DOI: 10.1111/cgf.14646
ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2022
D. L. Michels and S. Pirk
(Guest Editors)

Volume 41 (2022), Number 8

MP-NeRF: Neural Radiance Fields for Dynamic Multi-person synthesis from Sparse Views

X. J. Chao[†1] and H. Leung[1]

[1]City University of Hong Kong, Hong Kong

Abstract

Multi-person novel view synthesis aims to generate free-viewpoint videos for dynamic scenes of multiple persons. However, current methods require numerous views to reconstruct a dynamic person and only achieve good performance when only a single person is present in the video. This paper aims to reconstruct a multi-person scene with fewer views, especially addressing the occlusion and interaction problems that appear in the multi-person scene. We propose MP-NeRF, a practical method for multi-person novel view synthesis from sparse cameras without the pre-scanned template human models. We apply a multi-person SMPL template as the identity and human motion prior. Then we build a global latent code to integrate the relative observations among multiple people, so we could represent multiple dynamic people into multiple neural radiance representations from sparse views. Experiments on multi-person dataset MVMP show that our method is superior to other state-of-the-art methods.

Keywords: dynamic human, multi-person, view synthesis, volume rendering, 3D deep learning

1. Introduction

In this paper, we propose MP-NeRF that aims to achieve free-viewpoint rendering of multi-person video from sparse views, which has promising potential applications in AR/VR, immersive gaming, interactive telepresence, movie productions, and sports broadcasting. Despite significant successes brought by traditional computer graphics techniques showing impressive performance, synthesizing photorealistic video content still requires expensive capture equipment such as dense cameras, pre-scanned templates, commercial RGBD sensors, and repetitive human labor for 3D content creation.

Among previous free-viewpoint video generation of human-centric dynamic scenes, reconstruction solutions [MKGH16, CCS*15] model high-fidelity structure and render texture with novel views by a multi-view dome-based setup. However, the synthesis results depend on the reconstructed mesh resolution and fail to capture large-scale captured scenes. On the other hand, image-based rendering (IBR) techniques [GGSC96, CTMS03, ZKU*04] require densely captured viewpoints to interpolate textures in novel views. The free-view results, however, are vulnerable to occlusions and suffer from limited views that can be interpolated from the densely captured views, leading to abnormal texture artifacts. Benefiting from the improvement of neural rendering techniques [TFT*20], recent data-driven methods [MGK*19a, TZT*18, WWHY20a, MST*20, BMSR20] are able to generate photorealistic free-viewpoint video for static scenes without reliance on 3D template. Specifically, recent work [PSB*20, TTG*20, RJY*20, OMT*20] also reconstruct the dynamic scene via neural radiance field .

Despite the remarkable performance in the free-viewpoint synthesis of the dynamic scenes, the above solutions still suffer from several drawbacks when extending to human dynamic scenarios. Recent work Peng et al. [PZX*21] solve the above problems by incorporating smpl-guided latent codes over video frames to achieve human synthesis and non-rigid manipulation from sparse views. However, this method is limited to single-person synthesis. When extended to multiple people, this method requires a cumbersome process for each person to be handled separately, and at the same time, it failed to handle the problem of occlusion between people. Recent work ST-NeRF [ZLY*21] synthesizes multiple entities including human performers by neural layered representation. However, ST-NeRF still requires 16 cameras and is unable to manipulate the reconstructed human to act in new poses.

Generating free views containing multiple people from sparse views with occlusion handling is a challenging task. To tackle the above problems, we propose MP-NeRF which is the first approach to synthesize photorealistic free-viewpoint videos of multi-person scenes from sparse views, utilizing only 4 cameras to cover a view range up to 360 degrees. Compared with the extension of the traditional single-person method to multiple people, each person needs to be processed individually and the interaction between people is difficult to simulate naturally. The key idea of our work is to build

† Chairman Eurographics Publications Board

a global latent code that illustrates the relative location and person identification to control the spatial location of latent codes on different human SMPL surfaces [LMR*15]. To this end, our method firstly adopts a multi-way matching algorithm for multiple persons' 3D pose estimation stage, which intuitively reflects the spatial information from the coarse relative distance and SMPL parameter among multiple people. At the same time, to use the corresponding temporal information we also need to track the same person in long videos containing multiple people. Instead of learning each person separately, we build an identity-aware global latent code as the multi-person implicit representations by anchoring a set of latent codes to the vertices of multi-person SMPL templates, which are utilized to express the multiple human body surfaces and relative distances.

To address the occlusion among multiple persons from sparse views, a novel occlusion-aware neural radiance field is proposed. We construct multiple volumes with an affine transformation to represent multiple people in a scene. Specifically, given point X and viewing direction D, we design a combination of different volumes by selecting the maximum value of color and density among all volumes when occlusion appears. Representing the multi-person scene as compositional neural radiance fields is conducive to extending our multi-person synthesis to various content creations by editing each person.

To summarize, the main contributions of our work include:

- We are the first to build dynamic multi-person free-view synthesis tasks from sparse views. This work can achieve photorealistic multi-person static scene synthesis and generalization of new actions in dynamic scenes.
- We introduce a global implicit code with identity information to encode the relative location for multiple people in a frame and the identity for each person among frames. Thus, a series of implicit codes are mapped to the implicit field of density and color in different frames which naturally integrates multi-frame and multi-person observation.
- We propose a novel occlusion-aware neural radiance field that combines multiple people in a scene with an affine transformation. The density and color of each point obtained by the neural network regression are then extended between multiple people.

2. Related Work

2.1. Image-based rendering.

Image-based rendering aims to synthesize free views without detailed 3D geometry reconstruction. To obtain novel views from densely sampled images, some methods [GGSC96, DLD12] utilize light field interpolation. While the render results look impressive, the range of views that could be rendered is limited. To tackle this problem, [CDSHD13, PZ17] regard depth maps from input images as proxy geometries. They adopt the depth maps to warp observed images into the novel view and conduct image blending. However, these approaches are sensitive to the accuracy of 3D proxy geometries. Instead of hand-crafted parts of the image-based rendering pipeline, [KWR16, HPP*18, CGT*19] adopt the learnable counterparts to improve the robustness.

2.2. Implicit neural representation

With the development of differentiable rendering, recent methods apply the deep neural networks in scene representations learning from 2D RGB images with differentiable renderers, such as voxels [STH*19, LSS*19], textured meshes [TZN19, LXZ*19, LSMG20], point clouds [WWHY20b, ASK*20], multi-plane images [ZTF*18, FBD*19], and implicit functions [SZW19, LZP*20, NMOG20, MST*20, LGZL*20]. Specifically, SRN [SZW19] introduces an implicit neural representation that maps 3D coordinates to feature vectors. In a further step, a differentiable ray marching algorithm is adopted to render 2D feature maps, which are then interpreted into images with a pixel generator. Moreover, NeRF [MST*20] represents the whole scenes with implicit fields of color and density, which achieves photorealistic view synthesis results from dense views. Instead of utilizing dense view camera settings and representing the whole scene with a single implicit function, our approach introduces global latent codes for multiple people, which are used with a network to record the multi-person identification and multi-person geometry and appearance. Moreover, our method adopts multi-person neural radiance fields and compositions to represent each person more flexibly and handle the problem of multi-person occlusion.

2.3. Human performance capture

Existed approaches [NFS15, CCS*15, DKD*16, GLD*19] adopt the traditional pipeline to synthesize novel views of humans. The limitation is that previous methods depend on either RGBD sensors [CCS*15, DKD*16, SXZ*20] or a dense view camera setting [DHT*00, GLD*19] to obtain the high quality reconstruction. [MBPY*18, MGK*19b, WWHY20b] replace the traditional rendering methods with neural networks to deal with the geometric artifacts. To obtain human models in sparse multi-view cameras, template-based methods [CTMS03, DAST*08, GSDA*09, SGDA*10] utilize the pre-scanned human models as the guidance. Given the deformed template shapes, these methods could reconstruct dynamic humans. However, the deformed geometries tend to be distorted, and pre-scanned human shapes are uneasy to obtain. Furthermore some monocular methods [NSH*19, SHN*19, ZYW*19, SSSJ20] utilize the data-driven network to obtain the human prior, which makes it possible for them to reconstruct 3D human geometry and appearance from only a single image. Unfortunately, these methods fail to achieve photorealistic view synthesis and multi-person capture from sparse views.

3. Multi-person Neural Radiance Field

3.1. Overview

Our task aims to generate a free-viewpoint video of multiple performers by utilizing sparse multi-view pre-calibrated RGB cameras. The multi-view video is denoted as $\{\mathcal{I}_t^i | i = 1, ..., N_i, t = 1, ..., N_t\}$, where i is the camera index, N_i is the number of cameras, t is the frame index, and N_t is the number of frames. For each view, we utilize [GLL*18] to obtain the foreground multi-person mask and set the values of the background image pixels as zero similar to [PZX*21]. Figure 1 illustrates the pipeline of our method. Firstly,

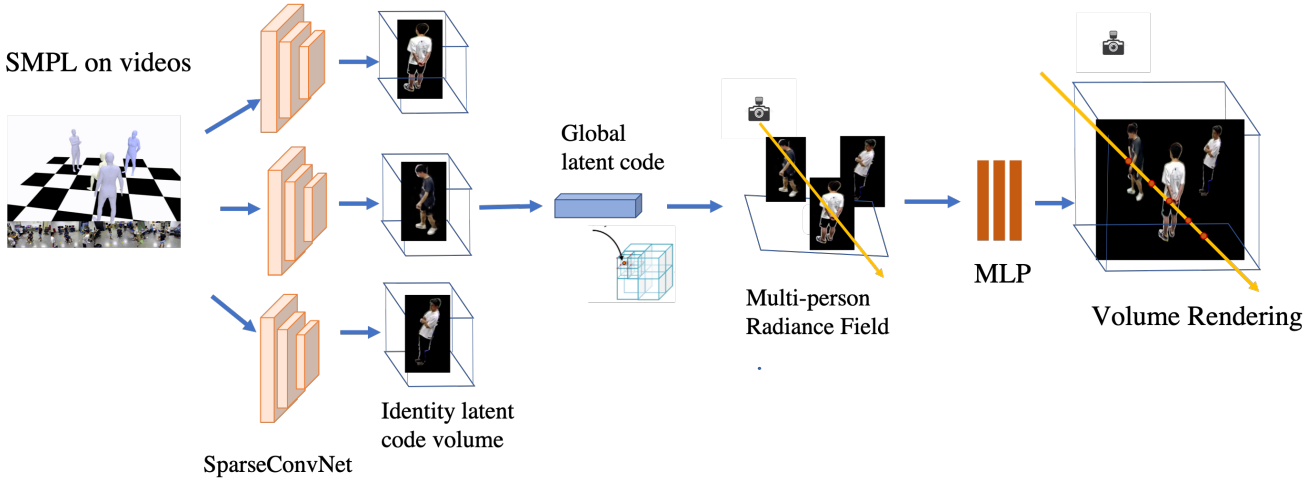

Figure 1: *Pipeline of our MP-NeRF approach.*

an off-the-shelf 3D multi-person pose estimation is adopted to predict the 3D keypoints of each person from the multi-view video. To obtain the identification of each person among the video frame, we track the different persons and interpolate missing frames. As a result, we could fit the SMPL model on the tracked 3D keypoints to the identity-aware multi-person SMPL template. Given the multi-person SMPL template, we build the global latent variable model. The obtained SMPL template with identity information needs to go through the code diffusion process to obtain the global position latent codes near the surface (Section3.2). Then we extend the density and color of each point in the 3D space by neural network regression among multiple people (Section 3.3). Finally, the image of any view can be obtained by volume rendering (Section 3.4). The whole network is obtained by minimizing the rendered image and input images for training. (Section 3.5)

3.2. Identity-aware structured latent codes

To control the spatial location of latent codes on different human body surfaces, we need to obtain human body templates (SMPL) with identity information in multi-person scenarios. Different from the previous multi-person performance capture methods which independently predict each person's SMPL and then combine them via post-processing, we apply a multi-way matching algorithm [DJH*19] to estimate the multi-person 3D keypoints together. This method could better reflect the relative positions and actions of multiple people. At the same time, compared with the previous method which requires multiple inferences corresponding to multiple people [ZSZ*21], our method only needs to infer once. This is especially advantageous when the number of people increases. To obtain the multi-person 3D pose estimation, we firstly utilize the Cascaded Pyramid Network [CWP*18] to get the 2D keypoints location of each people. Then we match the detected 2D poses in different views by clustering the 2D bounding boxes which belong to the same person. Moreover, multi-person 3D keypoints position could be effectively inferred by the resulting cluster which encodes the consistent correspondence across 2D observations among multiple views. On the other hand, we also need to track the same person among multiple people in different video frames. Thus we not only obtain the coarse position and shape of the multiple people in each frame, but we also obtain the corresponding identities of each person between frames. Then, we could build the identity-aware structured latent codes. Structure latent codes for a single person are similar to the local implicit representations [JSM*20, CLI*20, GCS*20] which are used to constrain the codes to express the dynamic human body. For multi-person scenarios, we build a global latent codes $\mathcal{Z} = \{Z_1, Z_2, ..., Z_i\}$ on vertices of the multi-person SMPL model, where i represents the number of person and $Z_i = \{z_1, z_2, ..., z_{6890}\}$. However, the structured latent codes are only applied on the vertices of the SMPL model, which are too sparse that most 3D points are zero vectors after interpolating for continuous 3D locations. We apply the SparseConvNet [GEVDM18] to diffuse the sparse structure latent codes define on the SMPL vertices to nearby 3D space for better trilinear interpolation. Different from single structure latent code for single person [PZX*21], our global latent code integrates the relative location of multi-person in a frame and the identity of each person among frames. Thus, a series of implicit codes are mapped to the implicit field of density and color in different frames which naturally integrates multi-frame and multi-person observation.

3.3. Multi-person neural radiance fields and compositions

The core of our method is a novel multiple neural volume representation for multi-person photo-realistic novel view synthesis. Neural Radiance Fields (NeRFs) [MST*20] parameterize the continuous function f with a multi-layer perceptron(MLP) which maps the 3D point $x \in R^3$ and a viewing directon $d \in S^2$ to an RGB color value $c \in R^3$ and a volume density $\sigma \in R^+$:

$$f_\theta : R^{L_\mathbf{x}} \times R^{L_\mathbf{d}} \rightarrow R^+ \times R^3 \quad (1)$$

$$(\gamma(\mathbf{x}), \gamma(\mathbf{d})) \mapsto (\sigma, \mathbf{c}) \quad (2)$$

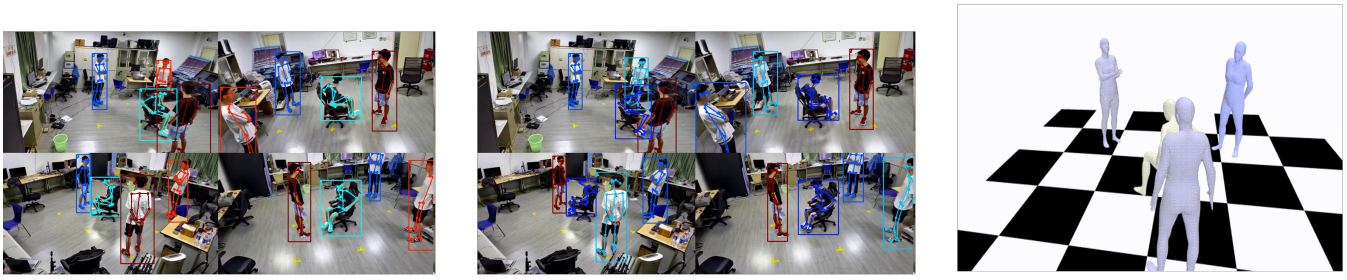

Figure 2: *Multi-person SMPL reconstruction.*

where θ illustrates the network parameters, $\gamma_{\mathbf{d}}$ and $\gamma_{\mathbf{x}}$ are positional encoding functions for viewing direction and spatial location. $L_{\mathbf{x}}, L_{\mathbf{d}}$ are the output dimensionalities of the positional encodings applied element-wise to each component of x and d:

$$\gamma(t,L) = (\sin(2^0 t\pi), \cos(2^0 t\pi), \ldots, \sin(2^L t\pi), \cos(2^L t\pi)) \quad (3)$$

where t represents a scalar input, a component of $\mathbf{x}$ or $\mathbf{d}$, and L the number of frequency octaves. However, the pure NeRFs [MST*20,PZX*21] require dense views to synthesize the free-view scene and fail to generalize to the dynamic scenes such as new poses for the person. To tackle this problem, we apply the identity-aware structure latent code vector as the input of MLP networks to predict the color and density of points in 3D space.

For the frame t, the volume density at point $\mathbf{x}$ is predicted as a function of only the latent code $\psi(\mathbf{x}, \mathcal{Z}, S_t)$, which is defined as:

$$\sigma_t(\mathbf{x}) = M_\sigma(\psi(\mathbf{x}, \mathcal{Z}, S_t)), \quad (4)$$

where M_σ represents an MLP network with four layers.

Similar to [LSS*19, MST*20], we take both the latent code $\psi(\mathbf{x}, \mathcal{Z}, S_t)$ and the viewing direction $\mathbf{d}$ as input for the color regression. To model the location-dependent incident light, the color model also takes the spatial location $\mathbf{x}$ as input. We observe that temporal variation factors affect the human appearance, such as secondary lighting and self-shadowing. Inspired by the auto-decoder [PFS*19], we assign a latent embedding ℓ_t for each video frame t to encode these temporal variation factors.

Specifically, for the frame t, the color at $\mathbf{x}$ is predicted as a function of the latent code $\psi(\mathbf{x}, \mathcal{Z}, S_t)$, the viewing direction $\mathbf{d}$, the spatial location $\mathbf{x}$, and the latent embedding ℓ_t. Following [RBA*19, MST*20], we apply the positional encoding to both the viewing direction $\mathbf{d}$ and the spatial location $\mathbf{x}$, which enables better learning of high frequency functions. The color model at frame t is defined as:

$$\mathbf{c}_t(\mathbf{x}) = M_{\mathbf{c}}(\psi(\mathbf{x}, \mathcal{Z}, S_t), \gamma_{\mathbf{d}}(\mathbf{d}), \gamma_{\mathbf{x}}(\mathbf{x}), \ell_t), \quad (5)$$

where $M_{\mathbf{c}}$ represents an MLP network with two layers, and $\gamma_{\mathbf{d}}$ and $\gamma_{\mathbf{x}}$ are positional encoding functions for viewing direction and spatial location, respectively. We set the dimension of ℓ_t to 128 in experiments.

At the same time, the pure NeRFs meet another limitation that they could only represent the entire scene but not disentangle different entities in the scene [NG21]. Compared with the method of reconstructing the whole scene by a single volume, our method is able to disentangle each person by constructing multiple volumes inspired by [NG21]. Then the synthesis of different actions of characters can be realized, which can be further extended to the editing of multiple people, thus providing a wider range of application scenarios for our method. Specifically, we represent each person using a separate volume in combination with an affine transformation

$$\mathbf{T} = \{\mathbf{t}, \mathbf{R}\} \quad (6)$$

where $\mathbf{t} \in R^3$ represent translation parameters, and $\mathbf{R} \in SO(3)$ a rotation matrix. Using this representation, we transform points from object to scene space as follows: $\mathrm{k(x)} = \mathrm{R} \cdot \mathbf{x} + \mathbf{t}$ This allows us to arrange multiple people in a scene. Since the multi-person scene will encounter the problem of occlusion between people, if only each person is reconstructed separately, it will not reflect the influence of other people on the person. To solve this problem, we query the color and density of each point in each volume and then we combine them to get the final color and density of the point. Compared with a single-person method, our method not only considers the color and density state of a single person in space but also considers the influence of other people, so it can obtain better free-view synthesis especially when there exit occlusion situation among multiple people.

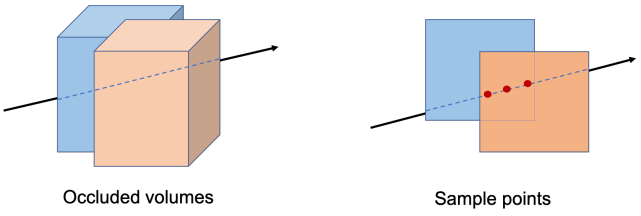

Figure 3: *Occlusion-aware neural radiance field.*

3.4. Multi-person Rendering

Through volume rendering technology, multi-person free-view 2D images could be rendered with 2D supervision. Therefore, there is

no need to use multi-person 3D scan data as supervision. The pixel colors are calculated via the volume rendering integral equation [KVH84] which accumulates volume colors and densities along the corresponding camera ray. Specifically, the integral could be approximated via numerical quadrature [Max95, MST*20]. Given a pixel, we first compute its camera ray $\mathbf{r}$ using the camera parameters and sample N_k points $\{\mathbf{x}_k\}_{k=1}^{N_k}$ along camera ray $\mathbf{r}$ between near and far bounds. The scene bounds are calculated from the SMPL model. Moreover, MP-NeRF predicts volume densities and colors at these points. For the video frame t, the rendered color $\tilde{C}_t(\mathbf{r})$ of the corresponding pixel is given by:

$$\tilde{C}_t(\mathbf{r}) = \sum_{k=1}^{N_k} T_k(1-\exp(-\sigma_t(\mathbf{x}_k)\delta_k))\mathbf{c}_t(\mathbf{x}_k), \tag{7}$$

$$where \quad T_k = \exp(-\sum_{j=1}^{k-1} \sigma_t(\mathbf{x}_j)\delta_j), \tag{8}$$

where $\delta_k = ||\mathbf{x}_{k+1} - \mathbf{x}_k||_2$ is the distance between adjacent sampled points. We set N_k as 64 in all experiments. With volume rendering, our model is optimized by comparing the rendered and observed images.

3.5. Training

Through the volume rendering techniques, we optimize the MP-NeRF to minimize the rendering error of observed images $\{\mathcal{I}_t^c | c = 1,...,N_c, t = 1,...,N_t\}$:

$$\{\ell_t\}_{t=1}^{N_t}, \mathcal{Z}, \Theta minimize \sum_{t=1}^{N_t}\sum_{c=1}^{N_c} L(\mathcal{I}_t^c, P^c; \ell_t, \mathcal{Z}, \Theta), \tag{9}$$

where Θ means the network parameters, P^c is the camera parameters, and L is the total squared error that measures the difference between the rendered and observed images. The corresponding loss function is defined as:

$$L = \sum_{\mathbf{r}\in\mathcal{R}} \tilde{C}(\mathbf{r}) - C(\mathbf{r}), \tag{10}$$

where $\mathcal{R}$ is the set of camera rays passing through image pixels, and $C(\mathbf{r})$ means the ground-truth pixel color. In contrast to frame-wise reconstruction methods [SF16,MST*20], our method optimizes the model using all images in the video and has more information to recover the 3D structures.

We adopt the Adam optimizer [KB14] for training the MP-NeRF. The learning rate starts from $5e^{-4}$ and decays exponentially to $5e^{-5}$ along the optimization. The training on a four-view video of 300 frames typically takes around 200k iterations to converge (about 14 hours).

4. Experiments

4.1. Implementation detail

We evaluate our method on a multi-view dataset called MVMP [DJH*19]. This dataset captures 4 dynamic human indoor videos via a multi-camera system which has only 8 synchronized cameras arranged around the 4 people to cover the viewing range up to 360 degrees. Each camera is calibrated and provides the RGB video at 640 × 360 resolution and 25 frames-per-second. Our multi-person model is implemented in PyTorch and trained on an NVIDIA Tesla V100 GPU.

Table 1: *Quantitative results on seen pose among 300 frames*

	training views			testing views		
model	psnr	mse	ssim	psnr	mse	ssim
NB	25.0683	0.0016	0.9045	14.6731	0.0610	0.5569
Ours	**27.7243**	**0.0015**	**0.9598**	**19.5329**	**0.0114**	**0.8461**

Table 2: *Quantitative results on new pose among 300 frames*

	training views			testing views		
model	psnr	mse	ssim	psnr	mse	ssim
NB	23.6890	0.0015	0.9187	10.3782	0.091	0.5817
Ours	**25.0206**	**0.0014**	**0.9350**	**19.2073**	**0.0127**	**0.8205**

4.2. Metrics

Similar to other novel view synthesis work [ZLY*21, MST*20, PZX*21], we calculate the corresponding peak signal-to-noise ratio (PSNR), mean-square error (MSE), and structural similarity index (SSIM) as quantitative metrics. Moreover, we provide the visualizations as a qualitative comparison.

4.3. Comparsions

Since we are the first to do multi-person novel view synthesis under sparse views, we compare our method with NeuralBody [PZX*21] which is a single-person novel view synthesis method. For NeuralBody, multi-person reconstruction needs to segment each person to obtain single-person SMPL and then decode the single-person local latent code volume to color and density values. As shown in Table 1, we compare our method with NeuralBody(NB) of synthesizing seen poses on training views and testing views. The results show that our method has higher performance with all three metrics. Moreover, we also compare the results of synthesizing new(unseen) poses on both training views and testing views as illustrated in Table 2. For all metrics, our method achieves better performance. The reason is that the training process of each person for NeuralBody is separate. In contrast, our method allows sharing parameters among different people and training multiple people together to improve the representation learning of the network. We also provide qualitative results of our method and NeuralBody(NB). As shown in Fig. 3, we compare the reconstruction results of NB and our method under training viewpoints. In the part where multiple people interact closely with each other, our method outperforms NB. At the same time, as shown in Fig. 4, we compare the synthesis results of NB and our method under the testing view. Our method also outperforms the single-person reconstruction method.

5. Ablation studies

We conduct ablation studies to explore the performances of our methods in a different number of camera views. Moreover, we also compare the effects of different video lengths for the multi-person novel view synthesis.

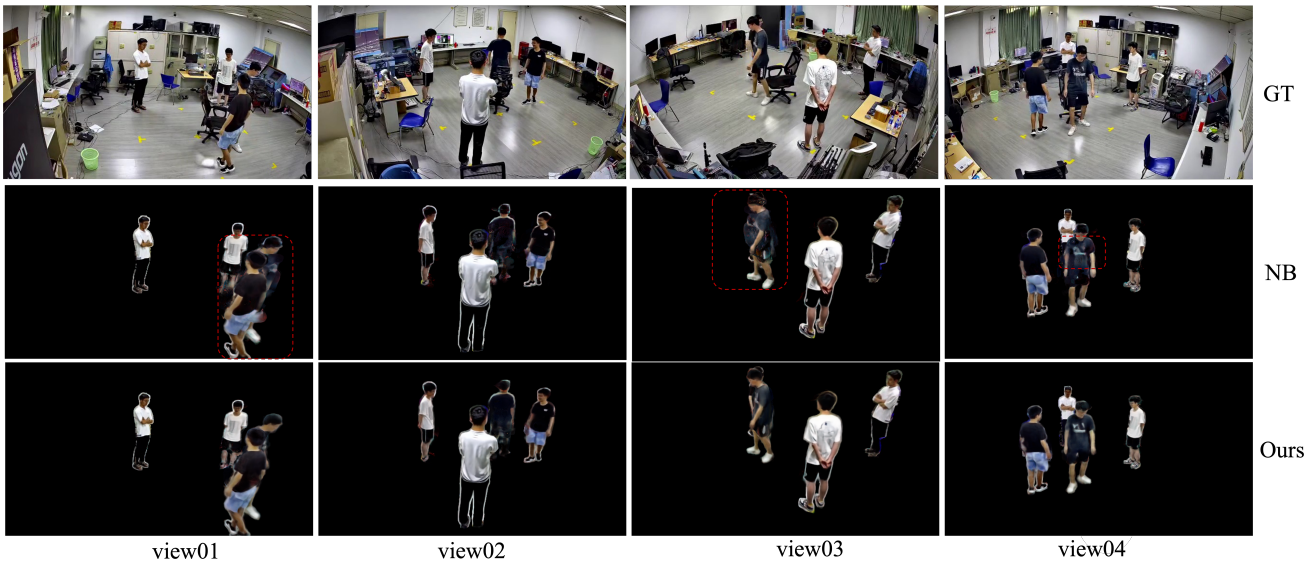

Figure 4: *Comparisons on seen poses.*

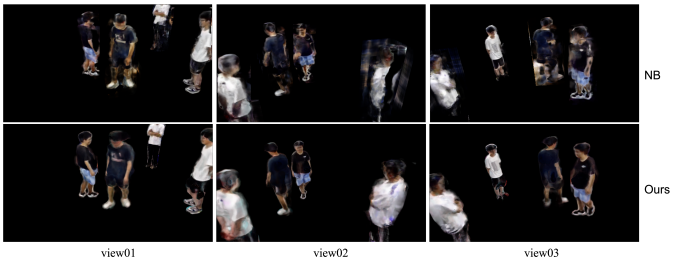

Figure 5: *Comparisons on new poses.*

Table 3: *Ablation study on different camera views on seen pose*

	training views			testing views		
	2 views	4 views	6 views	2 views	4 views	6 views
PSNR	18.4573	25.6401	**27.7242**	14.3911	19.1092	**19.5328**

Table 4: *Ablation study on different number of frames*

	training views			testing views		
	100	300	500	100	300	500
PSNR	11.2891	**27.7242**	24.3402	7.3480	**19.5328**	18.6229

5.1. Impact of the number of camera views.

As shown in Table 3, we respectively train our model on 2, 4, and 6 views. The results show that increasing the number of views improves the multi-person synthesis in both training views and testing views. Our method achieves multi-person novel view synthesis just from sparse views' images compared to previous methods which required dense views [ZLY*21] or scanning 3D human template as supervision [ZSZ*21].

5.2. Impact of the video length.

We train our models on different video lengths and choose 100, 300, and 500 frames respectively. The results in Table 4 show that increasing the video frames improves the performance of our method. However, the performance decreases when the frames are too long. The reason is that it is difficult for our model to fit very long videos which have also been reported in [PZX*21].

6. Conclusions

We present MP-NeRF, a novel approach for dynamic multi-person synthesis from sparse multi-view videos. The key idea of our work is to propose an identity-aware global latent code, which is able to incorporate the relative location and identification among multiple people. The global latent code is anchored on the vertices of multi-person SMPL templates obtained by multi-way matching and tracking algorithms among multi-person RGB videos. In a further step, we construct occlusion-aware multiple volumes with an affine transformation to represent multiple people in a scene and obtain the color and density of points in the occlusion region by a self-designed combination operator. The experimental results on the multi-person dataset demonstrate the effectiveness of our approach in multi-person free-view synthesis from sparse multi-view videos. We believe that our approach extends the presence of free-viewpoint to more persons, with abundant potential application in VR/AR, interactive telepresence, and immersive sports broadcasting. In the future, we plan to improve our work to achieve editing in the 3D content and increase clarity.

7. Acknowledgments

The work described in this paper was fully supported by a grant from City University of Hong Kong (Project No. 9678139).

References

[ASK*20] ALIEV K.-A., SEVASTOPOLSKY A., KOLOS M., ULYANOV D., LEMPITSKY V.: Neural point-based graphics. In *European Conference on Computer Vision* (2020), Springer, pp. 696–712. 2

[BMSR20] BEMANA M., MYSZKOWSKI K., SEIDEL H.-P., RITSCHEL T.: X-fields: implicit neural view-, light-and time-image interpolation. *ACM Transactions on Graphics (TOG) 39*, 6 (2020), 1–15. 1

[CCS*15] COLLET A., CHUANG M., SWEENEY P., GILLETT D., EVSEEV D., CALABRESE D., HOPPE H., KIRK A., SULLIVAN S.: High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG) 34*, 4 (2015), 69. 1, 2

[CDSHD13] CHAURASIA G., DUCHENE S., SORKINE-HORNUNG O., DRETTAKIS G.: Depth synthesis and local warps for plausible image-based navigation. *ACM Transactions on Graphics (TOG) 32*, 3 (2013), 1–12. 2

[CGT*19] CHOI I., GALLO O., TROCCOLI A., KIM M. H., KAUTZ J.: Extreme view synthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 7780–7789. doi:10.1109/ICCV.2019.00787. 2

[CLI*20] CHABRA R., LENSSEN J. E., ILG E., SCHMIDT T., STRAUB J., LOVEGROVE S., NEWCOMBE R.: Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *European Conference on Computer Vision* (2020), Springer, pp. 608–625. 3

[CTMS03] CARRANZA J., THEOBALT C., MAGNOR M. A., SEIDEL H.-P.: Free-viewpoint video of human actors. *ACM transactions on graphics (TOG) 22*, 3 (2003), 569–577. 1, 2

[CWP*18] CHEN Y., WANG Z., PENG Y., ZHANG Z., YU G., SUN J.: Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 7103–7112. 3

[DAST*08] DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H.-P., THRUN S.: Performance capture from sparse multi-view video. In *ACM SIGGRAPH 2008 papers*. 2008, pp. 1–10. 2

[DHT*00] DEBEVEC P., HAWKINS T., TCHOU C., DUIKER H.-P., SAROKIN W., SAGAR M.: Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (2000), pp. 145–156. 2

[DJH*19] DONG J., JIANG W., HUANG Q., BAO H., ZHOU X.: Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 7792–7801. 3, 5

[DKD*16] DOU M., KHAMIS S., DEGTYAREV Y., DAVIDSON P., FANELLO S. R., KOWDLE A., ESCOLANO S. O., RHEMANN C., KIM D., TAYLOR J., ET AL.: Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG) 35*, 4 (2016), 1–13. 2

[DLD12] DAVIS A., LEVOY M., DURAND F.: Unstructured light fields. *Comput. Graph. Forum 31*, 2pt1 (May 2012), 305–314. URL: https://doi.org/10.1111/j.1467-8659.2012.03009.x, doi:10.1111/j.1467-8659.2012.03009.x. 2

[FBD*19] FLYNN J., BROXTON M., DEBEVEC P., DUVALL M., FYFFE G., OVERBECK R., SNAVELY N., TUCKER R.: Deepview: View synthesis with learned gradient descent. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 2362–2371. doi:10.1109/CVPR.2019.00247. 2

[GCS*20] GENOVA K., COLE F., SUD A., SARNA A., FUNKHOUSER T.: Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 4857–4866. 3

[GEVDM18] GRAHAM B., ENGELCKE M., VAN DER MAATEN L.: 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 9224–9232. 3

[GGSC96] GORTLER S. J., GRZESZCZUK R., SZELISKI R., COHEN M. F.: The lumigraph. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1996), SIGGRAPH '96, Association for Computing Machinery, p. 43–54. URL: https://doi.org/10.1145/237170.237200, doi:10.1145/237170.237200. 1, 2

[GLD*19] GUO K., LINCOLN P., DAVIDSON P., BUSCH J., YU X., WHALEN M., HARVEY G., ORTS-ESCOLANO S., PANDEY R., DOURGARIAN J., ET AL.: The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG) 38*, 6 (2019), 1–19. 2

[GLL*18] GONG K., LIANG X., LI Y., CHEN Y., YANG M., LIN L.: Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 770–785. 2

[GSDA*09] GALL J., STOLL C., DE AGUIAR E., THEOBALT C., ROSENHAHN B., SEIDEL H.-P.: Motion capture using joint skeleton tracking and surface estimation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), IEEE, pp. 1746–1753. 2

[HPP*18] 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 (2018), 1–15. 2

[JSM*20] JIANG C., SUD A., MAKADIA A., HUANG J., NIESSNER M., FUNKHOUSER T., ET AL.: Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 6001–6010. 3

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 5

[KVH84] KAJIYA J. T., VON HERZEN B. P.: Ray tracing volume densities. *ACM SIGGRAPH computer graphics 18*, 3 (1984), 165–174. 5

[KWR16] KALANTARI N. K., WANG T.-C., RAMAMOORTHI R.: Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG) 35*, 6 (2016), 1–10. 2

[LGZL*20] LIU L., GU J., ZAW LIN K., CHUA T.-S., THEOBALT C.: Neural sparse voxel fields. In *Advances in Neural Information Processing Systems* (2020), Larochelle H., Ranzato M., Hadsell R., Balcan M. F., Lin H., (Eds.), vol. 33, Curran Associates, Inc., pp. 15651–15663. URL: https://proceedings.neurips.cc/paper/2020/file/b4b758962f17808746e9bb832a6fa4b8-Paper.pdf. 2

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG) 34*, 6 (2015), 1–16. 2

[LSMG20] LIAO Y., SCHWARZ K., MESCHEDER L., GEIGER A.: Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 5871–5880. 2

[LSS*19] LOMBARDI S., SIMON T., SARAGIH J., SCHWARTZ G., LEHRMANN A., SHEIKH Y.: Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751* (2019). 2, 4

[LXZ*19] LIU L., XU W., ZOLLHOEFER M., KIM H., BERNARD F., HABERMANN M., WANG W., THEOBALT C.: Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG) 38*, 5 (2019), 1–14. 2

[LZP*20] LIU S., ZHANG Y., PENG S., SHI B., POLLEFEYS M., CUI Z.: Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 2019–2028. 2

[Max95] MAX N.: Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics 1*, 2 (1995), 99–108. 5

[MBPY*18] MARTIN-BRUALLA R., PANDEY R., YANG S., PIDLYPENSKYI P., TAYLOR J., VALENTIN J., KHAMIS S., DAVIDSON P., TKACH A., LINCOLN P., ET AL.: Lookingood: Enhancing performance capture with real-time neural re-rendering. *arXiv preprint arXiv:1811.05029* (2018). 2

[MGK*19a] MESHRY M., GOLDMAN D. B., KHAMIS S., HOPPE H., PANDEY R., SNAVELY N., MARTIN-BRUALLA R.: Neural rerendering in the wild. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 6871–6880. doi:10.1109/CVPR.2019.00704. 1

[MGK*19b] MESHRY M., GOLDMAN D. B., KHAMIS S., HOPPE H., PANDEY R., SNAVELY N., MARTIN-BRUALLA R.: Neural rerendering in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 6878–6887. 2

[MKGH16] MUSTAFA A., KIM H., GUILLEMAUT J.-Y., HILTON A.: Temporally coherent 4d reconstruction of complex dynamic scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4660–4669. doi:10.1109/CVPR.2016.504. 1

[MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision* (2020), Springer, pp. 405–421. 1, 2, 3, 4, 5

[NFS15] NEWCOMBE R. A., FOX D., SEITZ S. M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 343–352. doi:10.1109/CVPR.2015.7298631. 2

[NG21] NIEMEYER M., GEIGER A.: Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 11453–11464. 4

[NMOG20] NIEMEYER M., MESCHEDER L., OECHSLE M., GEIGER A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3504–3515. 2

[NSH*19] NATSUME R., SAITO S., HUANG Z., CHEN W., MA C., LI H., MORISHIMA S.: Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 4480–4490. 2

[OMT*20] OST J., MANNAN F., THUEREY N., KNODT J., HEIDE F.: Neural scene graphs for dynamic scenes. 1

[PFS*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 165–174. 4

[PSB*20] PARK K., SINHA U., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., SEITZ S. M., BRUALLA R.-M.: Deformable neural radiance fields. *arXiv preprint arXiv:2011.12948* (2020). 1

[PZ17] PENNER E., ZHANG L.: Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics (TOG) 36*, 6 (2017), 1–11. 2

[PZX*21] PENG S., ZHANG Y., XU Y., WANG Q., SHUAI Q., BAO H., ZHOU X.: 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* (2021), pp. 9054–9063. 1, 2, 3, 4, 5, 6

[RBA*19] RAHAMAN N., BARATIN A., ARPIT D., DRAXLER F., LIN M., HAMPRECHT F., BENGIO Y., COURVILLE A.: On the spectral bias of neural networks. In *International Conference on Machine Learning* (2019), PMLR, pp. 5301–5310. 4

[RJY*20] REBAIN D., JIANG W., YAZDANI S., LI K., YI K. M., TAGLIASACCHI A.: Derf: Decomposed radiance fields. *arXiv preprint arXiv:2011.12490* (2020). 1

[SF16] SCHONBERGER J. L., FRAHM J.-M.: Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 4104–4113. 5

[SGDA*10] STOLL C., GALL J., DE AGUIAR E., THRUN S., THEOBALT C.: Video-based reconstruction of animatable human characters. *ACM Transactions on Graphics (TOG) 29*, 6 (2010), 1–10. 2

[SHN*19] SAITO S., HUANG Z., NATSUME R., MORISHIMA S., KANAZAWA A., LI H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 2304–2314. 2

[SSSJ20] SAITO S., SIMON T., SARAGIH J., JOO H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 84–93. 2

[STH*19] SITZMANN V., THIES J., HEIDE F., NIESSNER M., WETZSTEIN G., ZOLLHOFER M.: Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 2437–2446. 2

[SXZ*20] SU Z., XU L., ZHENG Z., YU T., LIU Y., FANG L.: Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *European Conference on Computer Vision* (2020), Springer, pp. 246–264. 2

[SZW19] SITZMANN V., ZOLLHÖFER M., WETZSTEIN G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems 32* (2019). 2

[TFT*20] TEWARI A., FRIED O., THIES J., SITZMANN V., LOMBARDI S., SUNKAVALLI K., MARTIN-BRUALLA R., SIMON T., SARAGIH J., NIESSNER M., PANDEY R., FANELLO S., WETZSTEIN G., ZHU J.-Y., THEOBALT C., AGRAWALA M., SHECHTMAN E., GOLDMAN D. B., ZOLLHÖFER M.: State of the art on neural rendering. *Computer Graphics Forum 39*, 2 (2020), 701–727. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14022, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14022, doi:https://doi.org/10.1111/cgf.14022. 1

[TTG*20] TRETSCHK E., TEWARI A., GOLYANIK V., ZOLLHÖFER M., LASSNER C., THEOBALT C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a deforming scene from monocular video. *arXiv preprint arXiv:2012.12247* (2020). 1

[TZN19] THIES J., ZOLLHÖFER M., NIESSNER M.: Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG) 38*, 4 (2019), 1–12. 2

[TZT*18] THIES J., ZOLLHÖFER M., THEOBALT C., STAMMINGER M., NIESSNER M.: Ignor: Image-guided neural object rendering. *arXiv preprint arXiv:1811.10720* (2018). 1

[WWHY20a] WU M., WANG Y., HU Q., YU J.: Multi-view neural human rendering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 1679–1688. doi:10.1109/CVPR42600.2020.00175. 1

[WWHY20b] WU M., WANG Y., HU Q., YU J.: Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 1682–1691. 2

[ZKU*04] ZITNICK C. L., KANG S. B., UYTTENDAELE M., WINDER S., SZELISKI R.: High-quality video view interpolation using a layered representation. *ACM transactions on graphics (TOG) 23*, 3 (2004), 600–608. 1

[ZLY*21] ZHANG J., LIU X., YE X., ZHAO F., ZHANG Y., WU M., ZHANG Y., XU L., YU J.: Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG) 40*, 4 (2021), 1–18. 1, 5, 6

[ZSZ*21] ZHENG Y., SHAO R., ZHANG Y., YU T., ZHENG Z., DAI Q., LIU Y.: Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 6239–6249. 3, 6

[ZTF*18] ZHOU T., TUCKER R., FLYNN J., FYFFE G., SNAVELY N.: Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. 37*, 4 (July 2018). URL: https://doi.org/10.1145/3197517.3201323, doi:10.1145/3197517.3201323. 2

[ZYW*19] ZHENG Z., YU T., WEI Y., DAI Q., LIU Y.: Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 7739–7749. 2