I-DACS: Always Maintaining Consistency between Poses and the Field for Radiance Field Construction without Pose Prior

Tianjun Zhang, Lin Zhang, *Senior Member, IEEE,* Fengyi Zhang, Shengjie Zhao, *Senior Member, IEEE,* and Yicong Zhou, *Senior Member, IEEE*

***Abstract*—The radiance field, emerging as a novel 3D scene representation, has found widespread application across diverse fields. Standard radiance field construction approaches rely on the ground-truth poses of key-frames, while building the field without pose prior remains a formidable challenge. Recent advancements have made strides in mitigating this challenge, albeit to a limited extent, by jointly optimizing poses and the radiance field. However, in these schemes, the consistency between the radiance field and poses is achieved completely by training. Once the poses of key-frames undergo changes, long-term training is required to readjust the field to fit them. To address such a limitation, we propose a new solution for radiance field construction without pose prior, namely I-DACS (Incremental radiance field construction with Direction-Aware Color Sampling). Diverging from most of the existing global optimization solutions, we choose to incrementally solve the poses and construct a radiance field within a sliding-window framework. The poses are unequivocally retrieved from the radiance field, devoid of any constraints and accompanying noise from other observation models, so as to achieve the consistency of poses to the field. Besides, in the radiance field, the color information is much higher-frequency and more time-consuming to learn compared with the density. To accelerate training, we isolate the color information to a distinct color field, and construct the color field based on an innovative direction-aware color sampling strategy, by which the color field can be derived directly from images without training. The color field obtained in this way is always consistent with the poses, and intricate details of training images can be retained to the utmost extent. Extensive experimental results evidently showcase both the remarkable training speed and the outstanding performance in rendering quality and localization accuracy achieved by I-DACS. To make our results reproducible, the source code has been released at https://cslinzhang.github.io/I-DACS-MainPage/.**

***Index Terms*—Radiance field, pose prior, incrementally, consistency, direction-aware sampling.**

I. Introduction

THE synthesis of photo-realistic novel views from a sequence of RGB frames is a pivotal challenge with broad applications in various domains [1]–[3], such as virtual/augmented reality [4]–[6], video editing [7], [8] and 3D reconstruction [9]–[11]. Presently, one mainstream solution to the challenge involves representing the scene as a 3D model [12], [13] and then synthesizing novel views based on it. In recent years, the radiance field representation [14] has garnered considerable attention from researchers, owing to its exceptional performance and great potential for freely rendering novel views with great fidelity. Furthermore, it has gradually found application within industrial sectors.

During training of the radiance field, accurate camera poses of the training video sequence are imperative. One common approach to obtain them involves the use of traditional SFM schemes [15], such as COLMAP [16], [17]. While this constitutes a well-established solution, the associated preprocessing often proves laborious and may struggle to yield reliable estimations in scenes lacking textures. By contrast, a more streamlined and elegant way is to parameterize the poses and estimate them jointly with the radiance field. Such an idea does not require the preprocessing step and holds the potential to recover more consistent poses to the radiance field. Most of the recent studies [18]–[22] follow this idea. However, the speed and the accuracy of them are still unsatisfactory due to inherent design limitations, which are mainly manifested in twofolds:

1) **Most of these methods primarily adhere to the global optimization framework, with limited incorporation of sequential trajectory information.** Within such a global mode, a substantial number of parameters undergo simultaneous adjustments. Consequently, the system's convergence often displays oscillatory patterns and tends to get trapped in local optima.
2) **The consistency between the poses and the radiance field in existing methods is upheld completely by training.** Specifically, when jointly optimizing the poses and the field, once the poses change, the radiance field needs to be refitted accordingly to be consistent with the current pose state, which inevitably results in significant time consumption during training.

As an attempt to overcome the limitations of existing schemes, we propose a novel framework to efficiently construct the radiance field without pose prior, namely I-DACS

This work was supported in part by the National Natural Science Foundation of China under Grant 62272343 and Grant 61936014; in part by the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission under Grant 21SG23; and in part by the Fundamental Research Funds for the Central Universities. (Corresponding author: Lin Zhang.)

Tianjun Zhang, Lin Zhang, Fengyi Zhang and Shengjie Zhao are with the School of Software Engineering, Tongji University, Shanghai 201804, China, and also with the Engineering Research Center of Key Software Technologies for Smart City Perception and Planning, Ministry of Education, China (email: {1911036, cslinzhang, zzfff, shengjiezhao}@tongji.edu.cn).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yicongzhou@um.edu.mo).

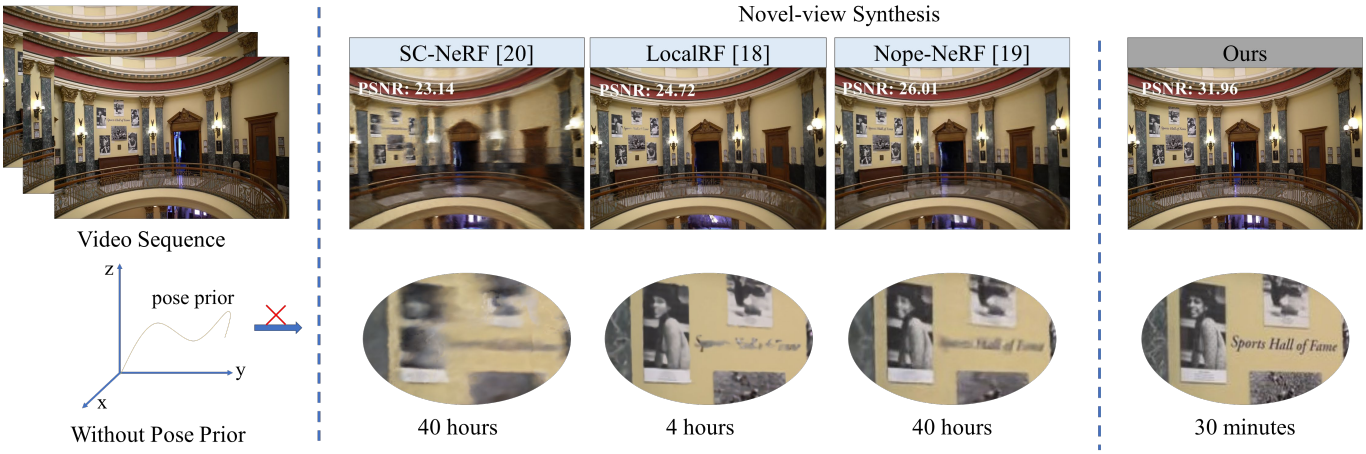


Fig. 1. **High-quality novel-view synthesis results from our method.** Focusing on the radiance field construction without pose prior, our scheme achieves a superior speed with high-quality novel-view synthesis performance, obviously outperforming other competitors.

(Incremental radiance field construction with Direction-Aware Color Sampling), which leverages a sliding-window-based incremental optimization framework seamlessly integrated with our novel direction-aware color sampling strategy. The incremental optimization framework can effectively recover poses consistent with the field, and in turn, the color sampling strategy ensures that the field is highly consistent with the poses. The superior accuracy and speed performance of I-DACS are shown in Fig. 1, and our contributions can be summarized as follows:

1) We propose an incremental sliding-window based optimization strategy to jointly estimate the radiance field and camera poses, and offer solid derivation to prove the theoretical equivalence of our strategy to the global one. Such an incremental strategy can effectively harness the local sequential information of frames without further introducing constraints from observation models inconsistent to the radiance field, thereby gradually and steadily constructing the field and recovering poses.
2) We present a direction-aware color sampling scheme to represent the color information of the radiance field as the color field separately. Applying our sampling scheme, the color field can be built directly from training images without training and naturally maintains consistency with the poses of key-frames, thereby ensuring fast convergence in pose-field optimization.
3) We design a new framework for radiance field construction without pose prior, namely I-DACS. I-DACS is implemented under our incremental joint optimization framework and models the density field and the color field individually. The density field is modeled in the hash-table form, whereas the higher-frequency color field is modeled using our direction-aware color sampling scheme, achieving excellent convergence speed and outstanding novel-view synthesis accuracy. The architecture of I-DACS is illustrated in Fig. 2.
4) Extensive experimental results corroborate the superiority of I-DACS in both the speed and the accuracy. Compared with existing SOTA schemes, I-DACS exhibits significant accuracy advantages in both localization and novel-view synthesis with much faster speed.

II. Related Work

A. 3D scene representations

The realm of 3D scene representation, a classic and pivotal domain, has witnessed the emergence of various solutions over recent decades. Among these solutions, three primary representations stand out: point cloud, voxel-grid, and neural field. Next, related work about these three kinds of representations will be introduced one by one.

Point cloud representations depict scenes as a collection of 3D points, often acquired directly through the scanning of LiDAR or RGBD cameras. Renowned for their simplicity and ease of acquisition, point clouds impose no specific constraints on scene topology and find extensive utility across diverse tasks such as scene understanding [23]–[25] and reconstruction [26]–[28]. However, as a discrete representation, point clouds lack explicit connectivity information among points, resulting in a relatively coarse depiction of both geometry and textures.

The voxel-grid representation, an extension of 2D pixel grids, simplifies the portrayal of 3D scenes into discrete volumetric grids. Occupancy grid [29]–[31], TSDF [32]–[34], and ESDF [35]–[37] are all prominent examples of voxel-grid representations widely utilized in various applications. Despite its regular geometric structure, voxel-grids necessitate the storage of the entire voxelized 3D scenes, often resulting in considerable storage overhead, especially for high-resolution grids. Consequently, researchers have explored integrating structures like octrees [38], [39] and hash tables [40], [41] with voxel-grid representations to alleviate the storage load problem.

In recent years, the neural field representation [42]–[44] has garnered significant attention for its capability to generate objects or scenes with arbitrary topologies and infinite resolution. This representation views the 3D scene as a function that takes the 3D position as input and can output any necessary values, such as occupancy probabilities, TSDF values, or

radiance values. As the neural field representation describes the mapping function rather than the attributes of the scene, it is also referred to as the "implicit representation". The neural representation has a descriptive advantage that traditional point cloud or voxel-grid representations cannot match, but it usually relies on training to be obtained, which leads to a relatively slow construction speed.

B. Radiance field representations

The radiance field has received widespread attention in recent years due to its powerful ability and potential to render photo-realistic novel views. As a milestone work, Mildenhall *et al.* proposed NeRF (Neural Radiance Field) [14] in 2020. In [14], the static scene is modeled as a radiance field, which is a continuous 5D function that outputs the radiance emitted in each direction at each point in space. Representing the radiance field as an MLP (Multi-Layer Perceptron), novel views can be synthesized based on classical volume rendering techniques, and such a differentiable rendering process can be implemented within the modern deep learning frameworks. A lot of subsequent researches are based on NeRF, and significant improvements were made in different aspects like parameterization [45]–[48], regularization [49]–[51], supervision [52]–[55] and dynamicity [56]–[58]. In the standard NeRF representation, the properties of the radiance field are not explicitly stored, but implicitly predicted by the neural network. Thus, the representations in these work are considered as implicit representations.

Different from the aforementioned implicit representations, there are also many methods choosing to represent the radiance field in an explicit or hybrid way. Among them, one effective explicit solution is to store the properties of the radiance field in voxel-grids. Plenoxel [59] uses the sparse 3D grid with spherical harmonic to model the radiance field explicitly. Such a representation can be optimized from calibrated images via gradient methods and regularization without any neural components, achieving two orders of magnitude faster training speed with no loss in visual quality compared with the standard NeRF representation. In [60], Sun *et al.* proposed DVGO that chooses a hybrid grid-neural representation. DVGO stores radiance field features with a voxel-grid and then decodes the features using neural networks. Compared with implicit representations, the training speeds of radiance fields in voxel-grid representations are much faster, but such representations also bring additional storage burdens.

To balance among rendering quality, training speed and space complexity, many researchers tried to compress the voxel-grids to more complete representations. In [61], Chen *et al.* proposed TensoRF and modeled the radiance field of a scene as a 4D tensor, which represents a 3D voxel-grid with per-voxel multi-channel features. After that, the 4D scene tensor can be decomposed into multiple compact low-rank tensor components via traditional CP decomposition, which leads to a significantly lower memory footprint in comparison to previous grid-based explicit representations of radiance field. Instant-NGP [62] is another work that effectively solves the storage problem of voxel-grid representations. In [62], representing the radiance field, features in the multi-resolution voxel-grids are maintained and stored as a feature vector, which can be accessed through the hash encoding on the space. Besides, an MLP is also used to decode the features and compensate for the hash conflict problem. To generalize across scenes, pixelNeRF [63] directly samples deep features from images and then sends them to a neural decoder to predict radiance values. PixelNeRF [63] supports the feature sampling on multiple images, yet it overlooks the observation directions of various sampling views, leading to synthesized novel views that are relatively blurry and exhibit noticeable artifacts.

C. Radiance field without accurate poses

For the construction of the radiance field, accurate camera poses are usually indispensable. Current mainstream solutions to obtain them involve employing SFM [16], [64], [65] or monocular SLAM methods [66]–[69]. While these methods have reached a level of maturity, they introduce additional preprocessing steps and increase the overall time consumption. Moreover, poses are solved from the sparse point cloud instead of the radiance field, and such inconsistency may encumber the radiance field from converging to the global optima.

A more elegant approach to construct the radiance field is to abandon the pre-processing SFM step, and directly optimize both poses and the radiance field in a joint manner. As a pioneering work, BARF [21] formulates poses in Lie algebraic form [70] and jointly optimizes the field and inaccurate poses in a coarse-to-fine manner. Apart from poses, NeRFmm [22] can also optimize the intrinsics of the camera along with the training evolvement. SC-NeRF [20] further models the distortion coefficients of the camera as optimizable parameters, and proposes a geometric consistency loss to force sampled rays from corresponding pixels on images to be close to each other. These three joint optimization methods can recover the radiance field without accurate poses, but they still highly rely on pose priors and can only adjust camera poses in limited ranges.

To achieve better universality, some researchers have further explored the problem of radiance field construction without pose prior. LocalRF [18] chooses to add frames progressively to the global optimization pool, and uses multiple local radiance fields to represent the global one. Thanks to its multi-field mechanism, LocalRF performs relatively satisfactorily in large-scale outdoor scenes under stable motion. However, LocalRF is highly dependent and sensitive to the optical flow estimation [71], which may introduce constraints that are inconsistent with the radiance field. Nope-NeRF [19] introduces the point cloud loss and surface-based photometric loss to constrain the relative poses of adjacent frames, which to some extent alleviates the convergence oscillation problem of global-optimization schemes. GNeRF [72] employs the generative adversarial network to recover the true distribution of key-frame poses. CF-NeRF [84] is a camera-parameter-free framework for the problem of neural radiance construction. CF-NeRF [84] can recover camera parameters without any prior during the incremental training, but the frequently conducted global optimization severely limits the efficiency and the robustness of it.

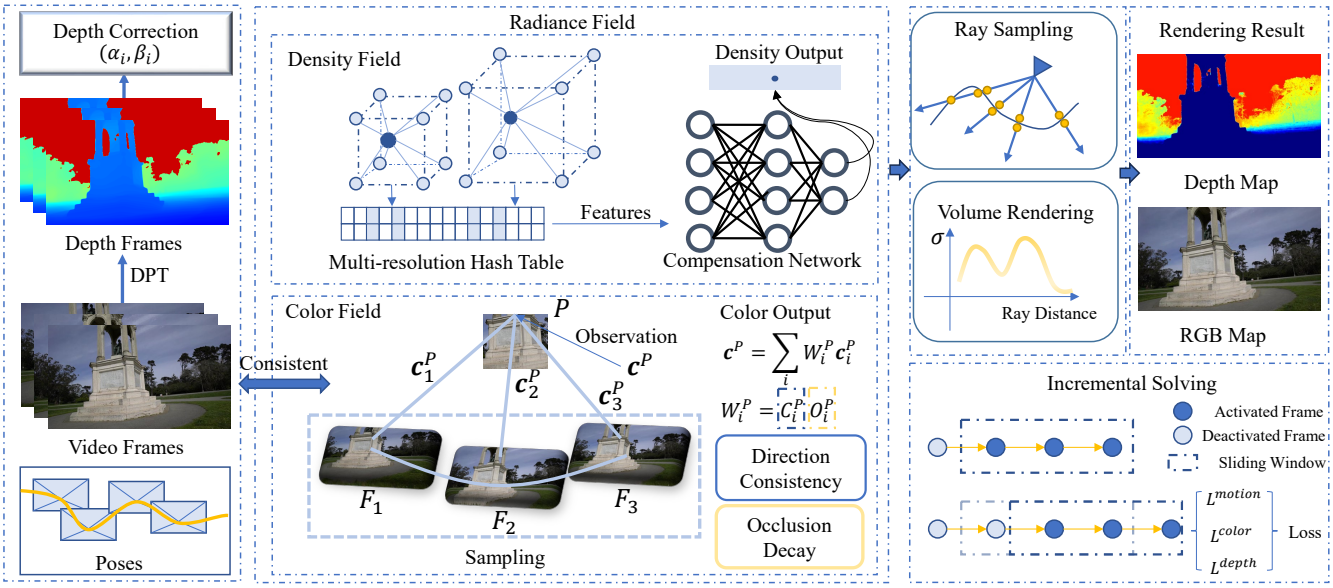


Fig. 2. **The architecture sketch of I-DACS.** In I-DACS, the depth maps of an input video sequence are extracted via monocular depth estimation, and further corrected by trainable correction coefficients. Then all trainable parameters are optimized jointly in an incremental manner. Besides, the radiance field is divided into two components: the density field and the color field. The density field is represented as a multi-resolution hash-table, while the color field is derived based on our direction-aware color sampling strategy.

Except for the aforementioned methods, some NeRF-based SLAM methods can also build the model of the environment without pose prior, such as NICE-SLAM [73], NICER-SLAM [74], Point-SLAM [75] and DIM-SLAM [76]. These NeRF-based SLAM methods mainly solve the problem of simultaneous localization and mapping, focusing more on the correctness of the mapping geometry rather than rendering quality. In these methods, the environment is usually modeled as TSDF maps in which the color information is direction-independent. Thus, such NeRF-based SLAM methods cannot achieve satisfactory novel-view synthesis quality. Besides, some studies [77]–[79] are dedicated to the localization of key-frames from known radiance fields, which motivate us and lay the theoretical foundation for our incremental optimization pipeline.

III. Incremental Joint Optimization Framework

A. Radiance field preliminaries

When the scene is represented as a radiance field, each point in the scene has two attributes: volume density and view-dependent color. Thus, the radiance field can be modeled as a mapping function,

$$F_{\boldsymbol{\Theta}}(\boldsymbol{x}, \boldsymbol{d}) \rightarrow (\boldsymbol{c}, \sigma), \tag{1}$$

where $\boldsymbol{\Theta}$ is the parameter of the field, $\boldsymbol{x}$ and $\boldsymbol{d}$ are the position and observing direction, respectively, and $\boldsymbol{c}$ and σ are the yielded color and density, respectively. The radiance field can be represented as an MLP, a voxel-grid or in a hybrid form, etc, and the representation utilized in I-DACS will be introduced in detail in Sec. IV.

To render a novel-view $\hat{\boldsymbol{I}}$, for each pixel $\boldsymbol{p}$ on the image, a ray $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$ can be cast from the camera origin $\boldsymbol{o}$ to the space, and the color $\boldsymbol{c_p}$ of $\boldsymbol{p}$ can be determined by accumulating the color of the field weighted by the corresponding density along the ray as,

$$\boldsymbol{c_p} = \int_{t_n}^{t_f} T(t)\sigma(\boldsymbol{r}(t))\boldsymbol{c}(\boldsymbol{r}(t), \boldsymbol{d})dt, \tag{2}$$

where t_n and t_f are the nearest and the farthest distances of the observation, respectively, and $T(t)$ is the accumulated opacity, which can be given as,

$$T(t) = exp(-\int_{t_n}^{t} \sigma(\boldsymbol{r}(s))ds). \tag{3}$$

It's worth mentioning that, since the integral function in continuous form is difficult to solve, $\boldsymbol{c_p}$ is usually computed as a discrete approximation.

Currently, the radiance field is mostly constructed via training. Given a sequence of images $\boldsymbol{\mathcal{I}} = \{\boldsymbol{I}_1, ..., \boldsymbol{I}_N\}$ and corresponding camera poses $\boldsymbol{\mathcal{T}} = \{\boldsymbol{T}_1, ..., \boldsymbol{T}_N\}$, a commonly utilized straightforward way to estimate the parameters of the radiance field $\boldsymbol{\Theta}$ is optimizing $\boldsymbol{\Theta}$ to minimize the rendering loss, which is usually the norm-loss between the rendered images of the radiance field and the corresponding ground-truth images. Besides, to achieve satisfactory accuracy and robustness, except for the RGB rendering loss, the rendering loss usually also contains other loss terms, such as the depth loss and the regularization loss.

B. Probabilistic theoretical foundation

In the optimization framework of I-DACS, depth supervision is incorporated to ensure the geometric correctness of the trained radiance field. Given the original image sequence $\boldsymbol{\mathcal{I}} = \{\boldsymbol{I}_1, ..., \boldsymbol{I}_N\}$, we use the powerful monocular depth estimation framework, namely DPT [80], to extract corresponding depth maps $\boldsymbol{\mathcal{D}} = \{\boldsymbol{D}_1, ..., \boldsymbol{D}_N\}$. Besides,

since the scale and the shift of depth maps are unobservable for monocular depth estimation schemes, following the mainstream solution, we built an optimizable scale-and-shift sequence $\boldsymbol{\mathcal{C}} = \{\alpha_1, \beta_1, ..., \alpha_N, \beta_N\}$ and utilize it to undistort depth maps as,

$$\hat{\boldsymbol{D}}_i = \alpha_i \boldsymbol{D}_i + \beta_i, \tag{4}$$

where $\hat{\boldsymbol{D}}_i$ is the undistorted depth map corresponding to $\boldsymbol{D}_i$. Finally, the goal of our optimization framework can be represented as,

$$\boldsymbol{\Theta}^*, \boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{C}}^* = \underset{\boldsymbol{\Theta}, \boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{C}}}{\arg\max}\, P(\boldsymbol{\Theta}, \boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{C}} | \boldsymbol{\mathcal{I}}, \boldsymbol{\mathcal{D}}). \tag{5}$$

Existing pose-field joint optimization frameworks are usually designed under the global mode, wherein parameters including poses of all frames and the global radiance field are optimized simultaneously. However, in such a straightforward way, the optimization often exhibits oscillatory patterns and tends to get trapped in local optima due to the considerable amount of optimizable parameters. To harness the full potential of the sequential information inherent in the camera trajectory, we opt to depart from the global mode and, instead, establish an incremental framework.

The process of the incremental optimization can be roughly divided into N stages. In the K_{th} stage of the incremental optimization, we aim to estimate the radiance field, camera poses and depth maps corresponding to the first K frames in the sequence, while the latter $N-K$ frames won't be considered. We use $\hat{P}_{1:K}$ to represent the corresponding posterior distribution in stage K given in Eq. 5 as,

$$\hat{P}_{1:K} = P(\boldsymbol{\Theta}_{1:K}, \boldsymbol{\mathcal{T}}_{1:K}, \boldsymbol{\mathcal{C}}_{1:K} | \boldsymbol{\mathcal{I}}_{1:K}, \boldsymbol{\mathcal{D}}_{1:K}), \tag{6}$$

where the subscript $1:K$ represents the first K elements in the set. It's worth mentioning that, $\boldsymbol{\Theta}_{1:K}$ is a set of K radiance fields consisting of $\boldsymbol{\Theta}_1$ to $\boldsymbol{\Theta}_K$, and $\boldsymbol{\Theta}_K$ is the radiance field parameters trained with the first K frames in the sequence. Though we only require the final radiance field, history fields are also incorporated here to ensure the rigor of theoretical derivation. Further, we have,

$$\begin{aligned}
\hat{P}_{1:K} &\propto \hat{P}_{1:K-1} \cdot \tilde{P}^p_K \cdot \tilde{P}^l_K \\
\tilde{P}^p_K &= P\left(\boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K | \boldsymbol{\Theta}_{1:K-1}, \boldsymbol{\mathcal{T}}_{1:K-1}, \boldsymbol{\mathcal{C}}_{1:K-1}\right) \\
\tilde{P}^l_K &= P\left(\boldsymbol{I}_K, \boldsymbol{D}_K | \boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K\right),
\end{aligned} \tag{7}$$

where $\boldsymbol{C}_K$ is a two-dimensional vector composed of α_K and β_K. Assuming the history state of $\boldsymbol{\Theta}$, $\boldsymbol{\mathcal{T}}$ and $\boldsymbol{\mathcal{C}}$ are accurate, the optimization goal in stage K can be given as,

$$\begin{aligned}
\boldsymbol{\Theta}^*_K, \boldsymbol{T}^*_K, \boldsymbol{C}^*_K &= \underset{\boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K}{\arg\max}\, \hat{P}_{1:K-1} \cdot \tilde{P}^p_K \cdot \tilde{P}^l_K \\
&= \underset{\boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K}{\arg\max}\, \tilde{P}^p_K \cdot \tilde{P}^l_K.
\end{aligned} \tag{8}$$

From Eqs. 6 $\sim$ 8, the global posterior of the whole sequence in Eq. 5 can be finally obtained and maximized recursively. Our derivation corroborates that our incremental optimization mode are theoretically equivalent to the global mode under proper approximations, while the variable amount that needs to be jointly optimized has been greatly reduced, implying the potential for faster and more stable training.

C. Motivation of loss design

To offer guidance to the loss function design, further reformulations on $\tilde{P}^p_K$ and $\tilde{P}^l_K$ in Eq. 8 are necessary. Specifically, based on valid conditional independent assumptions, $\tilde{P}^p_K$ can be decomposed as,

$$\begin{aligned}
\tilde{P}^p_K &= \tilde{P}^T_K \cdot \tilde{P}^C_K \cdot \tilde{P}^\Theta_K \\
\tilde{P}^T_K &= P\left(\boldsymbol{T}_K | \boldsymbol{\Theta}_{1:K-1}, \boldsymbol{\mathcal{T}}_{1:K-1}, \boldsymbol{\mathcal{C}}_{1:K-1}\right) = P\left(\boldsymbol{T}_K | \boldsymbol{\mathcal{T}}_{1:K-1}\right) \\
\tilde{P}^C_K &= P\left(\boldsymbol{C}_K | \boldsymbol{\Theta}_{1:K-1}, \boldsymbol{\mathcal{T}}_{1:K-1}, \boldsymbol{\mathcal{C}}_{1:K-1}\right) = P\left(\boldsymbol{C}_K | \boldsymbol{\mathcal{C}}_{1:K-1}\right) \\
\tilde{P}^\Theta_K &= P\left(\boldsymbol{\Theta}_K | \boldsymbol{\Theta}_{1:K-1}, \boldsymbol{\mathcal{T}}_{1:K-1}, \boldsymbol{\mathcal{C}}_{1:K-1}\right).
\end{aligned} \tag{9}$$

Among three terms in Eq. 9, $\boldsymbol{\Theta}_K$ is conditioned on $\boldsymbol{\Theta}_{1:K-1}$ in $\tilde{P}^\Theta_K$, while the scheme of directly constraining the parameters of the current radiance field and history ones is quite hard to be designed. In fact, the history radiance fields can be considered as a compressed representation of all utilized training images and depth maps. Thus, we have,

$$\tilde{P}^\Theta_K \approx P\left(\boldsymbol{\Theta}_K | \boldsymbol{\mathcal{I}}_{1:K-1}, \boldsymbol{\mathcal{D}}_{1:K-1}, \boldsymbol{\mathcal{T}}_{1:K-1}, \boldsymbol{\mathcal{C}}_{1:K-1}\right). \tag{10}$$

As we have assumed that history state is accurate, and $\boldsymbol{\mathcal{I}}_{1:K-1}$ and $\boldsymbol{\mathcal{D}}_{1:K-1}$ are irrelevant to the current state $\boldsymbol{T}_K$ and $\boldsymbol{C}_K$, Eq. 10 can be further reformulated as,

$$\begin{aligned}
\tilde{P}^\Theta_K &\approx P\left(\boldsymbol{\Theta}_K | \boldsymbol{\mathcal{I}}_{1:K-1}, \boldsymbol{\mathcal{D}}_{1:K-1}\right) \\
&\propto P\left(\boldsymbol{\mathcal{I}}_{1:K-1}, \boldsymbol{\mathcal{D}}_{1:K-1} | \boldsymbol{\Theta}_K\right) \\
&= P\left(\boldsymbol{\mathcal{I}}_{1:K-1}, \boldsymbol{\mathcal{D}}_{1:K-1} | \boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K\right).
\end{aligned} \tag{11}$$

Finally, the product of $\tilde{P}^p_K$ and $\tilde{P}^l_K$ can be given as,

$$\tilde{P}^p_K \cdot \tilde{P}^l_K = \tilde{P}^T_K \cdot \tilde{P}^C_K \cdot \tilde{P}^\Theta_K \cdot \tilde{P}^l_K = \tilde{P}^m_K \cdot \tilde{P}^o_K, \tag{12}$$

where $\tilde{P}^m_K$ is given as,

$$\tilde{P}^m_K = \tilde{P}^T_K \cdot \tilde{P}^C_K = P(\boldsymbol{T}_K | \boldsymbol{\mathcal{T}}_{1:K}) P(\boldsymbol{C}_K | \boldsymbol{\mathcal{C}}_{1:K}), \tag{13}$$

and $\tilde{P}^o_K$ is given as,

$$\begin{aligned}
\tilde{P}^o_K = \tilde{P}^\Theta_K \cdot \tilde{P}^l_K &= P\left(\boldsymbol{\mathcal{I}}_{1:K-1}, \boldsymbol{\mathcal{D}}_{1:K-1} | \boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K\right) \\
&\quad \cdot P\left(\boldsymbol{I}_{K-1}, \boldsymbol{D}_{K-1} | \boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K\right) \\
&= P\left(\boldsymbol{\mathcal{I}}_{1:K}, \boldsymbol{\mathcal{D}}_{1:K} | \boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K\right).
\end{aligned} \tag{14}$$

Since we mainly focus on the update of the local information of the radiance field in each stage of the incremental optimization, we only utilize frames in a local sliding window with size W instead of all frames. Thus, $\tilde{P}^o_K$ can be reformulated as,

$$\tilde{P}^o_K \approx P\left(\boldsymbol{\mathcal{I}}_{K-W:K}, \boldsymbol{\mathcal{D}}_{K-W:K} | \boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K\right). \tag{15}$$

Besides, the rendering of RGB images and depth maps can also be considered as conditionally independent to each other, and $\tilde{P}^o_K$ can be further decomposed as,

$$\begin{aligned}
\tilde{P}^o_K &= \tilde{P}^c_K \cdot \tilde{P}^d_K \\
\tilde{P}^c_K &= P\left(\boldsymbol{\mathcal{I}}_{K-W:K} | \boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K\right) \\
\tilde{P}^d_K &= P\left(\boldsymbol{\mathcal{D}}_{K-W:K} | \boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K\right).
\end{aligned} \tag{16}$$

Finally, by merging Eq. 9 $\sim$ 16, we have,

$$\begin{aligned}
\tilde{P}^p_K \cdot \tilde{P}^l_K &\approx \tilde{P}^m_K \cdot \tilde{P}^c_K \cdot \tilde{P}^d_K \\
\tilde{P}^m_K &= P(\boldsymbol{T}_K | \boldsymbol{\mathcal{T}}_{1:K}) P(\boldsymbol{C}_K | \boldsymbol{\mathcal{C}}_{1:K}) \\
\tilde{P}^c_K &= P\left(\boldsymbol{\mathcal{I}}_{K-W:K} | \boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K\right) \\
\tilde{P}^d_K &= P\left(\boldsymbol{\mathcal{D}}_{K-W:K} | \boldsymbol{\Theta}_K, \boldsymbol{T}_K, \boldsymbol{C}_K\right).
\end{aligned} \tag{17}$$

Defining $\tilde{P}_K = \tilde{P}_K^p \cdot \tilde{P}_K^l$, the goal of stage K of the incremental optimization is to maximize $\tilde{P}_K$ as given in Eq. 8, and the loss function is designed according to the principle that maximizing the distribution $\tilde{P}_K$ equals to minimizing the loss function. Thus, the goal of the stage-K optimization can be summarized as,

$$\boldsymbol{\Theta}_K^*, \boldsymbol{T}_K^*, \boldsymbol{C}_K^* = \underset{\boldsymbol{\Theta}_K, \boldsymbol{\mathcal{T}}_K, \boldsymbol{\mathcal{C}}_K}{\arg\min} L_K^{motion} + L_K^{color} + L_K^{depth}, \tag{18}$$

where the motion loss L_K^{motion}, the color loss L_K^{color} and the depth loss L_K^{depth} correspond to $\tilde{P}_K^m$, $\tilde{P}_K^c$ and $\tilde{P}_K^d$, respectively.

D. Loss function

In this subsection, we define the three loss terms given in Eq. 18 in detail one by one. For the motion loss, we constrain pose $\boldsymbol{T}_K$ with the constant velocity motion model, while the depth correction parameters $\boldsymbol{C}_K$ are constrained with the constant value model. Thus, the motion loss is given as,

$$\begin{aligned} L_K^{motion} &= \lambda_m(\|\overline{\boldsymbol{T}}_K, \boldsymbol{T}_K\|_Q + \|\boldsymbol{C}_{K-1}, \boldsymbol{C}_K\|) \\ \overline{\boldsymbol{T}}_K &= \boldsymbol{T}_{K-1}\boldsymbol{T}_{K-2}^{-1}\boldsymbol{T}_{K-1}, \end{aligned} \tag{19}$$

where λ_m is the motion loss weight, $\|*\|$ represents the smooth l_1-loss, and $\|\overline{\boldsymbol{T}}_K, \boldsymbol{T}_K\|_Q$ can be given as,

$$\|\overline{\boldsymbol{T}}_K, \boldsymbol{T}_K\|_Q = \|Q(\overline{\boldsymbol{R}}_K^T \boldsymbol{R}_K), Q_{\boldsymbol{I}}\| + \|\overline{\boldsymbol{R}}_K^T(\boldsymbol{t}_K - \bar{\boldsymbol{t}}_K)\|, \tag{20}$$

where $\boldsymbol{R}_K$ ($\overline{\boldsymbol{R}}_K$) and $\boldsymbol{t}_K$ ($\bar{\boldsymbol{t}}_K$) are the rotation matrix and the translation vector of $\boldsymbol{T}_K$ ($\overline{\boldsymbol{T}}_K$), respectively, $Q(*)$ is the quaternion representation of the inner rotation matrix, and $Q_{\boldsymbol{I}}$ is the vector representation of the identity quaternion, which can be given as $Q_{\boldsymbol{I}} = [1, 0, 0, 0]^T$.

For the color loss L_K^{color}, we straightforwardly use the smooth l_1-loss between the rendered images and associated training ones in the local sliding window. The depth loss L_K^{depth} is similar to the color loss in form. However, directly computing the norm-loss between the rendered depth maps and training ones brings an overfitting to the deep background regions. Thus, for the depth loss, we use a combination of the depth norm distance and the inverse depth norm distance, which can be given as,

$$L_K^{depth} = \lambda_d \sum_{i=K-W}^{K} (\|\boldsymbol{D}_i^{rend}, \hat{\boldsymbol{D}}_i\| + \|1/\boldsymbol{D}_i^{rend}, 1/\hat{\boldsymbol{D}}_i\|), \tag{21}$$

where $\boldsymbol{D}_i^{rend}$ and $\hat{\boldsymbol{D}}_i$ are the rendered depth map and the corresponding ground-truth, respectively, and λ_d is the weight of the depth loss. In I-DACS, λ_d is determined in an adaptive way, which is set to the ratio between the sum of RGB values of sampled pixels on the training image and the sum of corresponding depth values.

E. Framework implementations

In the optimization process of I-DACS, we first set the local sliding window to the beginning of the sequence, and jointly optimize poses, depth correction coefficients and the radiance field of the first W frames for N_i iterations, finally completing system initialization. After that, we incrementally estimate the trainable parameters of subsequent frames, and gradually construct the radiance field meanwhile. The frame rate of the sequence is downsampled to define key-frames. For a new frame, the radiance field is first considered to be the same as the previous state and thus fixed during the optimization, and only the pose of the current frame is adjusted for N_s iterations. The pose is completely solved from the radiance field thus it is highly consistent with the field. If such a frame is a key-frame, we further jointly optimize poses, depth correction coefficients, and the radiance field as given in Eq. 18 for additional N_s iterations. Besides, a global optimization of $2N_s$ iterations is conducted every sixteen frames, which jointly updates the poses and depth correction coefficients of all passed frames and the radiance field, in order to solve the history-forgetting problem of the radiance field. After the incremental optimization of all frames is completed, $10N_s$ iterations of the global optimization are conducted for the final fine-tuning.

F. Initialization for fast motion sequences

The standard initialization process outlined in Sec. III-E proves effective in most scenarios. However, for fast-motion or low-frame-rate sequences, such as those encountered in the Scannet dataset [81] which we utilized for evaluating our I-DACS and other competing methods, it does not work well. In such instances, we integrate the RAFT optical flow [71] into the initial phase to mitigate overfitting, as was done in [18]. Subsequently, our motion prior mechanism can operate normally, rendering the optical flow unnecessary. As the optical flow is exclusively employed in the initial $0.5N_i$ iterations of the initialization phase, its influence on the consistency between poses and the radiance field in I-DACS is quite negligible.

IV. Direction-aware Sampling based Color Field

A. Radiance field representation in I-DACS

Obviously, whether in I-DACS or other pose-field joint optimization schemes, poses of key-frames change frequently along with the training evolvement, which is quite different from the standard radiance field construction process. In traditional radiance field representations, poses of key-frames are incorporated into the system only by loss functions, or in other words, such representations rely solely on training to keep the consistency of the radiance field to poses. Once the poses change, training is required to readjust the radiance field to the corresponding consistent state, which is undoubtedly time-consuming.

To provide a more intuitive demonstration of the differing training difficulties between the color field and the density field, we conducted radiance field training for 2,000 iterations using data from the Scannet dataset [81]. We recorded the color loss and depth loss in l_2-norm throughout the training process. The results of a typical example sequence are depicted in Fig. 3. From the figure, it is evident that the convergence of depth loss occurs much more swiftly compared to color loss, thereby supporting our idea that learning the color field presents a significantly greater difficulty than learning the

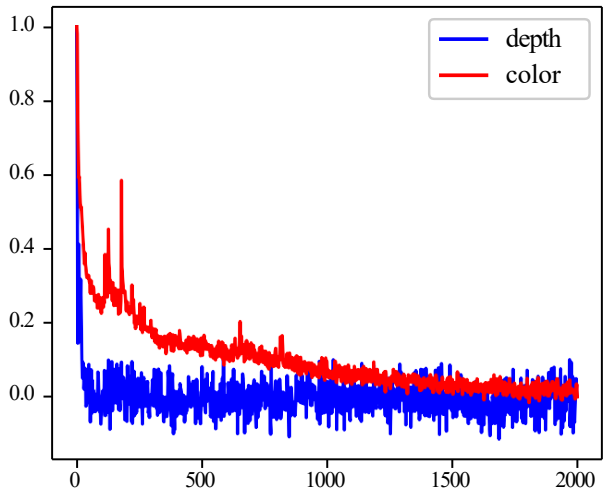


Fig. 3. **The illustration of the color loss and depth loss along with the training evolvement of the radiance field.** To make the comparison more intuitive, both the color loss and depth loss are transformed into the logarithmic domain and scaled by their initial and final values, respectively.

density field. In view of this, in I-DACS, we divide the radiance field to two separate fields, the density field and the color field, and design them individually as,

$$F_{\boldsymbol{\Theta}_D}(\boldsymbol{x}) \rightarrow \sigma \quad F_C(\boldsymbol{x}, \boldsymbol{d}) \rightarrow \boldsymbol{c}, \tag{22}$$

where $F_{\boldsymbol{\Theta}_D}$ and F_C are the density field and the color field, respectively, and $\boldsymbol{\Theta_D}$ consists of the trainable parameters of the density field. It's worth mentioning that our color field can be obtained directly from the images, depth-maps and poses of key-frames without training, thus no trainable parameters are required. For $F_{\boldsymbol{\Theta}_D}$, motivated by [62], we utilize a multi-resolution hash-table representation followed by a two-layer MLP. As a relatively mature solution, it won't be discussed in detail here. Instead, we focus on introducing the representation of F_C, which is modeled using our proposed direction-aware sampling strategy.

B. Sampling based color field

For a frame F^r to be rendered, the key-frame closest to it in physical distance and W_s frames before and after (totally $2W_s+1$ frames) are taken as reference frames. Noting reference frames as $\boldsymbol{\mathcal{F}}^r = \{F_0^r, ..., F_{2W_s}^r\}$, these frames can directly span the color field corresponding to the rendering frame according to their images and poses. Specifically, for a 3D point P at $\boldsymbol{x}^P$ with observation direction $\boldsymbol{d}^P$, the corresponding color $\boldsymbol{c}^P$ can be represented as,

$$\boldsymbol{c}^P = \sum_{i=0}^{2W_s} W(\boldsymbol{d}^P, \boldsymbol{d}_i^P) \cdot \boldsymbol{c}_i^P, \tag{23}$$

where $W(*)$ is our adaptive weight function, $\boldsymbol{c}_i^P$ is the sampled color of P on frame F_i^r, and $\boldsymbol{d}_i^P$ is the corresponding sampling direction. For $\boldsymbol{d}_i^P$, it is just the direction from P to the camera origin of frame F_i^r, which can be given as,

$$\boldsymbol{d}_i^P = \frac{\boldsymbol{x}^P - \boldsymbol{t}_i^r}{\|\boldsymbol{x}^P - \boldsymbol{t}_i^r\|_2}, \tag{24}$$

where $\boldsymbol{t}_i^r$ is the translation vector of pose $\boldsymbol{T}_i^r$ of frame F_i^r. To obtain $\boldsymbol{c}_i^P$, we can project P on F_i^r and further sample the corresponding RGB values as,

$$\boldsymbol{c}_i^P = \boldsymbol{I}_i^r(\boldsymbol{p}_i^P) = \boldsymbol{I}_i^r(\boldsymbol{K}\boldsymbol{T}_i^r\boldsymbol{x}^P), \tag{25}$$

where $\boldsymbol{I}_i^r$ is the RGB image of F_i^r, and $\boldsymbol{K}$ is the intrinsic matrix of the camera. In Eq. 25, the conversion from a homogeneous coordinate to a non-homogeneous one is ignored. It's worth mentioning that, in such a sampling operation, the bilinear interpolation is utilized and such an interpolation process guarantees the gradient backward of the loss. Next, we will introduce the definition of the adaptive weight $W(*)$ in Eq. 23 in detail.

C. Direction-aware sampling weight

In I-DACS, the weight of $\boldsymbol{c}_i^P$ is determined adaptively mainly according to the observed direction $\boldsymbol{d}^P$ and the sampling direction $\boldsymbol{d}_i^P$, which can be given as,

$$W(\boldsymbol{d}^P, \boldsymbol{d}_i^P) = C_i^P \cdot O_i^P, \tag{26}$$

where C_i^P is the direction weight to measure the consistency between the observation direction and the sampling one, and O_i^P is the occlusion decay ratio to reduce the weight of the sampling color in occluded regions. C_i^P is defined based on the cosine distance and can be given as,

$$C_i^P = \frac{1}{cos < \boldsymbol{d}^P, \boldsymbol{d}_i^P > + \epsilon}, \tag{27}$$

where $cos < * >$ represents the cosine distance and $\epsilon = 1e{-}5$ is used to prevent dividing by zero. C_i^P offers a smooth direction-aware interpolation among all sampled RGB values. However, sometimes the sampled color $\boldsymbol{c}_i^P$ is not the true observation of P due to occlusion. Only using such an interpolation may cause obvious artifacts in rendered views. Thus, we also introduce the occlusion decay ratio O_i^P, which can be given as,

$$O_i^P = \left(\frac{D^{gap}}{D^{gap} + Relu(D_i^{exc} - D^{gap})}\right)^2, \tag{28}$$

where D_i^{exc} is the ratio of the sampling distance exceeding the corresponding estimated depth, and D^{gap} is a hyper-parameter that ensures a certain degree of tolerance of the decay strategy to inaccuracies in depth estimations and poses. D_i^{exc} is given as,

$$D_i^{exc} = \frac{\|\boldsymbol{x}^P - \boldsymbol{t}_i^r\|_2 - \hat{\boldsymbol{D}}_i(\boldsymbol{p}_i^P)}{\hat{\boldsymbol{D}}_i(\boldsymbol{p}_i^P)}, \tag{29}$$

where $\hat{\boldsymbol{D}}_i$ is the undistorted depth estimation, and $\boldsymbol{p}_i^P$ is the projection of P on frame F_i^r.

Our color sampling scheme leverages RGB images and poses to directly span the local color field according to both observed directions and sampling ones without the need for explicit training, which provides robust support for novel-view synthesis while minimizing the occurrence of noticeable blurring and artifacts. The poses are elegantly modeled also in the color field representation instead of only in loss functions. That means once the poses change during training, the color field can naturally adapt to be consistent with their current

state, and only the density field needs to be readjusted, bringing the excellent training speed of I-DACS.

Actually, our sampling-based color field is somewhat similar to the idea in NeuralWarp [83]. Specifically, both NeuralWarp [83] and I-DACS utilize the accurate and high-frequency information sampled from the training images, choose to describe the geometry information and the color information in two separate fields, and handle the occlusion problems in sampling. However, there are also many differences between these two methods:

1) NeuralWarp [83] and I-DACS focus on different problems. NeuralWarp [83] focuses on generating the mesh of the scene. For comparison, I-DACS focuses on the task of novel-view synthesis.
2) NeuralWarp [83] and I-DACS have different motivations for using sampled colors to guide the training. The motivation of NeuralWarp [83] to sample color information from training images is to utilize the high-frequency color information to gain better geometric structure, while in our I-DACS, we further use the color sampling mechanism to keep the consistency between the radiance field and poses.
3) NeuralWarp [83] and I-DACS choose different representations to model the scene. NeuralWarp [83] chooses the SDF representation, in which color is irrelevant to the observation direction. Thus, NeuralWarp [83] samples color from a single frame. For comparison, our I-DACS selects a direction-aware color sampling strategy, in which the color information is sampled from multiple frames and then fused with direction-based smooth interpolation.
4) NeuralWarp [83] and I-DACS handle occlusion problems in different ways. NeuralWarp [83] handles the occlusion problem mainly by the density output of the network. For comparison, since in I-DACS the monocular depth estimation is adopted to offer depth supervision, we handle the occlusion problem by scale-corrected depth maps, which is a more efficient way.

D. Training details of the color field

As aforementioned, in I-DACS, the radiance field is decomposed into two parts, the density field and the color field. The density field is modeled using the hash-table representation, while the color field is based on our proposed direction-aware color sampling strategy. During the rendering process in the testing phase, color information is sampled from reference frames, including the nearest frame to the rendering pose and multiple frames within a local sampling window. However, in the training phase, if the supervision frame is a key-frame, this strategy would predominantly sample color information from the frame itself, potentially hindering training convergence. To address this issue, the supervision frame itself is excluded as a reference frame in such cases. Additionally, to bolster long-term data associations, older frames are incorporated into the sampling process during training. Specifically, aside from the reference frames within the sampling window, one frame is randomly selected from each of the three preceding subsequences: 5-10 frames, 10-15 frames, and 15-30 frames prior to the current frame.

TABLE I
QUALITATIVE COMPARISON WITH RELATED METHODS.

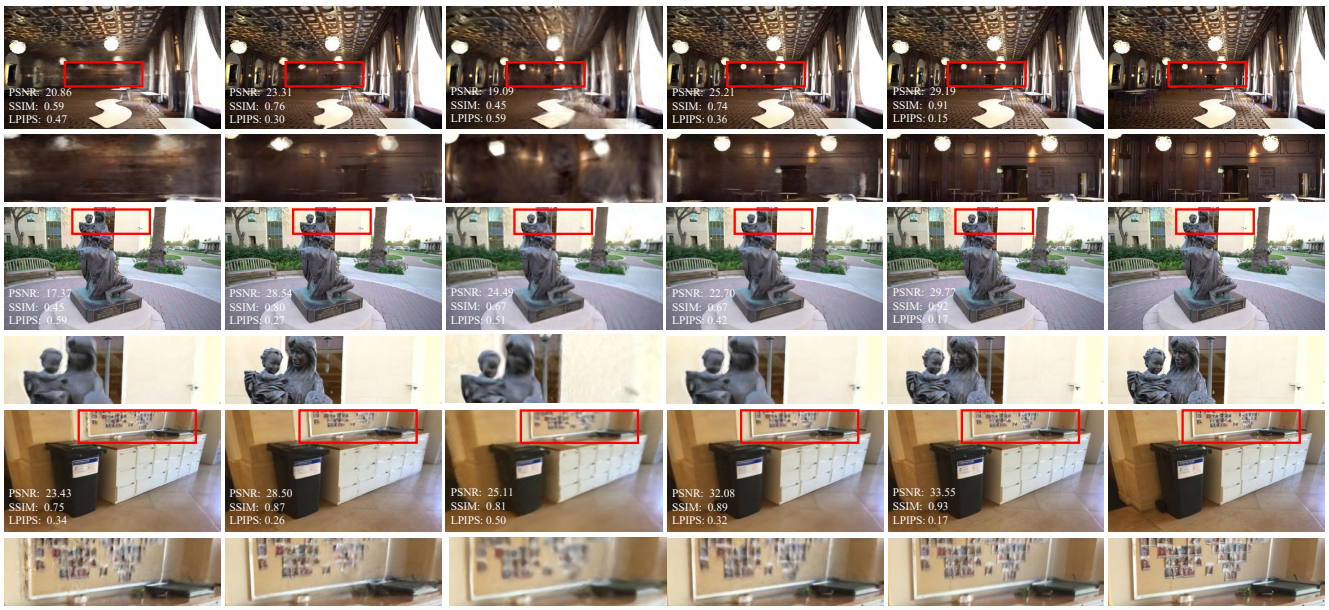
Method	Str	Con	Iter	Time
BARF [21]	Global	Training	200K	5 hours
SC-NeRF [20]	Global	Training	1M	40 hours
NeRFmm [22]	Global	Training	1M	20 hours
GNeRF [72]	Global	Training	1.2M	30 hours
Nope-NeRF [19]	Global	Training	1M	40 hours
LocalRF [18]	Incremental	Training	150K	4 hours
CF-NeRF [84]	Incremental	Training	180K	8 hours
I-DACS	**Incremental**	**Naturally**	**18K**	**30 minutes**

V. EXPERIMENTAL RESULTS

A. Experimental setup

Dataset. To guarantee the rationality of the comparison, we followed the experimental settings in [19] and conducted experiments on two real-world datasets, Tanks and Temples [82] and Scannet [81]. For Tanks and Temples, eight sequences were selected, including three indoor sequences and five outdoor sequences. The image resolution was downsampled to 960×540, which is half of the original resolution. In each sequence, $1/8$ of the images were chosen to construct the test set, while the remaining images were used for training. As for Scannet dataset, four sequences were utilized, each with consecutive 80-100 frames, and also $1/8$ of the frames were selected to build the test set. The images were also downsampled to half of the original resolution, 648×484. It is worth noting that there are dark regions on the boundaries of the images in Scannet, thus fifteen pixels on the boundaries were cropped off before the downsample. Except for the aforementioned two real-world dataset, we also use a synthetic dataset for evaluation, Replica [85]. For Replica, four sequences were chosen, each with consecutive 100 frames.

Implementation details. For the implementation of our incremental joint optimization framework, the motion loss weight λ_m was set to $1e-3$, the local sliding window size W was adaptively set to $4W_s+1$ to cover the sampling window and guarantee a satisfactory fitting for the local information of the color field, the initialization lasted for $N_i = 1200$ iterations, and the optimization for each frame lasted for $N_s = 100$ iterations. In each iteration of the training process, 2048 rays are selected randomly and no more than 128 points are sampled on each ray. As for our radiance field implementations, on the aspect of the density field, similar to [62], a 16-level multi-resolution hash-table representation was utilized and in each level the feature dimension was 2. Hash features were concatenated to a 32-dimension vector and then fed to a two-layer MLP with 64 hidden neurons to predict the density. On the aspect of the color field, the half-size W_s of the sampling window was set to 1, which is enough to achieve high-quality novel view synthesis. Besides, the tolerance parameter D^{gap} was set to 0.2.



(a) SC-NeRF [20] (b) LocalRF [18] (c) CF-NeRF [84] (d) Nope-NeRF [19] (e) I-DACS (f) Ground Truth

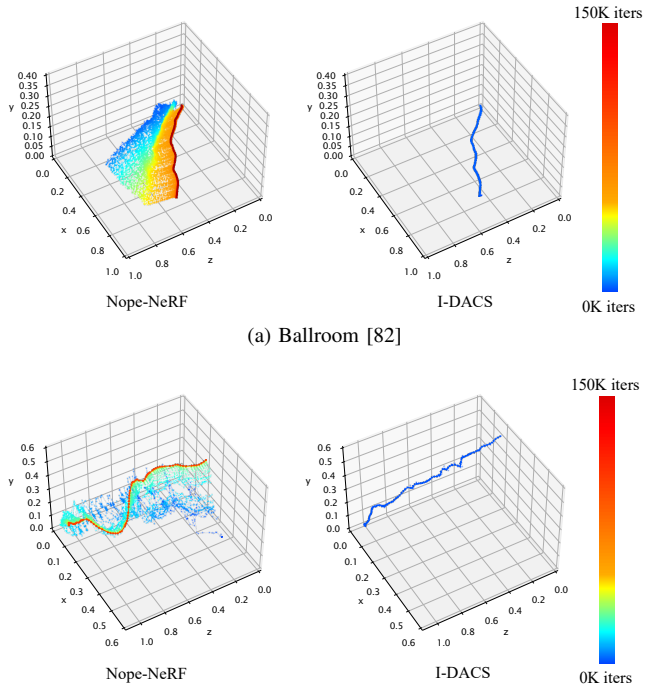
Fig. 4. **Typical samples of the novel-view synthesis results of compared methods.** In each group of data, the full synthesized views are shown on the top and the enlarged local ROIs are on the bottom. The results of the top two scenes are from the Tanks and Temples dataset [82], and the bottom one is from the Scannet dataset [81].

I-DACS was implemented using Python with PyTorch. All experiments were conducted on a workstation equipped with an Intel Xeon(R) CPU E5-2678 V3 processor and a TITAN RTX GPU.

B. Qualitative experiments

Traits of methods. As shown in Table I, we compared the SOTA and representative competitors in this field, including BARF [21], SC-NeRF [20], NeRFmm [22], GNeRF [72], Nope-NeRF [19], LocalRF [18] and CF-NeRF [84], and also our proposed I-DACS to demonstrate their characteristics more clearly in four aspects, including: 1) Which type of optimization strategy is utilized (Str)? 2) How to guarantee the consistency between the color field and poses (Con)? 3-4) How many iterations (Iter) and time (Time) are required to complete the training of a sequence including one hundred frames? From Table I, it is evident that our I-DACS can jointly optimize poses and the radiance field in an incremental manner. Besides, I-DACS achieves natural consistency between the color field and poses, implying the potential for enhanced efficiency in tasks involving joint pose-field estimation. This is further supported by the training iteration number and the time consumption. Specifically, compared with the current SOTA, Nope-NeRF [19] and LocalRF [18], I-DACS achieves an about 60-times and 8-times faster speed, respectively. Noticing that since the performance of some of the competitors [20]–[22], [72] are similar and relatively unsatisfactory, we only choose one representative [20] among them and also SOTA schemes [18], [19], [84] as baselines in subsequent comparative experiments.

Novel-view synthesis. So as to show the performance in novel-view synthesis of our I-DACS and other compared schemes more clearly, we trained all these models under the settings introduced in Sec. V-A. Subsequently, novel views were rendered and some representative samples are shown in Fig. 4. From Fig. 4, it can be seen that, thanks to our color sampling mechanism, our I-DACS can best preserve image details. Besides, our incremental optimization strategy offered



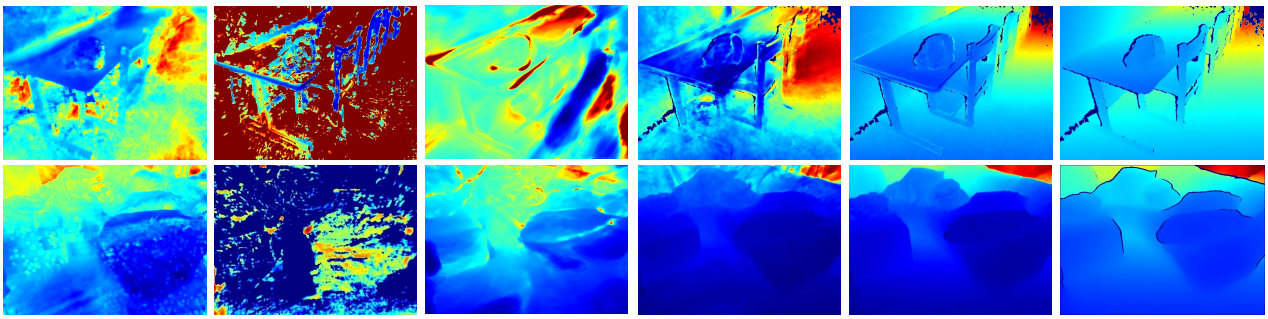
(a) Ballroom [82]

(b) Scene_0431 [81]

Fig. 5. **Illustrations of the trajectories along with the evolvement of training.** The trajectories corresponding to different iterations are plotted in different colors.

TABLE II
PERFORMANCE ON NOVEL-VIEW SYNTHESIS OF COMPARED JOINT OPTIMIZATION SCHEMES.

	Scenes	SC-NeRF [20]			LocalRF [18]			CF-NeRF [84]			Nope-NeRF [19]			**I-DACS**		
		PSNR↑	SSIM↑	LPIPS↓	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Scannet	0079_00	31.33	0.82	0.46	29.66	0.80	0.38	24.95	0.76	0.55	**32.47**	**0.84**	**0.41**	**34.78**	**0.88**	**0.21**
	0301_00	29.05	0.75	0.43	**30.30**	**0.81**	**0.30**	26.06	0.79	0.49	29.83	0.77	0.36	**32.10**	**0.82**	**0.23**
	0418_00	32.57	0.90	0.40	31.17	**0.82**	**0.30**	27.46	0.79	0.48	**31.33**	0.79	0.34	**32.10**	**0.82**	**0.21**
	0431_00	32.77	0.90	0.41	28.23	0.86	0.29	24.75	0.82	0.47	**33.83**	**0.91**	**0.39**	**34.19**	**0.94**	**0.15**
	Mean	30.60	0.81	0.41	29.84	0.82	**0.32**	25.81	0.79	0.50	**31.86**	**0.83**	0.38	**33.00**	**0.87**	**0.20**
Tanks and Temples	Church	21.96	0.60	0.53	22.37	0.68	**0.39**	22.23	0.60	0.51	**25.17**	**0.73**	**0.39**	**27.67**	**0.88**	**0.14**
	Barn	23.26	0.62	0.51	**28.11**	**0.82**	**0.25**	24.60	0.64	0.50	26.35	0.69	0.44	**31.75**	**0.91**	**0.12**
	Museum	24.94	0.69	0.45	24.43	**0.76**	0.32	20.93	0.53	0.57	**26.77**	**0.76**	**0.35**	**30.78**	**0.92**	**0.13**
	Family	22.60	0.63	0.51	**27.46**	**0.86**	**0.23**	23.60	0.64	0.52	26.01	0.74	0.41	**31.96**	**0.95**	**0.08**
	Horse	25.23	0.76	0.37	27.55	**0.87**	**0.19**	24.27	0.78	0.40	**27.64**	0.84	0.26	**31.85**	**0.95**	**0.08**
	Ballroom	22.64	0.61	0.48	23.02	**0.73**	**0.32**	19.41	0.42	0.57	**25.33**	0.72	0.38	**30.59**	**0.94**	**0.07**
	Francis	26.46	0.73	0.49	**29.57**	**0.85**	**0.30**	27.07	0.73	0.48	29.48	0.80	0.38	**33.43**	**0.92**	**0.15**
	Ignatius	23.00	0.55	0.53	**25.59**	**0.77**	**0.31**	20.30	0.43	0.64	23.96	0.61	0.47	**28.93**	**0.91**	**0.11**
	Mean	23.76	0.65	0.48	26.01	**0.79**	**0.29**	22.80	0.60	0.52	**26.34**	0.74	0.39	**30.87**	**0.92**	**0.11**



(a) SC-NeRF [20] (b) LocalRF [18] (c) CF-NeRF [84] (d) Nope-NeRF [19] (e) I-DACS (f) Ground Truth

Fig. 6. **Typical depth rendering results on the Scannet dataset [81] of compared methods.** The upper group of the rendered depth maps are from Scene_0079 sequence and the lower results are from Scene_0301

accurate poses consistent with the field, thereby minimizing blurring and artifacts in the synthesized results.

Optimization process of poses. To illustrate the fast and steady convergence of our incremental optimization strategy in I-DACS, we plot some typical samples of the trajectories along with the training evolvement in Fig. 5. Besides, the corresponding results of Nope-NeRF [19], which is the only scheme that can achieve somewhat comparable localization performance to I-DACS (please refer to Sec. V-C), are also offered for reference. From Fig. 5, it can be seen that in our I-DACS, the training of poses converges much faster and more stable compared with Nope-NeRF [19], corroborating the outstanding efficiency and accuracy of our I-DACS in localization qualitatively.

Depth rendering quality. Compared with existing radiance field representations, in our I-DACS the direction-aware sampling strategy offers higher-frequency color information and also benefits the fitting of the density field. Thus, I-DACS can show superior performance in recovering the geometric structure of the scene. To qualitatively support our analysis, we offer some typical samples of the rendered depth maps of compared methods on the Scannet dataset [81] in Fig. 6. From Fig. 6, it can be seen that, among all competitors, LocalRF [18] performs quite terribly on geometry in the indoor environment since it highly relies on the optical flow estimation and chooses a relative inverse depth supervision mechanism. SC-NeRF [20] and CF-NeRF [84] do not utilize depth supervision, thus they also performs unsatisfactorily. Nope-NeRF [19] and our I-DACS can recover relatively accurate geometric structures of the scene, while compared with Nope-NeRF, our I-DACS performs obviously better in describing details.

C. Quantitative experiments

Novel-view synthesis performance. Under the experimental settings introduced in Sec. V-A, we trained the models of both our I-DACS and other three typical competitors, and evaluated the quality of the synthesized novel views quantitatively. Three metrics, including PSNR, SSIM and LPIPS, were utilized, and the backbone network of the LPIPS metric is the VGG network. Experimental results are summarized in Table II. The experimental results unequivocally demonstrate the remarkable performance advantages exhibited by our I-DACS, corroborating the superiority of our direction-aware sampling strategy in describing intricate image details and the outstanding novel-view synthesis accuracy of I-DACS.

Localization accuracy. The localization accuracy of compared methods was evaluated by three metrics, the absolute trajectory error ATE, the relative translation error RPE_t and the relative rotation error RPE_r. Experimental results are offered in Table III. In the table, the unit of ATE and RPE_t are

TABLE III
LOCALIZATION ACCURACY IN BOTH ATE AND RPE OF COMPARED JOINT OPTIMIZATION SCHEMES.

	Scenes	SC-NeRF [20]			LocalRF [18]			CF-NeRF [84]			Nope-NeRF [19]			**I-DACS**		
		ATE↓	RPE_t ↓	RPE_r ↓	ATE	RPE_t	RPE_r	ATE	RPE_t	RPE_r	ATE	RPE_t	RPE_r	ATE	RPE_t	RPE_r
Scannet	0079_00	0.115	2.064	0.664	0.100	1.122	0.457	0.029	**0.421**	0.582	0.023	0.752	0.204	**0.019**	0.563	**0.179**
	0301_00	0.056	1.133	0.422	0.068	0.694	0.322	0.038	0.632	0.319	**0.013**	**0.399**	**0.123**	0.018	0.459	0.130
	0418_00	0.016	1.528	0.502	0.020	0.398	0.147	0.045	0.609	0.306	0.015	0.455	0.119	**0.013**	**0.367**	**0.100**
	0431_00	0.205	4.110	0.499	0.099	1.222	0.391	0.077	0.978	0.443	0.069	1.625	0.274	**0.034**	**0.804**	**0.166**
	Mean	0.098	2.209	0.522	0.072	0.859	0.329	0.047	0.660	0.413	0.030	0.808	0.180	**0.021**	**0.548**	**0.144**
Tanks and Temples	Church	0.108	0.836	0.187	0.039	0.324	0.137	0.018	0.053	0.052	0.008	**0.034**	**0.008**	**0.006**	0.044	0.010
	Barn	0.157	1.317	0.429	0.013	0.067	0.054	0.016	0.086	0.152	**0.004**	**0.046**	0.032	0.006	0.051	**0.021**
	Museum	0.316	8.339	1.491	0.021	0.221	0.308	0.021	0.585	0.280	0.020	0.207	0.202	**0.017**	**0.179**	**0.170**
	Family	0.142	1.171	0.499	0.003	0.055	0.017	0.018	0.074	0.123	**0.001**	0.047	0.015	**0.001**	**0.046**	**0.008**
	Horse	0.019	1.366	0.438	0.003	0.189	0.034	0.011	0.277	0.216	0.003	**0.179**	**0.017**	**0.002**	0.187	0.020
	Ballroom	0.012	0.328	0.146	0.020	0.246	0.086	0.017	0.637	0.314	0.002	**0.041**	0.018	**0.001**	0.048	**0.016**
	Francis	0.192	1.233	0.483	**0.003**	0.049	0.048	0.028	0.104	0.213	0.005	0.057	**0.009**	0.005	**0.045**	0.035
	Ignatius	0.085	0.533	0.240	0.007	0.067	0.029	0.042	0.349	0.328	**0.002**	**0.026**	**0.005**	0.003	0.065	0.006
	Mean	0.129	1.890	0.489	0.014	0.152	0.089	0.021	0.271	0.210	0.006	**0.080**	0.038	**0.005**	0.083	**0.036**

TABLE IV
LOCALIZATION ACCURACY AND RENDERING QUALITY OF COMPARED METHODS ON REPLICA DATASET.

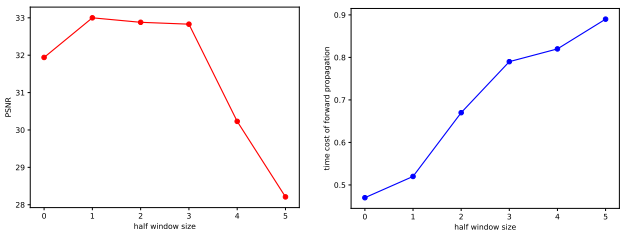
	Loc			NVS		
	ATE↓	RPE_t ↓	RPE_r ↓	PSNR↑	SSIM↑	LPIPS↓
CF-NeRF [84]	0.046	0.834	0.371	27.51	0.84	0.40
Nope-NeRF [19]	0.077	0.954	0.238	34.30	0.93	0.21
I-DACS	**0.012**	**0.167**	**0.036**	**37.66**	**0.97**	**0.12**

TABLE V
THE ACCURACY OF DEPTH RENDERING OF COMPARED METHODS ON SCANNET DATASET.

	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	σ_1↑	σ_2↑	σ_3↑
SC-NeRF	0.417	0.642	1.079	0.476	0.362	0.658	0.832
LocalRF	1.023	1.115	1.378	3.547	0.183	0.388	0.577
CF-NeRF	0.389	0.299	0.851	0.317	0.562	0.833	0.968
Nope-NeRF	0.141	0.137	0.568	0.176	0.828	0.970	0.987
I-DACS	**0.098**	**0.090**	**0.151**	**0.144**	**0.925**	**0.984**	**0.996**

TABLE VI
EVALUATION ON THE DESIGN VALIDNESS OF SEPARATE COLOR FIELD AND DENSITY FIELD IN I-DACS.

	Loc			NVS		
	ATE↓	RPE_t ↓	RPE_r ↓	PSNR↑	SSIM↑	LPIPS↓
NS-DACS	0.026	0.652	0.153	30.13	0.82	0.33
I-DACS	**0.021**	**0.548**	**0.144**	**33.29**	**0.87**	**0.20**



(a) Rendering quality (b) Rendering time cost

Fig. 7. **The rendering quality and rendering speed of I-DACS under different settings of the half window size W_s.** The rendering quality measured by PNSR is plotted in (a) and the rendering speed measured by the average forward-propagation time cost is plotted in (b).

both meters, and RPE_r is in unit of degrees. Noting that the values of RPE_t are all multiplied by 100 for better comparison. Table III shows that I-DACS performs best in the localization accuracy among all counterparts, and only Nope-NeRF [19] manages to achieve comparable performance on the Tanks and Temples dataset [82]. Since the speed performance of I-DACS overwhelmingly outperforms other methods (about 80 times faster than Nope-NeRF [19]), we can say that our I-DACS shows SOTA localization performance, excelling in both speed and accuracy.

Performance on synthetic dataset. For a more comprehensive evaluation, we also conducted quantitative experiments on the synthetic dataset, Replica [85]. Two representative competitors, CF-NeRF [84] and Nope-NeRF [19], and also our I-DACS were evaluated. Quantitative evaluation results were summarized in Table IV. From the results, it can be seen that our I-DACS shows an overwhelming performance advantage in both localization accuracy and rendering quality.

Depth rendering performance. To evaluate the geometric accuracy of I-DACS and its competitors, we conducted quantitative experiments for depth rendering on Scannet [81] dataset. Seven commonly utilized metrics were chosen to evaluate the accuracy of the yielded depth maps, including Abs Rel, Sq Rel, RMSE, σ_1, σ_2 and σ_3. Relevant experimental results were summarized in Table V. From Table V, it can be seen that our I-DACS performs obviously better than all competitors, corroborating its superior geometric accuracy.

Window size analysis. To determine the half window size W_s in color sampling, we evaluated the performance of I-DACS on Scannet dataset [81] under different W_s's settings. The evaluation was conducted mainly in two aspects: rendering quality and rendering speed. The rendering quality was measured by the average PSNR, while the rendering speed was measured by the average time cost for a single time forward propagation in I-DACS. Relevant experimental results were summarized in Fig. 7. From the experimental results, it can be clearly seen that when W_s is within 1-3, the rendering performance of I-DACS is similarly satisfactory, while larger W_s settings bring

TABLE VII
PERFORMANCE ON NOVEL-VIEW SYNTHESIS OF COMPARED NERF-BASED SLAM METHODS AND OUR I-DACS.

Method	Scene	PSNR↑	SSIM↑	LPIPS↓
NICER-SLAM [74]	0079_00	24.44	0.67	0.58
	0301_00	23.02	0.62	0.51
	0418_00	25.18	0.61	0.51
	0431_00	20.92	0.66	0.55
	Mean	23.39	0.64	0.54
DIM-SLAM [76]	0079_00	26.01	0.71	0.46
	0301_00	25.01	0.69	0.42
	0418_00	26.26	0.66	0.39
	0431_00	21.65	0.74	0.46
	Mean	24.73	0.70	0.43
I-DACS	0079_00	**34.78**	**0.88**	**0.21**
	0301_00	**32.10**	**0.82**	**0.23**
	0418_00	**32.10**	**0.82**	**0.21**
	0431_00	**34.19**	**0.94**	**0.15**
	Mean	**33.00**	**0.87**	**0.20**

more time cost for rendering. One important reason for this phenomenon may be that the frames far from the rendering view contribute little to the color field and may even make the occlusion problem more serious. Thus, in our implementations, we just set W_s to 1.

Validness of the radiance field representation. As aforementioned, in our I-DACS, the radiance field is divided into two separate fields, the density field and the color field, and the color field is modeled by our proposed direction-aware color sampling strategy. To corroborate the validness of our utilized radiance field representation, we further compared the performance of I-DACS with a variant baseline, NS-DACS. In NS-DACS, the radiance field is represented as a general hash-table. The quantitative evaluation was conducted on Scannet dataset [81], and the performance of I-DACS and NS-DACS in both localization accuracy and rendering quality were evaluated. It's worth mentioning that, without the sampling-based color field representation, NS-DACS took about three times longer training time to converge compared with I-DACS. Relevant experimental results are given in Table VI. From Table VI, it can be seen that I-DACS performs obviously better than NS-DACS in both localization and rendering, strongly corroborating the effectiveness of our currently utilized radiance field representation.

Comparison with NeRF-based SLAM. Existing NeRF-based SLAM methods can also model the scene without pose prior under the radiance field framework. However, as aforementioned, in these methods, the environment is usually modeled as TSDF maps in which the color information is direction-independent. Compared with the radiance field representations, such TSDF representations are more lightweight and can provide better support to the mesh exportation task, but they cannot achieve satisfactory novel-view synthesis quality. To verify our analysis, we evaluated the novel-view synthesis performance of our I-DACS and two typical monocular NeRF-based SLAM methods, including NICER-SLAM [74] and DIM-SLAM [76]. The quantitative evaluation results are summarized in Table VII. From Table VII, it can be obviously seen that the rendering quality of these NeRF-based SLAM methods is actually incomparable to our I-DACS.

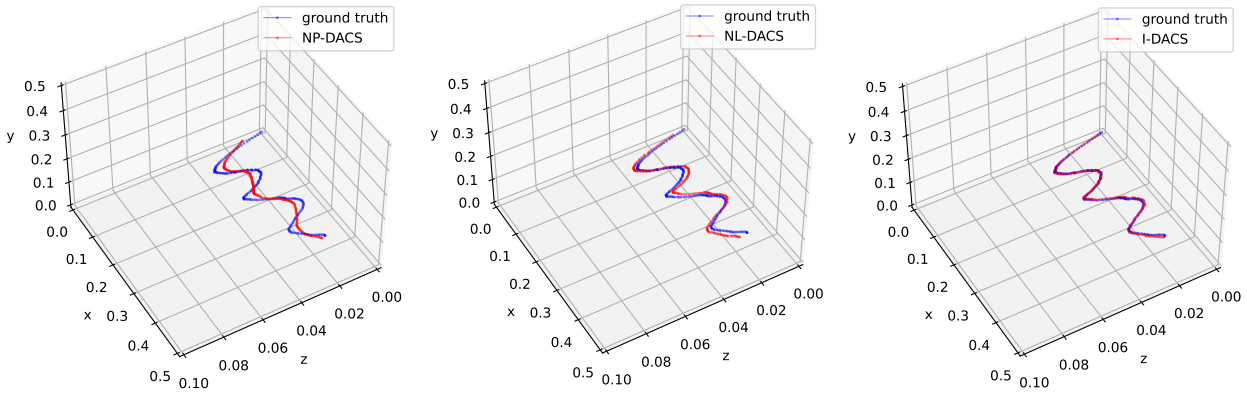
TABLE VIII
ABLATION STUDY OF I-DACS IN BOTH NOVEL-VIEW SYNTHESIS (NVS) AND LOCALIZATION (LOC) ON SCANNET DATASET.

	Loc			NVS		
	ATE↓	RPE$_t$ ↓	RPE$_r$ ↓	PSNR↑	SSIM↑	LPIPS↓
ND-DACS	0.030	0.627	0.176	32.96	**0.87**	0.21
NP-DACS	0.060	0.981	0.239	31.42	0.85	0.23
NI-DACS	0.026	0.602	0.160	31.20	0.86	0.22
NO-DACS	0.023	0.604	0.160	32.41	0.86	0.21
NL-DACS	0.030	0.574	0.158	33.27	0.86	**0.20**
I-DACS	**0.021**	**0.548**	**0.144**	**33.29**	**0.87**	**0.20**

Fig. 8. **Typical rendering results of compared baseline variants and I-DACS.** From top to bottom, the rendering results of NI-DACS, NO-DACS and I-DACS are offered, respectively. In each group of data, the synthesized image is given on the left and local enlarged ROIs are given on the right.

D. Ablation studies

The performance of our I-DACS was evaluated mianly in twofolds: the novel-view synthesis and the localization. To verify the superior performance of module configurations in I-DACS currently employed, we compared I-DACS with other five baseline variants on the Scannet dataset [81], which were: 1) ND-DACS: Training without the depth supervision; 2) NP-DACS: Training without the pose prior guidance; 3) NI-DACS: The direction-aware interpolation was substituted by a simple averaging; 4) NO-DACS: The occlusion decay mechanism was deactivated; 5) NL-DACS: The long-term data association of the color field (introduced in Sec. IV-D) was deactivated. Detailed quantitative experimental results are summarized in Table VIII. From the results, it can be seen that I-DACS outperforms all other variants in terms of both novel-view synthesis and localization, implying that our design is crucial in guaranteeing the performance of I-DACS. Next, to show the



(a) NP-DACS (b) NL-DACS (c) I-DACS

Fig. 9. **Qualitative comparison on the localization accuracy of NP-DACS, NL-DACS and I-DACS.** The ground truth trajectories are plotted as blue curves and the localization results of the evaluated methods are plotted as red curves.

performance gain brought by each module more intuitively, we further analyze them in detail with the help of qualitative evaluation results.

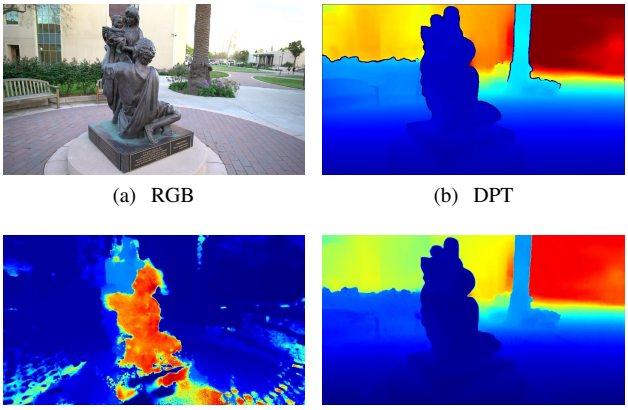
Rendering quality. Among all variants, NI-DACS and NO-DACS were used to evaluate the effectiveness of our direction-aware interpolation strategy and occlusion decay mechanism, respectively. These two mechanisms are utilized to directly generate the color field from training images and they directly affect the rendering quality of I-DACS. Typical rendering results of these two variants and I-DACS are given in Fig. 8. From Fig. 8, it can be seen that since NI-DACS ignores the observation direction and directly performs a simple average process on the sampled colors from multiple frames, obvious artifacts appear at the edge of the statue in its rendering results. Besides, NO-DACS does not handle the occlusion problem and thus the color may be sampled from incorrect positions in some of the frames, which also brings obvious artifacts in its rendering results. For comparison, our I-DACS can achieve high-quality rendering without obvious artifacts or blur. Through the ablation studies, it's corroborated that both the direction-aware sampling strategy and the occlusion decay mechanism are necessary for I-DACS.

Localization accuracy. NP-DACS and NL-DACS were about the evaluations on the motion prior supervision and the long-term data association mechanism, respectively. These two mechanisms are utilized in our incremental pose-field joint estimation framework and are directly related to the localization accuracy of I-DACS. To further qualitatively corroborate our analysis, typical samples of the localization trajectories yielded by these two variants and I-DACS are offered in Fig. 9. From Fig. 9, it can be clearly seen that, without motion prior, NP-DACS cannot show stable tracking performance. As for NL-DACS, without long-term data associations, the accumulated localization errors cannot be effectively eliminated. For comparison, the trajectory yielded by I-DACS is almost completely consistent with the ground truth, corroborating the necessity of the existing module configurations of our incremental framework.

Geometric accuracy. ND-DACS is used to verify the necessity of the depth supervision, which is most related to the geometry structure of the radiance field. To verify the necessity of the depth supervision, we offer quantitative and qualitative comparisons between ND-DACS and I-DACS in the accuracy of depth rendering. Specifically, quantitative depth rendering results of both ND-DACS and I-DACS on Scannet dataset [81] are offered in Table IX. Besides, typical samples of the rendered depth maps from both ND-DACS and I-DACS are offered in Fig. 10. From Table IX and Fig. 10, it can be found that I-DACS overwhelmingly surpass ND-DACS in the performance of geometric accuracy, strongly corroborating the effectiveness of the depth supervision.

TABLE IX
DEPTH RENDERING ACCURACY OF ND-DACS AND I-DACS ON SCANNET DATASET.

	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	σ_1↑	σ_2↑	σ_3↑
ND-DACS	0.205	0.308	0.669	0.257	0.798	0.934	0.969
I-DACS	**0.098**	**0.090**	**0.151**	**0.144**	**0.925**	**0.984**	**0.996**



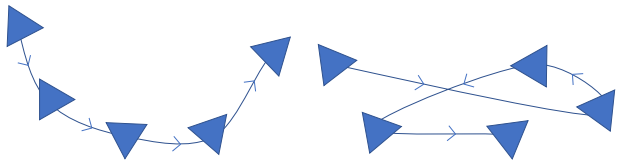
(a) RGB (b) DPT

(c) ND-DACS (d) I-DACS

Fig. 10. **The comparison between ND-DACS and I-DACS on the accuracy of geometric structure.** (a) and (b) are the ground truth RGB image and the corresponding depth map estimated by DPT, respectively. (c) is the depth map rendered by ND-DACS while (d) is the result of I-DACS.

VI. LIMITATIONS AND FUTURE WORK

Currently, the color field in I-DACS is implemented based on our proposed direction-aware color sampling strategy.



(a) Structured (b) Unstructured

Fig. 11. **The illustrations of the structured dataset and the unstructured dataset.** The poses of frames are plotted as blue triangles and frames are linked in order.

While our approach successfully enables rapid training and high-fidelity novel-view synthesis, storing images of keyframes in the memory may impose a certain degree of storage burden. Specifically, given a sequence from Tanks and Temples [82] or Scannet [81], the model in hash-table form occupies about 50.4MB space, while in I-DACS 7.2MB∼21.5MB additional space is necessary to store the RGB images of keyframes. Since I-DACS overwhelmingly surpasses existing similar methods in both rendering quality, localization accuracy and training speed, such a degree of memory overhead is usually worthy. This issue can potentially be mitigated through caching engineering design or image compression.

Besides, I-DACS can perform well on data sequences with structured frames, while it cannot handle the cases of training radiance fields from unstructured frames. The illustrations of structured and unstructured frames are given in Fig. 11. Actually, for the problem of radiance field construction on unstructured datasets without pose prior, as far as we know, not only our I-DACS but also all other existing similar methods cannot be usable. Specifically, BARF [21], NeRFmm [22] and SC-NeRF [20] can only refine the poses of frames in limited ranges and without pose prior they cannot normally work. LocalRF [18] and Nope-NeRF [19] rely highly on the optical flow or point cloud loss to offer inter-frame constraints, while on unstructured datasets such constraints cannot be established. The main cause to the failure of existing methods on the task of pose-free radiance field construction on unstructured dataset is that, there may be no common-view regions between adjacent frames since the data sequence is disordered. Thus, the solution space of poses is difficult to be narrowed down, which causes that the poses cannot be recovered. One possible feasible solution is that, clustering all frames in the unstructured dataset to establish their co-visible relationships. After that, the poses of all frames can be recovered in a proper order. Looking ahead, we remain committed to dedicating our efforts in these aspects to both enhance the robustness of our work and also expand its applicability.

VII. Conclusion

In this paper, we studied a practical problem, radiance field construction without pose prior, and proposed a novel solution, namely I-DACS. I-DACS chooses to track frames and construct the radiance field simultaneously in an incremental manner, and the poses are absolutely estimated from the radiance field, achieving consistent localization results to the field. The radiance field in I-DACS is decomposed into two distinct components: the density field and the color field. The density field is modeled in the commonly employed hash-table representation, while the color field, which is usually much more time-consuming to train, is described by our proposed direction-aware color sampling strategy. Such a representation can effectively preserve fine-grained image details and always keep the color field consistent to the keyframe poses, guaranteeing the fast and stable convergence of our I-DACS. One eminent feature of our I-DACS is that, keeping SOTA rendering quality and localization accuracy, it achieves an amazing speed performance in the task of pose-free radiance construction, which is about $8\times \sim 80\times$ faster compared with other existing competitors. The experimental results corroborate the superior performance of I-DACS.

References

[1] L. Li, Y. Huang, J. Wu, K. Gu, and Y. Fang, "Predicting the quality of view synthesis with color-depth image fusion," *IEEE Trans. Circuits and Syst. for Video Technol.*, vol. 31, no. 7, pp. 2509-2521, 2021.

[2] X. Song, Y. Dai, and X. Qin, "Deeply supervised depth map super-resolution as novel view synthesis," *IEEE Trans. Circuits and Syst. for Video Technol.*, vol. 29, no. 8, pp. 2323-2336, 2019.

[3] A. I. Purica, E. G. Mora, B. Pesquet-Popescu, M. Cagnazzo, and B. Ionescu, "Multi-view plus depth video coding with temporal prediction view synthesis," *IEEE Trans. Circuits and Syst. for Video Technol.*, vol. 26, no. 2, pp. 360-374, 2016.

[4] N. Deng, Z. He, J. Ye, B. Duinkharjav, P. Chakravarthula, X. Yang, and Q. Sun, "FoV-NeRF: Foveated neural radiance fields for virtual reality" *IEEE Trans. Visualization and Comput. Graph.*, vol. 28, no. 11, pp. 3854-3864, 2022.

[5] K. Li, T. Rolff, S. Schmidt, R. Bacher, S. Frintrop, W. Leemans, and F. Steinicke, "Bringing instant neural graphics primitives to immersive virtual reality," in *Proc. IEEE Conf. Virtual Reality and 3D User Interfaces*, 2023, pp. 739-740.

[6] C. Li, S. Li, Y. Zhao, W. Zhu, and Y. Lin, "RT-NeRF: Real-time on-device neural radiance fields towards immersive AR/VR rendering," in *Proc. IEEE/ACM Int'l Conf. Computer-Aided Des.*, 2022, pp. 1-9.

[7] Y. Yuan, Y. Sun, Y. Lai, Y. Ma, R. Jia, and L. Gao, "NeRF-Editing: Geometry editing of neural radiance fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 18353-18364.

[8] Z. Kuang, F. Luan, S. Bi, Z. Shu, G. Wetzstein, and K. Sunkavalli, "PaletteNeRF: Palette-based appearance editing of neural radiance fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 20691-2070.

[9] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "NICE-SLAM: Neural implicit scalable encoding for SLAM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 12786-12796.

[10] H. Wang, J. Wang, and L. Agapito, "Co-SLAM: Joint coordinate and sparse parametric encodings for neural real-time SLAM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 13293-13302.

[11] M. M. Johari, C. Carta, and F. Fleuret, "ESLAM: Efficient dense SLAM system based on hybrid representation of signed distance fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 17408-17419.

[12] T. Schops, T. Sattler, and M. Pollefeys, "Bad SLAM: Bundle adjusted direct RGB-D slam," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 134-144.

[13] Y. Chang, Y. Tian, J. P. How, and L. Carlone, "D-NeRF: Neural radiance fields for dynamic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 11210-11218.

[14] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, and R. N. R. Ramamoorthi, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 405–421.

[15] R. Hartley and A. Zisserman, "Multiple view geometry in computer vision" *Cambridge University Press.*, 2003.

[16] J. L. Schönberger and J. Frahm, "Structure-from-Motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 4104–4113.

[17] J. L. Schönberger, E. Zheng, M. Pollefeys, and J. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 501-518.

[18] A. Meuleman, Y. Liu, C. Gao, J. Huang, C. Kim, M. H. Kim, and J. Kopf, "Progressively optimized local radiance fields for robust view synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 16539-16548.

[19] W. Bian, Z. Wang, K. Li, J. Bian, and V. A. Prisacariu, "NoPe-NeRF: Optimising neural radiance field with no pose prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 4160-4169.

[20] Y. Jeong, S. Ahn, C. Choy, A. Anandkumar, M. Cho, and J. Park, "Self-calibrating neural radiance fields," in *Proc. IEEE Int'l Conf. Comput. Vis.*, 2021, pp. 5846-5854.

[21] C. Lin, W. Ma, A. Torralba, and S. Lucey, "BARF: Bundle-adjusting neural radiance fields," in *Proc. IEEE Int'l Conf. Comput. Vis.*, 2021, pp. 5741-5751.

[22] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, "NeRF: Neural radiance fields without known camera parameters" *arXiv preprint, arXiv:2102.07064*, 2021.

[23] L. Liu, Z. Zhuang, S. Huang, X. Xiao, T. Xiang, C. Chen, J. Wang, M. Tan, "CPCM: Contextual point cloud modeling for weakly-supervised point cloud semantic segmentation," in *Proc. IEEE Int'l Conf. Comput. Vis.*, 2023, pp. 18413-18422.

[24] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "PointCLIP: Point cloud understanding by CLIP," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 8552-8562.

[25] I. Lang, A. Manor, and S. Avidan, "SampleNet: Differentiable point cloud sampling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 7578-7588.

[26] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 628–644.

[27] S. Choi, A. D. Nguyen, J. Kim, S. Ahn, and S. Lee, "Point cloud deformation for single image 3D reconstruction," in *Proc. IEEE Int'l Conf. Image Process.*, 2019, pp. 2379-2383.

[28] L. Wei, S. Wan, X. Ding, F. Yang, and Z. Wang, "Adaptive geometry reconstruction for geometry-based point cloud compression," in *Proc. IEEE Int'l Conf. Multimedia and Expo*, 2023, pp. 1985-1990.

[29] M. Schreiber, V. Belagiannis, C. Glaser, and K. Dietmayer, "Dynamic occupancy grid mapping with recurrent neural networks," in *Proc. IEEE Int'l Conf. Robot. and Automat.*, 2021, pp. 6717-6724.

[30] C. Buerkle, F. Oboril, J. Jarquin, and K. U. Scholl, "Efficient dynamic occupancy grid mapping using non-uniform cell representation," in *Proc. IEEE Intell. Vehicles Symp.*, 2020, pp. 1629-1634.

[31] Q. Li, B. Dai, and H. Fu, "LIDAR-based dynamic environment modeling and tracking using particles based occupancy grid," in *Proc. IEEE Int'l Conf. Mechatronics and Automat.*, 2016, pp. 238-243.

[32] A. I. Boyko, M. P. Matrosov, I. V. Oseledets, D. Tsetserukou, and G. Ferrer, "TT-TSDF: Memory-efficient TSDF with low-rank tensor train decomposition," in *Proc. IEEE/RSJ Int'l Conf. Intell. Robots and Syst.*, 2020, pp. 10116-10121.

[33] H. Kim and B. Lee, "Probabilistic TSDF fusion using Bayesian deep learning for dense 3D reconstruction with a single RGB camera," in *Proc. IEEE/RSJ Int'l Conf. Intell. Robots and Syst.*, 2020, pp. 8623-8629.

[34] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and Andrew Fitzgibbon "Kinectfusion: Real-time dense surface mapping and tracking," in *Proc. IEEE Int'l Symp. Mixed and Augemented Reality*, 2011, pp. 127-136.

[35] X. Zhou, Z. Wang, H. Ye, C. Xu, and F. Gao, "EGO-Planner: An ESDF-free gradient-based local planner for quadrotors," *IEEE Robot. and Automat. Lett.*, vol. 6, no. 2, pp. 478-485, 2021.

[36] D. Zhu, T. Zhou, J. Lin, Y. Fang and M. Q. H. Meng, "Online state-time trajectory planning using timed-ESDF in highly dynamic environments," in *Proc. IEEE Int'l Conf. Robot. and Automat.*, 2022, pp. 3949-3955.

[37] S. Geng, Q. Wang, L. Xie, C. Xu, Y. Cao, and F. Gao, "Robo-Centric ESDF: A fast and accurate whole-body collision evaluation tool for any-shape robotic planning," in *Proc. IEEE/RSJ Int'l Conf. Intell. Robots and Syst.*, 2023, pp. 290-297.

[38] E. Vespa, N. Nikolov, M. Grimm, L. Nardi, P. H. J. Kelly, and S. Leutenegger, "Efficient octree-based volumetric SLAM supporting signed-distance and occupancy mapping," *IEEE Robot. and Automat. Lett.*, vol. 3, no. 2, pp. 1144-1151, 2021.

[39] E. Vespa, N. Funk, P. H. J. Kelly, and S. Leutenegger, "Adaptive-resolution octree-based volumetric SLAM," in *Proc. Int'l Conf. 3D Vis.*, 2019, pp. 654-662.

[40] M. Klingensmith, I. Dryanovski, S. Srinivasa, and J. Xiao, "Chisel: Real time large scale 3D reconstruction onboard a mobile device using spatially hashed signed distance fields," in *Proc. Robot. Sci. and Syst.*, 2015.

[41] T. Zhang, L. Zhang, Y. Chen, and Y. Zhou, "CVIDS: A collaborative localization and dense mapping framework for multi-agent based visual-inertial SLAM," *IEEE Trans. Image Process.*, vol. 31, pp. 6562-6576, 2022.

[42] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 165-174.

[43] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, "NeuralRecon: real-time coherent 3D reconstruction from monocular video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 15598-15607.

[44] H. Guo, S. Peng, H. Lin, Q. Wang, G. Zhang, H. Bao, and X. Zhou, "Neural 3D scene reconstruction with the manhattan-world assumption," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 5511-5520.

[45] A. Trevithick and B. Yang, "GRF: Learning a general radiance field for 3D representation and rendering," in *Proc. IEEE Int'l Conf. Comput. Vis.*, 2021, pp. 15182-15192.

[46] Q. Wang, Z. Wang, K. Genova, P. P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouse, "IBRNet: Learning multi-view image-based rendering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 4690-4699.

[47] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields," in *Proc. IEEE Int'l Conf. Comput. Vis.*, 2021, pp. 5855-5864.

[48] P. Wang, Y. Liu, Z. Chen, L. Liu, Z. Liu, T. Komura, C. Theobalt, and W. Wang, "F2-NeRF: Fast neural radiance field training with free camera trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 4150-4159.

[49] M. Kim, S. Seo, and B. Han, "InfoNeRF: Ray entropy minimization for few-shot neural volume rendering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 12912-12921.

[50] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. M. Sajjadi, A. Geiger, and N. Radwan, "RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 5480-5490.

[51] Z. Sheng, F. Liu, M. Liu, F. Zheng and L. Nie, "Open-set synthesis for free-viewpoint human body reenactment of novel poses," *IEEE Trans. Circuits and Syst. for Video Technol.*, early access.

[52] K. Deng, A. Liu, J. Zhu, D. Ramanan, "Depth-supervised NeRF: Fewer views and faster training for free," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 12882-12891.

[53] B. Roessle, J. T. Barron, B. Mildenhall, P. P. Srinivasan, M. Nießner, "Dense depth priors for neural radiance fields from sparse input views," in *Proc. IEEE Int'l Conf. Comput. Vis.*, 2022, pp. 12892-12901.

[54] S. Guo, Q. Wang, Y. Gao, R. Xie, L. Li, F. Zhu, and L. Song, "Depth-guided robust point cloud fusion NeRF for sparse input views," *IEEE Trans. Circuits and Syst. for Video Technol.*, early access.

[55] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, "MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction," in *Adv. Neural Inf. Process. Syst.*, 2022, pp. 25018-25032.

[56] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-NeRF: Neural radiance fields for dynamic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 10318-10327.

[57] S. Park, M. Son, S. Jang, Y. C. Ahn, J. Kim, and N. Kang, "Temporal interpolation is all you need for dynamic neural radiance fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 4212-4221.

[58] A. Cao and J. Johnson, "HexPlane: A fast representation for dynamic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 130-141.

[59] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 5501-5510.

[60] C. Sun, M. Sun, and H. Chen, "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 5459-5469.

[61] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "TensoRF: Tensorial radiance fields," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 333-350.

[62] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding" *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1-15, 2022.

[63] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelNeRF: Neural radiance fields from one or few images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 4578-4587.

[64] M. Farenzena, A. Fusiello, and R. Gherardi, "Structure-and-motion pipeline on a hierarchical cluster tree," in *Proc. IEEE Int'l Conf. Comput. Vis.*, 2009, pp. 1489-1496.

[65] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1851-1858.

[66] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834-849.
[67] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 40, no. 3, pp. 611-625, 2018.
[68] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147-1163, 2015.
[69] R. Mur-Artal, and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255-1262, 2017.
[70] W. C. Hoffman, "The Lie algebra of visual perception" *J. Math. Psychol.*, vol. 3, no. 1, pp. 65-98, 1966.
[71] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 402-419.
[72] Q. Meng, A. Chen, H. Luo, M. Wu, H. Su, L. Xu, X. He, J. Yu, "GNeRF: GAN-based neural radiance field without posed camera," in *Proc. IEEE Int'l Conf. Comput. Vis.*, 2021, pp. 6351-6361.
[73] Z. Zhu, S. Peng, V. Larsson, *et al.*, "NICE-SLAM: Neural implicit scalable encoding for SLAM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 12776-12786.
[74] Z. Zhu, S. Peng, V. Larsson, *et al.*, "NICER-SLAM: Neural implicit scene encoding for RGB SLAM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 42-52.
[75] E. Sandström, Y. Li, L. Van Gool and M. R. Oswald, "Point-SLAM: Dense neural point cloud-based SLAM," in *Proc. IEEE Int'l Conf. Comput. Vis.*, 2023, pp. 18387-18398.
[76] H. Li, X. Gu, W. Yuan, L. Yang, Z. Dong, and P. Tan, "Dense RGB SLAM with neural implicit maps," in *Int'l Conf. Learn. Representations*, 2023.
[77] Y. Lin, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T. Lin, "iNeRF: Inverting neural radiance fields for pose estimation," in *Proc. IEEE Int'l Conf. Intell. Robots and Syst.*, 2021, pp. 1323-1330.
[78] D. Maggio, M. Abate, J. Shi, C. Mario, and L. Carlone, "Loc-NeRF: Monte carlo localization using neural radiance fields," in *Proc. IEEE Int'l Conf. Robot. and Automat.*, 2023, pp. 4018-4025.
[79] Y. Lin, T. Müller, J. Tremblay, B. Wen, A. Evans, P. A. Vela, and S. Birchfield, "Parallel inversion of neural radiance fields for robust pose estimation," in *Proc. IEEE Int'l Conf. Robot. and Automat.*, 2023, pp. 9377-9384.
[80] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE Int'l Conf. Comput. Vis.*, 2021, pp. 12179-12188.
[81] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2432-2443.
[82] A. Knapitsch, J. Park, Q. Zhou, and V. Koltun, "Tanks and Temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1-13, 2017.
[83] F. Darmon, B. Bascle, J. C. Devaux, P. Monasse, and M. Aubry, "Improving neural implicit surfaces geometry with patch warping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 6250-6259.
[84] Q. Yan, Q. Wang, K. Zhao, J. Chen, B. Li, X. Chu, and F. Deng, "CF-NeRF: Camera parameter free neural radiance fields with incremental learning," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 6440-6448.
[85] J. Straub, T. Whelan, L. Ma, *et al.*, "The Replica dataset: A digital Replica of indoor spaces," *arXiv preprint, arXiv:1906.05797*, 2019.

Tianjun Zhang received his B.Sc. degree from the School of Software Engineering, Tongji University, Shanghai, China, in 2019. He is now pursuing his Ph.D. degree at the same department of Tongji University. His research interests include collaborative SLAM, computer vision, and sensor calibration.

Lin Zhang (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2003 and 2006, respectively. He received the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2011. From March 2011 to August 2011, he was a Research Associate with the Department of Computing, The Hong Kong Polytechnic University. In Aug. 2011, he joined the School of Software Engineering, Tongji University, Shanghai, China, where he is currently a Full Professor. His current research interests include environment perception of intelligent vehicle, pattern recognition, computer vision, and perceptual image/video quality assessment. He serves as an Associate Editor for IEEE Robotics and Automation Letters, and Journal of Visual Communication and Image Representation. He was awarded as a Young Scholar of Changjiang Scholars Program, Ministry of Education, China.

Fengyi Zhang received his B.Sc. degree from the School of Software Engineering, Shandong University, Jinan, China, in 2021. He is now pursuing his M.Sc. degree at the department of Tongji University. His research interests include image enhancement, neural network compression, 3D reconstruction, and machine learning.

Shengjie Zhao (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1988, the M.Sc. degree in electrical and computer engineering from the China Aerospace Institute, Beijing, China, in 1991, and the Ph.D. degree in electrical and computer engineering from Texas AM University, College Station, TX, USA, in 2004. He is currently the Dean of the College of Software Engineering and a Professor with the College of Software Engineering and the College of Electronics and Information Engineering, Tongji University, Shanghai, China. In previous postings, he conducted research at Lucent Technologies, Whippany, NJ, USA, and the China Aerospace Science and Industry Corporation, Beijing. He is a fellow of the Thousand Talents Program of China and an Academician of the International Eurasian Academy of Sciences. His research interests include artificial intelligence, big data, wireless communications, image processing, and signal processing.

Yicong Zhou (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from Hunan University, Changsha, China, and the M.Sc. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA. He is currently a Full Professor and the Director of the Vision and Image Processing Laboratory, Department of Computer and Information Science, University of Macau, Macau, China. His research interests include chaotic systems, multimedia security, computer vision, and machine learning. Dr. Zhou was a recipient of the Third Price of Macau Natural Science Award in 2014. He serves as an Associate Editor for Neurocomputing, Journal of Visual Communication and Image Representation, and Signal Processing: Image Communication. He is a Co-Chair of the Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society. He is a Senior Member of the International Society for Optical Engineering (SPIE).