TPD-NeRF: Temporally Progressive Reconstruction of Dynamic Neural Radiance Fields from Monocular Video

Yu-Jie Yuan
Institute of Computing Technology, CAS
University of Chinese Academy of Sciences

Leif Kobbelt
RWTH Aachen University

Jie Yang
Institute of Computing Technology, CAS

Yu-Kun Lai
Cardiff University

Lin Gao
Institute of Computing Technology, CAS
University of Chinese Academy of Sciences
gaolin@ict.ac.cn

Abstract

Due to their great performance in representing 3D scene geometry and appearance, Neural Radiance Fields (NeRF) have recently gained a lot of attention in applications like novel view synthesis. Some extensions of NeRFs to dynamic scenes have been proposed, but they either require synchronized multi-view video input or fail for faster motions or longer sequences. In this paper, we propose a novel dynamic NeRF framework, called TPD-NeRF, which takes a single monocular video as input and enables high quality synthesis of novel views for any time point even in highly dynamic scenes. The idea is to first establish local frame-to-frame consistency by training a sub-network that predicts short term offsets and hence generates frame-to-frame correspondences. Applying this network multiple times allows us to propagate correspondences from any frame of the input sequence to one global reference frame. Using the resulting global correspondences as supervision, we can train another sub-network to establish global consistency for the TPD-NeRF. This network effectively maps each dynamic state back to a canonical space, *i.e.* it captures the global motion in the scene. To further improve the visual quality, we introduce the space-time field network as the canonical NeRF to capture missing dynamic information of the two deformation networks. We extensively evaluate our method and compare it with previous work to demonstrate that our method outperforms existing dynamic NeRF methods.

***Keywords:** Monocular Dynamic Reconstruction, Temporal Local-to-Global, Dynamic NeRF*

1. Introduction

Novel view synthesis is a major task in computer vision and computer graphics, which can be used in various applications, such as electronic commerce, virtual/augmented reality (VR/AR), etc. Given a set of input images, traditional image-based rendering (IBR) methods typically blend several input images near the query view to generate a novel view image. However, they require dense input views and can struggle to deal with occlusion cases. Recently, implicit-based methods [55, 9, 49], especially neural rendering methods [77], have received a lot of attention. The images are directly rendered or synthesized by the neural network. Among them, Neural Radiance Field (NeRF) [50] is one of the representative works. It incorporates the multilayer perception (MLP) network with the traditional volume rendering [31], which forms an overall differentiable training framework supervised by a sparse set of input images. Its fascinating quality of novel view synthesis has stimulated a series of follow-up works, extending it to handle re-lighting [105], pose estimation [41], multi-scene generalization [8] and training and inference speedup [99, 51]. However, most of these works, including the original NeRF itself, focuses on static scenes, while dynamic scenes are crucial components in our real world. Therefore, it is necessary to bring the synthesis capabilities of NeRF to the novel view synthesis of dynamic scenes, especially in the challenging setting where only a monocular video is available, as it can be easily captured compared with costly multi-camera setup.

When reconstructing a dynamic 3D scene with NeRF, the effectiveness of training depends on a proper regularization, which promotes the local and global consistency of the scene's geometry and appearance over time and space.

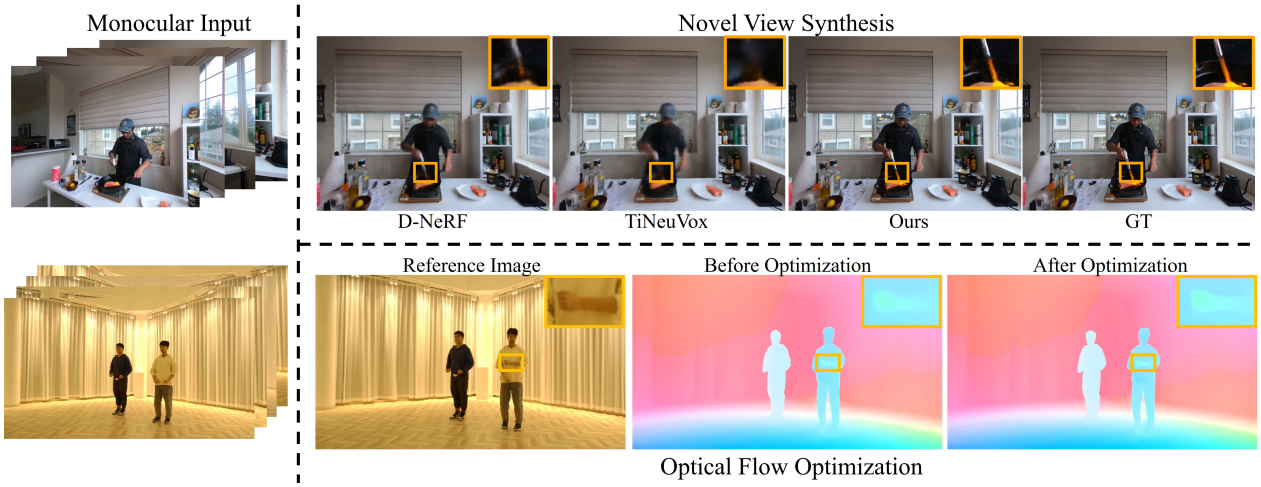


Figure 1. We propose a novel dynamic neural radiance field (NeRF) framework (TPD-NeRF), taking into account both local and global consistency of scene geometry and appearance. Our novel view synthesis results can better capture dynamic details, and the joint optimization strategy with the optical flow can in turn predict sharper and more accurate optical flow.

Here, *consistency* refers to the observation that the geometry and appearance in dynamic scenes are strongly correlated before and after motion. The term *local* refers to several subsequent frames and *global* refers to the whole dynamic sequence. It should be noted that the concepts of 'local' and 'global' are common in the spatial domain, we use these two terms in the temporal domain. To ensure consistency between different frames, it is necessary to establish correspondences between them. On establishing correspondence, existing dynamic NeRF works can be roughly divided into two categories, including space-time field-based methods and ray-bending-based methods. The approaches based on space-time fields extend the original static NeRF into a space-time field network that models space information and temporal motion simultaneously. The network takes the positionally encoded time or time-related vector as an additional input of NeRF. Additional supervision is then employed during training, such as depth [94] or optical flow [39]. Local consistency can be effectively promoted with the help of optical flow, but global consistency is difficult to formulate since there is no global reference. Moreover, they use a single MLP network to store all spatial and temporal information across frames, and may produce inaccurate detail reconstruction or have poor generalization ability when facing long-time image sequences. To address this, another approach based on ray bending leverages an additional MLP network to encode temporal motion information [59, 56]. This network, typically named as the deformation network or offset network, transforms the sampled points at different time stamps back to a canonical space (usually the first frame), and then a unified NeRF network is introduced to model density and color fields in the canonical space. The temporal and spatial information are modeled separately. The global mapping is learned by only color supervision [59] or the Jacobian matrix regularization for mapping fields [79, 56]. In this setting, global consistency is established automatically by mapping all frames into a common canonical space, but local consistency or inter-frame continuity is neglected as each frame is warped independently. On the contrary, our approach aims at including both local and global consistency in the training process by first learning local transformations between successive frames and then accumulating these local transformations in order to generate supervision for the training of the global network. Recall that in some NeRF-based SLAM method [106], the scene will be gradually reconstructed and ultimately optimized as a whole in the spatial domain. We transfer this local-to-global idea to the temporal domain.

In this paper, we propose a novel dynamic NeRF framework, named TPD-NeRF, which reconstructs dynamic scenes in a temporal local-to-global manner, helping to capture dynamic details, and achieve better novel view synthesis results. Our method first adopts the idea of ray bending, which maps the sampled point from the given frame to the canonical space. This mapping is equivalent to establishing correspondences between the given frame and the canonical space. For longer sequences and faster motions, finding appropriate correspondences gets increasingly hard and unreliable. For any two successive frames, however, these correspondences are relatively easy to establish via the optical flow since the short-term motion is usually small. Hence our key idea is to train a local deformation network to find frame-to-frame correspondences, and then generate fairly reliable frame-to-canonical space correspondences by concatenating sequences of frame-to-frame correspondences. A global network that predicts the frame-to-canonical correspondences is also trained with supervision from the local deformation network, which ensures global

consistency. This local-to-global framework is named as temporally progressive training (TPT). Further, we adopt the estimated optical flow between the adjacent frames to better supervise the training of the local deformation network. However, 2D estimation methods of optical flow typically rely on pixel differences, and their results may be inaccurate due to the lack of 3D perception, which may further affects the learning of our network. On the contrary, NeRF incorporates multi-view information into the training and reconstructs a 3D-aware model. Therefore, we propose a joint optimization learning strategy for the 3D scene flow and 2D optical flow. This joint optimization mechanism can help make local deformation learning more stable. Besides these, we observe that some dynamic details may still be lost by the deformation network even with our proposed temporally progressive training. So as an additional feature, we set up the canonical space itself as a space-time NeRF network such that it can take into account finer dynamic details in shape and appearance that are not captured by the deformation networks. Based on these, we propose a hybrid framework that incorporates the deformation network with space-time field network, which can boost each other and improve the performance. While a single space-time field network is unable to handle the long-time sequences, when most of the temporal information is encoded by the deformation network, the remaining dynamic details can well be encoded by it. It should be noted that different from the existing work [57, 15], both deformation networks in our method have supervisions which prevents possible degradation when combining the ray bending with the space-time field. In summary, our contributions are three-fold:

- We propose a temporally progressive training framework, named TPD-NeRF, which is able to better capture the deformation between each frame and the canonical frame, compared with previous direct global deformation learning.
- A joint optimization strategy of the 3D scene flow and 2D optical flow is introduced, which helps mutually improve the deformation learning with the supervision of optical flow in the early training stage and optimize the optical flow with the 3D-aware information of NeRF.
- A hybrid modeling strategy that combines a deformation network with a space-time field network both equipped with proper supervision to prevent possible degradation.

2. Related Work

Dynamic scene modeling is a fundamental research topic in 3D computer vision and graphics. In this section, we'll take a quick look at the new developments in dynamic NeRF and scene flow.

2.1. NeRF and General Dynamic NeRF

Neural Radiance Field (NeRF) [50] has achieved impressive results in novel view synthesis. It has been further extended for dynamic scenes [39, 94, 59, 79, 56, 57], better rendering effects [1, 82, 26], generalization on different scenes [65, 88, 8, 95, 104], faster training or inference speed [22, 99, 24, 62, 52, 7, 71, 16], re-lighting rendering [3, 70, 105, 91], geometry or appearance editing [44, 100, 27, 10, 30, 101], geometry reconstruction [87, 46, 74], 3D generation [21, 43], etc. For more comprehensive and detailed discussions and comparisons, we refer the readers to these surveys [11, 20, 78].

Our work focuses on general dynamic scenes, which are in contrast to specific dynamic objects, *e.g.*, dynamic human bodies [58, 54, 102] or faces [17, 61], where prior models such as SMPL [47] and 3DMM [2] are often used to help establish the correspondences between frames. General dynamic scenes lack such a prior.

As mentioned before, the space-time field-based methods directly take the positionally encoded time [39, 94] or learnable vectors [37] as one of the NeRF inputs, and use a single MLP network to encode spatial and temporal information simultaneously. These methods will additionally predict the per-point scene flow to adjacent frames, and utilize the estimated depth [94], optical flow [39] and cycle consistency loss [39, 13] as additional supervision. The per-point scene flow can be represented in the frequency domain via a discrete cosine transform [83], which can better conform to the inter-frame continuity. The static part and the dynamic part of the scene can be separately modeled by a static-dynamic mixed NeRF model [18]. Additional entropy loss function [93] is also introduced to promote more accurate static/dynamic segmentation. Those methods that use an additional MLP network to encode the temporal information and predict the offset [59, 79] or SE(3) transformation field [56] for each sampled point. Such transformations can be regularized by an an elastic energy constraint to constrain the Jacobian matrix of the transformation [56, 79]. Further, to handle topological changes, HyperNeRF [57] regards different topological states as hyperplanes of a high-dimensional space. NeRFPlayer [69] views a dynamic scene as the composition of three parts: dynamic, static and new, which are modeled by corresponding networks, followed by combining the output feature vectors with the predicted probabilities. [84] calculates the partial derivative of forward deformation to time and then integrates over a local interval to obtain local displacement, which is then supervised using the optical flow. The idea of this work is similar to ours, but we explicitly model local displacement and utilize a progressive training strategy to supervise the

global deformation network. Our method leverages the benefits of ray bending and the space-time field and proposes temporally progressive training to fully capture the dynamic details. Although TiNeuVox [15] also incorporates the positionally encoded time to the ray bending-based NeRF, its global offset network lacks necessary supervision to prevent possible degradation. Another type of dynamic NeRF aims to accelerate training and inference. They introduce an explicit voxel representation [36] or the combination of voxels and implicit networks [23, 66, 29, 85, 32] into dynamic modeling.

Recently, with the rise of 3D Gaussian Splatting (3DGS) [34, 92], dynamic modeling methods based on it have also been extensively proposed. Similar to dynamic NeRF methods, these works either utilize additional networks [96, 40, 67] or a feature representation [90] to encode dynamic information, or directly define dynamic properties on 3D Gaussian spheres [48, 97]. The explicit modeling nature of 3DGS allows it to use motion bases to fit dynamic information [73, 14, 33, 42, 38, 35]. Our temporally progressive modeling idea can be extended to the dynamic 3DGS.

2.2. Scene Flow and Optical Flow

Scene flow [81] is defined on the points in a dynamic scene and describes their motions. Many space-time field NeRFs [39, 18, 13] predict the frame-by-frame scene flow and use optical flow to supervise the training. While our method also predicts the frame-by-frame scene flow, different networks are used for prediction and modeling, and a global deformation network is introduced to ensure global-awareness. Optical flow can be obtained by projecting the scene flow onto the 2D image plane. It talks about the movement between the observer and the scene, including the movement of the scene and the camera. Optical flow is extensively used in action recognition [68] and video-related applications [86, 80, 6]. The traditional variational-based methods [25, 72, 4] estimate optical flow through energy optimization, whereas the neural-based methods [12, 28, 76] directly predict it from two input images via a neural network. There are also methods that estimate optical flow from point cloud sequences [45, 53] or scene flow from RGB-D images [60]. Based on the initial optical flow estimated by the 2D method [76], our method proposes to jointly optimize the scene flow and optical flow to supervise each other, which can help obtain better inter-frame correspondences.

3. Methodology

In this section, we will first briefly overview the preliminaries of our method, Neural Radiance Field (NeRF) [50] (Sec. 3.1). Then, we propose our novel dynamic NeRF network, TPD-NeRF, which reconstructs a dynamic scene from a monocular video in a temporal local-to-global fashion. We first introduce how to capture the local correspondences between adjacent frames which ensures local consistency of scene geometry and appearance (Sec. 3.2). This process will be supervised by the estimated optical flow and a joint optimization strategy is proposed to compensate for the estimation error. After the local training, we introduce a global network that ensures global temporal awareness and consistency of scene geometry and appearance. The local and global networks together develop a temporally progressive training mechanism, which is supervised step-by-step and aggregates information in the temporal domain to achieve temporally local-to-global learning (Sec. 3.3). Finally, to capture those missing dynamic information that cannot be modeled by ray bending, we combine the two deformation networks with the space-time field network (Sec. 3.4). Fig. 2 illustrates the whole pipeline of our method.

3.1. Preliminaries

Our method is based on the Neural Radiance Field (NeRF) [50], which represents scene geometry and appearance using a simple fully-connected network, given a set of input images. The multi-layer perceptron (MLP) network takes 3-dimensional spatial coordinates $\mathbf{p} = (x, y, z)$ and 2-dimensional viewing direction $\mathbf{d} = (\theta, \phi)$ as inputs and outputs the volume density σ and RGB values $\mathbf{c}$: $F(\Theta) : (\mathbf{p}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$, where Θ represents the trainable network weights. The camera intrinsics and extrinsics are assumed to be known or estimated by some methods, such as COLMAP [64, 63]. The camera rays are generated from the camera location to image pixels in the world coordinate system. Some points are sampled along the rays. The estimated color $\hat{C}(\mathbf{r})$ of each ray $\mathbf{r}(s)$ is calculated by the classical volume rendering method [31], and the continuous integral is approximated by quadrature:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)(1 - \exp(-\sigma_i \delta_i))\mathbf{c}_i, \quad (1)$$

where $\delta_i = s_{i+1} - s_i$ is the distance between adjacent samples on the ray, and N is the total number of sampled points on the ray. NeRF also adopts a stratified sampling strategy that samples uniformly in evenly-spaced bins. The network is trained under RGB supervision:

$$\mathcal{L}_{RGB} = \sum_{\mathbf{r} \in \mathcal{R}} \|\hat{C}(\mathbf{r}) - C(\mathbf{r})\|_2^2, \quad (2)$$

where $\mathcal{R}$ is a collection of rays sampled in a training batch, $\hat{C}(\mathbf{r})$ and $C(\mathbf{r})$ are the estimated color and the ground truth color of the pixel corresponding to the ray $\mathbf{r}$. To capture high-frequency details, NeRF also adopts positional encoding [75] on coordinates $\mathbf{p}$ and viewing direction $\mathbf{d}$.

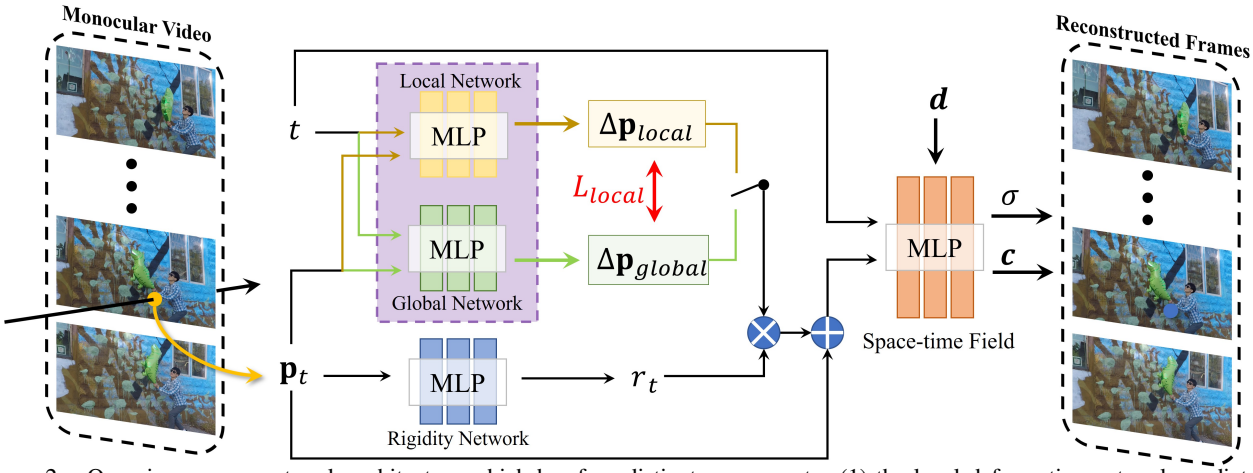


Figure 2. Overview on our network architecture which has four distinct components: (1) the local deformation network predicts the local offset between successive frames of the input video, (2) the global deformation network is trained with the supervision of the global correspondences between the frame t and the canonical frame that emerge from repeated application of the local deformation network (L_{local}). The network will switch between the cumulative results $\Delta\mathbf{p}_{local}$ of the local network and the outputs $\Delta\mathbf{p}_{glocal}$ of global network for transforming the sampled point $\mathbf{p}_t$, based on the training stage. This formulates our temporally progressive training. The mapping quality is enhanced by adding (3) a rigidity network that masks out the static part of the scene with the point-wise rigidity value r_t. The motion compensated position is passed on to (4) a space-time NeRF network that captures the fine details of the motion. $\mathbf{d}$ is the viewing direction, θ and $\mathbf{c}$ are the output density and color for volume rendering.

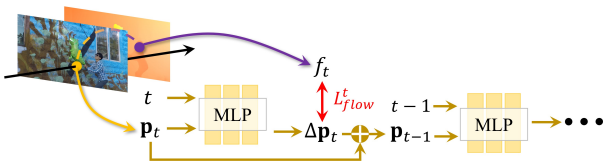


Figure 3. The specific process of our local deformation network. It predicts for each sampled point along a query ray its offset to the previous frame, conditioned on the current time. The sampled points are transformed back into the canonical space by querying the shared local network repeatedly. The accumulated offset $\Delta\mathbf{p}_{local}$ is obtained by accumulating $\Delta\mathbf{p}_t$ along the time dimension. The local offset $\Delta\mathbf{p}_t$ is supervised by L^t_{flow} with the optical flow under a joint optimization strategy.

3.2. Local Consistency by Local Correspondences

Our method is based on ray bending to model deformations. The existing ray bending-based NeRF methods directly predict point-wise correspondences between the observation frame and the canonical frame, which is hard to capture, especially for longer sequences and faster motions. Based on the observation that the motions between adjacent frames in a video sequence are much smaller than the motions between the observation frame and the canonical frame, we propose to first learn the local deformations between adjacent frames of dynamic sequences and then accumulate the local deformations to obtain the global deformations from the observation frame to the canonical frame.

We introduce an additional MLP network F_{local}, referred to as the local deformation network, to predict the local deformations of the sampled points from the current frame to the previous frame. Our local deformation network takes the current time step t and the coordinates $\mathbf{p}_t$ of the sampled point at time step t as inputs. Similar to the previous work [59], we use positional encoding [50] to embed $\mathbf{p}_t$ and t to the high-dimensional vectors for learning high-frequency details. The output is an offset vector $\Delta\mathbf{p}_t$ that is added to $\mathbf{p}$ and transforms the sampled point to the space of the previous frame. The transformed point $\Delta\mathbf{p}_{t-1}$ and the previous time step $t-1$ will be fed into F_{local} again to obtain the backward offset from $t-1$ to $t-2$. Note that, the local deformation networks for different frames are shared. When the time goes to 0, the sampled point is finally transformed into the canonical space. In summary, the sampled point $\mathbf{p}_t$ in time step t will be transformed to the position $\mathbf{p}_0$ in the canonical space step by step:

$$\begin{aligned}\mathbf{p}_{t-1} &= F_{local}(\mathbf{p}_t, t) + \mathbf{p}_t \\ \mathbf{p}_{t-2} &= F_{local}(\mathbf{p}_{t-1}, t-1) + \mathbf{p}_{t-1} \\ &\dots \\ \mathbf{p}_0 &= F_{local}(\mathbf{p}_1, 1) + \mathbf{p}_1\end{aligned} \tag{3}$$

Then, the transformed point $\mathbf{p}_0$ will be fed to the canonical NeRF together with the viewing direction $\mathbf{d}$ to obtain the volume density σ and color values $\mathbf{c}$. Finally, the pixel color corresponding to the ray is calculated by Eq. 1. The process of the local deformation network is illustrated in Fig. 3.

Although the local deformation between successive frames is easier to learn, the local deformation network still needs to transform the sampled points to the canonical space

step-by-step, which remains challenging when the training relies solely on RGB supervision (Eq. 2). Another advantage of local deformation network is that we can introduce optical flow to provide supervision. It should be noted that some space-time field-based methods [39] have adopted the optical flow as 2D supervision of the predicted forward and backward offsets. However, for the ray bending-based dynamic NeRF, since the predicted offset is from the current frame to the canonical frame, the optical flow between adjacent frames cannot be directly used for supervision. If one accumulates the optical flow, the estimation errors are also accumulated, which is not desirable. So on the one hand, we decompose the global deformation into local ones, which can directly be supervised by optical flow. On the other hand, a joint optimization of the local deformation network and optical flow is further introduced, which prevents the above issues.

Specifically, we adopt RAFT [76] to estimate the initial optical flow between every pair of adjacent frames. Denote the local offset obtained from time step t to time step $t-1$ of the sampled point $\mathbf{p}_{ti}$ as $\Delta\mathbf{p}_{ti}$, we can accumulate the overall offsets $\Delta\mathbf{p}_t(\mathbf{r})$ along the ray $\mathbf{r}$ in a similar way as Eq. 1 by replacing the color $\mathbf{c}_i$ with the offset $\Delta\mathbf{p}_{ti}$. We can further project the ray offset $\Delta\mathbf{p}_t(\mathbf{r})$ to the image plane of time step $t-1$ to obtain the translation $\Delta\mathbf{p}_t(u)$ of the corresponding pixel u by applying the camera projection matrix P_{t-1}. Then the backward translation $\Delta\mathbf{p}_t(u)$ of the pixel u can be supervised by the ground truth optical flow $f_t(u)$ under the following loss:

$$L^t_{flow} = \|\Delta\mathbf{p}_t(u) - f_t(u)\|^2_2. \tag{4}$$

The above loss function can be used to supervise the training of the local deformation network. However, the estimation of the optical flow only considers the color information between two adjacent frames, which may be inaccurate due to a lack of 3D perception. As a result, existing methods such as NSFF [39] decay the weight of optical flow loss after certain iterations. On the other hand, the training of the local deformation network relies on multi-view images that contain 3D consistency. Therefore, the learned local deformation network can help optimize the estimated optical flow and integrate 3D aware multi-view information and geometry information into the optical flow. To this end, different from NSFF [39], we propose a joint optimization strategy of 3D offset and 2D optical flow, which ensures that the offset learning is not affected while optimizing the optical flow results. Concretely, at the initial stage of training, we use optical flow to help with the learning of local offsets based on Eq. 4. After a certain number of iterations, the local deformation network has been optimized under the guidance of optical flow in the coarse level. We then mainly rely on RGB supervision to optimize the local network, and use the predicted scene flow to optimize the optical flow based on Eq. 4. The local offset and optical flow are optimized alternately at this stage. Note that only 2D optical flow maps are optimized here, excluding network parameters for predicting optical flow.

3.3. Temporally Progressive Training

The proposed local deformation network transforms the sampled points back to the canonical space step-by-step, but it lacks the ability to perceive the whole motion sequence. Therefore, we formulate a temporally progressive training strategy and introduce a global deformation network F_{global} after the convergence of the local deformation network F_{local}. F_{global} directly predicts the global offset $\Delta\mathbf{p}_{global}$ of the sampled point $\mathbf{p}$ from the current frame to the canonical frame. As we have the local deformation network which provides local correspondence between adjacent frames, in addition to color supervision, the global deformation network in our work is supervised by the local deformation network. Specifically, the accumulated offsets $\Delta\mathbf{p}_{local}$ of the local deformation networks are used to supervise the global deformation network via the local loss:

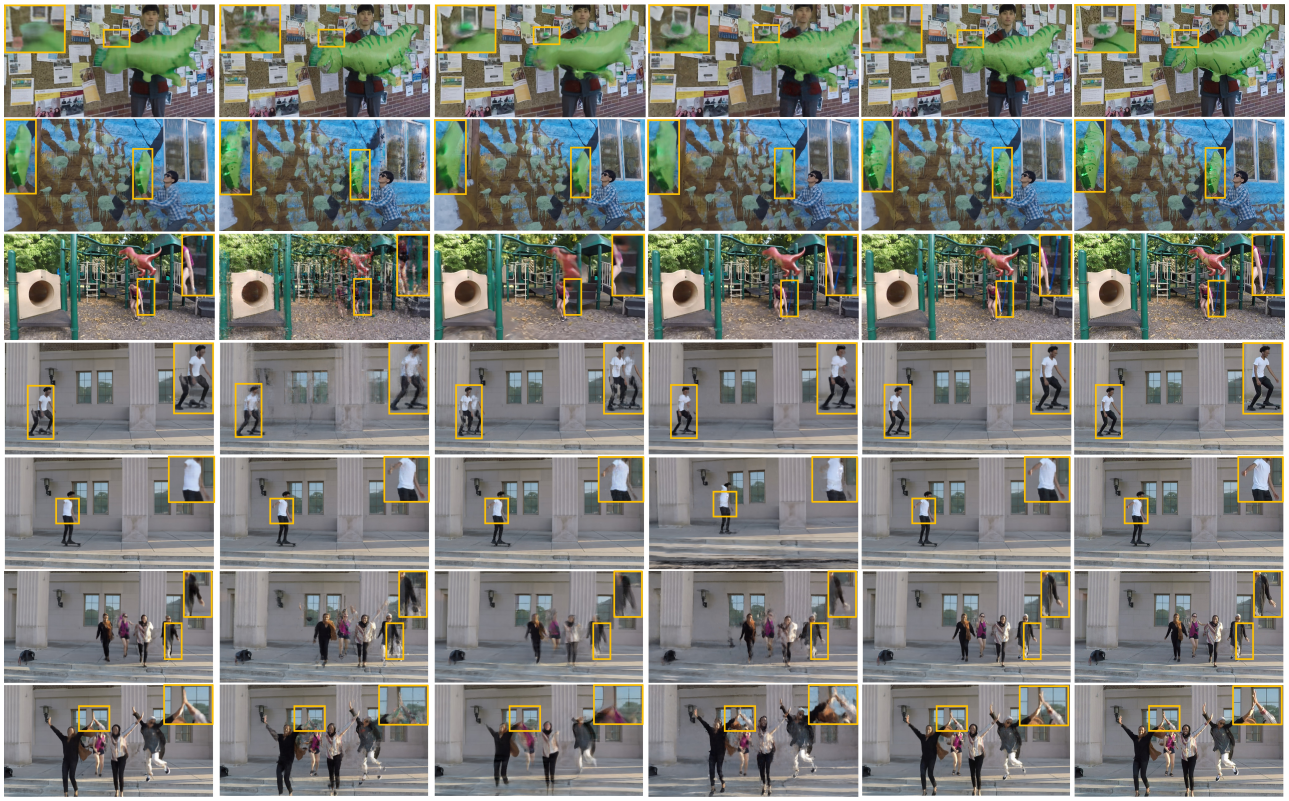
$$L_{local} = \|\Delta\mathbf{p}_{global} - \Delta\mathbf{p}_{local}\|^2_2. \tag{5}$$

Our local and global deformation networks are optimized in two stages, which is more conducive to the stability of training (see the experiments in Sec. 4.2.4). In this way, the global offset is additionally supervised by the accumulated local offsets, while the local offset is additionally supervised by the estimated optical flow and a joint optimization strategy is proposed to alleviate estimation errors. Through this temporal local-to-global procedure, each step has a direct supervision. Compared with other work [59], our global deformation network is easier to converge.

Furthermore, in order to distinguish the static sampled points from the dynamic sampled points for better convergence of the deformation networks, we introduce a rigidity network [79] to predict a rigidity value $r(\mathbf{p})$ for each sampled point $\mathbf{p}$. The rigidity network takes the coordinates of the sampled point as input, and the output probability of rigidity will act as a mask (1 for dynamic and 0 for static) on the predicted offset of each sampled point to determine whether to use the offset to transform the point. The predicted offset is multiplied with the predicted probability of rigidity and then added to the coordinates of the sampled point. The rigidity network is shared across the local stage and the global stage and is trained together with the local deformation network and the global deformation network.

3.4. Hybrid Modeling with Space-time Field

After being transformed by the local or global deformation network, the sampled point will go through the canonical NeRF built in the canonical space, which ensures a



(a) NSFF (b) D-NeRF (c) NR-NeRF (d) TiNeuVox (e) Ours (h) Ground Truth

Figure 4. Comparisons of the novel view synthesis on the Nvidia dataset [98]. The synthesized images are computed for training camera poses at specific times not included for training. Our method can better reconstruct dynamic details and preserve the basic scene geometry.

consistent 3D representation of the video sequence. However, limited by the representation ability, both the local and global deformation networks may not be able to model the complete deformation, *i.e.*, some dynamic details can be lost. Also, the deformation networks only transform the spatial coordinates and the time-related dynamic details are not taken into consideration. To address these issues, we propose a hybrid framework that incorporates the deformation network with space-time field network, which can boost each other and improve the performance.

Specifically, we introduce positionally encoded time as the input to the canonical NeRF as a dynamic feature representation. Then, our canonical NeRF is turned into a space-time field network:

$$F(\Theta) : (\zeta_p(\mathbf{p}_0), \zeta_d(\mathbf{d}), \zeta_t(t)) \mapsto (\mathbf{c}, \sigma) \tag{6}$$

where $\zeta_p(\cdot)$, $\zeta_d(\cdot)$ and $\zeta_t(\cdot)$ are the positional encodings of the position, view direction $\mathbf{d}$, and time t respectively, and $\mathbf{p}_0$ is the transformed coordinates. Note that different from [15, 57], the training of both local and global deformation networks have corresponding direct supervisions, which ensures that the deformation networks do not degenerate (*i.e.*, a poor scenario with deformation networks doing nothing and the space-time field network modeling all dynamic information).

3.5. Loss Functions

We use several loss functions to supervise the two-stage training of our method. We have described the RGB loss L_{RGB}, the optical flow loss L_{flow}, and the local loss L_{local}. In addition to these three losses, based on the sparsity of the motion, we also add the regularization loss L_{reg} to the predicted local deformation, global deformation and rigidity value,

$$L_{reg} = \|\Delta\mathbf{p}\|_2^2 + r(\mathbf{p}), \tag{7}$$

where $r(\mathbf{p})$ is the predicted rigidity value of the sampled point $\mathbf{p}$, and $\Delta\mathbf{p}$ denotes the local deformation or global deformation according to the training stage. So the total loss function L_1 in the local stage is:

$$L_1 = L_{RGB} + w_{flow}L_{flow} + w_{reg}L_{reg}. \tag{8}$$

w_{flow} and w_{reg} are the coefficients of the losses. We set $w_{flow} = 0.02$, $w_{reg} = 0.01$ in all our experiments. The total loss function L_2 in the global stage is:

$$L_2 = L_{RGB} + w_{local}L_{local} + w_{reg}L_{reg}. \tag{9}$$

We set $w_{local} = 10$, $w_{reg} = 0.01$. w_{local} will gradually decay by a ratio of 0.01 every 1000 steps during training.

Please refer to the supplementary document for detailed network architecture and training details.

4. Experiments and Evaluations

In this section, we perform extensive experiments on monocular videos, which come from public datasets [98, 19]. We first compare with the state-of-the-art dynamic NeRF approaches to prove the superiority of our method. Then, we evaluate and validate the important role of each proposed design through the ablation studies. In addition to the qualitative evaluations, we also adopt three commonly used metrics to measure the quantitative performance of the synthesized images, including Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Image Metric (SSIM) [89] and Learned Perceptual Image Patch Similarity (LPIPS) [103]. Note that for PSNR and SSIM, the larger the better, while for LPIPS the smaller the better. For the details of implementation and dataset, please refer to the supplementary document.

4.1. Evaluations and Comparisons

After the reconstruction of the dynamic monocular video, we can change the observation view at a specific time and synthesize a novel view image. This is the common evaluation setting for the Nvidia Dataset [98], where the observation view is sampled from views in the training set. Similar to the Nvidia dataset, we also generate some monocular videos from the NeuralVideo dataset [37] and ST-NeRF dataset [102]. See the supplementary document for more details. Since these datasets are generated from multi-view video sequences, we can switch view positions on the fly ("teleporting cameras"), according to DyCheck [19]. For a more realistic scenario, we also adopt the real monocular video from the DyCheck iPhone dataset [19].

We compare our method to existing dynamic NeRF methods. Specifically, we compare it to a space-time field-based method, NSFF [39], which adopts optical flow as early supervision, to two ray bending-based methods, D-NeRF [59] and NR-NeRF [79], which only adopt a global deformation network, and TiNeuVox [15], which uses time-aware neural voxels to accelerate the training. It should be noted that D-NeRF inputs the positionally encoded time while NR-NeRF inputs the learnable per-frame embedding.

We first synthesize novel view images following the common setting of Nvidia dataset [98]. The qualitative comparisons on this dataset are shown in Fig. 4. It is clearly observed that compared to other methods, our method can better reconstruct dynamic details (the first two rows) and maintain the basic geometry of the scene (the third and fourth rows), owing to our local-to-global temporal framework. The quantitative comparison results are presented in the left part of Table 1 which also supports this claim.

Table 1. Quantitative comparisons on the novel view synthesis on Nvidia dataset [98] and Dycheck dataset [19]. Our method achieves best performance on all three metrics (PSNR, SSIM, LPIPS), quantitatively superior to the existing methods.

Methods	Nvidia Dataset			Dycheck Dataset		
	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑
NSFF [39]	24.17	0.1467	0.7370	25.54	0.3560	0.6226
D-NeRF [59]	20.16	0.2616	0.4974	27.76	0.2691	0.7262
NR-NeRF [79]	19.81	0.2973	0.4905	23.26	0.3723	0.6448
TiNeuVox [15]	20.45	0.3758	0.5274	24.16	0.4757	0.5573
Ours	**26.91**	**0.098**	**0.7963**	**29.87**	**0.1665**	**0.8204**

Table 2. Quantitative comparisons on the novel view synthesis on Nvidia dataset [98] with 4DGaussians [90]. Our method achieves better performance on all three metrics (PSNR, SSIM, LPIPS).

Methods	PSNR↑	LPIPS↓	SSIM↑
4DGaussians [90]	23.35	0.2169	0.7724
Ours	**26.91**	**0.098**	**0.7963**

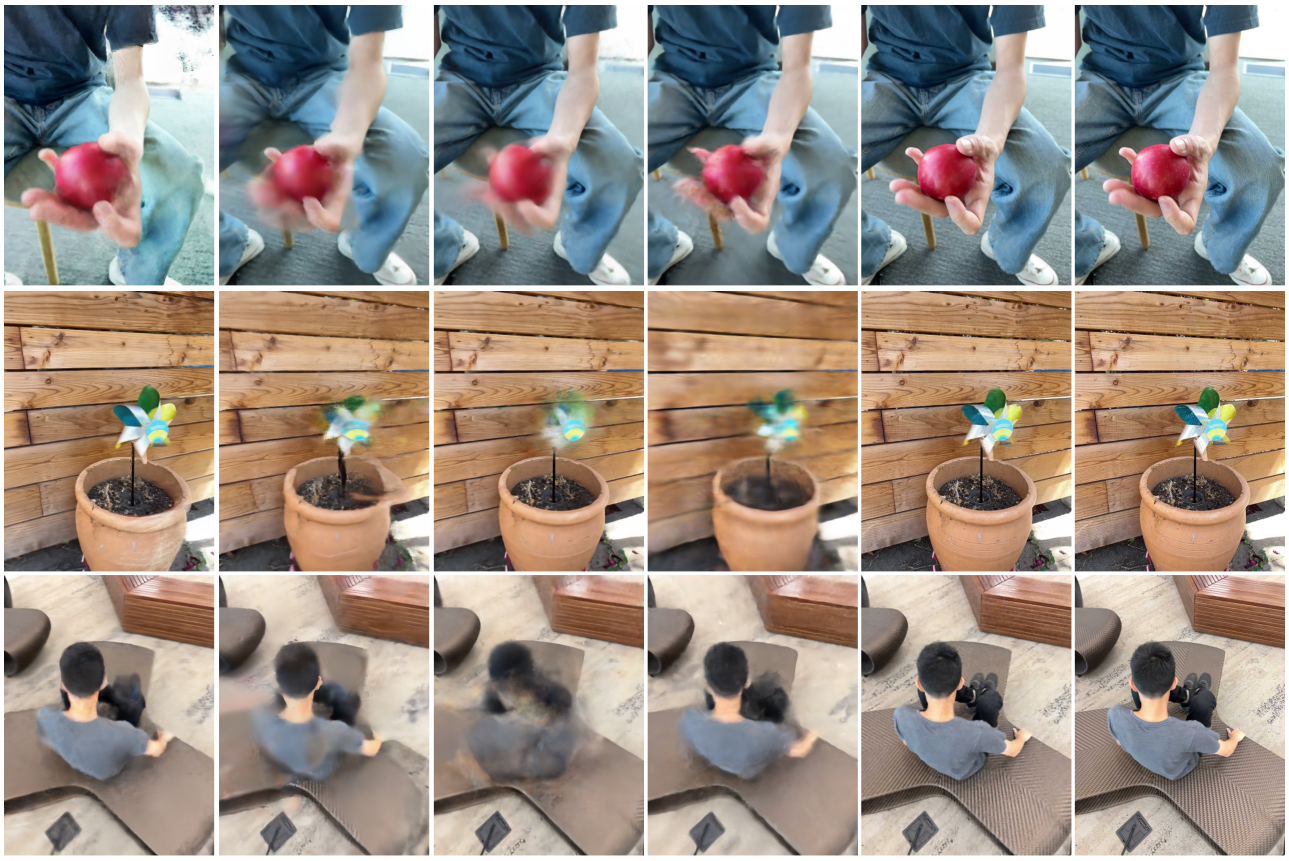
Our approach obtains the best scores on all three metrics. We show more results on the self-generated monocular videos from the ST-NeRF dataset [102] and NeuralVideo dataset [37] in the supplementary document.

As pointed out by Dycheck [19], the above datasets have a somewhat artificial setup of "teleporting cameras". So we additionally evaluate our TPD-NeRF and other methods on the Dycheck iPhone dataset which contains real-captured monocular videos. The qualitative comparisons are shown in Fig. 5. In this case, our method still beats other methods and can synthesize clear hand movement details, which shows good generalization ability. The quantitative comparison results are shown in the right part of Table 1, and we can see that our method achieves the best performance on all three metrics. We also leave a camera out when generating the monocular video from [37, 102] and show the corresponding qualitative comparisons in Fig. 6. The synthesized images come from the left-alone view that does not appear in the training set. Compared with other methods, our method can synthesize clearer dynamic details. For more results, please see the supplementary document.

Moreover, we also compare with a 3DGS-based dynamic modeling method, 4DGaussians [90], which uses the HexPlane feature representation [5] to encode dynamic information. The qualitative and quantitative comparison results are shown in Fig. 7 and Table 2, respectively. These results demonstrate the advantages of our method.

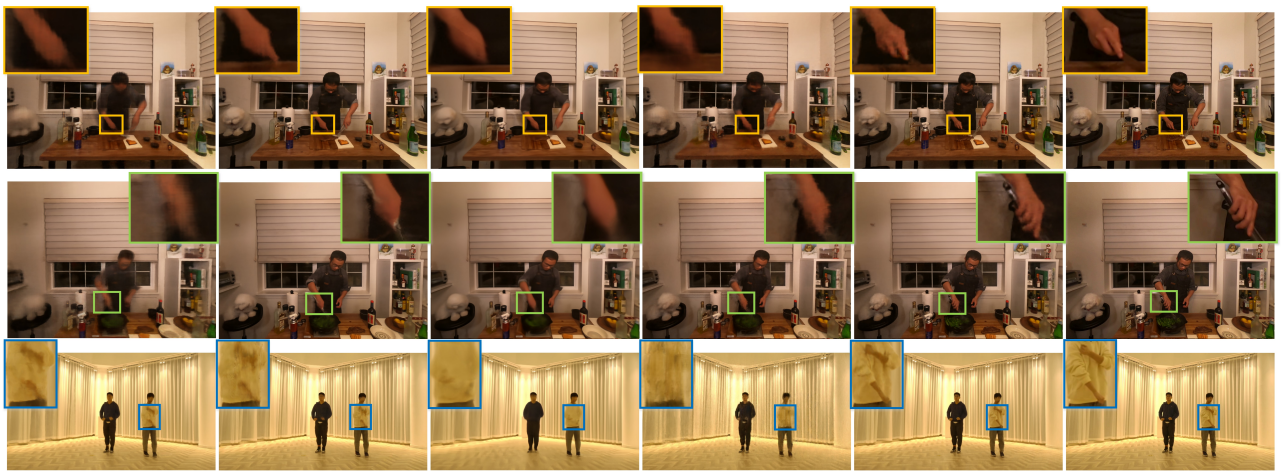
4.2. Ablation Studies

To validate and prove the important role of each key design, we perform several ablations to illustrate the effectiveness of each component in our method, including temporally progressive training (TPT), joint optimization of optical flow, and hybrid modeling with space-time field. The qualitative and quantitative results are shown in Fig. 8 and



(a) NSFF (b) D-NeRF (c) NR-NeRF (d) TiNeuVox (e) Ours (h) GT

Figure 5. Comparisons of the novel view synthesis on the DyCheck dataset [19]. The synthesized images come from validation views that do not appear in the training set. Compared with other methods, our method can synthesize clearer dynamic details.



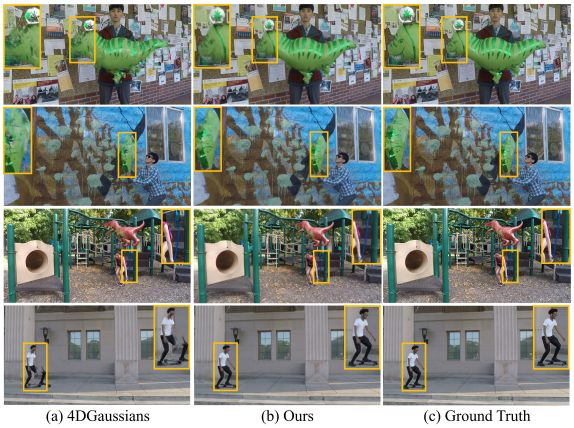
(a) NSFF (b) D-NeRF (c) NR-NeRF (d) TiNeuVox (e) Ours (h) GT

Figure 6. Comparisons of the novel view synthesis with unseen cameras. The synthesized images come from the left-alone view that does not appear in the training set. Compared with other methods, our method can synthesize clearer dynamic details.

Table 3, respectively. From the results, we can see that all our key designs indeed help improve the performance for dynamic NeRF reconstruction from monocular video.

4.2.1 Temporally progressive training

We propose to use both local and global deformation networks and adopt a temporally progressive training (TPT)



(a) 4DGaussians (b) Ours (c) Ground Truth

Figure 7. Comparisons of the novel view synthesis with 4DGaussians [90] on Nvidia dataset [98]. The results show that our method can synthesize clearer details in novel views.

Table 3. Ablation studies on the Dycheck dataset. From left to right, we remove one component from our full model respectively. 'TPT' stands for temporally progressive training.

Methods	w/o TPT	w/o joint	w/o hybrid	Full Model
PSNR↑	25.30	26.03	25.77	**29.82**
LPIPS↓	0.1945	0.2012	0.1426	**0.1045**
SSIM↑	0.7694	0.7532	0.8077	**0.8609**

strategy to ensure both local and global consistency of the scene geometry and appearance. This strategy can help the network capture more dynamic details. We compare the TPT strategy with a ablated strategy that directly uses a global deformation network which is supervised by the accumulated optical flow and adopts the joint optimization mechanism. The results are presented in Figs. 8 (a) and (d). The ablated version will lead to blurry synthesized images. The quantitative comparisons are shown in the first and last columns of Table 3, and it can be clearly seen that our method outperforms the ablated version by a large margin.

4.2.2 Joint optimization of optical flow

We use optical flow to supervise the training of our local deformation network and propose a joint optimization mechanism of 3D offsets and 2D optical flow, intending to make extensive use of optical flow to help with the learning of 3D offset and alleviate estimation errors. As shown in Fig. 8, compared with our full model without optimizing optical flow (column (b)), the joint optimization of 3D offset and optical flow (column (d)) achieves better performance. The quantitative comparison is presented in the second and the last columns in Table 3.

4.2.3 Hybrid modeling with space-time field

To compensate for the dynamic details not captured by the deformation networks, as well as some environmental changes, we incorporate the ray bending with the space-time field network to introduce time-related features into the canonical NeRF. As shown in Fig. 8, compared with the ablated version without hybrid modeling (column (c)), the synthesis quality has been improved and enhanced after hybrid modeling with the space-time field (column (d)), which is also reflected in the improvement of quantitative evaluation in the last column of Table 3.

Table 4. Ablation studies on training strategy.

Methods	PSNR↑	LPIPS↓	SSIM↑
Joint	26.29	0.1539	0.8180
Progressive (Ours)	**29.82**	**0.1045**	**0.8609**

4.2.4 Training strategy

When training local and global deformation networks, we adopt a progressive training strategy, which trains the two networks in separate stages. There is another training strategy, which trains two networks simultaneously. In practical operation, in order to ensure the stability of L_{local}, these two networks are trained in an alternating optimization manner. We compare our progressive training strategy with the joint training strategy. The qualitative comparison is shown in Fig. 9. It can be seen that the joint training is not stable which causes blurry results. The quantitative results are shown in Table 4 which also illustrates that progressive training is better than joint training.

5. Discussions and Conclusions

In this paper, we propose a novel method (TPD-NeRF) for modeling monocular dynamic video. Based on the observation that the motion between adjacent frames is smaller and easier to learn, we propose a temporally progressive learning strategy that captures the correspondences in a local-to-global manner by considering both local and global consistency of the scene geometry and appearance. This strategy can help better capture dynamic details. In order to provide additional supervision for the local deformation network, we employ a joint optimization of the 3D offset and estimated optical flow during the training, which is able to compensate for estimation errors in the 2D optical flow. Finally, we incorporate the ray bending with the space-time field, and introduce a hybrid modeling strategy. Extensive experiments demonstrate that our method is superior to the existing monocular dynamic video NeRF reconstruction methods. However, our method still has some shortcomings. The biggest issue is that our network is still based

(a) w/o TPT (b) w/o joint optimization (c) w/o hybrid modeling (d) Full Model (e) Ground Truth

Figure 8. Ablation study. The different alternatives are formed by removing one component from our full model respectively. 'TPT' stands for temporally progressive training. It can be seen that each component is essential.

(a) Joint training (b) Progressive training (c) Ground Truth

Figure 9. Ablation study on training strategy. We compare our progressive training with the joint training. It shows that the joint training leads to worse results.

on pure MLP networks, which suffers from the burden of expensive training computation, compared to recent work based on voxel representations [52]. The timing bottleneck in training and inference could be overcome by leveraging these explicit and implicit mixed representations, such as [7]. Another issue is that our method cannot handle complex object motions from a unconstrained camera motions.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 62322210), the Innovation Funding of ICT, CAS (No. E461020) and Beijing Municipal Science and Technology Commission (No. Z231100005923031).

References

[1] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 3

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 3

[3] M. Boss, R. Braun, V. Jampani, J. T. Barron, C. Liu, and H. Lensch. NeRD: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 3

[4] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:500–513, 2011. 4

[5] 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*, pages 130–141, 2023. 8

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 4

[7] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 11

[8] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su. MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 1, 3

[9] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 1

[10] Chong Bao and Bangbang Yang, Z. Junyi, B. Hujun, Z. Yinda, C. Zhaopeng, and Z. Guofeng. NeuMesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[11] F. Dellaert and L. Yen-Chen. Neural volume rendering: Nerf and beyond. *arXiv preprint arXiv:2101.05204*, 2020. 3

[12] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional net-

works. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 4

[13] Y. Du, Y. Zhang, H.-X. Yu, J. B. Tenenbaum, and J. Wu. Neural radiance flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14304–14314. IEEE Computer Society, 2021. 3, 4

[14] Y. Duan, F. Wei, Q. Dai, Y. He, W. Chen, and B. Chen. 4d gaussian splatting: Towards efficient novel view synthesis for dynamic scenes, 2024. 4

[15] 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. 3, 4, 7, 8

[16] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 3

[17] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 3

[18] C. Gao, A. Saraf, J. Kopf, and J.-B. Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 3, 4

[19] H. Gao, R. Li, S. Tulsiani, B. Russell, and A. Kanazawa. Monocular dynamic view synthesis: A reality check. In *NeurIPS*, 2022. 8, 9

[20] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022. 3

[21] L. Gao, F.-L. Liu, S.-Y. Chen, K. Jiang, C. Li, Y. Lai, and H. Fu. Sketchfacenerf: Sketch-based facial generation and editing in neural radiance fields. *ACM Transactions on Graphics*, 42(4), 2023. 3

[22] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin. FastNeRF: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021. 3

[23] X. Guo, G. Chen, Y. Dai, X. Ye, J. Sun, X. Tan, and E. Ding. Neural deformable voxel grid for fast optimization of dynamic view synthesis. In *Proceedings of the Asian Conference on Computer Vision*, pages 3757–3775, 2022. 4

[24] P. Hedman, P. P. Srinivasan, B. Mildenhall, J. T. Barron, and P. Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021. 3

[25] B. K. P. Horn and B. G. Schunck. Determining optical flow. In *Other Conferences*, 1981. 4

[26] W. Hu, Y. Wang, L. Ma, B. Yang, L. Gao, X. Liu, and Y. Ma. Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19774–19783, 2023. 3

[27] Y.-H. Huang, Y. He, Y.-J. Yuan, Y.-K. Lai, and L. Gao. StylizedNeRF: consistent 3D scene stylization as stylized NeRF via 2D-3D mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18342–18352, 2022. 3

[28] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1647–1655, 2016. 4

[29] H. Jang and D. Kim. D-tensorf: Tensorial radiance fields for dynamic scenes. *arXiv preprint arXiv:2212.02375*, 2022. 4

[30] K. Jiang, S.-Y. Chen, F.-L. Liu, H. Fu, and L. Gao. Nerffaceediting: Disentangled face editing in neural radiance fields. In *ACM SIGGRAPH Asia 2022 Conference Proceedings*, SIGGRAPH Asia'22, New York, NY, USA, 2022. Association for Computing Machinery. 3

[31] J. T. Kajiya and B. P. Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH Computer Graphics*, 18(3):165–174, 1984. 1, 4

[32] M. Kappel, V. Golyanik, S. Castillo, C. Theobalt, and M. Magnor. Fast non-rigid radiance fields from monocularized data. *arXiv preprint arXiv:2212.01368*, 2022. 4

[33] K. Katsumata, D. M. Vo, and H. Nakayama. An efficient 3d gaussian representation for monocular/multi-view dynamic scenes. *arXiv preprint arXiv:2311.12897*, 2023. 4

[34] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. 4

[35] A. Kratimenos, J. Lei, and K. Daniilidis. Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. *arXiv preprint arXiv:2312.00112*, 2023. 4

[36] L. Li, Z. Shen, Z. Wang, L. Shen, and P. Tan. Streaming radiance fields for 3d video synthesis. *arXiv preprint arXiv:2210.14831*, 2022. 4

[37] T. Li, M. Slavcheva, M. Zollhoefer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, and Z. Lv. Neural 3D video synthesis. 2021. 3, 8

[38] Z. Li, Z. Chen, Z. Li, and Y. Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis, 2024. 4

[39] 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*, pages 6498–6508, 2021. 2, 3, 4, 6, 8

[40] Y. Liang, N. Khan, Z. Li, T. Nguyen-Phuoc, D. Lanman, J. Tompkin, and L. Xiao. Gaufre: Gaussian deformation fields for real-time dynamic novel view synthesis. *arXiv preprint arXiv:2312.11458*, 2023. 4

[41] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[42] Y. Lin, Z. Dai, S. Zhu, and Y. Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle, 2024. 4

[43] F.-L. Liu, H. Fu, Y.-K. Lai, and L. Gao. Sketchdream: Sketch-based text-to-3d generation and editing. *ACM Transactions on Graphics (TOG)*, 43(4):1–13, 2024. 3

[44] S. Liu, X. Zhang, Z. Zhang, R. Zhang, J.-Y. Zhu, and B. Russell. Editing conditional radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5773–5783, 2021. 3

[45] X. Liu, C. R. Qi, and L. J. Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 529–537, 2019. 4

[46] Y.-T. Liu, L. Wang, J. Yang, W. Chen, X. Meng, B. Yang, and L. Gao. Neudf: Leaning neural unsigned distance fields with volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 237–247, 2023. 3

[47] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 3

[48] J. Luiten, G. Kopanas, B. Leibe, and D. Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis, 2024. 4

[49] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 1

[50] 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. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 1, 3, 4, 5

[51] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 1

[52] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 3, 11

[53] M. Niemeyer, L. M. Mescheder, M. Oechsle, and A. Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5378–5388, 2019. 4

[54] A. Noguchi, X. Sun, S. Lin, and T. Harada. Neural articulated radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5762–5772, 2021. 3

[55] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 1

[56] 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*, pages 5865–5874, 2021. 2, 3

[57] 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 Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 3, 7

[58] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 3

[59] 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*, pages 10318–10327, 2021. 2, 3, 5, 6, 8

[60] Y.-L. Qiao, L. Gao, Y.-K. Lai, F.-L. Zhang, M. Yuan, and S. hong Xia. Sf-net: Learning scene flow from rgb-d images with cnns. In *BMVC*, 2018. 4

[61] A. Raj, M. Zollhofer, T. Simon, J. Saragih, S. Saito, J. Hays, and S. Lombardi. Pixel-aligned volumetric avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11733–11742, 2021. 3

[62] C. Reiser, S. Peng, Y. Liao, and A. Geiger. KiloNeRF: Speeding up neural radiance fields with thousands of tiny MLPs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021. 3

[63] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 4

[64] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016. 4

[65] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[66] R. Shao, Z. Zheng, H. Tu, B. Liu, H. Zhang, and Y. Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. *arXiv preprint arXiv:2211.11610*, 2022. 4

[67] R. Shaw, J. Song, A. Moreau, M. Nazarczuk, S. Catley-Chandar, H. Dhamo, and E. Perez-Pellitero. Swags: Sampling windows adaptively for dynamic 3d gaussian splatting. *arXiv preprint arXiv:2312.13308*, 2023. 4

[68] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 4

[69] 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. *arXiv preprint arXiv:2210.15947*, 2022. 3

[70] P. P. Srinivasan, B. Deng, X. Zhang, M. Tancik, B. Mildenhall, and J. T. Barron. NeRV: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 3

[71] C. Sun, M. Sun, and H.-T. Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 3

[72] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2432–2439, 2010. 4

[73] J. Sun, H. Jiao, G. Li, Z. Zhang, L. Zhao, and W. Xing. 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos, 2024. 4

[74] J.-M. Sun, T. Wu, L.-Q. Yan, and L. Gao. Nu-nerf: Neural reconstruction of nested transparent objects with uncontrolled capture environment. *ACM Transactions on Graphics (TOG)*, 43(6):1–14, 2024. 3

[75] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 4

[76] Z. Teed and J. Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 4, 6

[77] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. M. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B. Goldman, and M. Zollhofer. State of the art on neural rendering. *Computer Graphics Forum*, 39, 2020. 1

[78] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pages 703–735. Wiley Online Library, 2022. 3

[79] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhofer, C. Lassner, and C. Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 2, 3, 6, 8

[80] Y.-H. Tsai, M.-H. Yang, and M. J. Black. Video segmentation via object flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3899–3908, 2016. 4

[81] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 722–729. IEEE, 1999. 4

[82] D. Verbin, P. Hedman, B. Mildenhall, T. Zickler, J. T. Barron, and P. P. Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5491–5500, 2022. 3

[83] C. Wang, B. Eckart, S. Lucey, and O. Gallo. Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994*, 2021. 3

[84] C. Wang, L. E. MacDonald, L. A. Jeni, and S. Lucey. Flow supervision for deformable nerf. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21128–21137, 2023. 3

[85] F. Wang, S. Tan, X. Li, Z. Tian, and H. Liu. Mixed neural voxels for fast multi-view video synthesis. *arXiv preprint arXiv:2212.00190*, 2022. 4

[86] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An. Deep video super-resolution using hr optical flow estimation. *IEEE Transactions on Image Processing*, 29:4323–4336, 2020. 4

[87] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, 2021. 3

[88] Q. Wang, Z. Wang, K. Genova, P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser. IBRNet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 3

[89] Z. Wang, A. Bovik, H. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004. 8

[90] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang. 4d gaussian splatting for real-time dynamic scene rendering, 2024. 4, 8, 10

[91] T. Wu, J.-M. Sun, Y.-K. Lai, and L. Gao. De-nerf: Decoupled neural radiance fields for view-consistent appearance editing and high-frequency environmental relighting. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 3

[92] T. Wu, Y.-J. Yuan, L.-X. Zhang, J. Yang, Y.-P. Cao, L.-Q. Yan, and L. Gao. Recent advances in 3d gaussian splatting. *Computational Visual Media*, 10(4):613–642, 2024. 4

[93] T. Wu, F. Zhong, A. Tagliasacchi, F. Cole, and C. Oztireli. D^2-nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. *arXiv preprint arXiv:2205.15838*, 2022. 3

[94] W. Xian, J.-B. Huang, J. Kopf, and C. Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021. 2, 3

[95] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 3

[96] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction, 2024. 4

[97] Z. Yang, H. Yang, Z. Pan, X. Zhu, and L. Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023. 4

[98] J. S. Yoon, K. Kim, O. Gallo, H. S. Park, and J. Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5335–5344, 2020. 7, 8, 10

[99] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 1, 3

[100] Y.-J. Yuan, Y.-T. Sun, Y.-K. Lai, Y. Ma, R. Jia, and L. Gao. Nerf-editing: geometry editing of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18353–18364, 2022. 3

[101] Y.-J. Yuan, Y.-T. Sun, Y.-K. Lai, Y. Ma, R. Jia, L. Kobbelt, and L. Gao. Interactive nerf geometry editing with shape priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[102] J. Zhang, X. Liu, X. Ye, F. Zhao, Y. Zhang, M. Wu, Y. Zhang, L. Xu, and J. Yu. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)*, 40(4):1–18, 2021. 3, 8

[103] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 8

[104] X. Zhang, S. Bi, K. Sunkavalli, H. Su, and Z. Xu. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5449–5458, 2022. 3

[105] X. Zhang, P. P. Srinivasan, B. Deng, P. Debevec, W. T. Freeman, and J. T. Barron. NeRFactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021. 1, 3

[106] 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*, pages 12786–12796, 2022. 2