

TivNe-SLAM: Dynamic Mapping and Tracking via Time-Varying Neural Radiance Fields

Chengyao Duan¹ and Zhiliu Yang^{1, 2 *}

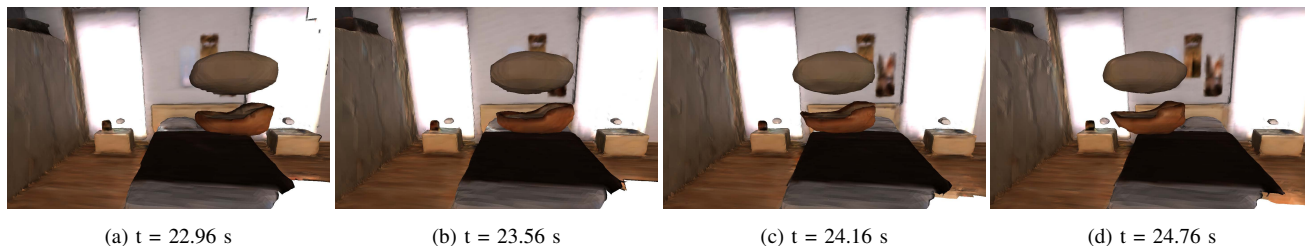


Fig. 1: Illustration of the reconstruction of a dynamic object from our TivNe-SLAM. We introduce a dynamic SLAM system capable of camera tracking and reconstruction of moving objects. Here, we show our 3D reconstruction result, at different time stamps, of a scene from the Room4 dataset [1], a flying ship is moving from the right to the left side of the room. The status of the flying ship is successfully captured and precisely reconstructed.

Abstract—Previous attempts to integrate Neural Radiance Fields (NeRF) into the Simultaneous Localization and Mapping (SLAM) framework either rely on the assumption of static scenes or require the ground truth camera poses, which impedes their application in real-world scenarios. This paper proposes a time-varying representation to track and reconstruct the dynamic scenes. Firstly, two processes, a tracking process and a mapping process, are maintained simultaneously in our framework. In the tracking process, all input images are uniformly sampled and then progressively trained in a self-supervised paradigm. In the mapping process, we leverage motion masks to distinguish dynamic objects from the static background, and sample more pixels from dynamic areas. Secondly, the parameter optimization for both processes is comprised of two stages: the first stage associates time with 3D positions to convert the deformation field to the canonical field. The second stage associates time with the embeddings of the canonical field to obtain colors and a Signed Distance Function (SDF). Lastly, we propose a novel keyframe selection strategy based on the overlapping rate. Our approach is evaluated on two synthetic datasets and one real-world dataset, and the experiments validate that our method achieves competitive results in both tracking and mapping when compared to existing state-of-the-art NeRF-based dynamic SLAM systems.

I. INTRODUCTION

Reconstructing an accurate dense map is crucial for tasks such as autonomous vehicle navigation, robot operation, and virtual reality. Existing dense visual Simultaneous Localization and Mapping (SLAM) frameworks are able to track camera poses and reconstruct complete indoor scenes. However, these methods have always struggled with feature

This work was supported by the Program of Yunnan Key Laboratory of Intelligent Systems and Computing under Grant 202405AV340009, and Yunnan Science Foundation of Yunnan Provincial Department of Science and Technology under Grant 202301AU070200.

¹ School of Information Science and Engineering, Yunnan University, Kunming, Yunnan 650500, China.

² Yunnan Key Laboratory of Intelligent Systems and Computing, Yunnan University, Kunming, Yunnan 650500, China.

* Corresponding author, zhiliu.yang@ynu.edu.cn

extraction and data association, which cause a serious perceptual aliasing problem [2]. Recently, SLAMs leverage Neural Radiance Fields (NeRFs) [3] to operate directly on raw pixel values without designing hand-crafted feature extraction, which omits the above difficulties of traditional methods.

NeRF has recently attracted a lot of research interest, which can obtain more accurate color and precise details by enrolling Multilayer Perceptrons (MLPs). Recently, a massive number of works adapt NeRF to SLAM domain, e.g. iMAP [4] applies a neural implicit representation to traditional dense reconstruction. NICE-SLAM [5] adopts a hierarchical and grid-based neural implicit encoding. However, it is required to define the size of reconstruction scenes in advance and to provide a pre-trained CNN model. VoxFusion [6] exploits octree-based representation and achieves scalable implicit scene reconstruction. The aforementioned methods can reconstruct high-quality maps. However, all these works assume the scenarios are static or deem dynamic objects as outliers. Thus, they are not able to reconstruct the dynamic objects. Naturally, NeRF-based methods are also extended to dynamic scenes marked by the algorithm named D-NeRF [7], then HyperNeRF [8] addresses the topological deformation problem in dynamic scenes. However, these works only focus on the mapping side, and directly utilize ground truth camera poses provided by the dataset.

To address these limitations, we propose a novel time-varying implicit representation to track camera poses and reconstruct moving objects in the dynamic scenes, which is named TivNe-SLAM. The inputs of our system are RGB-D image sequences with timestamps. Inspired by D-NeRF [7], our work extends 3D positions of objects to 4D positions by enrolling time information. Then, we transform points of dynamic objects from the deformation field to the canonical field. Next, colors and Signed Distance Function (SDF) of dynamic scenes are regressed by an MLP. The entire framework simultaneously maintains two processes, including a

tracking process and a mapping process, and executes the two processes in turn. To summarize, the contributions of our paper are as follows:

- We propose a novel 4D SLAM framework to simultaneously reconstruct the dynamic scenes and estimate the camera poses via Neural Radiance Fields (NeRFs).
- We introduce a time-varying representation to capture the position offsets of objects' movement, which enables our framework to eliminate holes and ghost trails that reside in traditional dynamic SLAM frameworks.
- We validate a novel overlap-based keyframe selection strategy to reconstruct dynamic objects more completely.

The rest of paper is organized as follows: An overview of related work is discussed in Section II. We provide a detailed explanation of our method in Section III. In Section IV, we demonstrate our experimental results including camera tracking, reconstruction quality, object completion ability, and ablation study. In the end, we summarize the experimental results in Section V.

II. RELATED WORK

Traditional Dense SLAM and Dynamic SLAM: Classic dense SLAM systems have developed rapidly in the past decades. KinectFusion [9] introduces a dense reconstruction and tracking system based on Truncated Signed Distance Function (TSDF) with an RGB-D camera. ElasticFusion [10] is a surface-based SLAM system and proposes two states, active and inactive, to control the activation state of voxels.

However, the above systems solely focus on static scenes, which are impractical in real-world applications. To this end, Co-Fusion [1] maintains a background model and multiple dynamic foregrounds, detecting moving objects through motion filtering and semantic segmentation. MaskFusion [11] utilizes Mask R-CNN [12] to achieve more precise segmentation of dynamic objects. MID-Fusion [13] introduces the Volume TSDF for dense mapping and tracking, but it involves at least four rounds of ray-castings, leading to slow processing. EM-Fusion [14] proposes a tracking method based on a probabilistic expectation maximization (EM) formulation for reconstructing dynamic objects. TSDF++ [15] advocates using a single large reconstruction volume to store the entire dynamic scene. RigidFusion [16] leverages motion priors and treats all dynamic objects as one rigid body. ACEFusion [17] adopts a hybrid representation including octrees and surfels. However, all these methods are prone to generating ghost trail effects.

NeRF-based Dynamic Scene Reconstruction: NeRF is first introduced for static scenes and has recently been extended to dynamic reconstruction. D-NeRF [7] associates time with 3D position and establishes a transformation between the deformation field and the canonical field. Nerfies [18] and HyperNeRF [8] associate a latent deformation code and an appearance code to each image, realizing a similar transformation to D-NeRF. NSFF [19] is proposed for handling varieties of in-the-wild scenes, including thin structures, view-dependent effects, and complex degrees of

motion. D²NeRF [20] creatively decouples dynamic objects and a static background for the monocular video. NeRF-Player [21] proposes a decomposition of dynamic scenes based on their temporal characteristics. Park et al. [22] utilizes an additional time parameter to execute temporal feature interpolation. PlenOctrees [23] presents a novel octree-based 3D representation to achieve real-time reconstruction, while Fourier-PlenOctrees [24] extends it to dynamic scenes. TiNeuVox [25] represents dynamic scenes with optimizable explicit data structures. FFDNeRF [26] leverages a forward flow field to better represent object motions. TensorRF [27] decomposes the 4D scene tensor into multiple low-rank tensors, and HexPlane [28] decompose 4D volumes in the same way, and represents dynamic scenes as a series of planes. However, all these methods train the models using input images with ground truth camera poses, which impedes their applications in real-world tasks. The method most closely resembling that in our paper is RoDynRF [29], which proposes a space-time synthesis algorithm from a dynamic monocular video and obtains accurate camera poses among high-speed moving objects, but it requires hours of training time.

NeRF-based Static SLAM: The power of NeRF in synthesizing photo-realistic novel views relies on accurate camera poses. Thus, some researchers integrate NeRF into SLAM. iNeRF [30] is the first work to obtain camera poses by leveraging a carefully trained NeRF. BARF [31] further improves iNeRF, by proposing a method to simultaneously train a NeRF and estimate the camera poses. iMAP [4] is the first dense real-time SLAM system based-on NeRFs and is able to estimate camera localization. DROID-SLAM [32] is only trained with monocular inputs and directly applied to stereo or RGB-D inputs, obtaining improved accuracy without retraining. NICE-SLAM [5] adopts a local-update strategy and proposes a hierarchical and grid-based neural implicit encoding, but it requires a pre-trained model. Vox-Fusion [6] adopts sparse voxel octree, and combines voxel embedding to mitigate artifacts of voxel borders. NICER-SLAM [33] is the RGB-only dense SLAM system and it proposes a local adaptive transformation for Signed Distance Functions (SDFs). Co-SLAM [34] designs a joint coordinate and sparse grid encoding. GO-SLAM [35] proposes a real-time global pose optimization system that considers the complete history information of input frames and incrementally aligns all poses. Despite the above algorithms achieving decent results for camera localization and scene reconstruction, they all rely on the assumption of a static scene.

III. METHODOLOGY

Given a stream of continuous RGB-D frames with color images, depths and corresponding timestamps, our TivNe-SLAM simultaneously maintains two processes, a tracking process and a mapping process, the overview is shown in Fig. 2. For each frame at time t , N pixels are sampled and cast as N rays, and are indexed by j . An unbalanced sampling strategy is exploited here, in which we leverage motion masks to distinguish dynamic objects from static

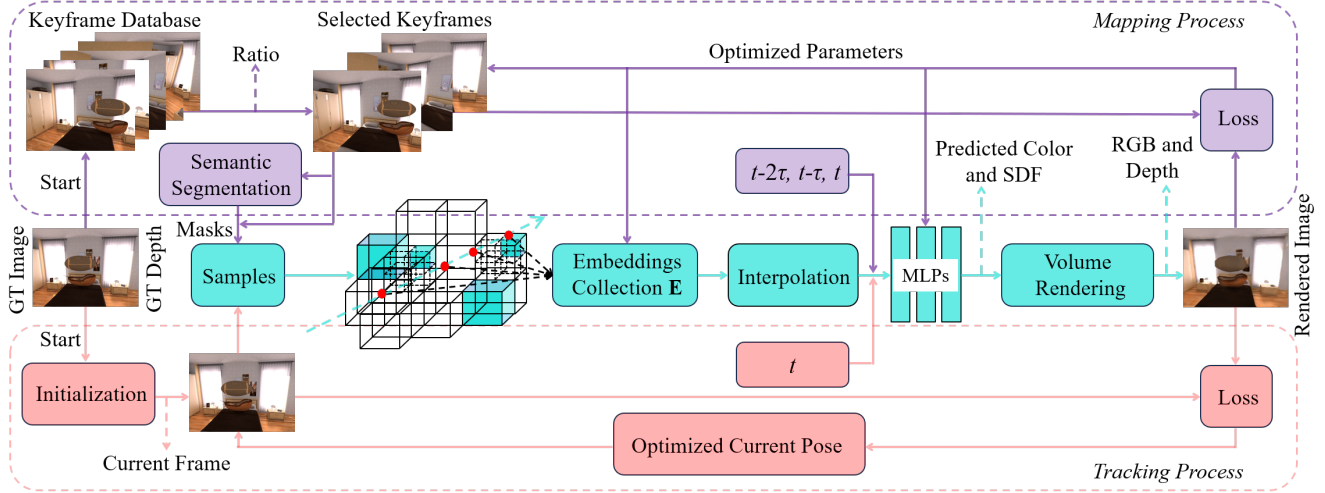


Fig. 2: Overview of our TivNe-SLAM framework. Our system simultaneously maintains two processes, a tracking process and a mapping process. **Tracking Process:** It firstly initializes a map utilizing the 1st frame and initializes the camera pose of each frame. Valid points of the current frame are sampled, encoded by a set of embeddings collection, and tri-linearly interpolated. The interpolated results are fed into two MLPs, and colors and SDF are predicted to render RGB images and depth images. Tracking loss is correspondingly constructed, and mapping parameters are frozen for tracking process, only the pose of the current frame is optimized. **Mapping Process:** After obtaining the camera pose of the current frame by the tracking process, we design a strategy to select target keyframes from an incrementally-growing database for reconstruction. Then it leverages Mask R-CNN [12] to obtain the mask segmentation of dynamic objects. As with the tracking process, points sampling, embedding, interpolation, and MLP regression are executed to obtain colors and SDF, and they are used to reconstruct the meshes. The poses, embedding parameters, and MLPs are optimized in the mapping process.

background and sample more pixels from dynamic areas. For each ray, M points are further sampled based on density and denoted as $\mathbf{x}_j^i = (x, y, z), i \in \{1, \dots, M\}, j \in \{1, \dots, N\}$, where i is the index for sampled 3D points along the j_{th} ray.

A. Deformation Field and Canonical Field

The detailed architecture of our neural fields transformation is further shown in Fig. 3. We define the scene at $t = 0$ as the canonical field and the scene at the other time as the deformation field. Inspired by D-NeRF [7], we leverage a deformation neural network Θ_d to transform positions from the deformation field to the canonical field. Specifically, Θ_d is trained to regress the offsets which represent moving distance of every point between the deformation field and the canonical field. To be noticed, the offset is zero only if $t = 0$. $\mathbf{e}_j^i(t)$ is tri-linearly interpolated result from embeddings collection \mathbf{E} and concatenated as $(\mathbf{e}_j^i(t), t)$ by adding a timestamp. The function of the deformation field can be expressed as Equation (1):

$$\Theta_d(\mathbf{e}_j^i(t), t) = \begin{cases} \Delta \mathbf{e}_j^i(t) & \text{if } t \neq 0, \\ 0 & \text{if } t = 0. \end{cases} \quad (1)$$

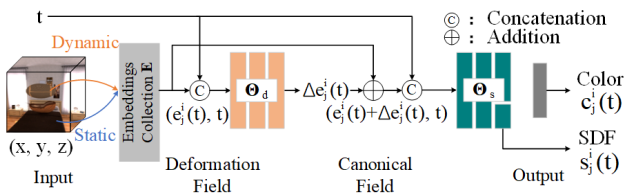


Fig. 3: Architecture of our neural deformation field and canonical field. We maintain a sparse embedding collection $\mathbf{E} \in \mathbb{R}^{H \times Q}$, which is a collection of Q -Dimensional vectors. We then encode positions $\mathbf{x}_j^i \in \mathbb{R}^3$ as $\mathbf{e}_j^i(t) \in \mathbb{R}^Q$. $\mathbf{e}_j^i(t)$ is associated with time t to regress offsets $\Delta \mathbf{e}_j^i(t)$ by Θ_d . Colors and SDFs are obtained by feeding $(\mathbf{e}_j^i(t) + \Delta \mathbf{e}_j^i(t), t)$ to Θ_s .

Accordingly, we employ another MLP, Θ_s , to regress colors and SDFs. Considering SDF values, instead of volume densities, are required in our task, the volume rendering formula utilized by the original NeRF [3] needs to be improved. Thus, we enroll a volume rendering technique via combining the network architecture in Azinović et al. [36] and Vox-Fusion [6].

In order to represent dynamic scenes, we further modify the formulas as follows:

$$\mathbf{e}_j^i(t) = \text{TriLerp}(\hat{\xi}(t)T(t-\tau)\mathbf{u}(t), \mathbf{E}). \quad (2)$$

$$\mathbf{c}_j^i(t), s_j^i(t) = \Theta_s(\Theta_d(\mathbf{e}_j^i(t), t) + \mathbf{e}_j^i(t), t). \quad (3)$$

where $\text{TriLerp}(\cdot, \cdot)$ denotes the tri-linear interpolation function which yields interpolated embeddings $\mathbf{e}_j^i(t)$. $\mathbf{u}(t) = \{(u, v)\}$ denotes pixels of current frame t . $T(t-\tau) \in SE(3)$ is the pose of the previous frame at time $t-\tau$, and $\hat{\xi}(t) \in SE(3)$ denotes the relative pose of the current frame and the previous frame. By following a zero-motion model, we always initialize the pose of the current frame with the pose of the last frame. Additionally, $\mathbf{c}_j^i(t)$ and $s_j^i(t)$ are the colors and SDFs of these points along each camera ray respectively.

$$\omega_j^i(t) = \sigma\left(\frac{s_j^i(t)}{tr}\right) \cdot \sigma\left(-\frac{s_j^i(t)}{tr}\right). \quad (4)$$

Where tr is a predefined parameter which presents truncated signed distance, $\sigma(\cdot)$ is the sigmoid activation function. Weight $\omega_j^i(t)$ gets the maximum value when $\sigma\left(\frac{s_j^i(t)}{tr}\right) = \sigma\left(-\frac{s_j^i(t)}{tr}\right)$, which means the point is around the object surface.

Lastly, $\mathbf{c}_j^i(t)$ and $d_j^i(t)$ stand for the colors and z-axis value of the i_{th} point along the j_{th} ray. The rendered color and rendered depth are calculated through the weighted

summation, as shown in $C_j(t)$ and $D_j(t)$:

$$C_j(t) = \frac{1}{\sum_{i=0}^{M-1} \omega_j^i(t)} \sum_{i=0}^{M-1} \omega_j^i(t) \cdot c_j^i(t). \quad (5)$$

$$D_j(t) = \frac{1}{\sum_{i=0}^{M-1} \omega_j^i(t)} \sum_{i=0}^{M-1} \omega_j^i(t) \cdot d_j^i(t). \quad (6)$$

B. Loss Function

Within each frame, N pixels are sampled and are further divided into B batches. Each batch is denoted as b . Each pixel is corresponding to a ray-casting. We redesign the loss function used in Azinović et al [36] and Vox-Fusion [6] as following:

$$\begin{aligned} \mathcal{L}^b(t) = & \sum_{b=0}^{B-1} (\lambda_{color} \mathcal{L}_{color}^b(t) + \lambda_{depth} \mathcal{L}_{depth}^b(t) \\ & + \lambda_{space} \mathcal{L}_{space}^b(t) + \lambda_{SDF} \mathcal{L}_{SDF}^b(t) \\ & + \lambda_{offset} \mathcal{L}_{offset}^b(t)). \end{aligned} \quad (7)$$

where $\mathcal{L}^b(t)$ is made up by five terms: color loss, depth loss, free-space loss, SDF loss, and offset loss. Finally, they are multiplied by their individual coefficients and are added together.

Specifically, color loss $\mathcal{L}_{color}^b(t)$ and depth loss $\mathcal{L}_{depth}^b(t)$ of sampling points are defined as:

$$\mathcal{L}_{color}^b(t) = \frac{1}{N} \sum_{j=0}^{N-1} \|C_j(t) - C_j^{gt}(t)\|. \quad (8)$$

$$\mathcal{L}_{depth}^b(t) = \frac{1}{N} \sum_{j=0}^{N-1} \|D_j(t) - D_j^{gt}(t)\|. \quad (9)$$

We predict color and depth at time t using Θ_s , and the loss between predicted images (RGB and depth) and the ground-truth images is constructed.

Then, free-space loss $\mathcal{L}_{space}^b(t)$ and SDF loss $\mathcal{L}_{SDF}^b(t)$ are defined as:

$$\mathcal{L}_{space}^b(t) = \beta_{space} \sum_{s \in S_{space}} (D_s(t) - tr)^2, \quad (10)$$

$$\text{where } \beta_{space} = \left(1 - \frac{P_{space}(t)}{P_{space}(t) + P_{tr}(t)}\right).$$

$$\mathcal{L}_{SDF}^b(t) = \beta_{SDF} \sum_{s \in S_{tr}} (D_s(t) - D_s^{gt}(t))^2, \quad (11)$$

$$\text{where } \beta_{SDF} = \left(1 - \frac{P_{tr}(t)}{P_{space}(t) + P_{tr}(t)}\right).$$

$P_{space}(t)$ and $P_{tr}(t)$ are the number of points in free space and near the surface of the object respectively. $s \in S_{space}$ indicate those points sampled in the free space (i.e. in front of the object surface) and $s \in S_{tr}$ are the points near to object surface. In $\mathcal{L}_{space}^b(t)$, $D_s(t)$, the depth of $s \in S_{space}$, is trained to be equal to truncated distance tr . In $\mathcal{L}_{SDF}^b(t)$, $D_s(t)$, the depth of $s \in S_{tr}$, is restricted within the truncation range and learns to be equal to ground truth value $D_s^{gt}(t)$.

Since the position of the background is supposed to remain static, the $\mathcal{L}_{offset}^b(t)$ is designed to minimize the offset of the background, as shown below:

$$\mathcal{L}_{offset}^b(t) = \sum_{s \in S_{bg}} \|\Delta e(t)\|^2. \quad (12)$$

$s \in S_{bg}$ are the sampled points from the background. $\Delta e(t)$ is the offset embeddings defined in Equation (1).

C. Camera Tracking

1) **Generation of Motion Mask:** In dynamic scenes, moving objects disturb the estimation of camera poses. Ours is able to estimate a precise camera pose and reconstruct the dynamic scenes by leveraging masks of moving objects. Our design is robust enough that it can work with a vanilla Instance Segmentation (IS) module as simple as Mask R-CNN [12].

2) **Estimation of Camera Pose:** In the tracking process, we freeze the parameters of embeddings collection \mathbf{E} and two MLPs (Θ_d and Θ_s), and only optimize the pose $T(t) = \hat{\xi}(t)T(t - \tau)$ of current frame. $\hat{\xi}(t) \in SE(3)$ is the relative pose from the last frame to the current frame which is defined in Equation (2). Specifically, we sample rays in the whole input image, then obtain the predicted colors and SDFs. In the end, we compute the loss described in Equation (7) between the predicted images and the ground-truth images.

D. Dynamic Mapping

1) **Keyframe Selection:** We maintain a keyframe database and insert new keyframes at a fixed interval. Vox-Fusion [6] randomly selects keyframes from the keyframe database to optimize the global map, which leads to incomplete reconstruction. We propose a novel strategy to select keyframes from the database by calculating the overlapping ratio between the current frame and each frame in the database. Specifically, we reproject the pixel points, $\mathbf{u}(t)$, sampled from the current frame t back into 3D space, and project them into the 2D plane of the target keyframe, as shown in Equation (13). Then we calculate whether $\mathbf{u}(t - q\tau)$ falls inside the pixel coordinate of the keyframe at $t - q\tau$, where q denotes the number of interval τ . We determine the overlapping ratio by computing the proportion of valid points to the total sampled points. Keyframes with a lower overlapping ratio are picked up for mapping optimization, causing the selected keyframes to cover as many novel views of scenes as possible.

$$\mathbf{u}(t - q\tau) = K \cdot T_{w2c}(t - q\tau) \cdot T_{c2w}(t) \cdot K^{-1} \cdot \mathbf{u}(t). \quad (13)$$

Where K represents the intrinsic matrix of the camera. T_{c2w} and T_{w2c} are the transformation matrix between the camera coordinate and the world coordinate.

2) **Optimization & Visualization:** Unlike the tracking process, the mapping process involves the overall optimization of parameters of embedding collections \mathbf{E} , MLPs (Θ_d and Θ_s), and the poses of target keyframes and the current frame. The loss function adopted for optimization is identical to Equation (7). After self-supervised training of each frame,

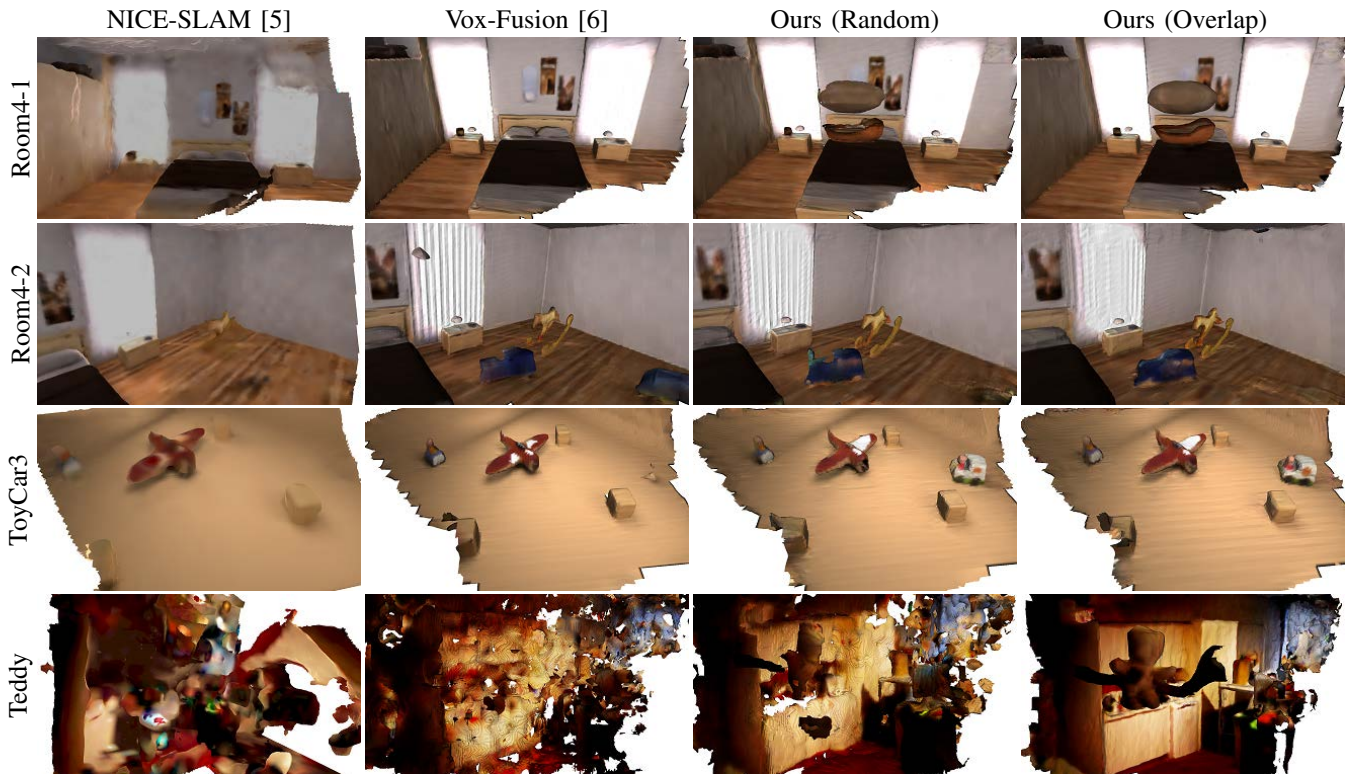


Fig. 4: Comparison of Mesh Quality among Two NeRF-based SLAMs and Two keyframe-selection Variations of Our TivNe-SLAM. Mapping results of different datasets are interpreted row by row. (1) **Room4-1:** We completely reconstruct the flying ship. However, NICE-SLAM [5] and Vox-Fusion [6] are unable to capture it. Randomly selecting keyframes generates unstable shapes and brings gaps into the reconstructed objects. Overlap-based TivNe-SLAM handles this problem well. (2) **Room4-2:** The results indicate that NICE-SLAM cannot reconstruct the blue car at all, and Vox-Fusion can only occasionally reconstruct the dynamic car, and fails to eliminate residual reconstruction in history positions. Similarly, our method with overlap-based strategy generates the best results. (3) **ToyCar3:** Only our method reconstructs the white car on the right side of the scene, but the others treat it as an outlier. Additionally, the reflected light on the plane’s wings is gradually shifted to the left wing in the input image, but our method still effectively captures this transitional process. (4) **Teddy:** It is clear that neither NICE-SLAM nor Vox-Fusion can reconstruct the scene of this real-world dynamic dataset. The method of exploiting randomly-selected keyframes yields poor results. However, our method based on overlapping selection fully reconstructs the dynamic teddy, arms of a person and the static background. (The brightness of images of the Teddy dataset is slightly adjusted for clearer visualization.)

we obtain an octree-based map up to the current frame, the newly trained MLP can regress the colors and SDF of every voxel in the scene. Then the mesh can be generated in real time using the marching cubes [37] algorithm.

IV. EXPERIMENTS

A. Experimental Setup

1) **Implementation Details:** We implement two tiny MLPs, Θ_d and Θ_s , in our experiment. Both of the MLPs consist of three hidden layers, and one Fully-Connected (FC) layer is attached for output of Θ_s . The SDFs are obtained from the third hidden layer of Θ_s and the color is obtained from the FC layer of Θ_s . The matrix size of embedding collection \mathbf{E} is configured as 20000×16 . Moreover, the tracking process of a single frame is configured to run for 30 iterations and the mapping process for 20 iterations. The camera pose of the first frame is initialized as an identity matrix. The size of input images for training are 640×480 . We down-sample the image to 320×240 for visualization of our rendering experiments. All experiments’ results are evaluated on an NVIDIA RTX 4090 GPU card with 24 GB memory.

2) **Datasets:** We evaluate our system on two different public synthetic datasets (i.e. Room4 and ToyCar3) and a

real-world dataset (i.e. Teddy) which are provided by Co-Fusion [1]. All three Datasets saved as RGB-D sequences with dynamic objects. Additionally, these datasets provide corresponding videos captured by a Kinect camera, from which we extract timestamps.

B. Evaluation of Camera Tracking

Absolute Trajectory Error (ATE) is adopted to evaluate the accuracy of camera tracking. We compare our system with other NeRF-based SLAM systems (i.e. NICE-SLAM and Vox-Fusion), dynamic NeRF (i.e. RoDynRF) and Structure-from-Motion technique (i.e. COLMAP [38]). As shown in TABLE I, best results are highlighted as **first** and **second**. We can clearly observe that our system predicts the most accurate camera pose in dynamic scenes compared to other methods on the Room4 dataset and the ToyCar3 dataset. That is to say, we successfully reconstruct the dynamic objects in the scenes, which is also beneficial for estimating more precise camera poses. Unfortunately, the real-world dataset lacks of ground truth camera trajectory, so that ATE of Teddy dataset is not evaluated.

C. Evaluation of Mapping Quality

1) **Qualitative Analysis:** Qualitative comparison of the reconstruction results between NICE-SLAM, Vox-Fusion,

TABLE I: Camera Tracking Results

Methods	ATE	Room4	ToyCar3
COLMAP [38]	RMSE [m](↓)	0.3448	0.1187
	Mean [m](↓)	0.2882	0.1039
	Std [m](↓)	0.1892	0.0574
NICE-SLAM [5]	RMSE [m](↓)	0.1305	0.0785
	Mean [m](↓)	0.0680	0.0617
	Std [m](↓)	0.1113	0.0486
Vox-Fusion [6]	RMSE [m](↓)	0.0164	0.0655
	Mean [m](↓)	0.0147	0.0543
	Std [m](↓)	0.0073	0.0366
RoDynRF [29]	RMSE [m](↓)	0.3903	0.2338
	Mean [m](↓)	0.3408	0.2051
	Std [m](↓)	0.1903	0.1123
Ours	RMSE [m](↓)	0.0131	0.0559
	Mean [m](↓)	0.0118	0.0444
	Std [m](↓)	0.0056	0.0339

RoDynRF and our system (TivNe-SLAM) is shown in Fig. 4. To be noticed, RoDynRF, Vox-Fusion and our TivNe-SLAM do not require pre-trained geometric models, while NICE-SLAM does. From the experimental results of all three datasets, obviously, we achieve better reconstruction quality for dynamic objects. Only our method successfully reconstructs all dynamic objects, including the flying ship, the rocking horse and the moving blue car of the Room4 dataset, the moving white car of the ToyCar3 dataset, and the moving teddy bear of the Teddy dataset. Especially in the Teddy dataset, other methods fail to reconstruct the scene when dynamic objects appear. NICE-SLAM is able to reconstruct scenes, but it still treats the dynamic objects as outliers and it tries to avoid reconstructing these dynamic objects as much as possible. Occasionally, Vox-Fusion only partially and inaccurately reconstructs dynamic objects and there are also some dynamic objects that are not reconstructed at all. We will further analyze the differences between the two different keyframe selection strategies in Section. IV-D.

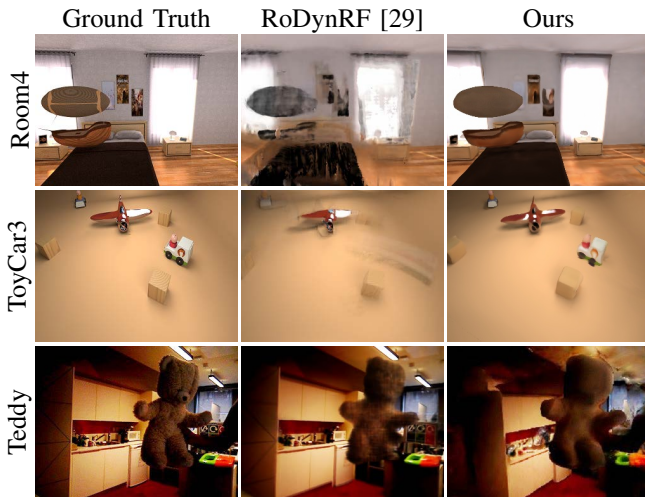


Fig. 5: Results Comparison between Rendered Images of Dynamic NeRF. Our TivNe-SLAM precisely renders images that closely resemble the input images on all three scenes. However RoDynRF fails to completely render dynamic objects in the first two scenes with fast-moving views. For the third scene, in which the camera moves more slowly, RoDynRF renders better result for both dynamic objects and the static background by adding the training time.

As shown in Fig. 5, we compare the images rendered by RoDynRF and our TivNe-SLAM. Our system achieves superior rendering results on synthetic datasets, because the camera moves rapidly in these scenarios, the views are switched at a fast speed. RoDynRF fails to accurately estimate camera poses and geometry because it is designed for scenes with slowly moving camera views. For the Teddy dataset, which is a real-world dataset, our system achieves competitive rendering results. Our method is able to reconstruct dynamic objects well, but it has difficulty rendering static backgrounds as well as the RoDynRF method. However, our method has an advantage in training time, the RoDynRF uses around 40 to 50 times as many training hours as ours. This is further reported and analyzed in Section IV-C.2.

Lastly, we also compare our mapping results with traditional SLAM methods which are capable of pose estimation. We demonstrate that these methods bring a series of reconstruction problems like ghost trails and empty holes for dynamic scenes. Results is shown in Fig. 6. This verifies the superiority of our method to enroll dynamic NeRF to a traditional SLAM.

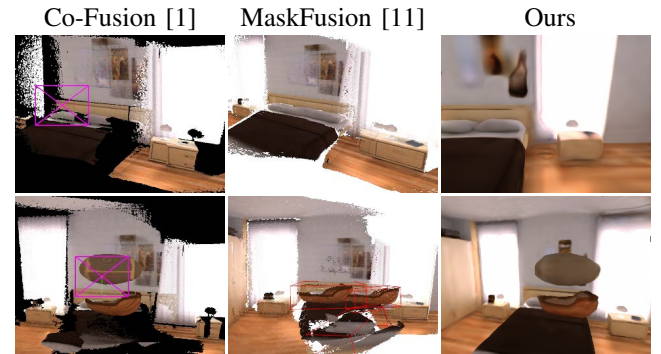


Fig. 6: Results Comparison with Traditional Dynamic SLAM Systems. We also compare our method with the Co-Fusion and MaskFusion, which both rely on a pre-trained CNN model. Both methods produce more artifacts (e.g., black bed and flying ship) and gaps (black and white regions) in the reconstructed scenes compared to our framework.

2) *Quantitative Analysis*: Quantitative results with four metrics, **MSE** (Mean Squared Error), **PSNR** (Peak Signal-to-Noise Ratio), **SSIM** (Structural Similarity Index) and **LPIPS** (Learned Perceptual Image Patch Similarity) [39] are shown in TABLE II with best results highlighted as **first** and **second**. It should be noted in advance that the scores of Vox-Fusion on the Teddy dataset are not reported, because scene reconstruction of Vox-Fusion crashes when dynamic objects appear in the scene.

As seen in TABLE II, our system achieves the best rendering results in comparison to the other three methods on two synthetic datasets. On the Teddy dataset, ours also obtains competitive results. The reason we cannot achieve the best results on real datasets is that we only used two light-weight MLPs to train the entire scene. When the scene is relatively large or complex, two tiny MLPs cannot completely regress the entire scene.

However, as mentioned in Section. IV-C.1, the decent rendering results of RoDynRF requires extensively time-

TABLE II: Novel View Synthesis Evaluation

Methods	Metrics	Room4	ToyCar3	Teddy
NICE-SLAM [5]	MSE(↓)	0.0114	0.0045	0.0148
	PSNR(↑)	22.1614	24.1507	18.4256
	SSIM(↑)	0.6755	0.8950	0.5483
	LPIPS(↓)	0.5983	0.2851	0.5420
Vox-Fusion [6]	MSE(↓)	0.0040	0.0033	N/A
	PSNR(↑)	25.5220	25.1758	N/A
	SSIM(↑)	0.7882	0.9158	N/A
	LPIPS(↓)	0.3969	0.1891	N/A
RoDynRF [29]	MSE(↓)	0.0095	0.0047	0.0012
	PSNR(↑)	21.0041	23.2258	29.5681
	SSIM(↑)	0.6146	0.8875	0.8673
	LPIPS(↓)	0.5352	0.2727	0.1764
Ours (overlap)	MSE(↓)	0.0020	0.0032	0.0069
	PSNR(↑)	27.2393	25.3456	22.6057
	SSIM(↑)	0.8103	0.9213	0.7099
	LPIPS(↓)	0.3711	0.1604	0.3802

consuming self-supervised training which prolongs the map building time. We report the approximated time required to train RoDynRF and TivNe-SLAM on all three datasets in TABLE III, and it is evident that our method trains much faster than RoDynRF on all datasets. Compared to RoDynRF, our system not only ensures a certain level of capture of dynamic scenes but also achieves real-time performance.

TABLE III: Comparison of Training & Mapping Time for Dynamic NeRF

Methods	Datasets	Training & Mapping Time (↓)
RoDynRF [29]	Room4	>11h
	ToyCar3	>10h
	Teddy	>12h
Ours	Room4	<0.3h
	ToyCar3	<0.2h
	Teddy	<0.3h

D. Object Completion

The keyframe selection strategy used in Vox-Fusion [6] is unstable, so its experimental results are different across multiple executions of experiments. Thus, we propose a novel keyframe selection strategy, calculating the overlap between current frame and the keyframes database. Then keyframes with a lower overlapping ratio are selected to reconstruct the dynamic scene. As expected, we achieve outstanding completeness in reconstructing dynamic objects.

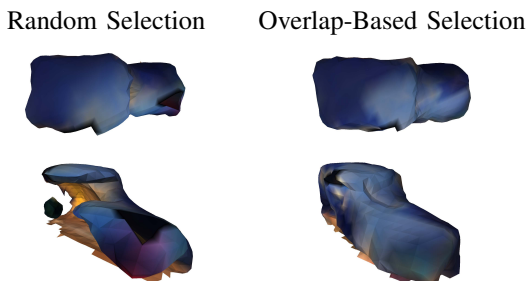


Fig. 7: Demonstration for Objects Completion. Random Keyframe selection strategy used in Vox-Fusion [6] is unable to reconstruct the 3D mesh that does not appear in the current view (Left column). Ours, based on overlap, however, fully reconstructs the blue car (Right column), even for some parts of the car that are not observed in the current camera view.

We visualize the reconstructed results of a blue car of two different keyframe-selection strategies in Fig. 7. It is evident

that random-selection strategy generates more gaps on the hood of the blue car, and the occluded area in the direction of the camera view is empty. However, the generative ability is out of the scope of this paper, as the part never observed by the camera is not completely filled by either strategy.

E. Ablation Study of Time and Mask

In this paper, we associate 3D positions with the time t and increase the number of sampling points within the area indicated by dynamic masks. To validate the effectiveness of two improvements, Fig. 8 and TABLE IV depict the results from ablation experiments. We conduct experiments where we separately omit masks and time t of dynamic objects, then compare their results with the proposed approach. It is evident that our method achieves the best results.

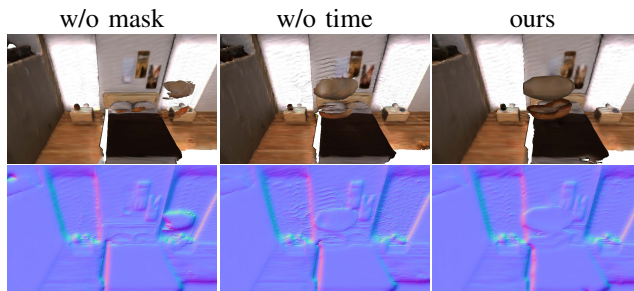


Fig. 8: Ablation Study for Time and Mask. The rendered images and their Norm images excluding the dynamic masks and the time t are reported. Rendering is executed with the 609_{th} frame of Dataset Room4. **w/o mask:** Reconstructed flying ship is incomplete and registered at incorrect positions. **w/o time:** Occasionally producing intermediate results, there are many holes on the ship and distorted lines on the wall.

TABLE IV: Quantitative Evaluation for Ablation Study

Datasets	Methods	MSE(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
Room4	w/o mask	0.0051	23.2877	0.7070	0.4384
	w/o time	0.0027	26.2984	0.7909	0.3948
	ours	0.0020	27.2393	0.8103	0.3711
ToyCar3	w/o mask	0.0033	24.9768	0.8868	0.1884
	w/o time	0.0033	24.9712	0.9173	0.1767
	ours	0.0032	25.3456	0.9213	0.1604

V. CONCLUSION

This paper introduces a novel dense SLAM system that leverages time-varying neural implicit representations to understand dynamic scenes. Our novelty lies in the introduction of a time-varying representation to extend 3D space positions to 4D space-temporal positions. Additionally, we introduce a more efficient keyframe selection strategy by calculating the overlapping ratio between the current frame and each frame in the keyframes database. Additionally, our keyframe selection strategy enables us to construct more complete dynamic objects. As opposed to existing dynamic SLAMs, our method does not rely on pre-trained models to reconstruct the dynamic objects.

Limits & Future Work: The trade-off between mapping quality of real-world scenes and complexity of DNNs, as well as camera pose estimation with high-speed moving objects are potential research directions for our future endeavors.

REFERENCES

- [1] M. Rünz and L. Agapito, "Co-fusion: Real-time segmentation, tracking and fusion of multiple objects," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4471–4478.
- [2] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.
- [3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [4] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "imap: Implicit mapping and positioning in real-time," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6229–6238.
- [5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 786–12 796.
- [6] X. Yang, H. Li, H. Zhai, Y. Ming, Y. Liu, and G. Zhang, "Vox-fusion: Dense tracking and mapping with voxel-based neural implicit representation," in *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2022, pp. 499–507.
- [7] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 318–10 327.
- [8] K. Park, U. Sinha, P. Hedman, J. T. Barron, S. Bouaziz, D. B. Goldman, R. Martin-Brualla, and S. M. Seitz, "Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields," *ACM Trans. Graph.*, vol. 40, no. 6, dec 2021.
- [9] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison *et al.*, "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 559–568.
- [10] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison, "Elasticfusion: Dense slam without a pose graph," in *Robotics: science and systems*, vol. 11. Rome, Italy, 2015, p. 3.
- [11] M. Runz, M. Buffier, and L. Agapito, "Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects," in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2018, pp. 10–20.
- [12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [13] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger, "Mid-fusion: Octree-based object-level multi-instance dynamic slam," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5231–5237.
- [14] M. Strecke and J. Stuckler, "Em-fusion: Dynamic object-level slam with probabilistic data association," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5865–5874.
- [15] M. Grinvald, F. Tombari, R. Siegwart, and J. Nieto, "Tsdf++: A multi-object formulation for dynamic object tracking and reconstruction," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 14 192–14 198.
- [16] Y.-S. Wong, C. Li, M. Niessner, and N. J. Mitra, "Rigidfusion: Rgb-d scene reconstruction with rigidly-moving objects," in *Computer Graphics Forum*, vol. 40, no. 2. Wiley Online Library, 2021, pp. 511–522.
- [17] M. Bujanca, B. Lennox, and M. Luján, "Acefusion-accelerated and energy-efficient semantic 3d reconstruction of dynamic scenes," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 11 063–11 070.
- [18] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5865–5874.
- [19] Z. Li, S. Niklaus, N. Snavely, and O. Wang, "Neural scene flow fields for space-time view synthesis of dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6498–6508.
- [20] T. Wu, F. Zhong, A. Tagliasacchi, F. Cole, and C. Oztireli, "D²nerf: Self-supervised decoupling of dynamic and static objects from a monocular video," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 653–32 666, 2022.
- [21] L. Song, A. Chen, Z. Li, Z. Chen, L. Chen, J. Yuan, Y. Xu, and A. Geiger, "Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2732–2742, 2023.
- [22] S. Park, M. Son, S. Jang, Y. C. Ahn, J.-Y. Kim, and N. Kang, "Temporal interpolation is all you need for dynamic neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4212–4221.
- [23] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, "Plenotrees for real-time rendering of neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5752–5761.
- [24] L. Wang, J. Zhang, X. Liu, F. Zhao, Y. Zhang, Y. Zhang, M. Wu, J. Yu, and L. Xu, "Fourier plenotrees for dynamic radiance field rendering in real-time," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 524–13 534.
- [25] J. Fang, T. Yi, X. Wang, L. Xie, X. Zhang, W. Liu, M. Nießner, and Q. Tian, "Fast dynamic radiance fields with time-aware neural voxels," in *SIGGRAPH Asia 2022 Conference Papers*, 2022, pp. 1–9.
- [26] X. Guo, J. Sun, Y. Dai, G. Chen, X. Ye, X. Tan, E. Ding, Y. Zhang, and J. Wang, "Forward flow for novel view synthesis of dynamic scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 022–16 033.
- [27] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "Tensorf: Tensorial radiance fields," in *European Conference on Computer Vision*. Springer, 2022, pp. 333–350.
- [28] A. Cao and J. Johnson, "Hexplane: A fast representation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 130–141.
- [29] Y.-L. Liu, C. Gao, A. Meuleman, H.-Y. Tseng, A. Saraf, C. Kim, Y.-Y. Chuang, J. Kopf, and J.-B. Huang, "Robust dynamic radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13–23.
- [30] L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T.-Y. Lin, "inertf: Inverting neural radiance fields for pose estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1323–1330.
- [31] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "Barf: Bundle-adjusting neural radiance fields," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5741–5751.
- [32] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," *Advances in neural information processing systems*, vol. 34, pp. 16 558–16 569, 2021.
- [33] Z. Zhu, S. Peng, V. Larsson, Z. Cui, M. R. Oswald, A. Geiger, and M. Pollefeys, "Nicer-slam: Neural implicit scene encoding for rgb slam," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 42–52.
- [34] H. Wang, J. Wang, and L. Agapito, "Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 293–13 302.
- [35] Y. Zhang, F. Tosi, S. Mattoccia, and M. Poggi, "Go-slam: Global optimization for consistent 3d instant reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3727–3737.
- [36] D. Azinović, R. Martin-Brualla, D. B. Goldman, M. Nießner, and J. Thies, "Neural rgb-d surface reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6290–6301.
- [37] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *Seminal graphics: pioneering efforts that shaped the field*, 1998, pp. 347–353.
- [38] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.
- [39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.