2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Combining Scene Coordinate Regression and Absolute Pose Regression for Visual Relocalization

Jiahao Ruan[1], Li He[2], Yisheng Guan[1*], Hong Zhang[2]

***Abstract*— Visual relocalization is a fundamental problem in computer vision and robotics. Recently, regression-based methods become popular and they can be categorized into two classes: absolute pose regression and scene coordinate regression. In this work, we present a combined regression network that jointly learns scene coordinate regression and absolute pose regression for single-image visual relocalization. The proposed network composes of a feature encoder and two regression branches with uncertainty modeling. In particular, we design a deep feature conditioning module, aiming at propagating the coarse pose information in absolute pose regression to inform the predictions in scene coordinate regression. The proposed network is trained in an end-to-end fashion to learn both regression tasks. Moreover, we propose an uncertainty-driven RANSAC algorithm that incorporates the predicted scene coordinates and their uncertainties to solve the camera pose during inference. To the best of our knowledge, this work is the first to combine scene coordinate regression and pose regression in a hierarchical framework for visual relocalization. Experiments on indoor and outdoor benchmarks demonstrate the effectiveness and the superiority of the proposed method over the state-of-the-art methods.**

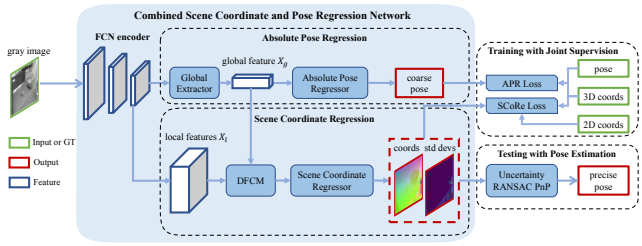

Fig. 1. Training and testing pipeline of the proposed method for visual relocalization. The proposed network composes of a fully convolutional encoder and two regression branches. Given the local feature produced by the encoder, the absolute pose regression branch extracts the global feature and regresses the camera pose directly. The scene coordinate regression branch takes local and global features as the input and then regresses pixel-wise scene coordinates. At training time, the network is trained in an end-to-end fashion and jointly supervised by the absolute pose regression loss and scene coordinate regression loss. At test time, an uncertainty-driven RANSAC-based PnP solver is adopted to estimate the precise pose with predicted coordinates and uncertainties.

I. INTRODUCTION

Visual relocalization aims at resolving the 6-DoF (degree of freedom) camera pose corresponding to a query image in a known scene. It is a fundamental and crucial component of autonomous robots to localize themselves globally for further accomplishing a navigation task.

Traditional visual relocalization methods establish data association between the query image and the known scene by feature matching. These methods can be categorized into two classes: structure-based methods [1], [2] which rely on local feature matching, and retrieval-bassed methods [3], [4] which search the most similar image by global descriptor matching. Instead of matching feature descriptors, absolute pose regression (APR) [5]–[10] and scene coordinate regression (SCoRe) [11]–[19] resort to deep neural networks to directly regress the camera pose or pixel-wise 3D scene coordinates. Compared with classic feature matching-based methods, regression-based methods encode the scene representation implicitly with model weights, and therefore do not require a database or a 3D point cloud and have high efficiency at test time.

Nevertheless, current APR methods simply leverage a network to map the image context to the camera pose and always fail to surpass the performance of retrieval baseline in terms of accuracy [20]. Although SCoRe methods can achieve better relocalization performance compared to other methods [13], they suffer from two main problems: the generalization w.r.t. viewpoint change, and ambiguous predictions caused by visual similarity between local image patches in texture-less or repetitive areas.

Inspired by the hierarchical relocalization methods [21], [22] that combine a retrieval-based method and a structure-based method to achieve accurate and efficient visual relocalization, our main argument is that APR is complementary to SCoRe, and combining the two in a hierarchical framework can also improve the performance of visual relocalization. However, APR methods and SCoRe methods have always been studied separately and how to effectively combine two methods in a framework remains unsolved.

In this paper, we propose a novel network for single-image visual relocalization, which jointly learns scene coordinate regression and absolute pose regression and combines two regression methods in a hierarchical relocalization framework. In this network, we design a Deep Feature Conditioning Module (DFCM) to feed back the coarse pose information in the APR branch to the SCoRe branch. Specifically, the DFCM transforms the local features with the guidance of

[1]J. Ruan and Y. Guan (corresponding author) are with the School of Electromechanical Engineering, Guangdong University of Technology, 510006 Guangzhou, China (e-mail: jhruan@foxmail.com; ysguan@gdut.edu.cn).

[2]L. He and H. Zhang are with Shenzhen Key Laboratory of Robotics and Computer Vision, and with the Department of Electrical and Electronic Engineering, Southern University of Science and Technology, 518055 Shenzhen, China (e-mail: {hel, hzhang}@sustech.edu.cn).

This work was supported in part by the The Pearl River Talent Recruitment Program (No. 2019QN01X761), Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515140044) and Guangdong Province Special Fund for Modern Agricultural Industry Common Key Technology R&D Innovation Team (No. 2019KJ129).

global feature used in APR, enhancing viewpoint invariance and discrimination of local features for SCoRe. Moreover, we use the robust KL loss derived from the KL divergence between the predicted distribution and target distribution to enable the network to learn scene coordinate regression with uncertainty awareness. Based on the uncertainty modeling, we propose an uncertainty-driven RANSAC-based PnP algorithm to resolve the camera pose with predicted scene coordinates and uncertainties at test time.

To validate the proposed method, we present relocalization results on the indoor dataset 7-Scenes [11] and the outdoor dataset Cambridge Landmarks [5]. The experimental results show the great advantages of our approach over the state-of-the-art methods.

In summary, our main contributions in this work are as follows: (1) We propose a combined scene coordinate and pose regression network for visual relocalization, which jointly regresses pixel-wise scene coordinates and absolute camera pose for an input image. (2) We design a deep feature conditioning module to propagate coarse pose information in absolute pose regression to inform scene coordinate regression. (3) Based on the uncertainty modeling of network predictions, we propose an uncertainty-driven RANSAC-based PnP algorithm to further improve the relocalization performance.

II. RELATED WORK

In the following, we discuss the main categories of methods for visual relocalization.

A. Structure-based Methods

The structure-based methods [1], [2] reconstruct the known environment as a sparse 3D point cloud by SfM [23], such that each 3D point has one or several visual feature descriptors. 2D-3D correspondences between the query image and 3D model are established by matching local features such as SIFT [24], and SuperPoint [25]. Typically, the camera pose can ultimately be solved by the PnP algorithm [26] in a RANSAC optimization framework [27] with the matched 2D-3D correspondences. The structure-based methods can achieve accurate relocalization but they may fail to find keypoints or match descriptors under certain circumstances, such as motion blur and texture-less areas.

B. Retrieval-based Methods

The retrieval-based methods represent the known environment as an image database consisting of a collection of reference images with known camera poses and global descriptors. Given a query image, the most similar database image can be retrieved by matching global image descriptors such as DenseVLAD [3] and NetVLAD [4], and used to approximate the pose of the query image. The retrieval-based methods are inferior to the structure-based methods in terms of accuracy. However, they can serve as an initialization process in a hierarchical relocalization framework and achieve precise pose combining with relative pose estimation [28], [29], structure-based methods [21], [22], [30], or feature alignment [31].

C. Absolute Pose Regression

Instead of matching descriptors to achieve data association, PoseNet [5] directly regresses the camera pose of the query image by training a neural network. The network encodes the relationship between image content and camera pose in its parameters, so that no 3D model or database is required at test time. Compared with feature matching-based methods, PoseNet has higher relocalization efficiency and a lower memory footprint. The variants of PoseNet have been proposed various improvement schemes, such as network architecture [6], [9], [10], loss function [7], and adding constraint [8]. However, experimental results demonstrated that current APR approaches are inherently more similar to image retrieval methods than to structure-based methods, and they fail to consistently outperform retrieval-based methods [20].

D. Scene Coordinate Regression

Unlike APR methods that directly regress camera pose, scene coordinate regression methods first establish 2D-3D correspondences by regressing 3D scene coordinates for image pixels and then calculate camera pose by RANSAC-based pose optimization, similar to structure-based methods.

Shotton *et al.* [11] proposed the first SCoRe method that utilizes random forest to regress pixel-wise 3D scene coordinates for RGB-D images. With the development of deep learning, random forests were replaced by neural networks which can only take RGB images as input in recent SCoRe works. Brachmann *et al.* [12] proposed a differential RANSAC (DSAC) method and extended variants [13], [14], [16] that enable the SCoRe pipeline to be trained in an end-to-end fashion. To extend the SCoRe method to solve temporal visual relocalization, KFNet [17] integrates the Kalman filter into a recurrent CNN network. HSC-Net [18] introduces a hierarchical joint classification-regression network architecture that can scale to large scenes for robust visual relocalization. However, it requires a dense 3D model and expensive data processing to partition the known scene. In order to achieve SCoRe with viewpoint invariance, SFT-CR [19] proposes a spatial feature transformation network and the CoordConv scheme was employed for enhancing discrimination of features, but it was only validated in the indoor environment.

III. METHODOLOGY

Current APR methods and SCoRe methods are popular solutions to visual relocalization. Although both methods make use of convolutional neural network (CNN), the former focuses on encoding the image contents into a compact global feature, whereas the latter extracts local features of image regions. In this work, we focus on jointly learning both regression in a hierarchical framework. Fig. 1 illustrates the training and testing pipeline of the proposed *combined scene coordinate and pose regression network* (CCP-Net) for visual relocalization.

Our CCP-Net consists of a feature encoder and two regression branches. Given an input grayscale image, the

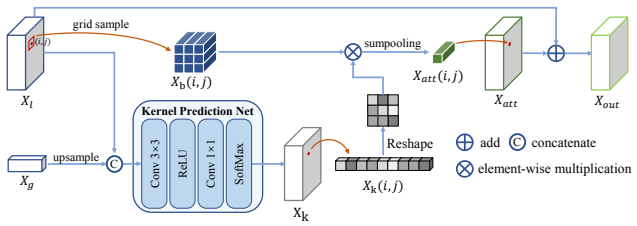

Fig. 2. Structure of the proposed deep feature conditioning module. This module takes the local and global features as input and transforms the local features with the guidance of pose information attached to the global feature.

network first extracts local features by the fully convolutional feature encoder. Then the absolute pose regression branch extracts the global feature and regresses the camera pose, such that the inherent relationship between the global feature and camera pose is established. Many previous works [18], [21], [22] show that a coarse-to-fine relocalization framework is useful to improve the relocalization performance. To this end, we design a deep feature conditioning module (DFCM) to transform local features with global-feature guidance, such that the pose generated from the absolute pose regression branch is fed back to the scene coordinate regression branch. The transformed local features are then used to predict scene coordinates and corresponding uncertainties. To enable end-to-end training of the proposed model, we introduce two task-specific loss functions and balance them with learnable parameters. At test time, the network predictions are used to solve the precise camera pose by a PnP solver within an Uncertainty-Driven RANSAC (UD-RANSAC) framework.

A. Feature Extraction

In feature extraction stage, the proposed network takes a single channel grayscale image with the size of $H \times W$ as input and produces a local feature map $X_l \in \mathbb{R}^{h \times w \times c}$ and a global feature $X_g \in \mathbb{R}^c$, where $h = H/8$, $w = W/8$ and $c = 512$. We use the ResNet [32]-based fully convolutional network (FCN) in DSAC* [14] as the backbone to extract local features, which are then fed into the global extractor for global feature encoding. For the global extractor, we adapt a 3×3 convolution layer and a global average pooling layer followed by the self-attention module [9] which enforces the extractor to focus on geometrically robust features.

B. Deep Feature Conditioning Module

The purpose of this module is to propagate camera pose in the APR branch to inform the predictions in the SCoRe branch. Using the viewpoint and position information contained in the global feature, this module decouples the local feature from viewpoints and alleviates the ambiguity problem. A simple and common way to fuse the global and local features is directly adding or concatenating these two features. However, using features with global receptive field to regress coordinates is susceptible to overfitting in the case of limited training data [18]. Inspired by SFT-CR [19], we design a deep feature conditioning module (DFCM) that transforms the local features by a local attention operation with the guidance of pose information attached to the global feature.

As shown in Fig. 2, the global feature X_g is firstly upsampled to match the size of the local X_l, and then both the local and the global features are concatenated as the input to the kernel prediction network to predict pixel-wise $n \times n$ attention kernels $X_k \in \mathbb{R}^{h \times w \times n \times n}$. Specifically, we first employ an $n \times n$ convolutional layer with a ReLU function rather than dot product to encode local features with their neighbors, and then the attention kernels are generated by a 1×1 convolutional layer and a softmax function. To transform each feature with its nearby features in an $n \times n$ local patch, we use a grid sample operation to generate the feature block $X_b \in \mathbb{R}^{h \times w \times n \times n \times c}$ from the local feature map X_l with n^2 times sampling. With the pixel-wise attention kernels, the attention feature X_{att} at position (i, j) can be obtained as follows:

$$X_{att}(i,j) = P(X_k(i,j) \otimes X_b(i,j)) \tag{1}$$

where $\otimes$ denotes the element-wise multiplication over the spatial domain and P represents the sum pooling operation.

At last, following the residual learning strategy, we directly add the attention feature map with the local feature map to obtain the output feature map, $X_{out} = X_l \oplus X_{att}$, where $\oplus$ denotes element-wise addition.

C. Regression and Loss Function

In order to jointly learn scene coordinate and absolute pose regression in an end-to-end fashion, we introduce task-specific regressors and loss functions for both regression tasks.

1) Scene Coordinate Regression: We leverage the same regressor in DSAC* [14], which contains three 1×1 convolution layers for scene coordinate regression. Instead of solely regressing 3D coordinates $\mathbf{x}^{3D}$, we introduce the aleatoric uncertainty to focus on minimizing coordinate errors of stable regions, making network predicts the probability distribution of scene coordinates. In particular, we use the multivariate Gaussian distribution with independent components to model 3D scene coordinates. For simplicity, we assume that x, y and z components have the same variances, and therefore the network outputs the expectations μ_x, μ_y, μ_z and a single standard deviation σ.

Inspired by [33], we design a robust Gaussian KL loss to learn the multivariate Gaussian distribution:

$$L_{\mathrm{rKL}} = \begin{cases} \frac{1}{2}e^2 + \sum_{k=1}^{d} \log \sigma_k, & e \leq \sqrt{2d}, \\ \sqrt{2d}e - d + \sum_{k=1}^{d} \log \sigma_k, & e \geq \sqrt{2d}. \end{cases} \tag{2}$$

where d is the dimensions of distribution, e denotes the weighted Euclidean distance $\sqrt{\sum_{k=1}^{d} \frac{(\mu_k - y_k)^2}{\sigma_k}}$, and y_k is the ground truth of kth component.

Therefore, the 3D-coordinate regression loss can be formulated with the 3D form of robust Gaussian KL loss (2):

$$\mathcal{L}_{coord} = \frac{1}{M} \sum_{i=1}^{h \times w} m_i L_{\mathrm{rKL},i}^{\mathrm{3D}} \tag{3}$$

where m_i is a Boolean variable indicating whether the ground truth of a 3D coordinate is available and M denotes the total number of available coordinates. In our implementation, we let the network predict $\log \sigma$ as an alternative to avoid gradient explosion.

Because there are certain pixels without ground truth 3D coordinates, we introduce the reprojection loss as an additional optimization objective to learn the 3D scene coordinate distributions in a self-supervised manner. We first transform 3D distributions of scene coordinates $p(\mathbf{x}^{3D}|\boldsymbol{\mu}^{3D}, \boldsymbol{\Sigma}^{3D})$ into 2D distributions of pixel coordinates with ground truth pose $\hat{\mathbf{T}}$ composed of rotation $\hat{\mathbf{R}}$ and translation $\hat{\mathbf{t}}$:

$$p(\mathbf{x}^{2D}|\mathbf{x}^{3D}, \hat{\mathbf{T}}) = \mathcal{N}(\pi(\hat{\mathbf{R}}\boldsymbol{\mu}^{3D} + \hat{\mathbf{t}}), \mathbf{J}_\pi \hat{\mathbf{R}} \boldsymbol{\Sigma}^{3D} \hat{\mathbf{R}}^T \mathbf{J}_\pi^T) \tag{4}$$

where π indicates the projection function and $\mathbf{J}_\pi$ indicates the Jacobian of projection function w.r.t. camera coordinates. Then we use ground truth pixel coordinates and transformed distributions to compute reprojection loss according to the 2D form of robust Gaussian KL loss (2):

$$\mathcal{L}_{reproj} = \frac{1}{h \times w} \sum_{i=1}^{h \times w} L_{\mathrm{rKL},i}^{\mathrm{2D}} \tag{5}$$

Since the distribution transformation is differentiable, the 3D scene coordinate distributions can be learned indirectly by minimizing the 2D reprojection loss. Finally, the overall loss for scene coordinate regression becomes:

$$\mathcal{L}_{score} = \mathcal{L}_{coord} + \mathcal{L}_{reproj} \tag{6}$$

2) Absolute Pose Regression: The APR regressor directly maps the global feature to rotation $\mathbf{R} \in$ *SO(3)* and translation $\mathbf{t} \in \mathbb{R}^3$ through Multilayer Perceptrons (MLPs). We employ a continuous 6-dimensional representation [34] for 3D rotation which is composed of two 3-dimensional vectors, and the 3D rotation can be recovered by vector normalization and cross product.

Previous APR works usually use separate loss functions for rotation and translation, and suffer from the difficulty of balancing two losses. Instead, we utilize the point matching loss [35] to learn camera pose. Given the ground truth pose $\hat{\mathbf{T}} = [\hat{\mathbf{R}}|\hat{\mathbf{t}}]$, predicted pose $\mathbf{T} = [\mathbf{R}|\mathbf{t}]$ and 3D points $\mathbf{X}$ which are visible in the image, this loss is formulated as:

$$\mathcal{L}_{\mathrm{PM}} = \frac{1}{|\mathbf{X}|} \sum_{\mathbf{x}_i \in \mathbf{X}} \|(\hat{\mathbf{R}}\mathbf{x}_i + \hat{\mathbf{t}}) - (\mathbf{R}\mathbf{x}_i + \mathbf{t})\|_1 \tag{7}$$

where $|\mathbf{X}|$ denotes the number of 3D points. Furthermore, we use a learnable parameter γ to normalize the scale of loss:

$$\mathcal{L}_{\mathrm{apr}} = \frac{1}{\gamma} \mathcal{L}_{\mathrm{PM}} + \log \gamma \tag{8}$$

Finally, we use two hyper-parameter λ_1 and λ_2 to weight two task-specific loss terms.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{score} + \lambda_2 \mathcal{L}_{apr} \tag{9}$$

D. Pose Estimation with Uncertainty-Driven RANSAC

Generally, the absolute pose regression can only predict a coarse camera pose, and therefore a pose estimation algorithm is still required to achieve accurate relocalization with predicted scene coordinates, which define dense 2D-3D correspondences between the image and the scene.

Typically, current SCoRe methods utilize a RANSAC-based PnP algorithm to recover the camera pose due to potential outliers in the predicted scene coordinates. Since our network also predicts coordinate uncertainties, we propose an uncertainty-driven RANSAC algorithm (UD-RANSAC) that integrates the predicted coordinate uncertainty into the RANSAC framework to further alleviate the influence of outliers. The UD-RANSAC estimates camera pose with the following steps:

1) Uncertainty-driven Pose Sampling: RANSAC samples correspondences uniformly randomly to generate multiple pose hypotheses with the PnP solver. In order to increase the chance of sampling an outlier-free correspondences subset, we use the coordinate uncertainty σ predicted by network to represent the noise level, and sample correspondences according to the multinomial distribution p_i defined as:

$$p_i = \frac{\sigma_i^{-2}}{\sum_{i=1}^{h \times w} \sigma_i^{-2}} \tag{10}$$

2) Hypothesis Selection: To select the optimal hypothesis pose, RANSAC usually sorts hypotheses by inlier counting. However, it is difficult to select the optimal pose as the ratio of outliers increases. In our work, we replace inlier counting with the maximum logarithmic likelihood of all 2D-3D correspondences to select the hypothesis:

$$\mathbf{T}^* = \arg\max_{\mathbf{T}_j} \sum_{i=1}^{h \times w} \log p(\mathbf{x}_i^{2D}|\mathbf{x}_i^{3D}, \mathbf{T}_j) \tag{11}$$

where $p(\mathbf{x}_i^{2D}|\mathbf{x}_i^{3D}, \mathbf{T}_j)$ is the transformed 2D distribution similar to (4).

3) Pose Refinement: Following the DSAC pipeline [14], we refine the chosen pose $\hat{\mathbf{T}}$ by minimizing the reprojection error of the inlier set $\mathcal{I}$. The proposed UD-RANSAC determines inlier according to the negative logarithmic likelihood rather than the reprojection error, such that both the reprojection error and the uncertainty are taken into account:

$$\mathcal{I} = \{\mathbf{x}_i^{3D}| - \log p(\mathbf{x}_i^{2D}|\mathbf{x}_i^{3D}, \mathbf{T}^*) < \tau\} \tag{12}$$

where τ denotes the likelihood threshold. We perform iterative refinement until the inlier set $\mathcal{I}$ converges. The final camera pose is then obtained.

IV. EXPERIMENTS

A. Datasets and Metrics

Datasets: The **7-Scenes** dataset [11] is a widely used RGB-D dataset recorded using a KinectV1, and consists of seven indoor scenes with challenging conditions such as motion blur, repeated structures, reflective surfaces and textureless areas. Ground truth poses of images and dense 3D

TABLE I

THE 5CM-5° ACCURACY AND MEDIAN POSE ERROR (M, °) ON 7-SCENES DATASET OF DIFFERENT METHODS

7-Scenes	Active Search [2]		PixLoc [31]		HSC-Net [18]		SCoordNet [17]		DSAC* [14]		DSAC*(re) [14]		SFT-CR [19][1]		Ours	
	Acc.	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**	**Med. Err.**
Chess	—	0.04, 1.96	—	0.02, 0.80	97.5	0.021, 0.68	—	0.019, 0.63	97.5	0.018, 1.10	97.5	0.018, 0.63	—	0.021, 0.70	**99.4**	**0.015, 0.47**
Fire	—	0.03, 1.53	—	0.02, 0.73	**96.7**	0.022, 0.87	—	0.023, 0.91	93.5	0.019, 1.24	93.5	0.019, 0.89	—	0.020, 0.78	95.2	**0.017, 0.68**
Heads	—	0.02, 1.45	—	**0.01**, 0.82	**100**	0.012, 0.86	—	0.018, 1.26	99.8	0.011, 1.82	99.8	0.011, **0.68**	—	0.011, 0.81	99.7	0.011, 0.70
Office	—	0.09, 3.61	—	0.03, 0.82	86.5	0.027, 0.79	—	0.026, 0.73	90.0	0.025, 1.15	90.0	0.025, 0.73	—	0.024, 0.66	**93.3**	**0.022, 0.61**
Pumpkin	—	0.08, 3.10	—	0.04, 1.21	59.9	0.040, 1.02	—	0.039, 1.09	62.4	0.039, 1.34	62.4	0.039, 1.02	—	0.034, 0.98	**67.3**	**0.033, 0.93**
Redkitchen	—	0.07, 3.37	—	0.03, 1.20	65.5	0.040, 1.18	—	0.039, 1.18	65.3	0.038, 1.68	65.3	0.038, 1.24	—	0.034, 1.06	**75.5**	**0.030, 0.97**
Stairs	—	0.03, 2.22	—	0.05, 1.30	87.5	0.031, 0.82	—	0.037, 1.06	87.5	0.029, 1.16	87.5	0.029, 0.79	—	0.035, 0.97	**88.9**	**0.024, 0.66**
Average	—	0.05, 2.46	—	0.03, 0.98	84.8	0.028, 0.89	—	0.029, 0.98	85.2	0.026, 1.36	85.2	0.026, 0.85	86.1	0.026, 0.85	**88.5**	**0.022, 0.72**

[1] SFT-CR does not report accuracy for each scene.

models are also provided using KinectFusion. **Cambridge Landmarks** [5] is an outdoor relocalization RGB dataset that is composed of RGB images of six landmarks in Cambridge. Compared to 7-Scenes, each landmark occupies an area of several hundred or thousand square meters. In this dataset, ground truth poses and sparse 3D point cloud models are generated with a SfM tool. We follow previous works, using the sparse model to render sparse scene coordinate ground truth and omitting the street scene.

Evaluation Metrics: Following the practice in works [13] [14], we adopt two typical metrics: (1) the median pose error; and (2) the percentages of the test images with error below a specific threshold which is set to 5cm in translation and 5° in rotation for the 7-Scenes dataset.

B. Implementation Details

Our method employs the RGB + 3D Model setting of DSAC* [14], in which the 3D model or depth image is only available for training. For the training process, the network takes grayscale images rescaled to 480px height as input and data augmentation is applied according to [14]. The kernel size of DFCM is 3×3 and the weight of scene coordinate regression loss λ_1 and absolute pose regression loss λ_2 are set to 10 and 1 respectively. We train our network from scratch with a batch size of 1 image and 1M iterations for one scene. The AdamW optimizer is employed with an initial learning rate of 10^{-4} and a weight decay of 10^{-4}. We halve the learning rate every 50K iterations for the last 200K iterations with an exponential decay schedule. For pose estimation at test time, we sample 16 pose hypotheses and use a likelihood threshold $\tau = 6.5$. We implement our method based on PyTorch and run our code[1] on an Intel Xeon Gold 5218R@2.10 GHz CPU with a single NVIDIA RTX 3090 GPU.

C. Comparison with the State-of-the-Art Methods

Performance on 7-Scenes: Using the official training and test sets provided by the dataset, we evaluate the proposed method and compare it with the state-of-the-art methods. We report the relocalization results taken from original papers for all the competitors. For DSAC* [14], we also report the reproduction results as DSAC*(re) which are obtained by running the pre-trained model provided by the authors.

[1]The code of our method: https://github.com/JhRuan96/CCPNet

TABLE II

THE MEDIAN POSE ERROR (M, °) ON CAMBRIDGE LANDMARKS DATASET OF DIFFERENT METHODS

Cambridge	PixLoc [31]	NG-DSAC++ [16]	HSC-Net [18]	SCoordNet [17]	DSAC* [14]	Ours
GreatCourt	0.30, **0.14**	0.35, 0.2	**0.28**, 0.2	0.43, 0.20	0.49, 0.25	0.33, 0.25
KingsCollege	0.14, 0.24	**0.13**, **0.2**	0.18, 0.3	0.16, 0.29	0.15, 0.29	0.14, 0.26
OldHospital	**0.16**, 0.32	0.22, 0.4	0.19, 0.3	0.18, **0.29**	0.21, 0.41	0.20, 0.35
ShopFacade	**0.05**, **0.23**	0.06, 0.3	0.06, 0.3	**0.05**, 0.34	**0.05**, 0.25	**0.05**, 0.25
StMarysChurch	0.10, 0.34	0.10, 0.3	**0.09**, **0.3**	0.12, 0.36	0.13, 0.45	**0.09**, **0.30**
Average	**0.15**, **0.25**	0.17, 0.28	0.16, 0.28	0.19, 0.30	0.21, 0.34	0.16, 0.28

As is shown in Table I, our method achieves the best performance on most scenes in terms of both 5cm-5° accuracy and median pose error. Compared with DSAC* baseline, our method leverages the same fully convolutional network to extract local features but gets better results without an additional end-to-end training process. Although SFT-CR also uses a local attention module to decouple features from viewpoints, it is outperformed by our method in all scenes. We attribute the performance gain to using the APR branch for learning a pose-aware global feature. Furthermore, compared to HSC-Net, our method does not adopt hierarchical discrete location labeling but learns the absolute pose regression task, enabling our method to avoid expensive data preprocessing while following a coarse-to-fine relocalization pipeline.

Performance on Cambridge Landmarks: Table II illustrates the comparison of CCP-Net with existing end-to-end methods in terms of median pose error on Cambridge Landmarks dataset. Overall, the hierarchical method PixLoc achieves the best results but it requires a 3D point cloud and a image database at test time. For most scenes, the proposed method outperforms the DSAC* baseline and achieves comparative performance with other methods despite it only being trained with a sparse 3D model.

D. Effectiveness of Absolute Pose Regression

In order to demonstrate that the global feature contains coarse pose information by learning absolute pose regression, we also compare our absolute pose regression results with other APR methods and the retrieval baseline on 7-Scenes and Cambridge Landmarks datasets, as shown in Table III.

TABLE III

THE AVERAGE MEDIAN POSE ERROR (M, °) AND THE RESPECTIVE RANKING (TRANS. / ROTATION) ON 7-SCENES AND CAMBRIDGE LANDMARKS DATASETS OF DIFFERENT APR METHODS

Method	7-Scenes		Cambridge	
	Average	Rank	Average	Rank
DenseVLAD+Inter. [20]	0.24, 11.7	6/9	1.67, 4.87	7/5
PoseNet [5]	0.45, 9.94	9/8	2.09, 6.84	8/8
LSTM-PN [6]	0.31, 9.85	8/7	1.30, 5.52	3/7
PoseNet-Learnable [7]	0.24, 7.87	6/5	1.43, 2.85	4/3
GeoPoseNet [7]	0.23, 8.12	5/6	1.63, 2.86	5/4
MapNet [8]	0.21, 7.77	4/4	1.63, 3.64	5/5
AtLoc [9]	0.20, 7.56	3/3	—	—
MS-Transfomer [10]	0.18, 7.28	2/2	1.28, 2.73	2/2
Ours	0.16, 5.19	1/1	0.98, 2.42	1/1

TABLE IV

MODEL CAPACITY AND EFFICIENCY

	DSAC*	SFT-CR	Reg-only	HSC-Net	CCP-Net*	CCP-Net
Model Size	28MB	105MB	104MB	165MB	48MB	48MB
Runtime per frame	37.5ms	—	71.2ms	92.8ms	39.7ms	75.2ms

We report the average median pose error and the respective ranking (translation / rotation) of different methods in each dataset. Note that for Cambridge Landmarks, the result of Great Court Landmark is not counted because some methods did not report their results in original papers.

From Table III, the APR branch of our method achieves the highest relocalization accuracy among contemporary APR solutions on both indoor and outdoor datasets, indicating that the absolute pose regression branch of our network is effective.

E. Model Capacity and Efficiency

We further investigate the model capacity and efficiency of the proposed CCP-Net method. With the platform of an Intel Xeon Gold 5218R@2.10 GHz CPU with a single NVIDIA RTX 3090 GPU, the model size in terms of the number of parameters size and the average relocalization time per frame of different methods are reported in Table IV. Note that we also report a variant of our method that estimates camera pose by a RANSAC-based PnP, denoted as CCP-Net*. As shown in Table IV, DSAC* is the most efficient method with the smallest model capacity. Compared to DSAC*, our method has more model parameters because of the employment of the APR branch and the deep feature conditioning module. However, the model size of our method is still much smaller than many methods. In terms of runtime, CCP-Net* is almost the same as DSAC* but CCP-Net consumes more time owing to the uncertainty transformation in UD-RANSAC.

F. Ablation Studies

To fully validate the different components proposed in this work, we evaluate its four different variants CCP-Net-I, CCP-Net-II, CCP-Net-III and CCP-Net-IV. These variants are defined according to whether uncertainty modeling, APR branch, deep feature conditioning module and UD-RANSAC are considered. Relocalization results of different variants on 7-Scenes dataset are presented in Table V.

TABLE V

ABLATION STUDIES OF DIFFERENT MODULES ON 7-SCENES DATASET

Method	Uncertainty	APR	DFCM	UD-RANSAC	Accuracy	Median Error (cm, °)
CCP-Net-I	—	—	—	—	79.0	2.98, 1.00
CCP-Net-II	✓	—	—	—	82.3	2.95, 0.92
CCP-Net-III	✓	✓	—	—	82.7	2.77, 0.91
CCP-Net-IV	✓	✓	✓	—	87.6	2.28, 0.75
CCP-Net	✓	✓	✓	✓	88.5	2.19, 0.72

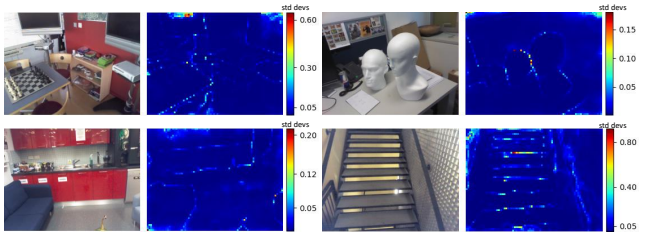

Fig. 3. Visualization of predicted uncertainties. Our network predicts large uncertainties at the regions with unstable measurement.

Comparing the results of CCP-Net-I and CCP-Net-II, the relocalization accuracy is significantly improved with the uncertainty modeling. The predicted uncertainty ought to reflect our degree of trust in the prediction of scene coordinates. To validate the significance of uncertainty modeling, we also visualize the predicted uncertainty in Fig. 3. It is shown that our network predicts large uncertainty values in unstable regions such as object boundaries, reflection surfaces and texture-less planes.

As seen in the results of CCP-Net-III, the APR branch enabling the network to be trained in an end-to-end fashion is also beneficial to relocalization. Moreover, the result of CCP-Net-IV indicates that the deep feature conditioning module can effectively boost the relocalization performance in terms of both the accuracy and the median pose error. Finally, the advantage of the proposed uncertainty-driven RANSAC algorithm can also be concluded by the comparison results.

V. CONCLUSION

In this work, we propose a novel hierarchical regression-based approach enabled by combining scene coordinate regression and pose regression in a compact network for visual relocalization. A deep feature conditioning module is designed to transfer pose information from absolute pose regression to scene coordinate regression. Based on the uncertainty modeling of network prediction, we further improve the relocalization accuracy by the uncertainty-driven RANSAC algorithm. Experiments on two benchmarks indicate that our method achieves a competitive or better performance against the state-of-the-art methods. In our future work, we shall extend our method to jointly learn multi-scene regression-based relocalization.

REFERENCES

[1] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2d-to-3d matching," in *International Conference on Computer Vision*, 2011, pp. 667–674.

[2] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1744–1756, 2016.

[3] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1808–1817.

[4] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

[5] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2938–2946.

[6] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using lstms for structured feature correlation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 627–637.

[7] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5974–5983.

[8] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, "Geometry-aware learning of maps for camera localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2616–2625.

[9] B. Wang, C. Chen, C. X. Lu, P. Zhao, N. Trigoni, and A. Markham, "Atloc: Attention guided camera localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 06, 2020, pp. 10 393–10 401.

[10] Y. Shavit, R. Ferens, and Y. Keller, "Learning multi-scene absolute pose regression with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2733–2742.

[11] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2930–2937.

[12] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac-differentiable ransac for camera localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6684–6692.

[13] E. Brachmann and C. Rother, "Learning less is more-6d camera localization via 3d surface regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4654–4662.

[14] E. Brachmann and C. Rother, "Visual camera re-localization from rgb and rgb-d images using dsac," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5847–5865, 2021.

[15] X. Li, J. Ylioinas, and J. Kannala, "Full-frame scene coordinate regression for image-based localization," *Robotics: Science and Systems Conference*, 2018.

[16] E. Brachmann and C. Rother, "Neural-guided ransac: Learning where to sample model hypotheses," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4322–4331.

[17] L. Zhou, Z. Luo, T. Shen, J. Zhang, M. Zhen, Y. Yao, T. Fang, and L. Quan, "Kfnet: Learning temporal camera relocalization using kalman filtering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4919–4928.

[18] X. Li, S. Wang, Y. Zhao, J. Verbeek, and J. Kannala, "Hierarchical scene coordinate classification and regression for visual localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 983–11 992.

[19] P. Guan, Z. Cao, J. Yu, C. Zhou, and M. Tan, "Scene coordinate regression network with global context-guided spatial feature transformation for visual relocalization," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5737–5744, 2021.

[20] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixe, "Understanding the limitations of cnn-based absolute camera pose regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3302–3312.

[21] P.-E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena, "Leveraging deep visual descriptors for hierarchical efficient localization," in *Conference on Robot Learning*, 2018, pp. 456–465.

[22] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.

[23] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4104–4113.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 224–236.

[26] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009.

[27] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[28] Q. Zhou, T. Sattler, M. Pollefeys, and L. Leal-Taixe, "To learn or not to learn: Visual localization from essential matrices," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3319–3326.

[29] M. Ding, Z. Wang, J. Sun, J. Shi, and P. Luo, "Camnet: Coarse-to-fine retrieval for camera re-localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2871–2880.

[30] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7199–7209.

[31] P.-E. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl, *et al.*, "Back to the feature: Learning robust camera localization from pixels to pose," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3247–3257.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[33] H. Chen, Y. Huang, W. Tian, Z. Gao, and L. Xiong, "Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 379–10 388.

[34] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5745–5753.

[35] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "Deepim: Deep iterative matching for 6d pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 683–698.