Neural Radiance Fields for Dynamic View Synthesis using Local Temporal Priors

Rongsen Chen[1,2], Junhong Zhao[1,2], Fang-Lue Zhang[1], Andrew Chalmers[1,2], and Taehyun Rhee[1,2]

[1] School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand
[2] Computational Media Innovation Centre (CMIC), Victoria University of Wellington, Wellington, New Zealand

Abstract. Neural Radiance Fields (NeRF) have demonstrated promising results in synthesizing novel view images from a set of unconstrained captured scenes. One important extension of NeRF is using it on non-rigid reconstruction. Although previous NeRF-based methods for dynamic scene reconstruction have presented visually appealing results, they still often show visual artifacts such as blurry or incorrect geometry of an object. One of the causes is that previous work performs reconstruction directly on the entire video sequence. The global temporal information over the video sequence introduces noise to the network, often leading to a non-optimal canonical space representation of the dynamic scene. In this paper, we present Local Temporal (LT) NeRF, a method to synthesize novel views of dynamic scenes using local temporal priors. Our novel LT module provides the local temporal priors using multi-view stereo sampling, and improves the deformation field reconstruction and hyper-space encoding. Our novel loss functions further supervise the NeRF for better optimization. We evaluate our method with dynamic scenes captured from monocular videos, outperforming the state-of-the-art.

Keywords: View Synthesis, Neural Radiance Fields.

1 Introduction

Synthesizing novel view images provides the freedom to navigate beyond the capturing locations. It is particularly challenging in dynamic scenes containing moving and deformable objects with complex geometry. Recent learning-based methods, highlighted by the advent of NeRF and its subsequent works, have achieved significant improvements to synthesize novel views in dynamic scenes.

The NeRF-based methods aim to model a scene's geometry and appearance as neural radiance fields using multi-layer perceptrons (MLPs), and use volumetric rendering to generate a novel view of the scene. A number of NeRF variants, including D-NeRF [33], Nerfies [31], and HyperNeRF [32], focused on addressing non-rigid reconstruction by learning a deformation field to map the observation coordinates of each input image into a canonical coordinate space. This is then used in a volume rendering process for novel view synthesis. Although these methods showed promising results,

there are still visual artefacts. This is because these deformation-based NeRF methods reconstruct the canonical space based on global temporal information (the entire video sequence). However, a video usually has less meaningful information about a given dynamic object between temporally distant frames. Moreover, in cases of building a NeRF for dynamic scenes, this information will introduce noise, making the training challenging.

In this paper, we propose a way to leverage the dynamic appearance from temporally-nearby images (local temporal priors) to complement the NeRF-based deformation field optimizations. We demonstrate the impact of learning with local temporal priors by introducing Local Temporal NeRF (LT-NeRF). Our novel local temporal (LT) module that uses a multi-view stereo (MVS) sampling to generate the local temporal priors. The local temporal priors are fed into an MLP which then supervises the deformation field and hyper-space encoding. LT module learns the relationship between the local temporal priors, density, and color, which effectively improves the reconstruction of the deformation field and hyper-space encoding. Our method shows promising results when synthesizeing dynamic scenes with moving subjects and subtle deformations, outperforming the stat-of-the-art. Our contributions are summarised as follows:

- We present LT-NeRF, a novel view synthesis approach from monocular videos containing dynamic objects using NeRF enhanced by local temporal priors.
- We propose a novel LT module providing local temporal priors to improve the deformation field reconstruction and hyper-space encoding.
- We introduce two loss functions that take into account the temporal local information to supervise the deformation field and hyper-space encoding optimization.

2 Related Work

View synthesis is a research topic that is closely related to 3D vision, and has been studied for decades. A naive solution to this task would be building explicit 3D scene geometry such as a point cloud and mesh, then rendering this geometry from a novel viewpoint [5, 12, 13, 17]. However, this has a limitation on the fidelity of the generated novel scene, as the reconstructed point cloud/mesh often suffers from visual artifacts such as floating geometry. Image-based rendering [4, 14, 15, 38] improves this, but the rendered scene can still have issues with scene-independent effects such as reflection on refraction. Alternatively, synthesis can also be achieved via implicit soft geometry via light field rendering [20, 22], however, this requires densely captured images.

With the emergence of deep learning, research has focused on using neural networks to learn a representation that is suitable for novel view synthesis. Early work in this direction aimed to convert the set of images into Multi-Plane Images (MPI) [3, 8, 25, 49]. More recently, NeRF [26] has demonstrated a superior view synthesis quality compared to other methods, with a series of subsequent researches that improve NeRF for various aspects such as anti-aliasing [1, 2], night scenes [24], rendering of large-scale outdoor/indoor scenes [35,39,46], acceleration of training [7,27], real-time rendering [46], and many others [23, 42, 45].

Non-rigid reconstruction [6, 16, 30, 40, 41, 50, 51] from monocular video (image sequence) is a challenging research problem for view synthesis. A key to solving this problem is to address the motion of the dynamic object. A solution for this is to find the correspondence between frames, which would then address the motion of the dynamic object. These can either be done via estimating the frame-to-frame optical flow fields [10], or estimating the long-term trajectories that are associated with each point in world space of an image sequence [36]. Another solution is to reconstruct a canonical space representation of the scene, and use an associated warp field to model per-frame deformation in the scene [28].

More recently, with the increasing popularity of NeRF [26], building neural radiance fields for non-rigid scenes has been of interest. The Neural Scene Flow Field (NSFF) [21] achieved non-rigid reconstruction via the use of optical flow to guide the warping of the point between time frames. Gao et al. [9] presents a dual ML model that handles the synthesis of static and dynamic scenes separately. D-NeRF [33] and Nerfie [31] propose to use MPL to reconstruct a canonical space representation of the scene, which can then be later used with the regular NeRF model. HyperNeRF [32] was extended from Nerfie, but focused on improving reconstruction on topological changes with an ambient network to encode hyper-space coordinates. There are also other works focusing on reconstruction for the human body [11, 44], and enable editability on the object [47] to allow users to manipulate the scene. TiNeuVox [7] has been introduced to speed up the training process while maintaining the network performance. In this paper, we focus on improving non-rigid reconstruction using local temporal priors.

3 Overview

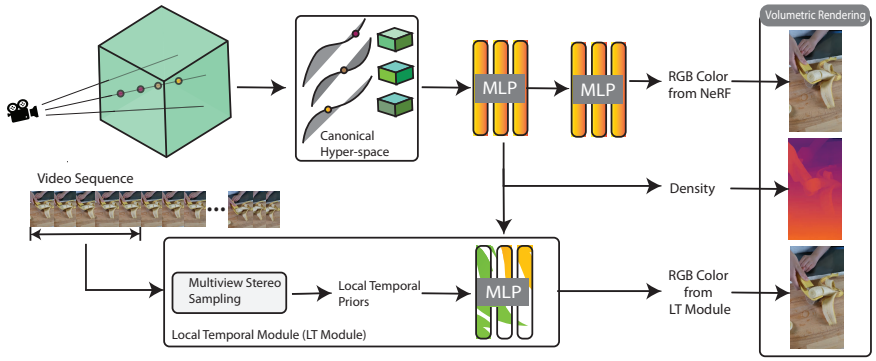

Fig. 1. An overview of our LT-NeRF. The local temporal priors are utilized to provide guidance on the geometric features for the hyper-space deformation encoder optimization.

Our goal is to reconstruct a dynamic scene from a video sequence $\{I_i\}(i = 1, .., N)$. Specifically, we will learn a NeRF-based representation that synthesizes novel views at

an arbitrary position and time. Our problem can be formulated as:

$$\mathbf{c}, \sigma = F(\mathbf{x}, \Psi, t), \tag{1}$$

where, $\mathbf{x}$ is a 3D location (x, y, z) in the viewing volume to synthesize, Ψ are the camera parameters of each reference view obtained using Structure-from-Motion [19, 29, 37], and t is the time. Our aim is to obtain volume density σ at $\mathbf{x}$, and the RGB color $\mathbf{c}$ at $\mathbf{x}$. When synthesizing a novel view, we use the corresponding camera parameter Ψ to obtain the ray direction of each pixel for volumetric rendering.

We propose LT-NeRF, a novel method that uses Local Temporal Priors (LTP) to achieve better quality view synthesis for dynamic scenes. The overall architecture of our LT-NeRF is shown in Figure 1. For each image I_i, we sample 3D points in the observation volume and build a hyper-space radiance field to associate 3D observation coordinates to hyper-space coordinates and predict their densities and colors (section 4). We then leverage an MLP to estimate the color of each spatial-temporal point by combining the color information from temporally nearby frames and the learned geometric features (subsection 5.1). Together with the original NeRF color prediction, the two color outputs provide temporally global and local appearance information that are complementary to each other, leading to better geometry reconstruction from the radiance field. For model training, we propose a local temporal loss to ensure the local temporal information can be effectively learned (subsection 5.2). Furthermore, we also use the auxiliary guided depth map from a CNN-based depth estimation method to improve the density prediction of the radiance field construction.

4 Dynamic Scene Representation

We use hyper-space neural radiance field (HyperNeRF) [32] as our base dynamic scene representation. HyperNeRF extends NeRF [26] by additionally learning a deformation module to map a 3D point in the observation space to a canonical space representation to obtain its density and color for volumetric rendering. The mapping is learned through a deformation field and an ambient field, which can be represented by: $F_d : (\mathbf{x}, \omega_t) \rightarrow (\mathbf{x}', w)$. Here, ω_t is the latent deformation code for the frame at time t and ω is the ambient coordinate. The hyper-space coordinates are fed to the MLPs for predicting color and density. For the final image reconstruction, the view direction $\mathbf{d}$ of each pixel is used to cast rays for volumetric rendering, which can be obtained using the camera parameter Ψ. A per-frame appearance code ϕ_t is also used to ensure the color changes observed between different frames are learned by the radiance field. Thus, the basic radiance field for the frame at time t is formulated as:

$$(\mathbf{c}', \sigma) = F_{\Theta_h}(\mathbf{x}, \Psi_t, \omega_t, \phi_t) \tag{2}$$

By optimizing the parameters $\Theta_h = \{\theta_{def}, \theta_d, \theta_c\}$ of the basic HyperNeRF, the model is capable of generating novel views of dynamic videos and handles movements of objects in the dynamic scene. Here, $\{\theta_{def}, \theta_d, \theta_c\}$ represents the parameters of the MLPs for predicting the deformed/ambient coordinates, density estimation, and color estimation respectively.

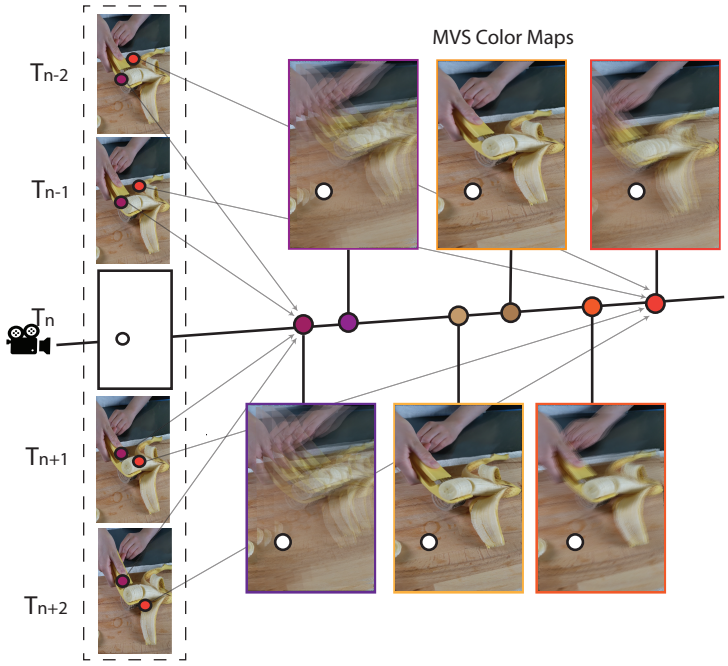

Fig. 2. Illustration of the Multiview Stereo (MVS) sampling process in our LT module.

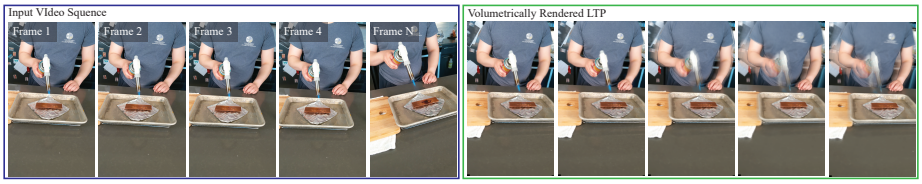

Fig. 3. An example of volumetric rendering with raw LTP reconstructed with the estimated depth. Given there are N frames from a video sequence (blue box), we isolate our deformation learning among n temporally local frame. The green box shows a volumetric rendering result with LTP, where $n = 2, n = 4, n = 20, n = 50, n = N$ respectively. For this example scene $N = 100$. The torch and person are dynamic objects in the scene.

5 Local Temporal NeRF

We hypothesize that the temporally nearby frames provide more reliable information for dynamic scene reconstruction including subtle deformation of non-rigid objects. We use n temporally closest neighboring frames for our view synthesis. The size of n was set experimentally to 4, because we found that smaller n often improved results for dynamic movement such as the torch scene in Figure 3.

We propose LT-NeRF, a novel method that utilizes local temporal information for better learning of dynamic scenes. We developed a novel LT module to leverage the stronger pixel-wise correlations in local temporal windows and provide local temporal priors to improve hyper-space encoding and deformation field optimization.

5.1 Local Temporal Module

Multiview Stereo (MVS) Sampling: The camera parameter ϕ_t for frame t is obtained in our preprocessing step. Given a 3D point $\mathbf{x}$ on a light ray, we sample Local Temporal Prior (LTP) $\hat{\mathbf{c}}(\mathbf{x})$ from its temporally nearest n images $\{I_i\}(i = 1, 2, ..., N)$ using MVS:

$$\begin{aligned} \hat{\mathbf{c}}(\mathbf{x}) &= \frac{1}{n} \sum_{(\hat{p},i)\in\Omega}^{n} I_i(\hat{\mathbf{p}}) \\ \Omega &= \{(\hat{\mathbf{p}}, m)|\hat{\mathbf{p}} = P(\phi_t, P^{-1}(\phi_i, \mathbf{x}))\} \end{aligned} \tag{3}$$

where P and P^{-1} represent the projection and re-projection between 3D space and 2D image space. Ω is the set of pixel coordinates $\hat{\mathbf{p}}$ for the corresponding pixels in the neighboring views. The illustration of our MVS sampling process is shown in Figure 2. Our LT-NeRF uses this LTP $\hat{\mathbf{c}}(\mathbf{x})$ to learn the canonical space representation of the scene with local temporal information.

LT-MLP: Our local temporal module has an MLP network with three fully-connected layers, and a ReLU activation function followed by the second layer as shown in Figure 1. Our LT module can be formulated as:

$$\mathbf{c'_{lt}} = F_{\theta_{lt}}(hF(\Psi_t, \phi_t, \hat{\mathbf{c}}(\mathbf{x}))) \tag{4}$$

where h is the feature output from the NeRF, and $\hat{\mathbf{c}}(\mathbf{x})$ is our LTP, and F is a single MLP layer for extract feature. As shown in Figure 1, Our LT module serves as an additional RGB prediction layer and outputs c'_{lt}.

NeRF with LT Module: Our LT module can effectively improve the hyper-space radiance field construction with our local temporal prior $\hat{\mathbf{c}}(\mathbf{x})$:

$$(\mathbf{c'}, \mathbf{c'_{lt}}, \sigma) = F_{\Theta_h}(x, \Psi_t, \omega_t, \phi_t, \hat{\mathbf{c}}(\mathbf{x})) \tag{5}$$

Here, Ψ_t, ω_t, ϕ_t are the same as defined in section 4. $\Theta_h = \{\theta_{def}, \theta_d, \theta_c, \theta_{lt}\}$ are the parameters of the sub-networks to be optimized: θ_{lt} for the LT module, θ_{def} for the

deformation/ambient coordinates prediction MLPs, and θ_d for the density estimation MLP. The rendered image from the NeRF-estimated color space will be used as the final LT-NeRF output.

5.2 Loss Functions

Given time t and camera Ψ_t, we use the volumetric rendering method used in NeRF [26] to render three output colors (C'_t, $\hat{C}_t$, $\bar{C}_t$). More specifically, we render our output using:

$$C = \int_{\mathbf{x}_f}^{\mathbf{x}_n} T(\mathbf{x})\sigma(\mathbf{x})\mathbf{c}(\mathbf{x})d\mathbf{x} \tag{6}$$

where $T(\mathbf{x}) = exp(-\int_{\mathbf{x}_n}^{\mathbf{x}} \sigma(s)ds)$, here s is the sample distance. σ is the density, $\mathbf{c} \in \{\mathbf{c}', \mathbf{c}'_{\mathbf{lt}}, \hat{\mathbf{c}}\}$ is the color on a light ray. $C \in \{C'_t, \hat{C}_t, \bar{C}_t\}$ is one of our outputs. x_n and x_f refers to the closest and furthest depths respectively. Although the three output colors are rendered similarly, they provide different information for the network to optimize.

The NeRF color C'_t is a neural rendering process that renders a color value without prior information. The LT module outputs colors $\hat{C}_t$ which are obtained by taking LTPs into account, where the scene's color is already established. Since our LTP is obtained via MVS sampling, the inherited MVS color information will be used in our LT module output. One characteristic of the MVS color map is that the closer the depth is to an optimal solution, the clearer the MVS color map will be. Therefore, our LT module output can be used to assess the quality of the reconstructed geometry in relation to an MVS color map. Similarly, $\bar{C}$ uses the raw LTP to assess the quality of the reconstructed geometry for the MVS operation.

We use an L2 norm as our rendering loss to uses our network to generate pixel colors using the radiance field for a novel viewpoint:

$$L_r = ||C_t(\mathbf{p}) - C'_t(\mathbf{p})||_2 \tag{7}$$

For our LT module, we also use L2 norm as our loss function to allow the LT module to regress the pixel color of the target view:

$$L_{\mathbf{lt}} = ||C_t(\mathbf{p}) - \hat{C}_t(\mathbf{p})||_2 \tag{8}$$

The L_{lt} loss is used to guide the depth estimation using the LT module output, with the idea that the correct depth value would yield an MVS color map more similar to the ground truth. However, since our LT module behaves similarly to an in-painting layer that matches the output to the ground truth by in-painting the LTP, some important information such as the clarity of the LTP might be overridden by the network. Thus, to ensure this information is still involved in the optimization stage, we also propose a raw LTP loss that directly uses the initial color of a 3D point estimated using multi-view stereo in the volumetric rendering:

$$L_{\mathbf{raw}} = ||C_t(\mathbf{p}) - \bar{C}_t(\mathbf{p})||_2 \tag{9}$$

where $\hat{C}$ is the integration of all the products of each sample point's raw LTP $\mathbf{r}$ and its estimated density σ along the ray for pixel $\mathbf{p}$.

To further improve the performance of our method, we employ the auxiliary depth guidance method proposed by Li et al. [21], where a standardized depth map provides an intuition of each object's relative position and can be used to guide the initial learning of the density estimation. In our experiment, we use the single depth estimation network MiDas [34] to estimate a depth guidance map. Denoting the standardized depth value for a point's depth d as $n(d)$, we apply an L2 depth guidance loss as:

$$L_d = ||n(d) - n(d')||_2 \tag{10}$$

where $n(d')$ is the standardized depth estimated from our network, obtained by performing volumetric rendering on the length of the light ray [26], and $n(d)$ is the standardized depth guidance estimated using MiDas [34].

With all of these losses combined together, our overall loss function is:

$$L = L_r + \alpha L_{\mathbf{lt}} + (1 - \alpha) L_{\mathbf{raw}} + \beta L_d \tag{11}$$

The contribution of the LT module reconstruction loss and the raw LTP loss is weighted using the hyperparameter α. We use $\alpha = 0.99$ in all our experiments as we did not want the network to place too much emphasis on $L_{\mathbf{raw}}$ since this will contain noise around the boundary of objects due to occlusion. The parameter β is set to decay exponentially, and thus just affects the early stage of the entire training process.

5.3 Implementation

We implement our network with python's PyTorch library. For the deformation module, we follow a similar setup as the original HyperNeRF [32], where we use 8 dimensions for both the apparent and deformation latent code, and 2 ambient dimensions. We use the Adam optimizer [18] to optimize our network, set the learning rate to 10^{-3}, and exponentially decay the learning rate alongside our training steps. We set $\alpha = 0.99$ and $\beta = 0.4$. Similar to the learning rate, β will also exponentially decay along with the training, since we only use it to resolve the spatial positioning of objects at the beginning of our training.

6 Results

We compare LT-NeRF with the state-of-the-art NeRF-based view synthesis method from dynamic scenes, including NSFF [21], Nerfie [31], and HyperNeRF [32]. We compare our LT-NeRF with them both quantitatively and qualitatively.

To compare our work with prior studies, we employ the same dataset as used by HyperNeRF [32]. We created two datasets, identified as A and B, using distinct methods for selecting training and testing data. In dataset A, we take one frame out of every 10 frames of each input video sequence as the test data and use the remaining frames for training. In contrast, for dataset B, we follow the same setup as in prior works [31, 32] by assigning the left view to the training data and the right view to the test data in

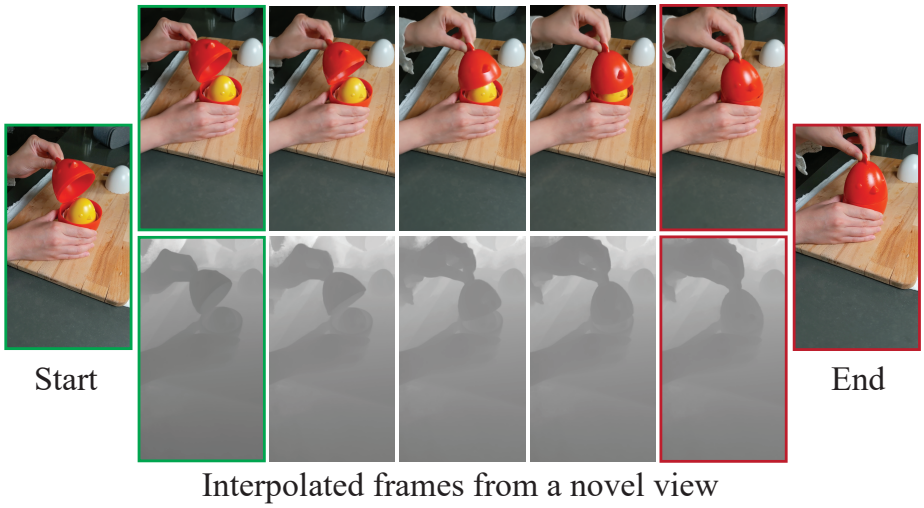

Fig. 4. Novel views synthesized by linearly interpolating the two frames of the chicken (left and right), showing smooth motion.

an alternating fashion. It is worth noting that dataset A has a video sequence with a frame rate of 15 fps, while dataset B has a 5 fps video sequence. This difference in fps implies that the dynamic object in dataset A will undergo fewer changes between different time frames. Therefore, dataset A will allow us to examine situations where the object's motion is relatively slow, whereas dataset B will showcase situations with relatively faster object motion.

For quantitative evaluation, we adopted metrics LPIPS [48], MS-SSIM [43], and PSNR, just as what have been used in other related works [26, 32]. In the comparisons, we use images with 268×480 resolution. For network training, we set the batch size to 2048 and the number of sampling rays to 128. Moreover, all methods were trained with 250,000 iterations.

6.1 Quantitative Evaluation

Our quantitative analysis results are presented in Table 1, Table 2, and Table 3, which clearly show that our method outperforms others for most test scenes. Specifically, our method achieved an average PSNR of 31.33 in dataset A, representing a significant improvement from the 27.53 achieved by HyperNeRF. Moreover, our method's PSNR is also higher than NSFF and Nerfie. In terms of MS-SSIM and LPIPS, our approach consistently outperforms other methods, indicating that LT-NeRF can generate superior results when synthesizing novel views.

Regarding dataset B, our method showed better results compared to the previous method, although its performance dropped compared to dataset A. The drop in performance suggests that our method is more effective on objects' motion within a rea-

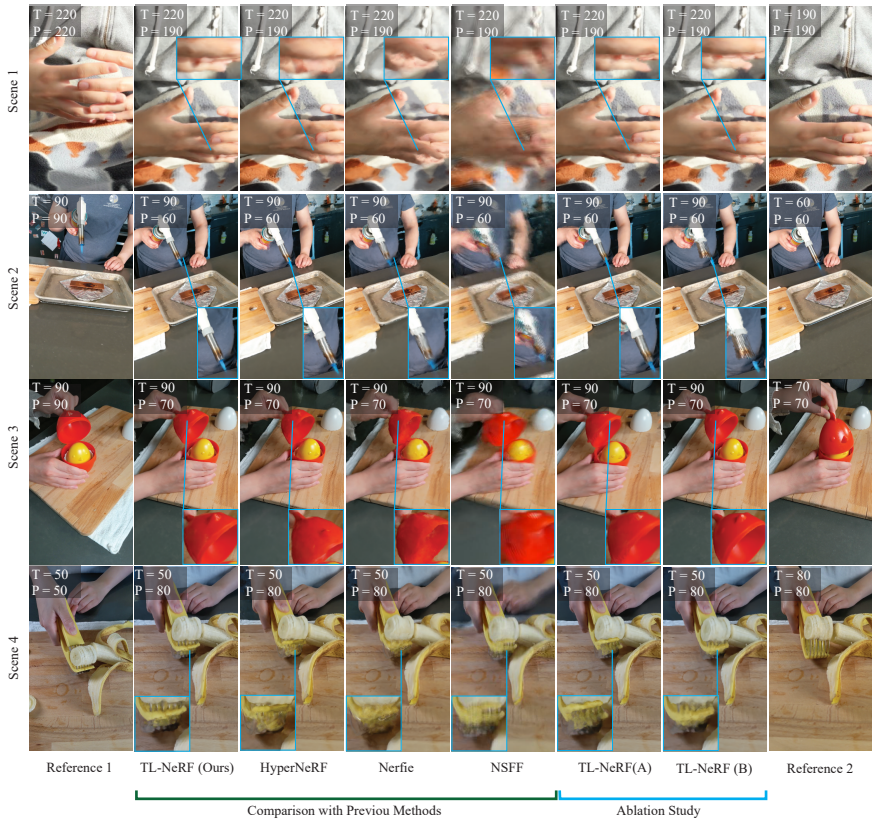

Fig. 5. Time-pose interpolation results of test scenes, where T is the time index and P is the camera pose index in the frame sequence. Reference 1 and 2 are ground truth frames (when $T = P$) for time T and pose P.

sonable range. We did observe an exception on the VRIG-chicken scene, which we will discuss in section 7. However, overall, our method still demonstrated better performance qualitatively. Notably, as illustrated in our supplementary, LT-NeRF outperforms other methods in reducing object shifting artifacts, thereby improving the accuracy and stability of object representation in the scene. These findings support that our method is effective in improving the image quality of synthesized novel views by using local temporal priors to guide network optimization.

6.2 Qualitative Evaluation

A comparison between our method and the other three previous methods is shown in Figure 5. We can see that our method can recover more details and maintain the object's shape with fewer blurring artifacts. NSFF has shown issues with blurring and ghosting, which may be caused by inconsistent geometry reconstruction across different time frames. HyperNeRF has shown improved quality results than NSFF and

Nerfie. However, it often has broken shapes on dynamic objects, as shown in scenes 1 and 4. HyperNeRF also has stretched objects and blurry surfaces, which can be found in scenes 2 and 3. Compared to these methods, our results show a more consistent object shape and sharper image fidelity. Based on the results at various times and camera poses for each test scene, we can see that our method outperforms the other methods in time-position interpolation.

	Chicken (100 frames)			Banana (100 frames)			Chocolate (100 frames)			Hand (226 frames)		
	PSNR↑	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
NSFF [21]	24.41	0.8915	0.1329	27.78	0.9195	0.0992	27.58	0.9651	0.0624	29.87	0.9766	0.074
Nerfie [31]	27.55	0.9485	0.1037	30.15	0.9609	0.0821	26.27	0.9528	0.0618	27.55	0.9594	0.0713
HyperNeRF [32]	26.89	0.9390	0.0867	29.07	0.9594	0.0668	28.09	0.9768	0.0413	27.07	0.9586	0.0658
LT-NeRF (Ours)	29.67	0.9701	0.0554	32.56	0.9802	0.0382	29.05	0.9815	0.0240	32.35	0.9888	0.0568
LT-NeRF (A)	25.55	0.9272	0.0983	31.35	0.9697	0.0492	28.09	0.9758	0.0351	30.65	0.9697	0.0615
LT-NeRF (B)	29.65	0.9695	0.0609	31.98	0.9767	0.0369	29.09	0.9815	0.0283	32.16	0.9888	0.0593

Table 1. The reconstruction quality for different methods on dataset A (4 scenes). We use text color green, blue, red to mark the best, the second best, and the third best results, respectively.

	VRIG-Chicken (164 frames)			VRIG-3D Printer (207 frames)			VRIG-Broom (197 frames)		
	PSNR↑	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
NeRF [21]	19.9	0.777	0.325	20.7	0.780	0.357	19.9	0.653	0.692
NV [21]	17.6	0.615	0.336	16.2	0.665	0.330	17.7	0.623	0.360
NSFF [21]	26.9	0.944	0.106	27.7	0.947	0.125	26.1	0.871	0.284
Nerfie [31]	26.7	0.943	0.078	20.6	0.830	0.108	19.2	0.567	0.325
HyperNeRF [32]	26.9	0.948	0.079	20.0	0.821	0.111	19.3	0.591	0.269
LT-NeRF (Ours)	26.0	0.904	0.102	23.1	0.885	0.097	21.9	0.742	0.217

Table 2. The reconstruction quality for different methods on dataset B (3 scenes). The color green, blue, red indicate best, second best, and third best, respectively.

6.3 Ablation Study

Our LT-NeRF addresses the dynamic scene reconstruction issues by using the LT module with specific loss functions to regularize the network optimization. To understand the contributions of each component, we conduct an ablation study to demonstrate how LT-NeRF performs when removing the LT module from the pipeline (see Figure 1, denoted as LT-NeRF (A)) as well as removing depth-related loss items L_d from the loss function (see Equation 11, denoted as LT-NeRF (B)). For LT-NeRF (A), we use loss $L = L_r + \beta L_d$ to regularize the network training. For LT-NeRF (B), we train the network use loss $L = L_r + \alpha L_{\mathbf{lt}} + (1-\alpha) L_{\mathbf{raw}}$.

As shown in Table 1, the results of LT-NeRF (A) without the LT module in the pipeline shows the worst results across all metrics in all test scenes when compared with

	DataSet A			DataSet B		
	PSNR↑	**SSIM↑**	**LPIPS↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**
NeRF [21]	-	-	-	20.2	0.737	0.458
NV [21]	-	-	-	17.2	0.634	0.342
NSFF [21]	27.41	0.9400	0.0891	26.9	0.921	0.172
Nerfie [31]	27.88	0.9554	0.0797	22.2	0.780	0.170
HyperNeRF [32]	27.53	0.9537	0.0662	22.01	0.77	0.153
LT-NeRF (Ours)	30.91	0.9802	0.0436	23.8	0.846	0.132
LT-NeRF (A)	28.91	0.9606	0.0610	-	-	-
LT-NeRF (B)	30.72	0.9806	0.0512	-	-	-

Table 3. The average reconstruction quality for different methods on dataset A and B. We use text color green, blue, red to mark the best, the second best, and the third best results, respectively.

both LT-NeRF and LT-NeRF (B), which has the LT module included. Adding the LT module increased the average PSNR from 28.91 to 30.91 and MS-SSIM from 0.9606 to 0.9802. For LILPS, adding the LT module reduced the error from 0.0610 to 0.0436. The consistent improvement of all metrics shows that our proposed LT module is effective in improving image quality. Qualitative results also show LT-NeRF (A) without LT module gives the worst result in most cases. For instance, scene 3 in Figure 5, we observed that LT-NeRF (A) showed blurrier results on the chicken toy. In scene 4, we observe that LT-NeRF (A) fails to reconstruct the correct shape of the banana cutter. Compared to this, LT-NeRF and LT-NeRF(B) are both able to demonstrate better results. This shows that our proposed LT module can improve the quality of novel view synthesis and generate more visually appealing results.

Removing depth-guidance regularization also affects the results but just to a very small degree in most cases. The performance on PSNR, MS-SSIM, and LILPS metrics are either slightly decreased or kept very close. The contribution of depth guidance may be limited by the fact that the depth used for guidance is produced by the CNN-based depth estimation method and thus doesn't have sufficient accuracy for supporting the view synthesis task. Nonetheless, employing depth guidance could improve the robustness of our method in general (see Figure 5). From the depth maps shown in Figure 6 we can see that the depth output from LT-NeRF trained without depth guidance (Figure 6 (A)) would sometimes give slightly inconsistent depth on the same region of an object (e.g., the top of the torch). The underlying reason could be that, although the LTP helps to enhance depth optimization by giving prior RGB information for the estimated point, it also introduces some noise to the network. Since not all pixels captured from various camera poses will be able to intersect perfectly due to intensity differences and position motion distortion in the image. The inconsistent depth in Figure 6 gives an indication of why artifacts appeared in the results of different time frames Figure 5 generated by LT-NeRF(B) in scene 2. Employing depth guidance addresses this issue.

6.4 Additional comparisons

It is noteworthy to mention that the volumetrically rendered output $\hat{C}_t$ from the LT module can also generate high-fidelity novel views. However, the generated novel views

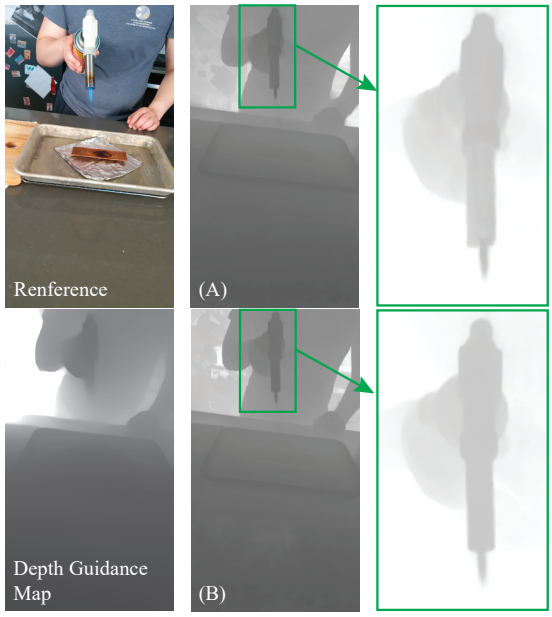

Fig. 6. Preview of the estimated depth for the chocolate scene. The top is depth optimized without depth guidance (LT-NeRF(B)), and the bottom is with depth guidance (LT-NeRF (ours)).

will only have appealing results closer to the trained location (i.e., index T = index P). When the camera position moves far away from the time and camera pose index used for training, visual artifacts will appear as shown in Figure 7. On the other hand, the synthesized novel view based on the NeRF color space is much more stable. Therefore, we choose to use the volumetrically rendered output from NeRF color space as our final output.

7 Limitations and Discussion

Our LT-NeRF approach may face challenges in certain extreme cases, such as videos with fast-moving objects, scenes with challenging camera angles, and scenarios where differentiating between the dynamic object and the background is difficult. Furthermore, while our approach outperforms HyperNeRF in cases where we achieve better-reconstructed geometry (i.e., better canonical space presentation), our results may exhibit some blurriness in cases where the geometry quality is suboptimal (as observed in Figure 8). We hypothesize that this blurriness may be due to the ambiguity of texture information in the MVS module. One feasible solution is to use subspace selection techniques [50] to identify and include only the most similar frames within a local distance, which may lead to improved texture reconstruction quality. Another limitation of our method is that training time will be significantly longer depending on the length and frame rate of the video sequence.

Fig. 7. An example comparing the output of our network's LT-module and NeRF components, where a) is at the training position, b) is moving slightly away, and c) is further away.

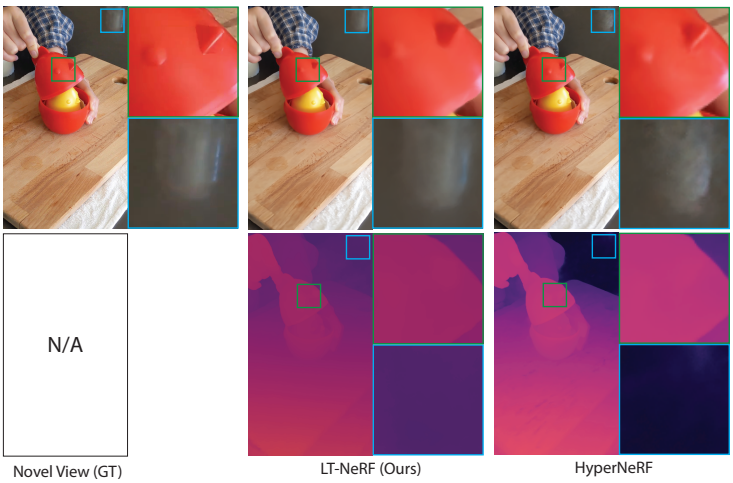

Fig. 8. Illustration of an existing issue of our method. When the reconstructed geometry is sub-optimal, our method tends to produce a blurrier result.

8 Conclusion

In this paper, we present LT-NeRF, a new approach using local temporal priors for synthesizing novel view images of dynamic scenes from monocular videos. Our novel LT module provides the local temporal priors using multi-view stereo sampling, and improves the deformation field reconstruction and hyper-space encoding of dynamic scenes. We further introduce two novel loss functions to account for our local temporal prior that improves the NeRF optimization. We tested our method with various dynamic scenes and compared the synthesized results against the state-of-the-art. The result shows that our approach with local temporal prior outperforms the prior works relying on global temporal information.

Our method still has room to improve to address the challenging cases discussed in our limitations section. Since we implemented our approach as a separate LT module, we believe it could easily be integrated into other NeRF architectures to test and improve the robustness against scenes with dynamic objects.

Acknowledgments. This work was supported by the Entrepreneurial University Programme from the Tertiary Education Commission of New Zealand.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021)
2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
3. Broxton, M., Flynn, J., Overbeck, R., Erickson, D., Hedman, P., Duvall, M., Dourgarian, J., Busch, J., Whalen, M., Debevec, P.: Immersive light field video with a layered mesh representation. ACM Transactions on Graphics (TOG) **39**(4), 86–1 (2020)
4. Buehler, C., Bosse, M., McMillan, L., Gortler, S., Cohen, M.: Unstructured lumigraph rendering. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 425–432 (2001)
5. Cohen-Steiner, D., Da, F.: A greedy delaunay-based surface reconstruction algorithm. The visual computer **20**(1), 4–16 (2004)
6. Dai, Y., Li, H., He, M.: A simple prior-free method for non-rigid structure-from-motion factorization. International Journal of Computer Vision **107**(2), 101–122 (2014)
7. Fang, J., Yi, T., Wang, X., Xie, L., Zhang, X., Liu, W., Nießner, M., Tian, Q.: Fast dynamic radiance fields with time-aware neural voxels. arXiv preprint arXiv:2205.15285 (2022)
8. Flynn, J., Broxton, M., Debevec, P., DuVall, M., Fyffe, G., Overbeck, R., Snavely, N., Tucker, R.: Deepview: View synthesis with learned gradient descent. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2367–2376 (2019)
9. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5712–5721 (2021)
10. Garg, R., Roussos, A., Agapito, L.: Dense variational reconstruction of non-rigid surfaces from monocular video. In: Proceedings of the IEEE Conference on computer vision and pattern recognition. pp. 1272–1279 (2013)

11. Habermann, M., Liu, L., Xu, W., Pons-Moll, G., Zollhoefer, M., Theobalt, C.: Hdhumans: A hybrid approach for high-fidelity digital humans. arXiv preprint arXiv:2210.12003 (2022)
12. Hedman, P., Alsisan, S., Szeliski, R., Kopf, J.: Casual 3d photography. ACM Transactions on Graphics (TOG) **36**(6), 1–15 (2017)
13. Hedman, P., Kopf, J.: Instant 3d photography. ACM Transactions on Graphics (TOG) **37**(4), 1–12 (2018)
14. Hedman, P., Philip, J., Price, T., Frahm, J.M., Drettakis, G., Brostow, G.: Deep blending for free-viewpoint image-based rendering. ACM Transactions on Graphics (TOG) **37**(6), 1–15 (2018)
15. Hedman, P., Ritschel, T., Drettakis, G., Brostow, G.: Scalable inside-out image-based rendering. ACM Transactions on Graphics (TOG) **35**(6), 1–11 (2016)
16. Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., Stamminger, M.: Volumedeform: Real-time volumetric non-rigid reconstruction. In: European conference on computer vision. pp. 362–379. Springer (2016)
17. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the fourth Eurographics symposium on Geometry processing. vol. 7 (2006)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Koenderink, J.J., Van Doorn, A.J.: Affine structure from motion. JOSA A **8**(2), 377–385 (1991)
20. Levoy, M., Hanrahan, P.: Light field rendering. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 31–42 (1996)
21. Li, Z., Niklaus, S., Snavely, N., Wang, O.: 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. pp. 6498–6508 (2021)
22. Lin, Z., Shum, H.Y.: A geometric analysis of light field rendering. International Journal of Computer Vision **58**(2), 121–138 (2004)
23. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7210–7219 (2021)
24. Mildenhall, B., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., Barron, J.T.: Nerf in the dark: High dynamic range view synthesis from noisy raw images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16190–16199 (2022)
25. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019)
26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
27. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. arXiv preprint arXiv:2201.05989 (2022)
28. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 343–352 (2015)
29. Özyeşil, O., Voroninski, V., Basri, R., Singer, A.: A survey of structure from motion*. Acta Numerica **26**, 305–364 (2017)
30. Paladini, M., Del Bue, A., Xavier, J., Agapito, L., Stošić, M., Dodig, M.: Optimal metric projections for deformable and articulated structure-from-motion. International journal of computer vision **96**(2), 252–276 (2012)

31. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021)
32. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. arXiv preprint arXiv:2106.13228 (2021)
33. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
34. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence (2020)
35. Rematas, K., Liu, A., Srinivasan, P.P., Barron, J.T., Tagliasacchi, A., Funkhouser, T., Ferrari, V.: Urban radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12932–12942 (2022)
36. Ricco, S., Tomasi, C.: Dense lagrangian motion estimation with occlusions. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1800–1807. IEEE (2012)
37. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
38. Sinha, S.N., Kopf, J., Goesele, M., Scharstein, D., Szeliski, R.: Image-based rendering for scenes with reflections. ACM Transactions on Graphics (TOG) **31**(4), 1–10 (2012)
39. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: Scalable large scene neural view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8248–8258 (2022)
40. Torresani, L., Hertzmann, A., Bregler, C.: Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. IEEE transactions on pattern analysis and machine intelligence **30**(5), 878–892 (2008)
41. Wand, M., Adams, B., Ovsjanikov, M., Berner, A., Bokeloh, M., Jenke, P., Guibas, L., Seidel, H.P., Schilling, A.: Efficient reconstruction of nonrigid shape and motion from real-time 3d scanner data. ACM Transactions on Graphics (TOG) **28**(2), 1–15 (2009)
42. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2021)
43. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003)
44. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16210–16220 (2022)
45. Wizadwongsa, S., Phongthawee, P., Yenphraphai, J., Suwajanakorn, S.: Nex: Real-time view synthesis with neural basis expansion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8534–8543 (2021)
46. Wu, X., Xu, J., Zhu, Z., Bao, H., Huang, Q., Tompkin, J., Xu, W.: Scalable neural indoor scene rendering. ACM Transactions on Graphics (TOG) **41**(4), 1–16 (2022)
47. Wu, Y., Deng, Y., Yang, J., Wei, F., Chen, Q., Tong, X.: Anifacegan: Animatable 3d-aware face image generation for video avatars. arXiv preprint arXiv:2210.06465 (2022)
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)

49. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
50. Zhu, Y., Huang, D., De La Torre, F., Lucey, S.: Complex non-rigid motion 3d reconstruction by union of subspaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1542–1549 (2014)
51. Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., et al.: Real-time non-rigid reconstruction using an rgb-d camera. ACM Transactions on Graphics (ToG) **33**(4), 1–12 (2014)