables the ST-PixLoc model to focus more on robust and high-confidence multi-scale feature points, allowing ST-PixLoc to adapt to different scenes and improve its generalization ability across various scenes.

In ST-PixLoc, the FE network extracts feature points and generates feature descriptors from the query image and the reference image. Subsequently, it establishes the feature residuals between them. The LM algorithm iteratively optimizes the pose based on these feature residuals. The Jacobian matrix in the LM algorithm is divided into two parts: the first part consists of the gradients of image features, and the second part consists of the derivatives of the camera pose with respect to the 2D coordinates of the image. The product of these two parts forms the Jacobian matrix. In many scenes with repetitive textures, the changes in the image are not significant, leading to small variations in gradients and causing biases in image alignment and feature recognition. In such cases, it is difficult for the Jacobian matrix to perform incremental calculations during iteration. As a result, subsequent iterations may fall into the same region, resulting in a local optimal solution.

To address this issue, an Image Gradient Optimization Network is designed to process the dataset, enhancing the weight of points with larger gradients in the image. This helps the neural network find the correct direction for gradient descent, thereby reducing pose calculation errors caused by falling into local optimal solutions. By using the Image Gradient Optimization Network to obtain image edge information, the obtained edge information is merged with the original image (concatenation on the channels) to enhance the image gradients.

The image gradient optimization network in this paper is based on a fully convolutional network designed from VGG16, whose structure is illustrated in the Figure 4. Compared to the original VGG16 network, all fully connected layers and the fifth pooling layer have been removed. The removal of fully connected layers enhances the effectiveness and computational efficiency of the VGG16 output. Pooling layers downsample the feature maps, which is not conducive to edge localization. Since extracting edge information requires recalculating pixel values, the convolutional layers in the image gradient optimization network first pass through a 1×1 convolutional layer to increase dimensionality. Then, they go through another 1×1 convolutional layer to add up the

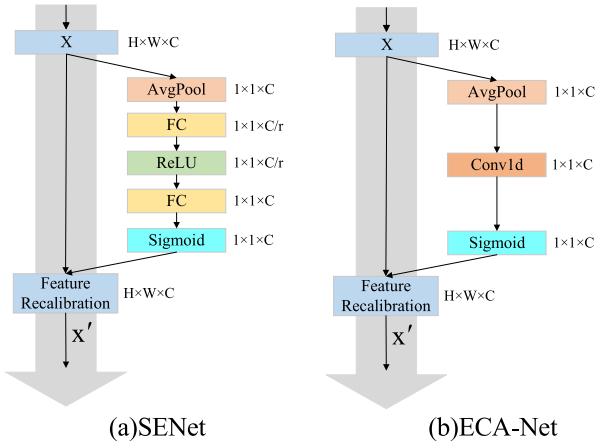


FIGURE 6. SENet and ECA-Net architectures.

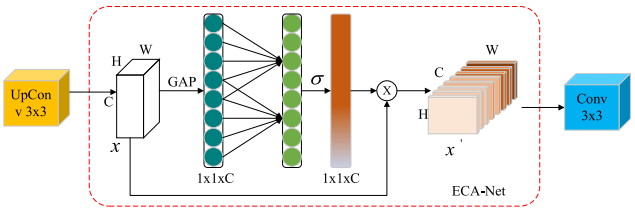


FIGURE 7. Detailed structure of ECA-Net module in ResUNet-E network upsampling.

output elements to obtain composite features. Subsequently, an upsampling layer (deconvolution) enlarges the size of the feature maps to restore the image size to its initial dimensions. Finally, the outputs of each layer are concatenated, and a 1×1 convolution is applied for fusion to achieve the ability to obtain various blended information.

The image gradient optimization network can enhance image gradients, helping to improve the initial estimated pose.

This improves the initial iteration quality of the LM algorithm, making it less likely to fall into local optima during the iterative optimization process. This enhances the performance of the LM algorithm and the generalization capability of ST-PixLoc.

B. ResUNet-E NETWORK

Before the LM algorithm iteratively optimizes the pose, it needs to calculate the feature residuals between the query image and the reference image. Obtaining robust multi-scale features from both the query image and the reference image is essential. The design concept of the proposed FE network, ResUNet-E, is the same as that of UNet; its structure is shown in Figure 5(a). The network consists of two parts—downsampling and upsampling—allowing for end-to-end training.

The downsampling part extracts representative features; it improves upon the UNet by replacing the VGG network with ResNet50 to enhance FE capabilities. The improved network removes the original network's final pooling and fully connected layers and retains the residual structure, thereby preserving UNet's fully convolutional nature. An image initially passes through a 7×7 convolutional layer and a max-pooling layer, followed by passing through four residual modules, each comprising three, four, six, and three basic residual units. In addition, the network weights for ResNet50 are pretrained on ImageNet, which not only helps prevent overfitting but also accelerates convergence. To ensure that the feature map size output by the network matches the input image size, padding strategies are adopted for all convolutional layers in the network.

In visual localization, continuous frame images are typically used as input. Such consecutive frame images exhibit subtle changes between each frame. As the number of convolutional layers increases, the image feature maps gradually reduce in size, causing minor details to be submerged by redundant information during convolution operations, thereby reducing localization accuracy. In addition, the importance of information varies across different channels and positions, but convolution operations indiscriminately blend this information. To address these issues, we employed a channel attention mechanism, known as ECA-Net [31]. This attention mechanism captures interchannel information and discerns the importance of various parts of the input features. Consequently, it assigns distinct weights to different channels, extracting more critical and discriminative information, thereby enabling the model to make more accurate judgments.

ECA-Net is a lightweight attention mechanism that builds upon SENet [32], alleviating the side effects of dimension reduction caused by fully connected layers. SENet and ECA-Net structures are shown in Figure 6 (a) and (b), respectively. Notably, ECA-Net removes the fully connected layer and ReLU activation function from SENet and uses one-dimensional convolution after average pooling to obtain inter channel information. This attention mechanism captures inter channel information and discerns the importance of various parts of the input features. Consequently, it assigns distinct weights to different channels, extracting more critical and discriminative information, thereby enabling the model to make more accurate judgments.

ECA-Net can be flexibly incorporated into various CNN architectures, as demonstrated through extensive experiments by Wang et al. [31]. revealing that integrating ECA-Net into CNNs enhances network performance, with ECA-Net outperforming SENet and CBAM [33] comprehensively. To amplify subtle features in images, emphasize scene details, enhance network performance, and improve visual localization accuracy, we incorporated ECA-Net into the FE network. The detailed structure is shown in Figure 7. ECA-Net is employed after the upsampling operations in UNet, allowing the network to focus on important features while ignoring less significant ones. The improved network can effectively extract critical scene information and fine details, enhancing visual localization precision in complex scenes.

Figure 5(b) shows the specific network structure of ResUNet-E. After extracting high-level semantic information from the image through the downsampling phase, it is necessary to use upsampling to restore the original image details. After undergoing four upsampling operations, the image is restored to its original scale, and the network outputs three feature maps at different scales F_1, F_2, F_3. Downsampling operations result in the loss of image information, making information recovery challenging during upsampling. Therefore, there are skip connections between upsampling and downsampling operations for the purpose of utilizing low-level information to aid image recovery. To emphasize detailed information, ECA-Net is employed after the upsampling operation to assist the network in achieving better re-covery of the original image.

In the ST-PixLoc model, the FE network ResUNet-E uses ResNet50 as the downsampling FE component. We are committed to improving the FE capability of the UNet network. Compared to the VGG network, the deeper ResNet50 network can extract more complex multi-scale features, which include more detailed semantic information at different levels. In the upsampling part, we added the channel attention ECA-Net to achieve information distribution across different channels and positions, making the features recovered during the upsampling stage more distinguishable and improving semantic quality. To obtain generalizable features and facilitate the upsampling decoding process, ECA-Net is a good choice. It allows ST-PixLoc to meet efficiency requirements while maximizing generalization capability.

C. LOSS FUNCTION OPTIMIZATION

In visual localization, it is common to use either the pose estimation error L_{em} or geometric reprojection error L_{re} as a standalone loss function. For instance, PixLoc solely employs L_{re} as the loss function. These loss functions are defined as follows:

$$L_{em} = \left\| x - \hat{x} \right\|_2 + \alpha \left\| q - \hat{q} \right\|_2 \tag{1}$$

$$L_{re} = \frac{1}{l} \sum_{l} \sum_{i} \left\| \prod (R_l P_i + t_l) - \prod (\hat{R}_l P_i + \hat{t}_l) \right\|_2 \tag{2}$$

where x and $\hat{x}$ denote the position coordinates of the ground truth and estimated locations, respectively; q and $\hat{q}$ denote the angular deviations of the ground truth and estimated values, respectively; α denotes the weighting coefficient used to balance rotation and translation errors; R_l and $\hat{R}_l$ denote the ground truth and estimated rotation matrices, respectively; P_i denotes 3D points within the scene; t_l and $\hat{t}_l$ denote the ground truth and estimated translation vectors, respectively; and l denotes the number of feature map layers (three layers are considered in this work).

In typical scenarios, L_{re} provides strict geometric constraints, resulting in high localization accuracy. However, in some complex scenes, the constraints imposed by L_{re} may become ineffective, leading to lower localization accuracy.

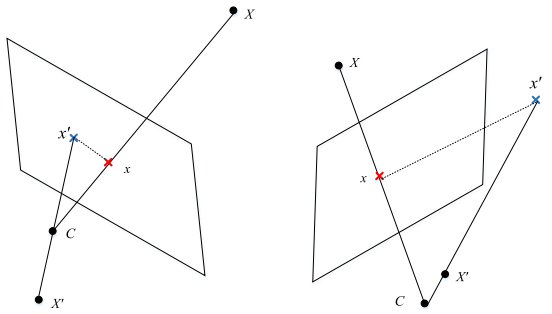


FIGURE 8. The case of reprojection error failure.

As shown in Figure 8, the case of reprojection error failure, where X is a 3D point in the scene, X' is the estimated coordinate, C is the center of the camera, x is the true projection, and x' is the projection of the estimated value. In the left figure of Figure 8, if the estimated value is located behind the center of the camera, there is a large gap between the true value and the estimated value, but the reprojection error is still very small. In this case, the constraining effect of the reprojection error is almost zero. In the right figure of Figure 8, the estimated value is very close to the center of the camera, and the point projected to the image coordinate system will greatly deviate from the true value. The initial reprojection error is very large, and the constraining effect of the reprojection error is small. This situation may cause the model to fall into an erroneous local minimum. Therefore, we chose to integrate L_{em} and L_{re} for joint model training, thereby reinforcing constraints in complex scenes and enhancing localization accuracy.

$$L = \frac{1}{l}\Big(\sum_{l} \left\| \prod (RP_i + t) - \prod (\hat{R}P_i + \hat{t}) \right\|_2 + \lambda\Big(\left\| t - \hat{t} \right\|_2 + \left\| \frac{trace(R^{-1}\hat{R})}{2} \right\|_2\Big)\Big) \tag{3}$$

where, λ is the fusion coefficient, ranging from 0 to 1. Through experimental comparison, the optimal fusion coefficient value is 0.6. After introducing the pose estimation loss into the loss function, in complex scenes, when the reprojection error fails, the geometric constraints are strengthened by fusing the pose estimation loss, effectively solving the problem of reprojection error loss failure in complex scenes.

IV. EXPERIMENTAL RESULTS

In this section, we describe the datasets used in the experiments, outline the experimental procedures, and analyze the results of the enhanced UNet on the 7-Scenes [25] and Cambridge Landmarks datasets [11]. Comparative experiments with other methods were also performed, involving Active Search (a traditional visual localization method) [4] and SANet (a scene-agnostic coordinate regression method).

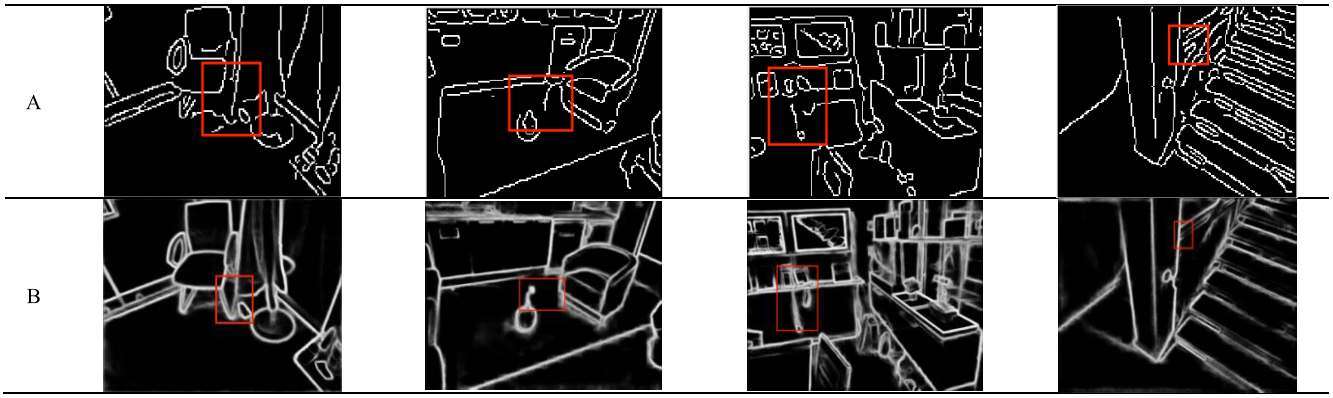


FIGURE 9. The extracted edge features from the traditional Canny operator and the edge features obtained from the image gradient optimization network proposed in this paper are compared. Column A represents the Canny operator, and Column B represents the image gradient optimization network proposed in this paper.

A. DATASETS AND EXPERIMENTAL SETUP

1) DATASET

We employ the MegaDepth dataset [34] to train network models. This dataset is an outdoor open-scene dataset, comprising network data from 196 well-known landmarks worldwide, with approximately 100,000 images. Subsequently, the pre-trained models are directly used for comparative experiments on two publicly available color scene datasets the indoor 7-Scenes dataset and the outdoor Cambridge Landmarks dataset to validate the proposed model's performance. These two datasets encompass complex scenes with features such as repetitive textures, low textures, and dynamic objects, providing a comprehensive reflection of the visual localization algorithm's performance.

The 7-Scenes dataset comprises Chess, Heads, Office, Fire, Pumpkin, Red Kitchen, and Stairs scenes. The dataset includes challenging images with repetitive textures (Stairs), low textures (Fire and Pumpkin), and object occlusion (Red Kitchen). The Cambridge Landmarks dataset comprises King's College, Old Hospital, Shop Façade, St Mary's Church (hereinafter, St M. Church), and Great Court scenes. The dataset includes a significant amount of interference from pedestrians, vehicles, etc. Data are collected at different time points, representing varying lighting and weather conditions.

2) EVALUATION

Localization accuracy is a key metric for assessing visual localization, encompassing both translation and rotation errors [10]. For comparability, we employ the median of localization errors to evaluate the performance of the proposed model. Further, to verify the model's stability, we calculate the percentage of successfully localized points with a threshold of 5 cm/5° (translation error/rotation error) on 7-Scenes. On Cambridge Landmarks, we record the percentage of successfully localized points with a threshold of 25 cm/2° to reflect the model's high-precision localization performance.

3) EXPERIMENTS

In the experiments, a learning rate of 0.00001 is used, with a batch size of six training samples, and the Adam optimizer is chosen. The training process continues for 20–100K iterations until the network converges. The selection of weight coefficient α for pose estimation error L_{em} is not the same for different datasets, and after tuning, values of α around 1 and 5 are suitable for 7-Scenes and Cambridge Landmarks, respectively. In the overall loss function, λ is set to 0.6. All experiments are conducted using the PyTorch library on a single NVIDIA GeForce RTX 3090Ti GPU.

B. EXPERIMENTAL DETAILS

In traditional edge detection methods, the Canny operator yields the best results. We compare the effectiveness of the Canny operator with our proposed image gradient optimization network. As depicted in the Figure 9, traditional edge detection methods tend to focus on regions with high gradients, which often leads to significant noise. Even the superior performance of the Canny edge detection method can result in missing edge lines. Conversely, DL-based edge detection networks prioritize the overall coherence of scene edges, generating stable and smooth edges by comprehending the scene.

The image gradient optimization network is designed based on VGG16 for image edge extraction. After obtaining the edge information from the image, the original image is fused with the binary edge image on the channel level to enhance the influence of high-gradient regions in the image on Camera Localization.

The ResUNet-E network in the article is highly adaptable, allowing for flexibility in adjusting the number of convolutional layers, channel counts, etc., as per specific requirements, enabling appropriate compression or expansion. By modifying the network architecture, various feature maps with different dimensions and scales can be obtained.

In this experiment, the scale l is set to 3, and the network ultimately outputs three different-scale feature maps. Here,

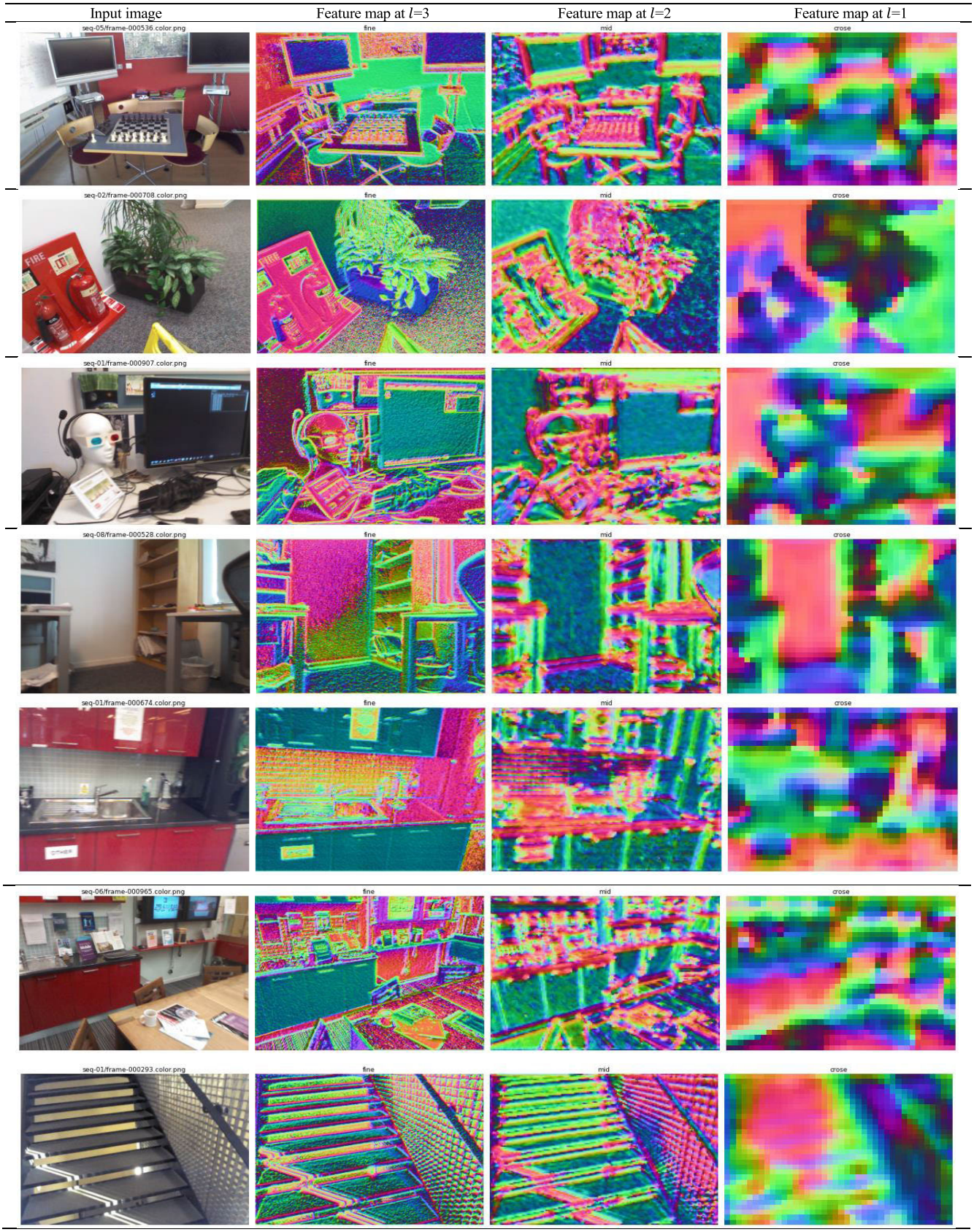


FIGURE 10. The figure shows the feature maps output at three different scales of the ResUNet-E network. Rows 1 to 7 represent the output scales for the seven scenes in the 7-Scenes dataset, while rows 8 to 12 represent the output scales for the five scenes in the Cambridge Landmarks dataset.

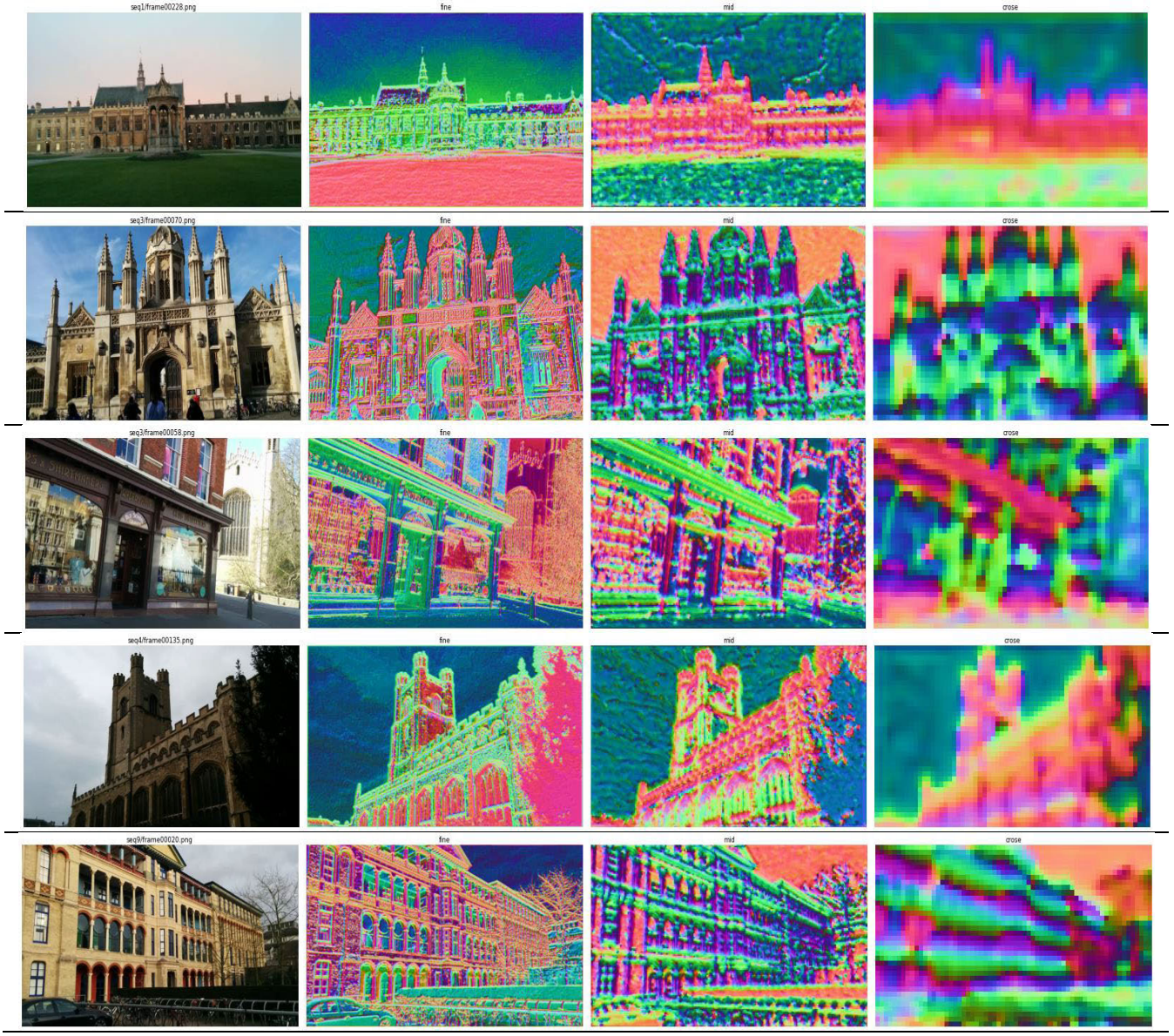


FIGURE 10. ***(Continued.)*** **The figure shows the feature maps output at three different scales of the ResUNet-E network. Rows 1 to 7 represent the output scales for the seven scenes in the 7-Scenes dataset, while rows 8 to 12 represent the output scales for the five scenes in the Cambridge Landmarks dataset.**

a smaller l indicates a smaller scale, with feature map dimensions for each scale being $D_l = 32, 128, 128$. respectively. When $l = 1$, the feature map F_1 is obtained by passing the output of the last layer of the encoder through a 3×3 convolutional layer. For $l = 2$, the output of the last layer of the encoder undergoes 2×2 upsampling, followed by an ECA-Net module, and then fused with the feature map corresponding to the scale of the encoder part. This process is repeated twice, followed by another 3×3 convolutional layer to obtain feature map F_2. Similarly, for $l = 3$, the steps to obtain feature map F_3 are the same as when $l = 2$. Pose optimization is sequentially performed on the feature maps of the three scales using the Levenberg-Marquardt algorithm. The three scale feature maps output by the network are shown in Figure 10. We display the image features at three different scales, mapped to RGB. The red features represent high-frequency edge features in their local direction, the green features represent texture features, and the blue features represent low-frequency features in smooth areas. As the scale of the feature maps increases, the textures and other details in the images also gradually increase.

In the pose optimization process, the LM algorithm is employed to find the optimal pose. The algorithm takes as input the feature residuals r_k^i between the multi-scale feature maps F_q^l of the query image and F_k^l of the reference image. By minimizing these feature residuals, the camera pose

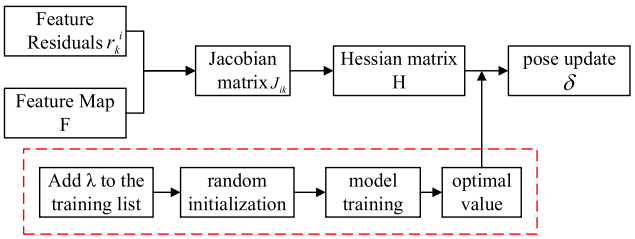


FIGURE 11. The process of pose optimization for separating model parameters.

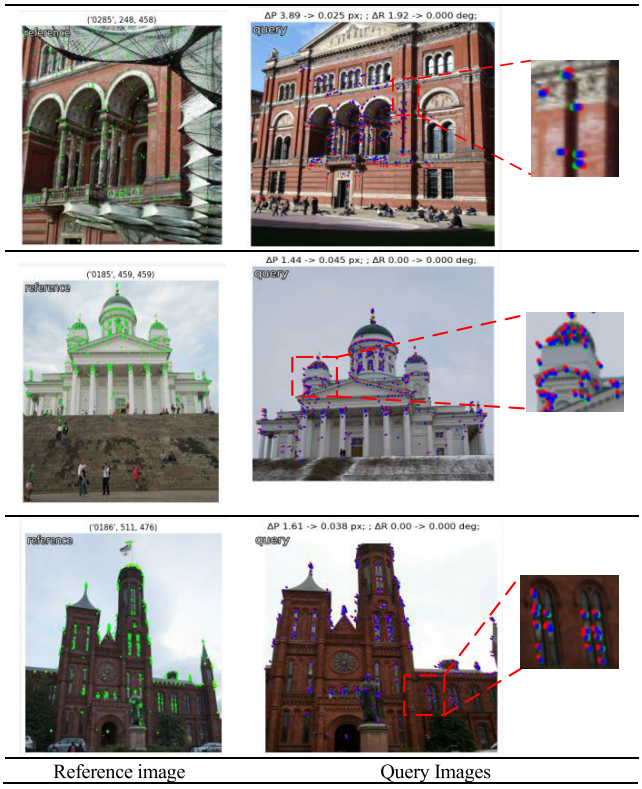


FIGURE 12. The figure shows the results of pose Optimization. The green dots represent ground truth value, the red dots represent values before Optimizatio, the blue dots represent values after Optimization.

TABLE 1. Comparison result of rotation errors on 7-Scenes.

Scene	Active Search	SANet	PixLoc	Ours
Chess	1.96°	0.88°	0.83°	0.81°
Fire	1.53°	1.08°	0.79°	0.73°
Heads	1.45°	1.48°	0.87°	0.84°
Office	3.61°	1.00°	0.83°	0.80°
Punpkin	3.10°	1.32°	1.19°	1.14°
Red Kitchen	3.37°	1.40°	1.27°	1.19°
Stairs	2.22°	4.59°	1.31°	1.10°
Average	2.46°	1.68°	1.01°	0.94°

(R, t) is optimized. This paper enhances the model's ability to generalize to new scenes by separating model parameters from the pose optimizer. The damping factor λ is treated as a fixed model parameter to decouple the optimizer from the training data. The workflow is illustrated in the Figure 11.

TABLE 2. Comparison result of translation errors on 7-scenes.

Scene	Active Search	SANet	PixLoc	Ours
Chess	4 cm	3 cm	2.5 cm	2.5 cm
Fire	3 cm	3 cm	1.9 cm	1.8 cm
Heads	2 cm	2 cm	1.3 cm	1.2 cm
Office	9 cm	3 cm	2.8 cm	2.7 cm
Punpkin	8 cm	5 cm	4.3 cm	4.1 cm
Red Kitchen	7 cm	4 cm	3.6 cm	3.4 cm
Stairs	3 cm	16 cm	5.1 cm	4.2 cm
Average	5.1 cm	5 cm	3.1 cm	2.8 cm

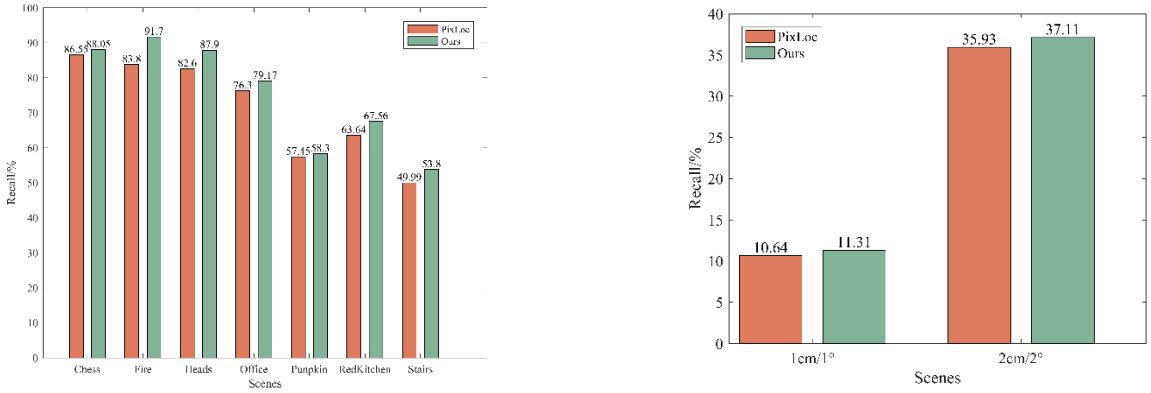
Taking reference from the setting of the weighting coefficients in the attention mechanism, we set λ as a trainable parameter of the model. For the original non-trainable tensor-type parameter λ, we convert it to a trainable parameter of the parameter type. At the same time, λ is bound to the parameter list of the ResUNet-E model, associated with the model. Once bound to the model's parameter list, the parameter λ will be updated iteratively along with the model training.

The results of pose optimization are shown in Figure 12, with the left image being the reference image and the right image being the query image. In the images, green pixels represent the ground truth values corresponding to pixels in the reference image, red pixels represent the initial values of pixels projected from the reference image to the query image, and blue pixels represent the optimized values. It can be observed that the initial pose before optimization deviates significantly from the ground truth values, whereas the pose after optimization is closer to the ground truth, with near overlap. After pose optimization, the error between the pixels in the reference image and those projected to the query image is reduced: for the first query image, the translational error decreases from 3.89 pixels to 0.025 pixels, and the rotational error decreases from 1.92 degrees to 0; for the second query image, the translational error decreases from 1.44 pixels to0.045 pixels; and for the third query image, the translational error decreases from 1.61 pixels to 0.038 pixels.

C. RESULTS ON 7-SCENES

The experimental results for 7-Scenes are shown in Table 1 and 2 for the com-parison of translation and rotation errors, respectively. As shown in Table 1 and Table 2, among the scene-agnostic methods, PixLoc exhibits the best performance. Compared with PixLoc, our model shows smaller localization errors in all scenes, except Chess and Pumpkin, with particularly better performance in Stairs with repetitive textures and Fire with low textures. In Chess, our model exhibited smaller rotation errors, while in Pumpkin, the localization errors of PixLoc and our model were com-parable.

Analyzing the rotation errors, our model outperformed PixLoc in all scenes, except Pumpkin, with an overall improvement of 6.9%. In terms of translation errors, our model outperformed PixLoc in all scenes, except Chess and Pumpkin, with an overall improvement of 9.7%.



(a) Comparison result of accuracy on 7-Scenes

(b) Average accuracy of our model and PixLoc

FIGURE 13. The comparison results of accuracy and recall between the model proposed in this paper and the PIXLOC model on the 7-Scenes dataset.

TABLE 3. Comparison result of rotation errors on Cambridge Landmarks dataset.

Scene	Active Search	SANet	PixLoc	Ours
K. College	0.6°	0.54°	0.26°	0.23°
Old Hospital	1.0°	0.53°	0.79°	0.63°
Shop Facade	0.4°	0.47°	0.23°	0.22°
St M. Church	0.5°	0.57°	0.36°	0.31°
Great Court	0.6°	1.95°	0.18°	0.12°
Average	0.62°	0.80°	0.36°	0.30°

TABLE 4. Comparison results of translation errors on cambridge landmarks.

Scene	Active Search	SANet	PixLoc	Ours
K. College	42 cm	32 cm	15.7 cm	15.4 cm
Old Hospital	44 cm	32 cm	48.9 cm	35.9 cm
Shop Facade	12 cm	10 cm	6.1 cm	5.1 cm
St M. Church	19 cm	16 cm	14.3 cm	12.5 cm
Great Court	120 cm	328 cm	42.5 cm	23.0 cm
Average	33.8 cm	83 cm	25.5 cm	18.3 cm

The positioning accuracy was calculated with a threshold of 5 cm/5° on 7-Scenes; the results are shown in Figure 13(a). Overall, our model outperformed PixLoc, with an improvement in positioning accuracy in all scenes, especially in the complex scenes Stairs and Fire, where the positioning accuracy increased by 4.80% and 7.90%, respectively. This indicates that our model exhibited stronger adaptability and better performance in complex scenes.

To further validate the performance of our model, we also calculated the average positioning accuracy with thresholds of 1 cm/1° and 2 cm/2° on 7-Scenes; the results are shown in Figure 13(b). Our model outperformed PixLoc, with a 3.58% increase in average accuracy when using a threshold of 2 cm/2°.

Figure 14 shows plots of camera localization trajectories on 7-Scenes. The localization trajectories represent a camera's motion path during image capture, with outliers indicating larger positioning errors and deviations from the original trajectory. Thus, fewer outliers in the localization trajectory indicate smaller positioning errors and better model performance. In Figure 14, the images in the first and third rows show the localization trajectory of PixLoc, while the images in the second and fourth rows show the localization trajectory of our model. It is evident that our model's localization trajectory is clearer and contains fewer outliers, indicating reduced positioning errors.

D. RESULTS ON CAMBRIDGE LANDMARKS

The experimental results on the Cambridge Landmarks dataset are presented in Table 3 for translational error comparison and Table 4 for rotational error comparison. As shown in Table 3 and 4, among the scene agnostic methods, PixLoc performs the best. Our model exhibits lower positioning errors than PixLoc in all scenes except King's College. In King's College, our model achieves a lower rotational error than PixLoc. Analyzing the rotational error, our model demonstrates improvements of 11.5%, 20.3%, 4.3%, 13.9%, and 33.3% in all scenes, resulting in an overall increase of 16.7%. Regarding translational error, our model showcases improvements of 1.9%, 26.6%, 16.4%, 12.6%, and 45.9% in all scenes, leading to an overall increase of 28.2%.

In addition, positioning accuracy was evaluated with a threshold of 25 cm/2° on Cambridge Landmarks. Overall, our model outperforms PixLoc. Particularly, in St M. Church and Great Court, our model achieves notable improvements of 3.58% and 18.16% in positioning accuracy, respectively, indicating that our model excels in large-scale outdoor scenes.

E. ABLATION EXPERIMENT

We also conducted ablation experiments to compare and analyze the contributions of each improvement point to our model's high-precision localization (threshold of 5 cm/5°). Recall rate was used as the performance metric in the evaluation to comprehensively assess the performance of the network's innovations. PixLoc was used as the baseline model, and the improvement points, namely, Image Gradient

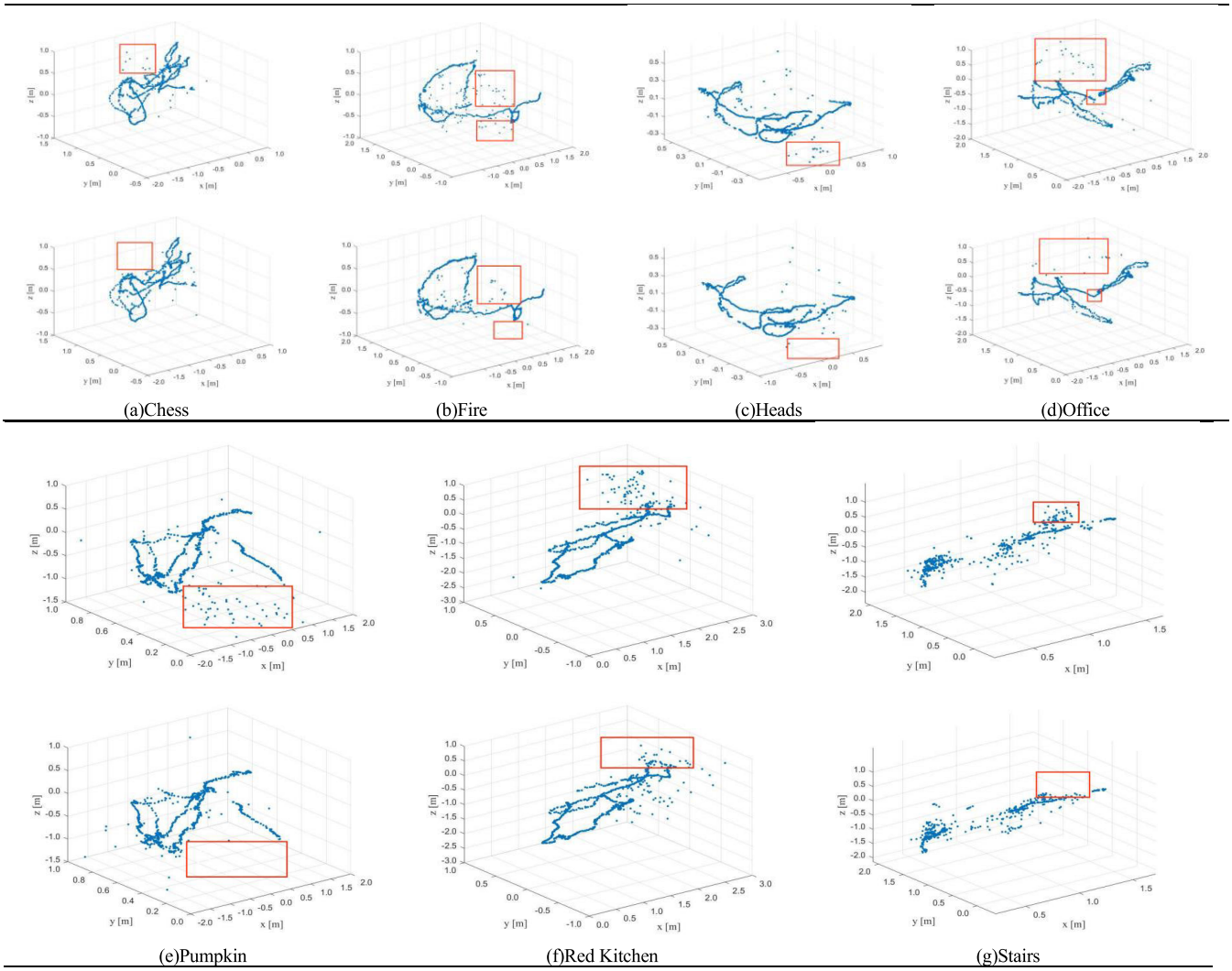


FIGURE 14. The trajectory plots of camera localization for the seven scenes in the 7-Scenes dataset are shown in the figure. The first and third rows depict the localization trajectories of the PixLoc model, while the second and fourth rows represent the localization trajectories of the method proposed in this paper.

TABLE 5. Ablation experiment.

Scene	PixLoc	PixLoc+Point1	PixLoc+Point2	PixLoc+Point3	PixLoc+Point123
Chess	86.55%	85.70%	85.90%	88.95%	88.05%
Fire	83.80%	88.15%	87.40%	84.70%	91.70%
Heads	82.60%	86.00%	86.30%	82.90%	87.90%
Office	76.30%	76.50%	77.68%	79.00%	79.17%
Punpkin	57.45%	56.95%	57.82%	58.65%	58.30%
Red Kitchen	63.64%	66.46%	66.70%	66.18%	67.56%
Stairs	49.00%	54.50%	54.80%	46.90%	53.80%
Average	71.33%	73.47%	73.80%	72.47%	75.21%

Optimization Network, ResUNet-E Network, and loss function optimization, corresponded to Point1, Point2, and Point3, respectively. From Table 5, By enhancing image gradients through the image gradient optimization network, biases in image alignment recognition features are mitigated, preventing the LM algorithm from easily falling into local optima during the iteration process. it is evident that changing the network to ResNet resulted in improved model accuracy in five different scenes, with the most significant improvements observed in Fire and Stairs, i.e., by 4.35% and 5.50%, respectively. Adding the channel attention, ECA-Net, further improved the model's accuracy in six different scenes, with the most notable improvement seen in Stairs, i.e, by 5.80%. Optimizing the loss function also led to enhanced model

performance, particularly in Office and Pumpkin, with accuracy improvements of 2.70% and 2.54%, respectively. The accuracy in Chess and Pumpkin even exceeded that of the model with all improvement points. Finally, incorporating all improvement points significantly increased localization accuracy in all scenes, especially in Stairs and Fire with low or repetitive textures, which showed improvements of 4.80% and 7.90%, respectively. These results indicate that the improved model is more suitable for use in low- or repetitive-texture scenes.

V. CONCLUSION

In response to the challenges faced by scene-agnostic methods in achieving accurate localization in complex scenes, we propose an improved visual localization method based on UNet. By incorporating image gradient optimization network, ResNet, ECA-Net, and optimizing the loss function, the proposed model overcomes the challenges of localization in complex scenarios. Experimental results conducted on two public datasets demonstrate that our model reduces the average rotation and translation errors by 5.9% and 6.5%, respectively, on the 7-Scenes dataset. In the Cambridge Landmarks dataset, it improves the average rotation and translation errors by 16.7% and 27.5%, respectively, enabling more precise predictions of camera poses. Our model, ST-PixLoc, exhibits more reliable localization performance when generalized to unknown scenes.

REFERENCES

[1] I. Abaspur Kazerouni, L. Fitzgerald, G. Dooly, and D. Toal, "A survey of state-of-the-art on visual SLAM," *Expert Syst. Appl.*, vol. 205, Nov. 2022, Art. no. 117734.

[2] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2D-to-3D matching," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 667–674.

[3] M. Donoser and D. Schmalstieg, "Discriminative feature-to-point matching in image-based localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 516–523.

[4] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1744–1756, Sep. 2017.

[5] L. Liu, H. Li, and Y. Dai, "Efficient global 2D-3D matching for camera localization in a large-scale 3D map," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2391–2400.

[6] E. Brachmann and C. Rother, "Visual camera re-localization from RGB and RGB-D images using DSAC," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5847–5865, Sep. 2022.

[7] S. Lynen, B. Zeisl, D. Aiger, M. Bosse, J. Hesch, M. Pollefeys, and T. Sattler, "Large-scale, real-time visual-inertial localization revisited," *Int. J. Robot. Res.*, vol. 39, no. 9, pp. 1061–1084, 2020.

[8] A. Bhowmik, S. Gumhold, C. Rother, and E. Brachmann, "Reinforced feature points: Optimizing feature detection and description for a high-level task," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 4947–4956.

[9] J. Li, G. Li, and T. H. Li, "Attention guided invariance selection for local feature descriptors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 23–28.

[10] S. Baker and I. Matthews, "Lucas–Kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, Feb. 2004.

[11] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 2938–2946.

[12] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu, "Image-based localization using hourglass networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Venice, Italy, Oct. 2017, pp. 870–877.

[13] T. Naseer and W. Burgard, "Deep regression for monocular camera-based 6-DoF global localization in outdoor environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Vancouver, BC, Canada, Sep. 2017, pp. 1525–1530.

[14] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6555–6564.

[15] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using LSTMs for structured feature correlation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 627–637.

[16] V. Balntas, S. Li, and V. Prisacariu, "RelocNet: Continuous metric learning relocalisation using neural nets," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 8–14.

[17] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "DSAC—Differentiable RANSAC for camera localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2492–2500.

[18] L. von Stumberg, P. Wenzel, Q. Khan, and D. Cremers, "GN-Net: The Gauss–Newton loss for multi-weather relocalization," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 890–897, Apr. 2020.

[19] P.-E. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl, and T. Sattler, "Back to the feature: Learning robust camera localization from pixels to pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 3246–3256.

[20] C. Zhang, X. Wang, R. Xiong, X. Fan, and D. Zhao, "Local–global dynamic filtering network for video super-resolution," *IEEE Trans. Comput. Imag.*, vol. 9, pp. 963–976, 2023.

[21] K. Ning, Y. Liu, Y. Su, and K. Jiang, "Point-voxel and bird-eye-view representation aggregation network for single stage 3D object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 3223–3235, Mar. 2023.

[22] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quart. Appl. Math.*, vol. 2, no. 2, pp. 164–168, Jan. 1944.

[23] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. Soc. Ind. Appl. Math.*, vol. 11, no. 2, pp. 431–441, 1963.

[24] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in RGB-D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 2930–2937.

[25] E. Brachmann and C. Rother, "Learning less is more—6D camera localization via 3D surface regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4654–4662.

[26] M. Cai, H. Zhan, C. S. Weerasekera, K. Li, and I. Reid, "Camera relocalization by exploiting multi-view constraints for scene coordinates regression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3769–3777.

[27] L. Yang, Z. Bai, C. Tang, H. Li, Y. Furukawa, and P. Tan, "SANet: Scene agnostic network for camera localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 42–51.

[28] S. Tang, C. Tang, R. Huang, S. Zhu, and P. Tan, "Learning camera localization via dense scene matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 1831–1841.

[29] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[30] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany, 2015, pp. 5–9.

[31] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 11531–11539.

[32] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[33] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 8–14.

[34] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from Internet photos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2041–2050.

JING WANG received the Ph.D. degree. She is a Master's Supervisor, and has been engaged in teaching and research in the fields of communication and information systems, signal and information processing, and digital image processing. She has led and participated in nearly 20 projects, including the National 863 Program, 973 Program, the National Natural Science Foundation of China, Shaanxi Provincial Natural Science Foundation, Shaanxi Provincial Industrial Research Project, Shaanxi Provincial Department of Education Special Project, Xi'an Beilin District Science and Technology Plan Project, and enterprise collaborative projects. Her project research achievements have earned her two second prizes in science and technology from Xi'an University of Science and Technology. Her research interests include radar signal processing, key technologies and applications of interferometric synthetic aperture radar (InSAR), and computer vision.

YIBO WANG was born in February 1999. He is currently pursuing the master's degree with Xi'an University of Science and Technology. His research interests include deep learning, computer vision, and FPGA technology.

YUCHU JIN was born in July 1997. She received the master's degree from Xi'an University of Science and Technology. She is currently a Technical Officer with the Electronic Research and Development Center, BYD Company. Her research interests include deep learning and camera pose estimation.

CHENG GUO was born in July 2000. He is currently pursuing the master's degree with Xi'an University of Science and Technology. His research interests include computer vision and object pose estimation.

XUHUI FAN was born in Shanxi, China, in 1989. She received the M.S. degree in communication and information systems and the Ph.D. degree in electronic science and technology from Northwestern Polytechnical University, Xi'an, China, in 2016 and 2021, respectively. Since 2021, she has been a Lecturer with the College of Communication and Information Technology, Xi'an University of Science and Technology, Xi'an. In 2023, she was elected as an Associate Professor with the College of Communication and Information Technology, Xi'an University of Science and Technology. Her research interests include array antenna synthesis, radar signal processing and its applications, and optimization algorithms.