DRSM: EFFICIENT NEURAL 4D DECOMPOSITION FOR DYNAMIC RECONSTRUCTION IN STATIONARY MONOCULAR CAMERAS

Weixing Xie[1,2,3,†], Xiao Dong[4,†], Yong Yang[1], Qiqin Lin[1], Jingze Chen[1], Junfeng Yao[1,2,3,], Xiaohu Guo[5]*

[1]Center for Digital Media Computing, School of Film, School of Informatics, Xiamen University
[2]National Institute for Data Science in Health and Medicine, Xiamen University
[3]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan, Ministry of Culture and Tourism
[4]Department of Computer Science, BNU-HKBU United International College
[5]Department of Computer Science, The University of Texas at Dallas

ABSTRACT

With the popularity of monocular videos generated by video sharing and live broadcasting applications, reconstructing and editing dynamic scenes in stationary monocular cameras has become a special but anticipated technology. In contrast to scene reconstructions that exploit multi-view observations, the problem of modeling a dynamic scene from a single view is significantly more under-constrained and ill-posed. Inspired by recent progress in neural rendering, we present a novel framework to tackle 4D decomposition problem for dynamic scenes in monocular cameras. Our framework utilizes decomposed static and dynamic feature planes to represent 4D scenes and emphasizes the learning of dynamic regions through dense ray casting. Inadequate 3D clues from a single-view and occlusion are also particular challenges in scene reconstruction. To overcome these difficulties, we propose deep supervised optimization and ray casting strategies. With experiments on various videos, our method generates higher-fidelity results than existing methods for single-view dynamic scene representation.

Index Terms— Single-view Reconstruction, Dynamic Scene Reconstruction, Neural Radiance Field

1. INTRODUCTION

In recent years, the popularity of short videos and live broadcasts has led to the generation of a lot of video data, most of which are dynamic content from a single perspective of a fixed camera. We try to efficiently reconstruct and realistically render dynamic scenes in single view videos. Dynamic scenes in the video may be disturbed or obscured by other objects, such as hands (Fig. 2) and wires (Fig. 3). Our goal is to accurately recover the entire static and dynamic scene of interest to the viewer while removing occluding objects. Neural Radiation Fields (NeRF) [1] tackles novel view synthesis of static scene by learning implicit representations of objects from multiple captured views. To model dynamic scenes, many works [2] propose the ray deformation paradigms that parameterizes a deformed scene as a NeRF in canonical space with a time-dependent deformation for dynamic reconstruction [3, 4, 5, 6, 7]. Other works learn the 4D scene representation by decoupling static and dynamic scenes with different NeRFs [8, 9]. For example, D^{2}NeRF [8] achieves dynamic and static decoupling, which can remove all dynamic objects in the scene. But this does not solve the problem where we want to reconstruct dynamic and static scenes simultaneously.

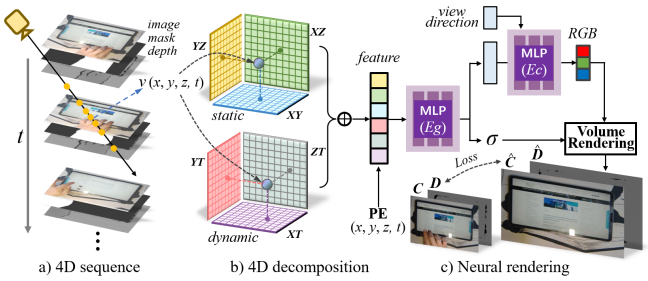


Fig. 1. Framework of the proposed DRSM.

Typically, dynamic NeRFs rely on video flow captured by multi-view cameras [10, 11] or one free-viewpoint camera [3, 4, 5, 6, 7] to get full view perception of dynamic scenes. Different with them, we aim to solve the problem of modeling dynamic scene in single view, which is ill-posed and challenging due to limited geometric perception. Many works exploit auxiliary information to help understand the structure of the scene, such as SMPL [12] prior to help constrain human motion space [13, 14] or depth prior to help recover geometry of objects [15, 16]. Among them, NDR [15] solves the geometric reconstruction of moving objects and can be modified as a background reconstruction technique to solve our problem.

Moreover, the optimization for dynamic NeRFs is com-

*Corresponding author. †Equal contribution. The research was supported by Natural Science Foundation of China (No.62072388) and public technology service platform project of Xiamen city (No.3502Z20231043).

putationally intensive since it requires multiple MLP evaluations. To avoid huge memory footprint of previous methods [17], we decouple spatial and temporal features via planar factorization [18] to model 4D field for single-view videos.

Overall, our technical contributions are as follows: 1) We propose an efficient 4D decomposition framework (DRSM) with planar factorization for fast **D**ynamic **R**econstruction in **S**tationary **M**onocular Cameras; 2) we address the inherent motion-appearance ambiguity for single-view using depth prior; 3) we propose an efficient importance sampling strategy (ISDM) based on dynamic and mask regions to improve the reconstruction quality for time-variant and occluded regions; 4) we demonstrate a convincing rendering quality and smooth point clouds on multiple short-form videos.

2. METHOD

The architecture of our network DRSM is shown in Fig. 1. We take a video $V = \{\mathbf{I}_i, \mathbf{D}_i, \mathbf{M}_i : i \in [1, T]\}$ from a single viewpoint as input, where $\mathbf{I}_i$ is the i-th frame image, $\mathbf{D}_i$ is the corresponding depth image and $\mathbf{M}_i$ is the mask of occluded objects to be removed. The object mask can be obtained by combining the Segment Anything Method (SAM) [19] with the OSTrack tracking model [20]. The video duration is normalized to [0, 1]. Thus, time of the i-th frame is i/T.

Our network starts by randomly picking a frame for training. We employ the ISDM sampling strategy to identify high-priority region and build casting rays. For sampling points along casting ray, we use bilinear interpolation to query their features on spatial and temporal tri-planes and construct the fused features, which are then passed to MLP decoders to predict color and density. We apply volume rendering to generate color and depth for each casting ray, and design rendering losses for supervision. After training, the network learns 4D representation and can reconstruct video, point cloud and synthesize novel views.

2.1. Preliminaries

NeRF [1] learns a regression function *F* that takes the encoded coordinates of a 3D point $\mathbf{x} = (x, y, z)$ observed from a view direction $\mathbf{d} = (\theta, \phi)$ as input, and outputs the corresponding radiance $\mathbf{c}$ and volume density σ: $F_{\text{NeRF}} : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$. The estimated color $\hat{C}(\mathbf{r})$ and depth $\hat{D}(\mathbf{r})$ of a pixel can be rendered by integrating the radiance by tracking a ray $\mathbf{r}(s) = \mathbf{o} + s\mathbf{d}$, cast from the camera toward the center of the pixel:

$$\hat{C}(\mathbf{r}) = \int_{s_n}^{s_f} T(s)\sigma(\mathbf{r}(s))\mathbf{c}(\mathbf{r}(s), \mathbf{d}), \tag{1}$$

$$\hat{D}(\mathbf{r}) = \int_{s_n}^{s_f} T(s)\sigma(\mathbf{r}(s))s\, ds, \tag{2}$$

$$T(s) = \exp\left(-\int_{s_n}^{s} \sigma(\mathbf{r}(p))\, dp\right). \tag{3}$$

$T(s)$ is the accumulated transmittance along the ray $\mathbf{r}$ up to s.

2.2. 4D decomposition for dynamic scenes

A dynamic scene could be naively represented as a 4D volume $\mathbf{V}$. Inspired by [21], we decompose $\mathbf{V}$ into a static volume and a dynamic volume by planar factorization:

$$\mathbf{V} = \{\mathbf{V}_s\{P_{XY}, P_{XZ}, P_{YZ}\}, \mathbf{V}_d\{P_{XT}, P_{YT}, P_{ZT}\}\}. \tag{4}$$

Here static volume $\mathbf{V}_s$ is projected to a tri-plane representing only spaces of xy, xz, and yz. The dynamic volume $\mathbf{V}_d$ is projected to a tri-plane representing spaces and time, xt, yt, and zt. Each plane has dimension $N \times N \times W$, where N is the resolution and W is the number of feature channels. This approach allows us to represent a 4D volume efficiently using six planes (Fig. 1b). For a 4D point $v = (x, y, z, t)$, we can query its features $\mathbf{f}(v)$ by projecting it onto these planes and use bilinear interpolation ψ to obtain the corresponding values:

$$\begin{aligned}\mathbf{f}(v) = \psi(P_{XY}, x, y) \odot \psi(P_{XZ}, x, z) \odot \psi(P_{YZ}, y, z) \\ \odot\psi(P_{XT}, x, t) \odot \psi(P_{YT}, y, t) \odot \psi(P_{ZT}, z, t),\end{aligned} \tag{5}$$

where $\psi(P_{XY}, x, y)$ means given regularly spaced feature plane P_{XY} and the x, y coordinates, using bilinear interpolation to calculate the plane feature of v. The $\odot$ represents Hadamard product to get fused features.

We use two small MLPs (Fig. 1c) to decode the fused features $\mathbf{f}(v)$ like Instant-NGP [22]. The features and positional encoding are concatenated and fed into the geometry MLP E_g to obtain density σ and high dimensional features $\mathbf{f}'(v)$:

$$\sigma(v), \mathbf{f}'(v) = E_g(\mathbf{f}(v), \gamma(v)). \tag{6}$$

Here, $\gamma(\cdot)$ is an encoding function [1]. Then, we concatenate the feature $\mathbf{f}'(v)$ with the positional encoding of view direction (θ, ϕ) and feed it into the color MLP E_c to obtain the radiance:

$$\mathbf{c}(r, g, b) = E_c(\mathbf{f}'(v), \gamma(\theta, \phi)). \tag{7}$$

2.3. ISDM sampling strategy

Previous scene representation methods [2] usually randomly sample a batch of pixels/rays on the whole input image for training. In our work, we focus on learning the representation for dynamic scene of interest while removing occluding objects. Uniform sampling is no longer suitable for our method because dynamic areas and occluded areas require higher sampling weights.

We propose the importance sampling strategy based on dynamic and mask regions. For the occlusion mask $\mathbf{M}_i$ of frame i (0 for occluded pixels), we ignore those pixels in occluded region in the ray selection. We create an importance map $\tilde{\mathbf{P}}_i$ to guide the pixel sampling, assigning higher probability for those regions with higher occlusion frequencies

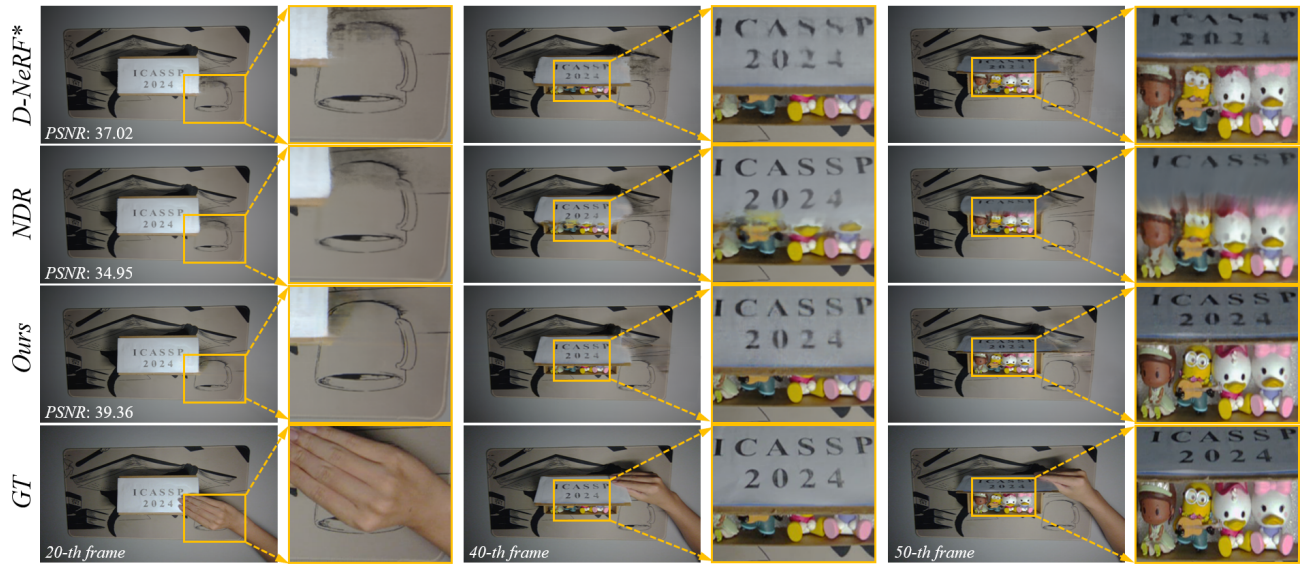


Fig. 2. Comparison of DRSM and other methods on dynamic reconstruction results. We remove the hand in video and show *PSNR* metric of each method.

across all frames. The sampling importance map is calculated according to element-wise division:

$$\tilde{\mathbf{P}}_i = \mathbf{M}_i T / (\sum_{k=1}^{T} \mathbf{M}_k + \varepsilon). \tag{8}$$

In addition to occlusion areas, we should also prioritize sampling dynamic areas. In uniform sampling, a large proportion of selected pixels may fall into the static background, which contributes less to the dynamic reconstruction. To identify the dynamic region, we calculate temporal difference of pixels on frames i and j [11]:

$$\mathbf{P}_i = \tilde{\mathbf{P}}_i \odot \min(\frac{1}{3} \left\| \mathbf{I}_i - \mathbf{I}_j \right\|_1, \alpha), j \in (i-\tau, i+\tau), \tag{9}$$

where α is a lower-bound parameter controlling the sampling weights of the dynamic region and τ is set to 25 in the experiment. ISDM sampling adjusts the sampling probability of time-varying and occlusion areas, which helps improve reconstruction quality and speed up training.

2.4. Optimization

We supervise scene reconstruction in terms of reconstructed image $\hat{C}$, depth $\hat{D}$, and regularization loss to optimize the parameters of feature planes and MLPs. For each batch of training data, there are R rays sampled by ISDM strategy on one frame. We first minimize the difference between the ground truth color and the predicted color, as shown in Eq. (1). To assist in scene representation for single-view input, we further optimize the geometry using depth supervision. The color loss and depth loss are shown in the following equations:

$$\mathcal{L}_{\text{color}} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} \left\| C(\mathbf{r}) - \hat{C}(\mathbf{r}) \right\|_2^2, \tag{10}$$

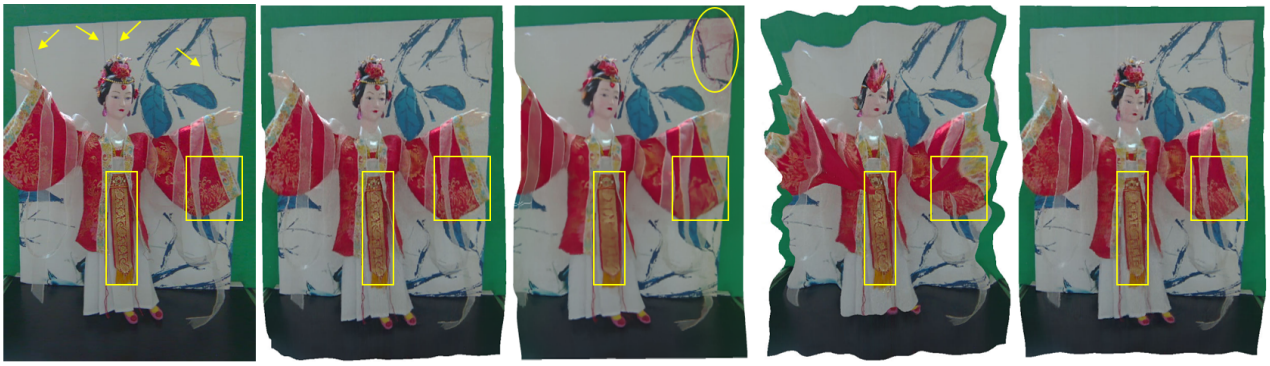
$$\mathcal{L}_{\text{depth}} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} \left\| D(\mathbf{r}) - \hat{D}(\mathbf{r}) \right\|_2^2. \tag{11}$$

Dynamic scene reconstruction in stationary monocular camera is a severely ill-posed problem. To achieve robust reconstruction, we apply strong regularizers. We adopt 2D total variation (TV) loss $\mathcal{L}_{\text{TV-2D}}$ for space planes in [22] and 1D TV loss $\mathcal{L}_{\text{TV-1D}}$ on the space axis for space-time planes and a similar smooth loss $\mathcal{L}_{\text{smooth}}$ on the time axis. The total optimization objective is:

$$\mathcal{L} = \mathcal{L}_{\text{color}} + \lambda_1 \mathcal{L}_{\text{depth}} + \lambda_2 \mathcal{L}_{\text{2D}} + \lambda_3 \mathcal{L}_{\text{1D}} + \lambda_4 \mathcal{L}_{\text{smooth}}. \tag{12}$$

3. EXPERIMENTS

Experimental settings. We normalize the scene into device coordinates (NDC) to handle monocular videos and then sample casting rays within the NDC space. We use a model with four symmetric spatial resolutions 64, 128, 256 and 512. The feature length W at each scale is set to 32. We set the frequencies of positional encoding $\gamma(\cdot)$ for sampling points and view direction to 4. In each training iteration, a batch contains $\mathcal{R} = 2048$ sampling rays. The loss weights in Eq.(12) are empirically set as $\lambda_1 = 1.0$, $\lambda_2 = 0.0002$, $\lambda_3 = 0.0001$, $\lambda_4 = 0.001$. Adam [23] optimizer is adopted for training, and the initial learning rate is set to 0.01. We train all scenes with $5k$ and $10k$ iterations on a single RTX 3090 GPU, which take around 15 and 35 minutes, respectively. We build a video dataset, including life videos related to box, marionette, web page, xiangqi, calligraphy and toy. We use the "Record3D" app on iPhone and RGBD camera to record videos. Each video lasts for $5 \sim 7$ seconds and we sample 10 frames per second for training.



Reference(*PSNR*) a) Complete model(*36.42*) b) w/o Planar factorization (*33.62*) c) w/o Depth supervision (*35.32*) d) w/o ISDM sampling (*35.87*)

Fig. 3. Ablation study on a marionette dancing video. We remove manipulating wires and show the reconstructed point clouds.

Table 1. Quantitative comparisons on our collected dataset. We report PSNR↑, SSIM↑, LPIPS↓ and training time (minutes).

Model	"Box"			"Marionette"			"Web page"			"Xiangqi"			"Calligraphy"			"Toy"			Time
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	
D-NeRF*	37.02	0.945	0.084	33.62	0.885	0.058	35.80	0.947	0.068	35.21	0.959	0.049	36.27	0.911	0.095	37.14	0.944	0.088	477min
NDR	34.95	0.942	0.094	33.74	0.909	0.053	35.37	0.944	0.086	35.24	0.955	0.062	35.02	0.890	0.139	35.03	0.940	0.097	666min
Ours-5K	38.07	0.945	0.071	34.12	0.918	0.044	37.79	0.954	0.047	35.20	0.930	0.048	37.20	0.928	0.073	37.05	0.942	0.089	**15min**
Ours-10K	**39.36**	**0.953**	**0.070**	**36.42**	**0.945**	**0.023**	**39.38**	**0.964**	**0.043**	**37.27**	**0.964**	**0.037**	**38.51**	**0.943**	**0.058**	**38.42**	**0.950**	**0.079**	35min

Comparison experiments. We compare our method with other dynamic scene reconstruction methods for monocular videos, such as D-NeRF [3] and NDR [16]. D-NeRF builds a deformable neural radiance field based on a canonical 3D representation and time-guided motion fields. However, the model performance of D-NeRF depending on a canonical frame suffers when objects exhibit long-distance translations [24]. NDR focuses on modeling dynamic foreground objects based on bijective motion map and implicit representations of MLPs. Using an MLP with a specific bandwidth to learn both spatial and temporal variations simultaneously results in suboptimal reconstruction of complex scenes.

As shown in Fig. 2, D-NeRF* and NDR are modified version of original models with depth supervision for fair comparison with our method. D-NeRF* failed to capture the deformation of long-distance moving objects, i.e., the characters on the box. The predicted color of the doll predicted by NDR is affected by the movement of the box. This is because NDR's bijective map focuses on learning the geometric changes of moving objects not the high-frequency details of static part. Our network is specifically designed for the reconstruction of combined static and dynamic scenes, resulting in better video appearance reconstructions.

In Table 1, we show the indicators such as PSNR, SSIM and LPIPS of 6 videos to quantitatively evaluation the reconstruction. Our model outperforms the existing methods on multiple aspects and requires shorter training time.

Ablation study. We present ablation experiments on network modules and the reconstructed point clouds in Fig. 3. The marionette dancing video contains some manipulating wires to be removed. Without planar factorization, our network failed to reconstruct high quality texture details in static region (color prediction error in yellow ellipse) as well as dynamic region (sleeves and decorations in yellow box). Furthermore, we observe severe distortions in the reconstructed point cloud when depth supervision is disabled, indicating that the network is unable to learn the correct geometry from single-view input without prior. Without ISDM sampling, predicted high-frequency textures also become blurry. We provide the corresponding PSNR indicator to further demonstrate the effectiveness of proposed modules. Our complete model produces high-fidelity reconstructions.

4. CONCLUSION

This paper presents a novel neural 4D decomposition for dynamic reconstruction from single-view videos. Without observation from multi-viewpoints, the problem of modeling dynamic scenes is typically quite challenging. We apply planar decomposition to static and dynamic scenes respectively to improve the model's modeling ability of 4D scenes. To address the ambiguous geometry, we utilize depth prior to constrain the motion space. The adaptive sampling strategies aid the reconstruction on moving objects and occluding regions. The ablation study demonstrates the effectiveness of proposed network. We conduct rich experiments to show the superiority of our network than existing methods on the special task of single-view dynamic scene construction.

5. REFERENCES

[1] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[2] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa, "Monocular dynamic view synthesis: A reality check," *NeurIPS*, 2022.

[3] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *CVPR*, 2021.

[4] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla, "Nerfies: Deformable neural radiance fields," *ICCV*, 2021.

[5] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang, "Dynamic view synthesis from dynamic monocular video," in *ICCV*, 2021.

[6] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang, "Neural scene flow fields for space-time view synthesis of dynamic scenes," in *CVPR*, 2021.

[7] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt, "Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video," in *ICCV*, 2021.

[8] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli, "D^2NeRF: Self-supervised decoupling of dynamic and static objects from a monocular video," *NeurIPS*, 2022.

[9] Boyu Zhang, Wenbo Xu, Zheng Zhu, and Guan Huang, "Detachable novel views synthesis of dynamic scenes using distribution-driven neural radiance fields," *arXiv preprint arXiv:2301.00411*, 2023.

[10] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou, "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *CVPR*, 2021.

[11] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al., "Neural 3d video synthesis from multi-view video," in *CVPR*, 2022.

[12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[13] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman, "Humannerf: Free-viewpoint rendering of moving people from monocular video," in *CVPR*, 2022.

[14] Hao Zhao, Jinsong Zhang, Yu-Kun Lai, Zerong Zheng, Yingdi Xie, Yebin Liu, and Kun Li, "High-fidelity human avatars from a single rgb camera," in *CVPR*, 2022.

[15] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim, "Space-time neural irradiance fields for free-viewpoint video," in *CVPR*, 2021.

[16] Hongrui Cai, Wanquan Feng, Xuetao Feng, Yan Wang, and Juyong Zhang, "Neural surface reconstruction of dynamic scenes with monocular rgb-d camera," *NeurIPS*, 2022.

[17] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa, "Plenoctrees for real-time rendering of neural radiance fields," in *ICCV*, 2021, pp. 5752–5761.

[18] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al., "Efficient geometry-aware 3d generative adversarial networks," in *CVPR*, 2022.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[20] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen, "Joint feature learning and relation modeling for tracking: A one-stream framework," in *ECCV*, 2022.

[21] Ang Cao and Justin Johnson, "Hexplane: A fast representation for dynamic scenes," in *CVPR*, 2023.

[22] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (ToG)*, vol. 41, no. 4, pp. 1–15, 2022.

[23] D Kinga, Jimmy Ba Adam, et al., "A method for stochastic optimization," in *ICLR*. San Diego, California;, 2015.

[24] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Anton Van Den Hengel, "Bali-rf: Bandlimited radiance fields for dynamic scene modeling," *arXiv preprint arXiv:2302.13543*, 2023.