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

NeRF-Loc: Visual Localization with Conditional Neural Radiance Field

Jianlin Liu[1], Qiang Nie[1], Yong Liu[1] and Chengjie Wang[1]

Abstract— **We propose a novel visual re-localization method based on direct matching between the implicit 3D descriptors and the 2D image with transformer. A conditional neural radiance field(NeRF) is chosen as the 3D scene representation in our pipeline, which supports continuous 3D descriptors generation and neural rendering. By unifying the feature matching and the scene coordinate regression to the same framework, our model learns both generalizable knowledge and scene prior respectively during two training stages. Furthermore, to improve the localization robustness when domain gap exists between training and testing phases, we propose an appearance adaptation layer to explicitly align styles between the 3D model and the query image. Experiments show that our method achieves higher localization accuracy than other learning-based approaches on multiple benchmarks. Code is available at `https://github.com/JenningsL/nerf-loc`.**

I. INTRODUCTION

Visual re-localization is the task of estimating camera's orientation and position in known scenes given a query image. It is an important module in simultaneous localization and mapping (SLAM) and structure from motion (SFM) systems, as well as being the prerequisite of various applications such as autonomous driving and augmented reality(AR).

As the mainstream of solving visual localization problem, structure-based methods consist of two steps, 1) find correspondence between 3D points and 2D pixels 2) compute camera pose by PnP solver with Ransac. 3D-2D point pairs are typically obtained by scene agnostic feature matching. In recent years, scene-specific localization methods attract increasing attention, where the scene is memorized in the neural network weights. Among these methods, direct pose regression[12][26] has fast inference speed but low precision. Scene coordinate regression directly predicts absolute 3D coordinates of image pixels and uses scene structure explicitly to improve accuracy. More recently, many efforts[28][38][22][17] have been made to use implicit neural representation to replace explicit 3D models in localization pipeline. Different from the commonly used discrete 3D models like point cloud and voxel grid, NeRF is one of the implicit 3D representations inferred from a sparse set of posed images, which models geometry and visual information in continuous 3D space. Photorealistic and differentiable rendering of NeRF enables many applications such as direct pose optimization and training set expansion. Nonetheless, existing methods using NeRF for pose estimation are either limited to data augmentation[8][7][20] or analysis by synthesis[37]. In these regards, we aim to directly localize the input image by matching it with a generalizable implicit neural 3D model.

Unlike matching-based methods that generalize well across scenes, scene-specific localization methods are limited to the training scene, although they usually perform better under texture-less indoor conditions utilizing scene prior (per-scene information). Generalizability and scene prior seem to be contradictory in learning, how can we add scene prior to a localization model pretrained across multiple scenes? In this paper, we show that by re-designing the 3D representation in the visual localization pipeline, generalizability and scene prior can be both leveraged to attain better direct 3D-2D matching and localization accuracy.

Inspired by the recent success of generalizable NeRF, our pipeline adopts a conditional NeRF model that can generate 3D features at arbitrary 3D locations. The generated 3D features are shared by both the subsequent matching and rendering. To keep the generalizability across scenes, our neural 3D model is conditioned on a support set that is composed of several posed reference images and depth maps. Based on the support set, continuous 3D descriptors are generated by aggregating multiview and point-wise features in a single forward pass. Similar to the training procedure of generalizable NeRF, our 3D model not only learns general matching during the joint training of multiple scenes as a good start but also memorizes coordinate-based scene prior in a residual way during per-scene optimization to boost the performance. With the learned conditional neural 3D model, efficient visual localization can be done by matching it with the image. To do this, some reference points are randomly sampled from any reconstructed 3D model to query 3D descriptors. Then, a transformer-based matcher is applied to estimate correspondences between the sampled 3D points and dense 2D pixels. When using our method with images in the wild, appearance changes between support images and target images are non-negligible for localization robustness and rendering quality. Therefore, we further propose an appearance adaptation layer to explicitly model the appearance factor in our 3D representation, which improves the localization robustness against domain changes.

Our contributions are three-fold:

- We propose a novel visual localization method based on matching between the conditional NeRF 3D model and 2D image, which formulates scene coordinate regression and feature matching in a unified framework. The proposed pipeline adopts a *multi-scenes pretraining then per-scene finetuning* paradigm to learn both shared knowledge and scene prior for localization task.

[1]Jianlin Liu, Qiang Nie, Yong Liu and Chengjie Wang are with Tencent, Shennan Boulevard, Nanshan District, Shenzhen, China {jenningsliu, stephennie, choasliu, jasoncjwang}@tencent.com

- In order to tackle appearance changes between training support images and the query image, we propose an appearance adaptation layer to align image styles between the query image and the 3D model before matching. Experiments show that the robustness against domain changes is improved.
- Extensive experiments on real-world localization benchmarks are conducted to demonstrate the effectiveness of the proposed method.

II. Related Work

a) Pose Estimation: In the context of visual relocalization, 3D-2D matching is the essential problem. In recent years, deep learning has boosted the performance of 2D feature matching[16][24]. To handle texture-less scene, detector-free 2D matching methods [29][11][6] has been proposed, which shows promising results. Acquiring 2D-3D correspondence from sparse 2D feature matching has its drawbacks, 1)[9] pointed out that 2D keypoints may not reproduce under dramatic appearance changes, leading to failure of matching. 2)exhaustive matching between image pairs from retrieval is less efficient than direct 3D-2D matching [30]. However, to the best of our knowledge, direct 3D-2D matching is seldom studied because cross-modality matching is difficult. Scene coordinate regression(SCR) takes a different path that directly regresses the dense 3D scene coordinates of input image. Dsac[2] propose a differentiable counterpart of Ransac that enable end-to-end training of the scene coordinates based pose estimator. In the follow-up work, Dsac++[3] implements Dsac without learnable parameters which increases generalization capabilities. More recently, [4] extents Dsac++ to support RGB-D input and improve the initialization procedure. However, SCR is scene-specific because of per-scene training. To achieve scene agnostic localization, [32] proposes to construct a correlation tensor between query image and some reference images to regress coordinate map. Different from the existing methods, our method unifies scene coordinate regression to a direct 3D-2D matching framework, 3D-2D correspondences are established by matching 2D image features with 3D features from a conditional NeRF model.

b) NeRF: NeRF is initially proposed in the seminal work [19] as an implicit 3D model for novel view synthesis. NeRF-W[18] extends NeRF to images in the wild with varying appearances and transient objects. Since NeRF memorizes the scene in the network weights, it does not generalize to other scenes. To mitigate the burden of per-scene optimization, researchers have proposed several methods to build generalizable NeRF[5][15][34][35], where a general model is pretrained across different scenes and per-scene finetuning is applied later. In these methods, the NeRF model is conditioned on the local scene structure represented either as posed images or cost volume. We notice that scene-specific localization methods, such as scene coordinate regression, share the same idea with NeRF as they also remember the scene in neural network. In this work, we propose a modified version of conditional NeRF model to better suit visual localization pipeline.

c) NeRF+Pose Estimation.: As a 3D representation and differentiable renderer, NeRF can help visual localization by synthesizing more training images or refining pose with photometric loss. [20] proposes an offline training data generation method based on NeRF to enhance camera pose regression performance. In [37], the authors propose to estimate pose by minimizing the difference between rendered image and query image. However, it is time-consuming and only validated on data without significant illumination changes. In [7], a histogram-assisted NeRF is used to mitigate the domain gap between rendered image and real image, so that a better direct alignment loss can be used to train pose regression. Our method focuses on designing a more practical localization pipeline based on the conditional NeRF, where 3D-2D matching is utilized to avoid slow analysis by synthesis process while keeping all desirable properties brought by NeRF.

III. Method

We choose conditional NeRF as 3D model in our pipeline, the motivation behind is two-fold: 1) to facilitate more extension, such as direct 3D model alignment and expanding train set by synthesizing novel images. 2) to make use of the feature aggregation for better 3D descriptors.

To learn both general knowledge and scene prior, our method consists of two training phases. Firstly, the scene-agnostic pose estimator is trained across multiple scenes. Secondly, per-scene optimization is applied to further improve localization accuracy. Specifically, the pose estimator is realized by predicting the 2D projection of 3D model points. Inspired by LoFTR[29], a coarse-to-fine matching scheme is adopted to directly estimate 3D-2D correspondences. Note that LoFTR performs dense-to-dense matching between image pair, while our method performs sparse-to-dense matching between 3D model and image. Since matching all 3D model points is computationally expensive and even infeasible sometimes, 3D keypoint sampling is applied to keep only a small portion of points before matching. It's worth mentioning that this sparse-to-dense matching avoids the potential poor repeatability of 2D keypoint detector under large viewpoint changes. After the 3D-2D matches are established, a basic PnP solver with Ransac is used to compute camera pose.

A. Conditional Neural 3D Model

In this section, we will introduce the 3D model representation in our method. To perform direct 3D-2D matching, the 3D model predicts the associated 3D descriptor for a given 3D location x as query. To facilitate more downstream tasks, a generalizable NeRF is chosen as the 3D model which is shared by the matching module and neural rendering module. As generalizability is desirable for localization, our method constructs a NeRF model on the fly, conditioned on several support images with known pose. While building a NeRF model in one forward pass is difficult. To make the

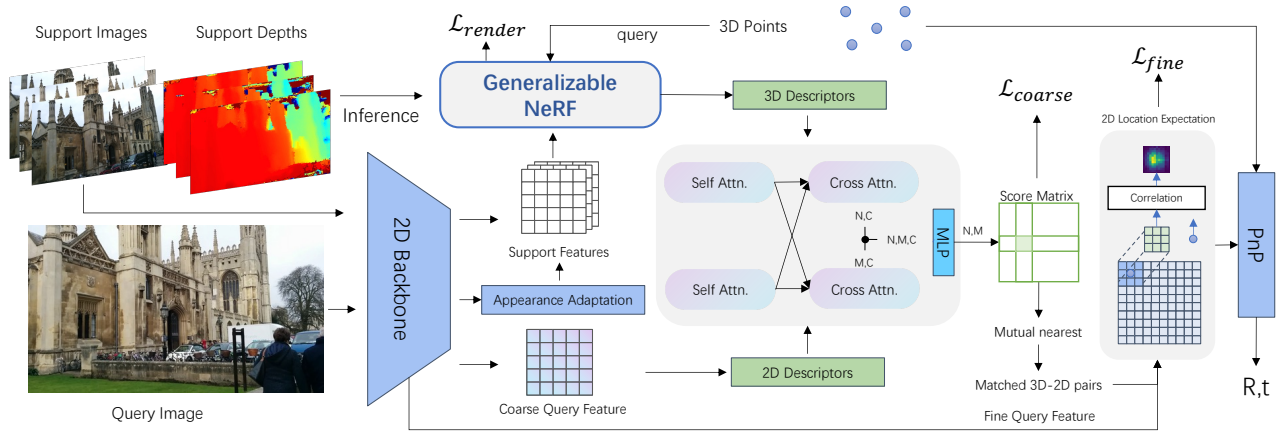

Fig. 1. System overview. In our localization pipeline, the scene is represented as generalizable NeRF which is conditioned on support images. 3D points are fed into the NeRF model to generate 3D descriptors. 3D-2D correspondences are obtained by direct matching between 3D and 2D descriptors, in order to compute camera pose via PnP solver.

problem easier, a noisy and incomplete depth map for each support frame is assumed to be available beforehand. Unlike [35] where source depth maps are generated at inference time, we generate these support depth maps offline with off-the-shelf MVS method. There are two benefits of this: 1) MVS time can be saved during inference, 2) It is compatible with depth maps from range sensors. Before computing 3D descriptors for query points, the support RGBD images need some processing to better represent the local scene. Given the support images and depths, image features are first extracted by a 2D backbone. Then, with known camera parameters, the image feature and raw depth are lifted to 3D neural support points $P = \{P_s, F_s, \Lambda_s, D_s\}$ as in [35], where P_s F_s Λ_s D_s represent location, feature, confidence and viewing direction of support points respectively. Inspired by [15], reference depth maps can also be used for visibility reasoning though they are noisy and incomplete. The raw depths are warped and fused in each support frame to be utilized by a visibility reasoner. The visibility reasoner estimates the likelihood of whether a certain 3D point is visible in a specific reference frame, which requires a complete depth of that reference frame. Hence, this visibility reasoning can be considered as a cross-frame depth validation and completion step. For more details, please refer to [15].

In order to compute the 3D feature of query points $X \in \mathbb{R}^{N\times 3}$, we combine the projected multiview features and the neural support points in a complementary way. Specifically, X are projected to each support frame for fetching multiview features, which are aggregated by computing the visibility weighted mean and variance. The visibility-aware aggregated multiview features $F_m \in \mathbb{R}^{N\times C}$ contain the geometry and visual information at location X, which are equivalent to the elements of MVS cost volume. To augment F_m with local scene structure, we use KNN to search k-nearest neighbors of X in neural support points P and take their point features $F_s \in \mathbb{R}^{N\times K\times C}$. F_m and F_s are passed into a multi-heads attention(MHA) block to get correlated support point feature $F^{'} \in \mathbb{R}^{N\times K\times C}$. In this way, multiview features and support point features interact with each other to get better feature aggregation. The MHA module also outputs the attention weights W_a for each local support point, which are multiplied by the inverse distance weights W_d and confidence Λ_s of local support points to get the final local weights $W = W_a * W_d * \Lambda_s \in \mathbb{R}^{N\times K}$. Finally, the 3D feature for X is calculated as,

$$\mathcal{M}(X) = \{f(x_i) = \sum_{k=0}^{K-1} \frac{w_{ik} * f^{'}_{ik}}{K}; x_i \in X, w_i \in W, f^{'}_i \in F^{'}\} \quad (1)$$

$\mathcal{M}(X)$ is scene agnostic 3D representation since it is computed from the support images, It works well for multi-scenes localization training. However, it does not utilize any scene prior explicitly for finding 3D-2D correspondences, which is valuable when per-scene finetuning is available. To this end, we propose a simple yet effective augmentation for this scene agnostic 3D representation. As indicated in Fig.2, the 3D coordinate of query points are passed to position encoding followed by an MLP to produce coordinate-based features. These coordinate-based features are added to the scene-agnostic 3D descriptors as residuals. Note that this operation is only used for per-scene optimization stage. We empirically found that scene prior features further boost the localization performance significantly.

B. Appearance Adaptation Layer

Our 3D model depends on support images, which may have different lighting conditions or exposures to the query image. The 3D-2D matching will deteriorate when the style of the query image changes. Other methods [25] usually tackle this problem by using more diverse training images to learn invariant features. Though the high-level features are usually quite robust to appearance changes, eliminating domain gaps can further improve the robustness. On the other hand, novel view synthesizer will have difficulty in rendering image that is similar to the target image, which leads to bad performance on direct model alignment. NeRF-W [18] proposes to learn the appearance embedding as

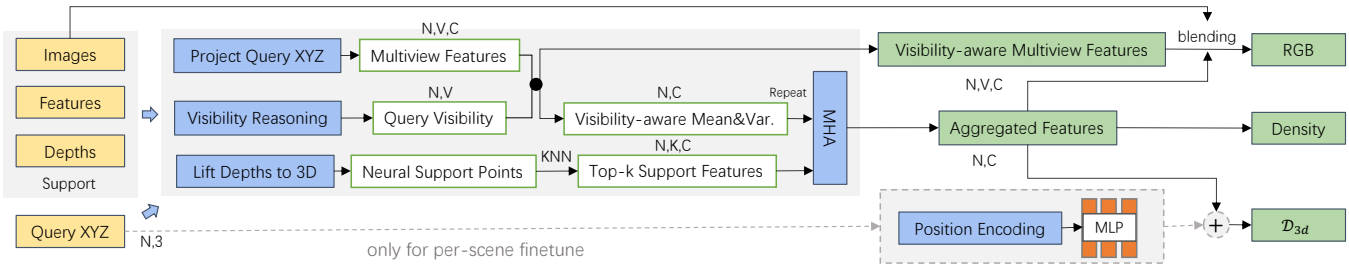

Fig. 2. Architecture of our conditional NeRF model, a feature generator for any 3D location is shared by novel view synthesis and 3D-2D matching. V refers to the number of support images. For more details about visibility reasoning, please refers to [15].

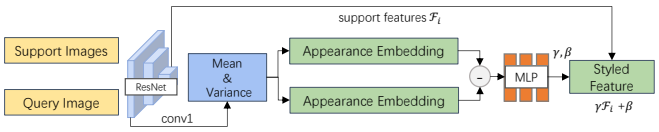

Fig. 3. Appearance adaptation.

parameters for each training image individually. However, for visual localization problem, the target appearance embedding should be computed from the query image. To obtain better robustness against low-level image statistics changes, we propose an appearance adaption layer that explicitly aligns the style of support images and target image. As shown in Fig. 3, we extract appearance embedding from both the target image and support images. Image feature pyramid is defined as $\mathcal{F}_i$, where $\mathcal{F}_0$ refers to the original RGB image. Following [10], channel-wise mean and variance of a fixed low-level feature map $\mathcal{F}_1$ is used as appearance embedding. Appearance embedding of the target image is subtracted from those embeddings of support images and then fed into an MLP to get appearance difference embedding δ. Given a specific feature level $\mathcal{F}_i \in \mathbb{R}^{H\times W\times C}$ of support images to be modulated, δ is decoded into a scaling factor γ and offset β of C channels. Finally, the global affine transformation defined by γ and β is applied to $\mathcal{F}_i$ so that it is aligned with the target style, which is formulated as $AD(\mathcal{F}_i) = \gamma\mathcal{F}_i + \beta$. Appearance adaptation layer is applied to modulate $\mathcal{F}_0$, $\mathcal{F}_2$ and $\mathcal{F}_3$ respectively.

C. Localization Pipeline

In this section, we will introduce our localization pipeline based on the conditional neural 3D model introduced in section III-A. Overall, our localization system consists of three steps: support images selection, 3D-2D matching and pose estimation by solving PnP.

As the first step, how to construct support set depends on the application scenario. We utilize two methods that will cover most of the applications.

1) Image retrieval: A general way is using image retrieval to select support frames. This method works well for the outdoor scene, which makes our method applicable to very large scene. Despite its simplicity and effectiveness, image retrieval performs poorly in indoor scenes with repetitive patterns and object-level scenes, and does not guarantee the spatial uniformity of the selected images.

2) Image coreset: For small scenes such as an object instance, an evenly down-sampled trajectory can be used as support images, which is named image coreset. To attain image coreset, we apply *farthest pose sampling* on all training frames, which repeatedly picks the next frame that has the most different viewing angle from any selected frame, until the max number of frames is reached.

After the support images are selected, a coarse to fine transformer-based matching module is leveraged to find 3D-2D correspondences directly. Specifically, we extract coarse and fine 3D descriptors from the neural 3D model to represent the scene geometry. These two level 3D models are denoted as $\mathcal{M}_c$ and $\mathcal{M}_f$, where $\mathcal{M}_c$ is based on $\mathcal{F}_0$ and $\mathcal{F}_3$, $\mathcal{M}_f$ uses $\mathcal{F}_0$ and $\mathcal{F}_2$.

In coarse level matching, 3D reference points $X \in \mathbb{R}^{N\times 3}$ are randomly sampled and fed into $\mathcal{M}_c$ to get 3D descriptors $\mathcal{D}_{3d} = \mathcal{M}_c(X) \in \mathbb{R}^{N\times C}$. Note that X can be sampled from any 3D point cloud such as the sparse points from SFM, as long as it reveals the correct scene structure. However, we choose to simply sample from neural support points. The coarse level target feature map is reshaped to get 2D descriptors $\mathcal{D}_{2d} \in \mathbb{R}^{HW\times C}$ for coarse matching. After that, $\mathcal{D}_{3d}$ and $\mathcal{D}_{2d}$ are transformed to $\mathcal{D}'_{3d}$ and $\mathcal{D}'_{2d}$ by cross and self-attention layers. The transformed descriptors are used to compute correlation tensor of shape (N, HW, C) which is converted to correlation matrix of shape (N, HW) with MLP later. Therefore, the coarse level matching score $\mathcal{S}$ is formulated as $\mathcal{S} = sigmoid(mlp(\mathcal{D}'_{3d} \odot \mathcal{D}'_{2d})) \in \mathbb{R}^{N\times HW}$, where $\odot$ refers to computing correlation tensor, as shown in Fig.1. Based on $\mathcal{S}$, coarse 3D-2D correspondences are generated by filtering with predefined score threshold τ and mutual nearest checking. To train the coarse matcher, we compute the ground truth score matrix $\mathcal{S}^*$ with the provided depth map. Focal loss[14] is adopted as the coarse matching loss function.

In fine level matching, for each matched 3D-2D pair from coarse matching, we construct the fine level 2D descriptors $\mathcal{D}^{fine}_{2d} \in \mathbb{R}^{M\times 7\times 7\times C}$ by taking a feature patch centered at the coarse 2D location, and query fine level 3D descriptors $\mathcal{D}^{fine}_{3d} = \mathcal{M}_f(X_{matched}) \in \mathbb{R}^{M\times C}$, where M is the number of valid coarse matches. Similar with the coarse matching, $\mathcal{D}^{fine}_{3d}$ and $\mathcal{D}^{fine}_{2d}$ are first transformed by a self/cross-attention layers. Then a correlation matrix is computed by

$\mathcal{S}^{fine} = softmax(mlp(\mathcal{D}_{3d}^{fine'} \odot \mathcal{D}_{2d}^{fine'}))$. Finally, 2D coordinates are refined to sub-pixel level by adding spatial expectation $\hat{\mathcal{E}}$ based on $\mathcal{S}^{fine}$. Following LoFTR, the fine level matching is supervised by L2 loss between predicted and ground truth location $\mathcal{L}_{fine} = ||\hat{\mathcal{E}} - \mathcal{E}||^2$. For more detail, please refer to [29].

To keep the benefits of NeRF as 3D model, an auxiliary rendering head is added upon the shared 3D feature from $\mathcal{M}_f$, which is supervised by L2 loss $\mathcal{L}_{render}$. Following [15], depth recovering loss $\mathcal{L}_{depth}$ is added to provide direct supervision on visibility reasoning, where the network is required to remove noises from support depth maps during training. In summary, the final training loss is defined as followed,

$$\mathcal{L} = \mathcal{L}_{coarse} + \mathcal{L}_{fine} + \mathcal{L}_{render} + \mathcal{L}_{depth} \tag{2}$$

IV. Experiments

A. Datasets

We evaluate our pipeline on four localization benchmarks, ranging from object level to outdoor scene.

a) Indoor: 12Scenes[33] and 7Scenes[27] are two room-size indoor localization datasets, where depth images are scanned by the structure light sensor and the ground truth poses are provided. There are no significant illumination changes or dynamic objects. 12Scenes contains 12 scenes and each contains several hundreds of frames. Each scene in 7Scenes has several thousands of frames. RGB images are registered to depth maps in 12Scenes but not in 7Scenes. Following [4], we register the images in 7Scenes first. Camera parameters of 7Scenes are less accurate, which brings challenges to fitting a global consistent 3D model.

b) Outdoor: Cambridge Landmarks[12] dataset scaling from $875m^2$ to $5600m^2$ is for evaluating localization algorithms in large-scale outdoor scenes. Illumination and exposure of camera are different between sequences. Transient objects commonly exist in this dataset.

c) Object: Onepose[30] is an object-level pose estimation dataset collected by hand-held smartphones. Several videos are captured around object instances with different backgrounds. Trajectories of different videos are aligned to a unified coordinate system.

B. Evaluation Protocol

Localization accuracy is defined as the ratio of correctly localized frames, given a criterion of success. For all datasets except Cambridge, a localized frame with rotation error below 5° and translation error below 5cm is considered as correct. As the scene scale of Cambridge dataset varies a lot, the translation threshold should be adaptively set for each scene. Here, the same translation thresholds with [4] are used (35cm for St. Mary's Church, 45cm for Great Court, 22cm for Old Hospital, 38cm for King's College and 15cm for Shop Facade). Besides localization accuracy, the commonly used median translation(cm) and rotation(°) error are also reported on Cambridge dataset. The same train/test split with [4] is used.

C. Implementation Details

For each dataset, the network is pretrained across all scenes for 30 epochs, and then optimized per scene for another 30 epochs, using Adam optimizer with learning rate of $5 * 10^{-4}$. Geometry augmentations (*random zooming* and *random rotation*) and Color jitter are applied to input images. Appearance adaptation is not used for Onepose, since the images are cropped by 2D detection box and padded with black background. To select support images, Image retrieval[1] is used for all datasets except Onepose, which uses image coreset. For image retrieval, 5 images are randomly selected from the top 20 retrieval images during training, while top-10 images are used during test time. For image coreset, 16 support images are kept for both training and testing. We use the MVS algorithm in Colmap to prepare depth maps for training images except for the indoor datasets. For matching in all experiments, 1024 3D points are randomly sampled. Localizing one frame takes 250ms on Nvidia V100 GPU, using 10 support images.

D. Comparison with Other Methods

As indicated in Table I, II and III, our method consistently achieves the best overall performance among learning-based visual localization methods on all four datasets, which shows the practical value of our system. Qualitative results can be found in Figure 4 and the supplementary video.

TABLE II

Localization accuracy on two indoor datasets. Results of other methods are from [4].

Method	ORB+PnP	DSAC	DSAC++	DSAC*	SCoCR	**Ours**
12Scenes	53.7	83.5	96.8	99.1	99.3	**99.8**
7scenes	40.7	60.2	74.4	85.2	84.8	**89.5**

TABLE III

Localization accuracy on object-level dataset Onepose.

Method	0447	450	0488	0493	0494	0594	Avg.
PVNet[21]	25.3	12.7	4.2	9.4	19.2	7.7	13.1
Onepose[30]	90.0	98.1	74.0	87.3	81.9	78.9	85.0
Ours	100	100	99.5	99.7	71.2	60.4	**88.5**

E. Ablation Study

1) Effectiveness of Scene Prior: To validate design choices of the proposed training scheme, we decouple it into three components: multi-scenes pre-training(*pre-train*), scene coordinate-based feature(*coord*) and per-scene optimization(*per-scene*). As shown in Table IV, for 7Scenes our method trained on multiple scenes jointly already outperforms the strong scene coordinate regression-based baseline method DSAC*[4]. Per-scene optimization injects valuable scene prior to aid localization, which further boosts the accuracy significantly(last two rows). Specifically, if the scene coordinate-based feature is explicitly used as we proposed, the network learns better. From the third row, we can see that

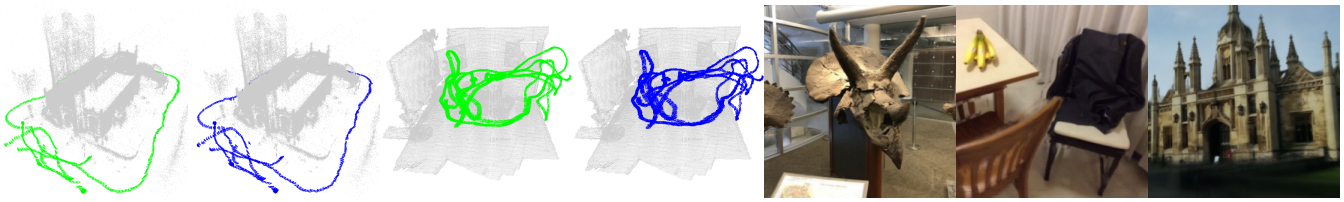

Fig. 4. Qualitative results of pose estimation(first four, Green: groundtruth, Blue: prediction) and rendering(last three).

TABLE I

EVALUATION ON OUTDOOR AND INDOOR LOCALIZATION BENCHMARKS. WE REPORT THE MEDIAN TRANSLATION(CM) AND ROTATION(°) ERROR FOR EACH SCENE. AVG. STANDS FOR THE AVERAGE MEDIAN ERROR, AND ACC. IS THE ABBREVIATION FOR AVERAGE ACCURACY.

	Cambridge Landmarks - outdoor						7scenes - indoor							
Method	Church	Court	Hospital	College	Shop	Avg.↓	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Acc.
SANet[36]	16/0.57	328/1.95	32/0.53	32/0.54	10/0.47	83.6/0.8	3/0.9	3/1.1	2/1.5	3/1.0	5/1.3	4/1.4	16/4.6	68.2
DSAC[2]	55/1.6	280/1.5	33/0.6	30/0.5	9/0.4	81.4/0.9	2/0.7	3/1.0	2/1.3	3/1.0	5/1.3	5/1.5	190/49.4	60.2
InLoc[31]	18/0.6	120/0.6	48/1.0	46/0.8	11/0.5	48.6/0.7	3/1.1	3/1.1	2/1.2	3/1.1	5/1.6	4/1.3	9/2.5	66.3
DSM[32]	12/0.4	44/0.2	24/0.4	19/0.4	7/0.4	21.2/0.4	2/0.7	2/0.9	1/0.8	3/0.8	4/1.2	4/1.2	5/1.4	78.1
DSAC++[3]	13/0.4	40/0.2	20/0.3	18/0.3	6/0.3	19.4/0.3	**2/0.5**	2/0.9	1/0.8	3/0.7	4/1.1	4/1.1	9/2.6	74.4
DSAC*[4]	13/0.4	49/0.3	21/0.4	15/0.3	5/0.3	20.6/0.3	2/1.1	2/1.2	1/1.8	3/1.2	4/1.3	4/1.7	3/1.2	85.2
HACNet[13]	9/0.3	28/0.2	19/0.3	18/0.3	6/0.3	16.0/0.3	2/0.7	2/0.9	1/0.9	3/0.8	4/1.0	4/1.2	3/0.8	84.8
PixLoc[25]	10/0.3	30/0.1	**16/0.3**	14/0.2	5/0.2	15/0.2	2/0.8	**2/0.7**	**1/0.8**	3/0.8	4/1.2	**3/1.2**	5/1.3	75.7
Ours	**7/0.2**	**25/0.1**	18/0.4	**11/0.2**	**4/0.2**	**13/0.2**	2/1.1	2/1.1	1/1.9	**2/1.1**	**3/1.3**	3/1.5	**3/1.3**	**89.5**

per-scene training from scratch does not achieve comparable performance as other settings. In summary, (1) multi-scenes training is important for learning general knowledge to avoid overfitting, (2) scene prior can be used to improve scene-agnostic localization, especially by adding learned scene coordinate features to the original 3D features.

TABLE IV

ABLATION STUDY ABOUT TRAINING SCHEME AND AGGREGATION.

pre-train	coord	per-scene	nerf-agg	7Scenes	Cambridge
✓	✗	✗	✗	84.8	79.0
✓	✗	✗	✓	85.8	79.5
✗	✓	✓	✓	81.2	78.3
✓	✗	✓	✓	88.2	80.3
✓	✓	✓	✓	**89.5**	**81.6**

2) Effectiveness of Conditional NeRF: By design, the rendering head in our pipeline is optional. We found that rendering loss does not contribute to the matching accuracy in our experiments. However, according to Table IV, the feature aggregation of conditional NeRF improves 3D-2D matching, especially on indoor data with more occlusions. When *nerf-agg* is removed, multiview feature aggregation is by simply averaging.

3) Effectiveness of Appearance Adaptation: To validate the effectiveness of the appearance adaptation layer, we change the style of test images for evaluation. Specifically, experiments are conducted on Cambridge dataset, where daytime images are transferred to nighttime with the method in [23]. Besides the median translation error and accuracy, matching IoU(intersection-over-union between predicted and ground truth 3D-2D pairs) is also reported as the direct measurement of matching quality, considering the performance gap may be reduced by the robust estimator in PnP solver. Peak Signal-to-Noise Ratio(PSNR) is the metric to evaluate novel view synthesis. From Table V, three conclusions can be drawn, 1) appearance changes of test images lead to a significant drop of localization accuracy(79.5 to 73.6) as well as the view synthesis quality, 2) adding stronger color jitter augmentation during training only helps improving localization in the day-to-night transfer setting but not novel view synthesis 3) localization and view synthesis performance are both better by adding the proposed appearance adaptation layer as shown in the third row of Table V.

TABLE V

ABLATION STUDY ABOUT APPEARANCE ADAPTATION.

Color Jitter	Appearance	Error↓	Accuracy	IoU	PSNR
✗	✗	20.0	73.6	0.254	14.4
✓	✗	19.1	75.1	0.286	15.5
✓	✓	**17.5**	**75.8**	**0.299**	**19.7**

V. CONCLUSION

In this paper, we present a novel visual localization pipeline based on direct 3D-2D matching between the NeRF model and images. By conditioning the NeRF model on support images, the proposed method is scene agnostic. To inject scene prior, a scene coordinate-based 3D feature module for per-scene optimization is proposed, which greatly improves the per-scene performance. Moreover, we propose an appearance adaptation layer to better handle the domain gap between the query image and the conditional 3D model, which enhances the system's robustness in terms of novel view synthesis and localization.

REFERENCES

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[2] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6684–6692, 2017.

[3] Eric Brachmann and Carsten Rother. Learning less is more-6d camera localization via 3d surface regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4654–4662, 2018.

[4] Eric Brachmann and Carsten Rother. Visual camera re-localization from rgb and rgb-d images using dsac. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021.

[6] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 20–36. Springer, 2022.

[7] Shuai Chen, Xinghui Li, Zirui Wang, and Victor Adrian Prisacariu. Dfnet: Enhance absolute pose regression with direct feature matching. *arXiv preprint arXiv:2204.00559*, 2022.

[8] Shuai Chen, Zirui Wang, and Victor Prisacariu. Direct-posenet: Absolute pose regression with photometric consistency. In *2021 International Conference on 3D Vision (3DV)*, pages 1175–1185. IEEE, 2021.

[9] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. S2dnet: learning image features for accurate sparse-to-dense matching. In *European Conference on Computer Vision*, pages 626–643. Springer, 2020.

[10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[11] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021.

[12] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015.

[13] Xiaotian Li, Shuzhe Wang, Yi Zhao, Jakob Verbeek, and Juho Kannala. Hierarchical scene coordinate classification and regression for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11983–11992, 2020.

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[15] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022.

[16] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Contextdesc: Local descriptor augmentation with cross-modality context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2527–2536, 2019.

[17] Dominic Maggio, Marcus Abate, Jingnan Shi, Courtney Mario, and Luca Carlone. Loc-nerf: Monte carlo localization using neural radiance fields. *arXiv preprint arXiv:2209.09050*, 2022.

[18] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021.

[19] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[20] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*, pages 1347–1356. PMLR, 2022.

[21] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019.

[22] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. *arXiv preprint arXiv:2210.13641*, 2022.

[23] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[24] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[25] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik 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*, pages 3247–3257, 2021.

[26] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura 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 (CVPR)*, June 2019.

[27] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013.

[28] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021.

[29] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[30] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022.

[31] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018.

[32] Shitao Tang, Chengzhou Tang, Rui Huang, Siyu Zhu, and Ping Tan. Learning camera localization via dense scene matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1831–1841, 2021.

[33] Julien Valentin, Angela Dai, Matthias Nießner, Pushmeet Kohli, Philip Torr, Shahram Izadi, and Cem Keskin. Learning to navigate the energy landscape. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 323–332. IEEE, 2016.

[34] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021.

[35] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022.

[36] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. Sanet: Scene agnostic network for camera localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 42–51, 2019.

[37] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021.

[38] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022.