HumanNeRF: Efficiently Generated Human Radiance Field from Sparse Inputs

Fuqiang Zhao[1]   Wei Yang[2]   Jiakai Zhang[1]   Pei Lin[1]   Yingliang Zhang[3]
Jingyi Yu[1]   Lan Xu[1,4]

[1] ShanghaiTech University   [2] Huazhong University of Science and Technology   [3] DGene
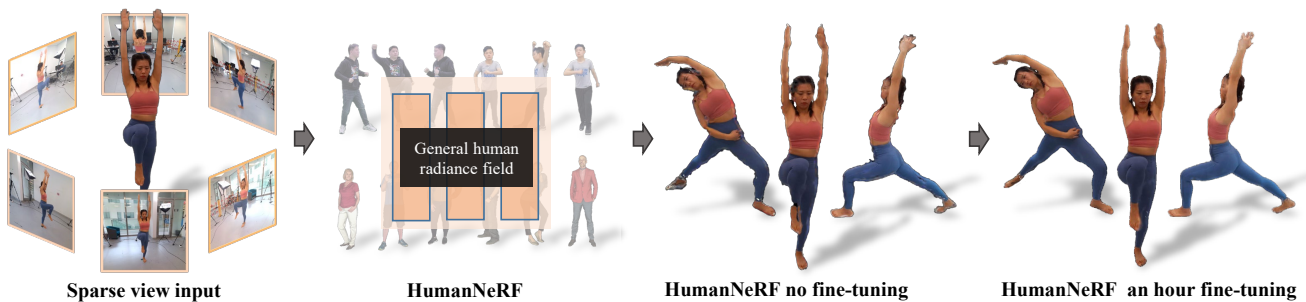[4] Shanghai Engineering Research Center of Intelligent Vision and Imaging

Figure 1. Our proposed HumanNeRF utilizes on-the-fly efficient general dynamic radiance field generation and neural blending, enabling high-quality free-viewpoint video synthesis for dynamic humans. Our approach only takes sparse images as input and uses a pre-trained network on large human datasets. Then we can effectively synthesize a photo-realistic image from a novel viewpoint. While these results contain artifacts, we fine-tune 300 frames for a specific performer using only an hour and generate improved results.

Abstract

Recent neural human representations can produce high-quality multi-view rendering but require using dense multi-view inputs and costly training. They are hence largely limited to static models as training each frame is infeasible. We present HumanNeRF - a neural representation with efficient generalization ability - for high-fidelity free-view synthesis of dynamic humans. Analogous to how IBRNet assists NeRF by avoiding per-scene training, HumanNeRF employs an aggregated pixel-alignment feature across multi-view inputs along with a pose embedded non-rigid deformation field for tackling dynamic motions. The raw HumanNeRF can already produce reasonable rendering on sparse video inputs of unseen subjects and camera settings. To further improve the rendering quality, we augment our solution with in-hour scene-specific fine-tuning, and an appearance blending module for combining the benefits of both neural volumetric rendering and neural texture blending. Extensive experiments on various multi-view dynamic human datasets demonstrate effectiveness of our approach in synthesizing photo-realistic free-view humans under challenging motions and with very sparse camera view inputs.

1. Introduction

View synthesis of human activities enables numerous applications in visual effects and telepresence, with unique and immersive viewing experiences. However, a convenient and high-quality solution from the light-weight capture setup remains a cutting-edge yet bottleneck technique.

Early solutions require a dome-based multi-view setup for accurate reconstruction [7, 10] and image-based rendering in novel views [3, 65]. Volumetric approaches [44, 58] enable light-weight reconstruction, but they still heavily rely on the depth sensors and are restricted by the limited mesh resolution. Recent neural rendering techniques have achieved significant progress [13, 27, 29, 47]. Remarkably, NeRF [29] and its dynamic extensions [24, 33, 35, 50, 53, 64] enable photo-realistic novel view synthesis for dynamic scenes without heavy reliance on the reconstruction accuracy. However, these solutions still require expensive dense capture views or suffer from tedious time-consuming per-scene training, which highly limits the practicality. Only recently, some approaches [5, 54, 62] enhance NeRF [29] with

image-conditioned features to break the per-scene training constraint for efficient radiance field generation of static scenes. But few researchers explore such generalizable NeRF representation under the complex dynamic human settings. The recent work [45] further enables generalizable human rendering from 6 RGB streams by combining texture blending with implicit geometry inference [37, 38] only in novel views. However, it suffers from severe artifacts near the occluded regions due to the lack of global inherent geometry and texture modeling.

In this paper, we present *HumanNeRF* – a practical and high-quality neural free-view synthesis approach for general dynamic humans using only sparse RGB streams. As illustrated in Fig.1, our approach enables photo-realistic human rendering by efficiently optimizing a more generalizable radiance field on-the-fly for unseen performers in an hour, favorably transcending previous long-term per-scene training approaches.

Our key idea is to marry the dynamic NeRF representation with neural image-based blending in a light-weight and two-stage framework. We extend the concept of general radiance field into the dynamic and temporal setting to break the per-scene constraint for efficient rendering. We also explore an effective implicit blending strategy to boost the texture result of volumetric rendering with the level of detail present in the sparse input images. Specifically, we first adopt an implicit scheme to aggregate image-conditioned features from our sparse input, which enables generalizable inference of motion and appearance in the dynamic NeRF framework. Then, we introduce a pose-embedded hybrid deformation scheme to enhance the generalization ability for unseen identities under various motions and garments. It combines explicit model-based warping with implicit subtle displacement modeling, so as to learn a reliable radiance field in an inherent canonical space. Note that our scheme also supports efficient per-performer fine-tuning with temporally sparse sampling, which significantly improves the rendering quality even on unseen poses. However, we observe that existing dynamic NeRF-based volumetric rendering still fails to generate high-frequency texture details, especially for challenging unseen identities and poses. To this end, we combine the image-based rendering with NeRF-based volume rendering into a novel neural blending scheme through implicit and occlusion-aware blending weight learning. It enables accurate appearance rendering in the target view with the level of texture detail in the adjacent input images. To summarize, our main contributions include:

- We present a high-quality performance rendering approach via efficient radiance field generation for arbitrary performers from sparse RGB streams, achieving significant superiority to existing the state of the art.

- We extend the generalizable NeRF into the new realm of dynamic and light-weight setting through implicit feature aggregation and hybrid deformation.

- We propose a novel implicit blending scheme to preserve the texture detail from the input images, providing photo-realistic appearance rendering.

2. Related work

Human Performance Capture. Markerless human performance capture techniques have been widely adopted to achieve human free-viewpoint video or reconstruct the geometry. Some recent work only relies on the light-weight and single view setup [6, 14, 59, 60], but these methods require the pre-scanned template or naked human model and it is difficult for them to achieve photo-realistic view synthesis. The high-end approaches [13, 24, 26, 41] are able to produce high-quality surface motion and appearance reconstruction, but they require dense cameras and a controlled imaging environment which is not easily accessible. Other monocular RGB-D based methods [12, 19, 30, 43, 57, 63] adopt the traditional modeling and rendering pipeline to synthesize novel views of humans. However, these methods still suffer from the inherent self-occlusion constraint and cannot capture the motions in occluded regions. The light-weight multi-view solutions [9, 10, 58] which is most similar to our method serve as a good compromise between over-demanding hardware setup and high-fidelity reconstruction but still rely on 3 to 8 RGBD streams as input.

Neural Rendering. Recently, a lot of work has shown significant progress on 3D scene modeling and photo-realistic novel view synthesis via differentiable neural rendering manner based on various data representations, such as point clouds [1, 55], voxels [27, 39, 61],or texture meshes [25, 48]. More recent implicit manner based work [23, 29, 32, 40, 42, 49, 56] achieves impressive results for novel view synthesis for a specific scene. However, dedicated per-scene training is required in these methods when applying the representation to a new scene. Some methods [5, 21, 36, 37, 45, 54] utilize pixel-aligned features from source images to enable generalizable human modeling without per-scene training constraint. However, the method [37] generates blur texture results due to the reliance on implicit texture representation, while the method [45] suffers from geometric discontinuity due to the lack of temporal information. Recently, Kwon *et al.* [21] utilized temporally aggregated features to compensate the sparse input views, achieving generalizable human radiance field generation. However, they still suffer from blur artifacts when generalizing unseen identities with complex motions due to self-occlusion. In contrast, we utilize generalizable human NeRF with occlusion-aware pixel-aligned features and adopt implicit blending, achieving high-quality novel view
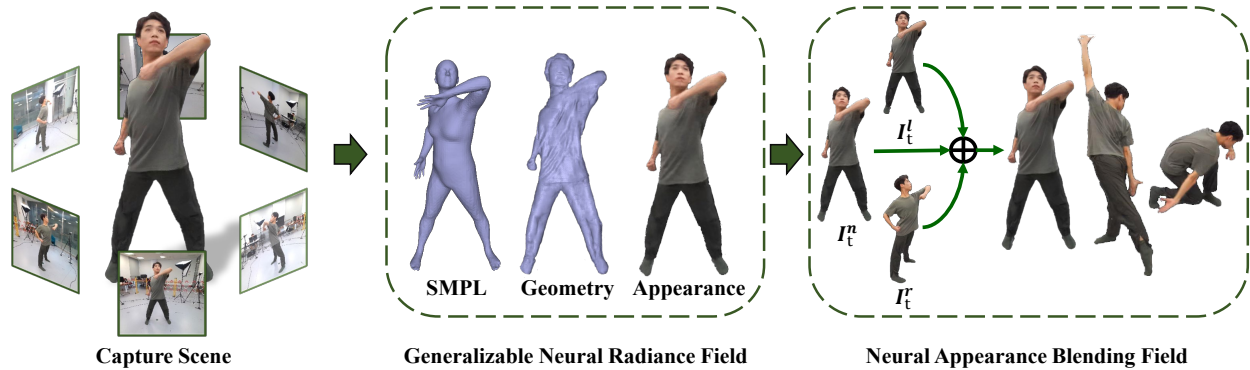
Figure 2. The overview of our HumanNeRF method. Assuming the video input from six RGB cameras surrounding the performer, our approach consists of a generalizable neural radiance field (Sec. 3.1), an optional fast per-scene fine-tuning scheme and a novel neural appearance blending field (Sec. 3.2).

synthesis with the level of texture detail present in the input images.

Image based rendering. Previous work of IBR [8, 11, 22] aims at synthesizing novel view from a set of source images through blending weights of reference pixels without recovering detailed 3D geometry. Blending weights are calculated based on ray space approximation [22] approximate proxy geometry [2, 8, 16]. Though their rendering results are impressive, the range of renderable viewpoints is limited. In recent work [4, 34, 65], researchers have proposed improved methods by inferring depth maps from input images as proxy geometries. For example, some work [15, 46] utilize two stages of multi-view stereo. First, they generate a grid surface that depends on the view and then there is a CNN to calculate the blending weights. While these methods can handle sparser views than other approaches and achieve promising results in some cases, they are sensitive to the quality of reconstructed proxy geometries [18, 37]. Comparably, our method embraces image blending into implicit representations pipeline under the light-weight multi-RGB, which enables photo-realistic appearance and geometry reconstruction in novel views.

3. The HumanNeRF Approach

We first introduce the problem formulation and overall scheme of our HumanNeRF method. Given K synchronized videos of a performer captured at different viewpoints (360°around preferably) with T frames, $\mathcal{I} = \{I^{k,t}\}$, in each video, our method aims to synthesize free-viewpoint videos of the performer and also generalize the motion to an arbitrary person with high fidelity. Fig. 2 illustrates the high-level components of our system. The core step of our approach is the efficient generalized Neural Radiance Field for dynamic humans, which adapts the NeRF [29] for dynamic human representation.

We leverage the parametric human body model SMPL [28] for estimating a basis model, and use an MLP network to learn subtle displacements of the human body. The output is then deformed into a canonical pose for NeRF optimization and rendering (Sec. 3.1). The efficient generalization ability comes from our aggregated pixel alignment features $\mathcal{F} = \{F_t\}$ from multi-view input images by projecting the 3D sample point into images and blending individual image features. Although our generated NeRF outputs human geometry with good quality, synthesized textures may contain artifacts and lack high-frequency details. Thus, we use a novel neural appearance blending scheme to refine texture details by aggregating colors from neighboring views. The final synthesized results exhibit a photo-realistic appearance with fine details (Sec. 3.2).

3.1. Generalizable Dynamic Neural Radiance Field

We retain NeRF's ability for novel view synthesis and geometry details rendering. However, NeRF assumes stationary subjects and performs per-scene optimization, which makes it not directly applicable to our problem. We make two main changes to NeRF to handle human dynamics and gain generalization ability. Specifically, we first warp the camera ray to account for human motion before sampling for NeRF and combine the viewing direction input with aggregated pixel alignment features for gaining generalizability.

Aggregated Pixel Alignment Feature. We propose an aggregated pixel alignment feature for NeRF generalization. Specifically, we use a U-Net network U to extract image feature maps representing local image appearance. Given an input image $I^k \in \mathbb{R}^{H \times W \times 4}$ with mask as the last channel, the output of U is a 2D feature map
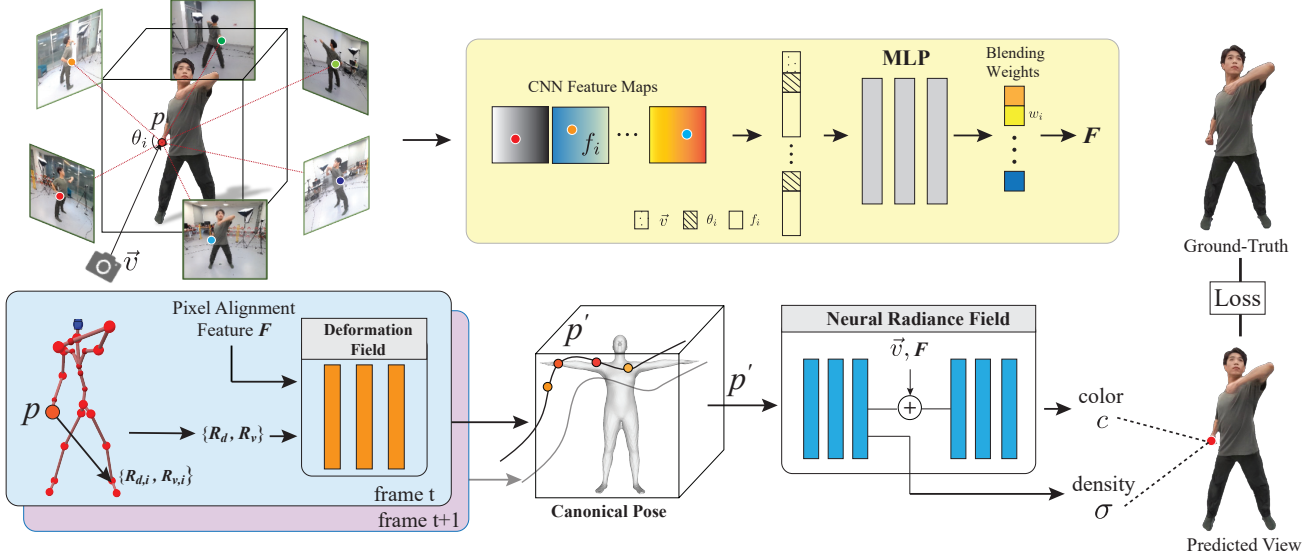
Figure 3. Illustration of our generalizable neural radiance field module. For feature extraction, we concatenate $\{f_i\}_{i=1}^{\mathcal{K}}$ with the viewing directions $\vec{v}$ and the angle $\{\theta_i\}_{i=1}^{\mathcal{K}}$ of viewing direction relative to each the query ray of the source view. We use R_d and R_v to model the distances and directions between sample point p and the 24 joints of the SMPL skeleton. We use the neural radiance field to regress volume density and RGB radiance at location p' in canonical space and its corresponding blending features F.

$f^k \in \mathbb{R}^{H \times W \times C}$, i.e.,

$$f^k = U(I^k) \qquad (1)$$

For each spatial point $p \in \mathbb{R}^3$ fed to NeRF, we first project it into view k at $q^k \in \mathbb{R}^2$ and fetch the corresponding feature vector f_q^k. The aggregated pixel alignment feature of p then is a weighted summation of image features as $F_p = \sum_{k=1}^{K} w^k f_q^k$ for k in $1...K$, and we use an MLP network to estimate the blending weights w_q^k as:

$$w_q^k = \mathbf{MLP}_{\mathcal{B}}(\vec{v}, \theta_q^k, f_q^k) \qquad (2)$$

where $\vec{v}$ is the view direction of p in camera view and θ_q^k is the angle of the viewing direction w.r.t. the sample ray from p to q^k.

Pose Embedded Non-rigid Human Deformation. To accommodate the human dynamics, we warp the human body from the current time frame to a common canonical pose so that NeRF receives static sampling queries, similar to [24, 31, 35, 51]. In practice, we find an MLP module tends to learn subtle displacements other than handle large deformations. To address this issue, we fit the SMPL model to a human body in the current time frame and deform the model to a common canonical pose using inverse-skinning transformation $\mathcal{S}$ [17, 24]. The resulting model usually exhibits inconsistencies with image observations. We further apply a pose-dependent non-rigid deformation field $\mathbf{MLP}_d$ to learn the subtle displacement. Our pose embedded non-

rigid deformation field can be formulated as:

$$p' = \mathcal{S}(p, \mathcal{M}, w^s) + \mathbf{MLP}_d(R_d, R_v, F_p) \qquad (3)$$

where $\mathcal{M}$ is the estimated motion, w^s is the corresponding skinning weight of sample point p. We use $R_d \in \mathbb{R}^{24}$ and $R_v \in \mathbb{R}^{72}$ to model the distances and directions between p and the 24 joints of SMPL skeleton. F_p is the aggregated pixel alignment feature.

Finally, we have our generalizable dynamic neural radiance field Φ, which takes the transformed 3D location p', view direction $\vec{v}$ and F_p as input and predicts the volume density σ and color c at point p before deformation as:

$$(\mathbf{c}, \sigma) = \Phi(p', \vec{v}, F_p) \qquad (4)$$

Fig. 3 shows the overview of our generalizable dynamic neural radiance field.

Dynamic Human Volume Rendering. We utilize the physically based volume rendering [20] technique to synthesize a new view image similar to the original NeRF. The only difference here is the query ray is bent by the deformation field before sending to NeRF. In particular, we compute a pixel's color at frame t by marching a corresponding ray and accumulating radiance at sampled points between near and far bounds, i.e.,

$$C_{\mathbf{r}} = \sum_{i=1}^{N} T(p_i)\big[(1 - e^{-\sigma_{p_i}\delta_{p_i}})c_{p_i}\big] \qquad (5)$$

where $T(p_i) = e^{-\sum_{k=1}^{i-1} \sigma_{p_i} \delta_{p_i}}$ and $\delta(p_i) = p_{i+1} - p_i$ is the distance between adjacent samples and N is the number of the sampled points on the ray.

Fast Per-subject Fine-tuning. Due to the limited training data along with diversity among different identities and scenes, artifacts and imperfections can still be observed for an unseen person when transferring motion. To address this problem, we adopt a fast fine-tuning solution as compensation to our original framework which treats the network optimized on the performer as an initialization state. Specifically, we first train our network on various subjects/performers and freeze the feature blending network $\mathbf{MLP}_{\mathcal{B}}$. And then when given an unseen subject, we optimize the network parameters of our deformation field $\mathbf{MLP}_d$ and the generalizable neural radiance field Φ.

3.2. Neural Appearance Blending

We observe that textures produced by NeRF rendering in the above section contain artifacts and lack high-frequency details sometimes due to the sparsity of input views. Inspired by image based rendering methods, we further propose a novel neural blending scheme for appearance refinement. Most of the texture information in a target view can be recovered by its only two adjacent input views in our multi-view setting. Considering a certain time frame, we first render a depth map $D^{\mathbf{v}}$ at the target view $\mathbf{v}$ from our generalizable neural radiance field at inference time. And then we back-project each point q in $D^{\mathbf{v}}$ with color $C_q^{\mathbf{v}}$ into neighboring two views $\mathbf{v}_1$ and $\mathbf{v}_2$ and fetch the colors $C_q^{\mathbf{v}_1}$ and $C_q^{\mathbf{v}_2}$, along with the corresponding visibility $O_q^{\mathbf{v}_1}$ and $O_q^{\mathbf{v}_2}$ which is determined by depth difference. At training time, the depth maps are rendered from the synthetic human model dataset, such as Twindom [52]. We further extract q's corresponding image feature $f_q^{\mathbf{v}_1}, f_q^{\mathbf{v}_2}$, and then feed feature and visibility information into our neural appearance blending network $\mathbf{MLP}_{\mathcal{A}}$,

$$W_q = \mathbf{MLP}_{\mathcal{A}}(f_q^{\mathbf{v}_1}, O_q^{\mathbf{v}_1}, f_q^{\mathbf{v}_2}, O_q^{\mathbf{v}_2}) \tag{6}$$

where $W_q \in \mathbb{R}^3$ is the appearance blending weight. The final color for q in $\mathbf{v}$ then is:

$$C_q^{\mathbf{v}*} = W_q \cdot C_q, \ \ C_q = [C_q^{\mathbf{v}_1}, C_q^{\mathbf{v}}, C_q^{\mathbf{v}_2}] \tag{7}$$

where $\cdot$ denotes the dot product.

3.3. Implementation Details

Here we describe the implementation details including the training scheme of our approach. Our generalizable NeRF module (including the feature extraction network U, feature blending network $\mathbf{MLP}_{\mathcal{B}}$, deformation network $\mathbf{MLP}_d$ and the adapted NeRF Φ) and the appearance blending network $\mathbf{MLP}_{\mathcal{A}}$ are independent and we train them separately.

Closest View / Left
Project Ray
Depth
Rendered Novel View
Closest View / Right
MLP
Color Blending Weights
□ / ■ Visibility: Yes / No
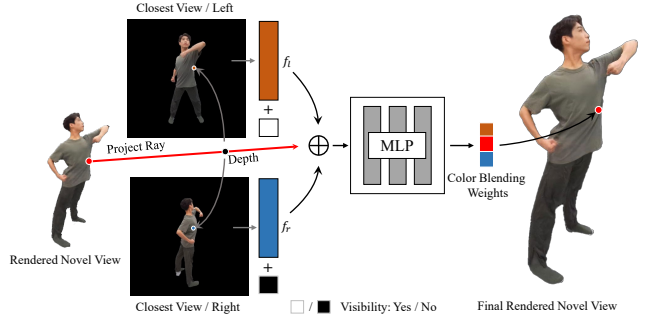Final Rendered Novel View

Figure 4. Illustration of our neural appearance refinement scheme. Our appearance blending network takes two adjacent image features f_r, f_l and corresponding occlusion information as input and then output three-dimensional weights to blend our rendering result with fine-detailed appearance information of the adjacent input views.

To optimize our network, we use a color loss $\mathcal{L}_c$ which measures the difference between the rendered color $C_{\mathbf{r}}$ and the ground truth color $\hat{C}_{\mathbf{r}}$ of camera ray $\mathbf{r}$:

$$\mathcal{L}_c = \sum_{\mathbf{r} \in \mathcal{R}} (||C_{\mathbf{r}} - \hat{C}_{\mathbf{r}}||_2^2) \tag{8}$$

and a silhouette loss $\mathcal{L}_m$ which is formulated as:

$$\mathcal{L}_m = \sum_{\mathbf{r} \in \mathcal{R}} \mathbf{BCE}(M_{\mathbf{r}} - \alpha_{\mathbf{r}}) \tag{9}$$

where $\alpha(\mathbf{r}) = \sum_{i=1}^{N} T(p_i)(1 - e^{-\sigma_{p_i} \delta_{p_i}})$ is the rendered mask for $\mathbf{r}$. The total loss is combination of $\mathcal{L}_c$ and $\mathcal{L}_m$:

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_m \tag{10}$$

where λ is the weight to balance the two losses. Specifically, we set λ=0.1 in our implementation. And we only use $\mathcal{L}_c$ for the neural appearance blending model.

Training Details. We train our models using Adam optimizer with a learning rate that decays from $1e^{-4}$ to $1e^{-5}$ during training. Besides, we sample $4,096$ camera rays for each mini-batch and sample 32 and 64 points from near to far following the hierarchical sampling strategy. We optimize all our networks on a PC with a single Nvidia GeForce RTX3090 GPU. The training time of our generalizable NeRF module is about 2 days. Depending on the number of video frames, the fine-tuning time ranges from 30 to 90 minutes, for input images with $1,080 \times 1,080$ resolution. Besides, we train our neural appearance refinement model for about 1 to 2 days.

Datasets. We train our generalizable NeRF on 1820 static scans from the Twindom [52] dataset, which consists of 120 camera views. And we collect 6 view videos for 26

Figure 5. The appearance results of our HumanNeRF method on several sequences, including "dance1", "yoga", "batman", "swing", "dance2", "ironman", "sport", "dance3" and "spiderman" from the upper left to lower right.

subjects with challenge motion, such as dancing and yoga. We also augment the data by rigging the pre-scanned 3D model and simulating challenging poses with 120 views to boost the generation ability of our networks. For the neural appearance blending module, we only train it on Twindom [52] dataset.

4. Experimental Results

In this section, we evaluate our HumanNeRF method on a variety of challenging scenarios. As demonstrated in Fig. 5, our approach generates high-quality appearance results and handles humans with rich textures, challenging poses and etc.

4.1. Comparison

We first compare our HumanNeRF method with per-scene optimization approaches including Neural Body [33], Neural Volumes [27] and ST-NeRF [64] both qualitatively and quantitatively. Furthermore, we also compare our method with generalizable methods, i.e., IBRNet [54] and

NeuralHumanFVV [45], in our sparse view input setting.

As shown in Fig. 6, compared with the per-scene optimization methods, our HumanNeRF achieves better results in just a short fine-tuning time. Results from our approach exhibit much better textures and the geometries are complete and accurate both for the "Taichi" from public ZJU-MoCap [33] and the "Batman" data collected by ourselves. When compared with generalizable methods, our method outperforms others and well addresses self-occlusions as shown in Fig. 7.

As for quantitative comparison, we show the PSNR, SSIM, LPIPS and MAE metrics of our approach and other methods on real testing data in Tab. 1. Specifically, we set reference camera images as ground truth, and calculate the metrics for synthesized images from methods for comparison.

As we can see from the table, our HumanNeRF outperforms other methods across all metrics. This demonstrates the generated views from our methods are closest to the real captured data. We also want to mention that even without
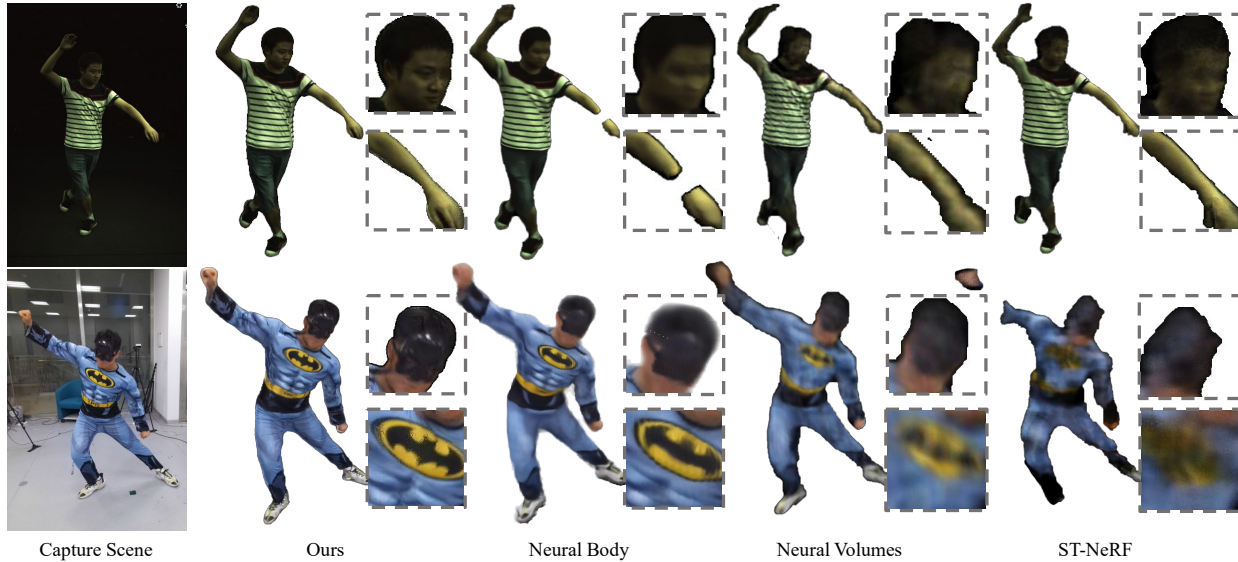
Figure 6. Qualitative comparison against per-scene training methods. We compare our method with Neural Body, Neural volumes, and ST-NeRF on "Batman" from our multi-view datasets and "Taichi" from ZJU-MoCap datasets. Our approach generalizes the most photo-realistic and finer detail.

Method	PSNR↑	SSIM↑	LPIPS↓	MAE↓
ST-NeRF	17.34	0.8547	0.1493	11.38
NeuralVolumes	27.32	0.9408	0.0999	2.905
NeuralBody	28.21	0.9544	0.0762	2.294
IBRNet	30.73	0.9817	0.0348	1.154
NeuralHumanFVV	27.86	0.9785	0.0440	1.237
$Ours_{wo\_bo}$	25.80	0.9456	0.0825	3.354
$Ours_{wo\_ft}$	29.51	0.9741	0.0461	1.521
$Ours_{wo\_rf}$	29.69	0.9620	0.0703	2.016
Ours	**33.01**	**0.9842**	**0.0334**	**0.9307**

Table 1. **Quantitative comparison against several methods in terms of rendering accuracy.** Compared with NeRF, ST-NeRF, Neural Volumes, NeuralBody, IBRNet and Neural-HumanFVV, our approach achieves the best performance in **PSNR,SSIM,LPIPS** and **MAE** metrics.
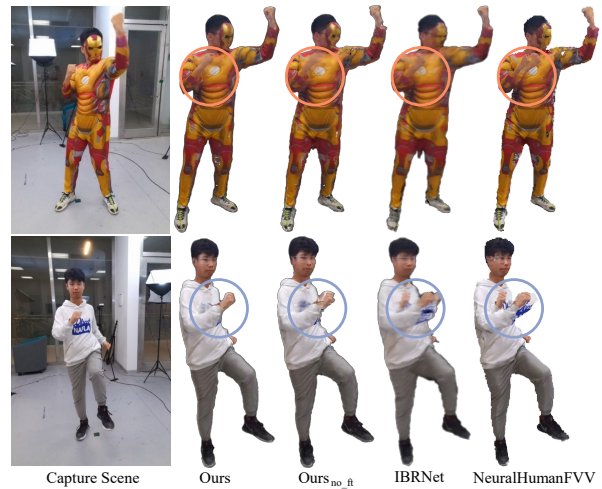

Figure 7. Qualitative comparison against generalizable methods. We compare our method with IBRNet, NeuralHumanFVV. Note that our approach generates better appearance results and well addresses the self-occlusion problem.

per-scene fine-tuning, our method still achieves comparable results.

4.2. Ablation Study

Appearance Blending and Fast Fine-tuning. Here, we evaluate the performances of different modules in our approach. We first demonstrate the effectiveness of our per-scene fine-tuning strategy by directly comparing the output of generalizable NeRF and results after fine-tuning. As we can see in Fig. 8, results without fine-tuning are low-detailed and blurry. While lack of our novel appearance refinement module leads to blurring rendering artifacts, especially around the boundaries. In contrast, our complete

approach achieves photo-realistic results with better decomposition for various entities.

Camera Number. To evaluate the impact of the number of input views on our framework, we compare the results of our method with various numbers of input camera views.

As shown in Fig. 9, the rendering results with views less than two suffer from severe geometric and rendering arti-
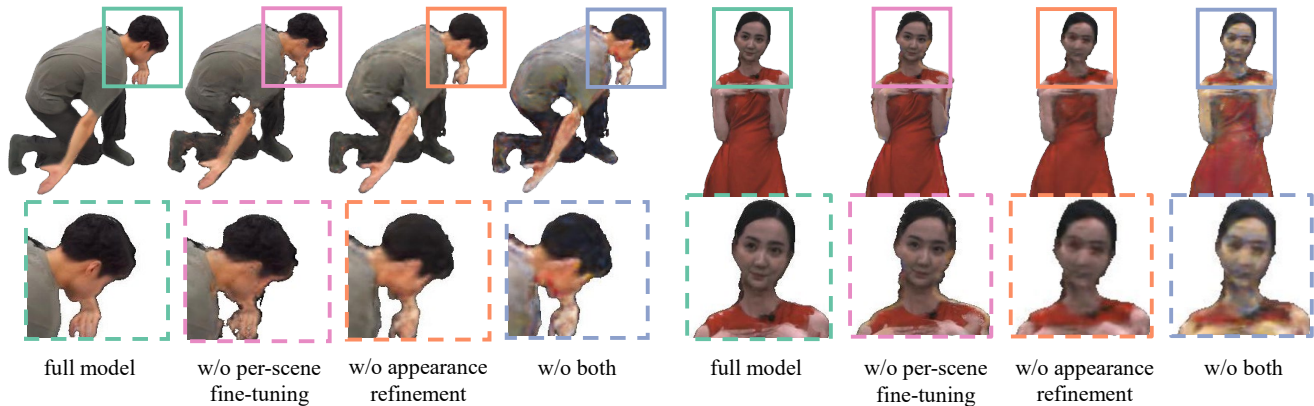
Figure 8. Qualitative evaluation of different variations in our method. This evaluation demonstrates the contribution and effectiveness of our algorithmic components.
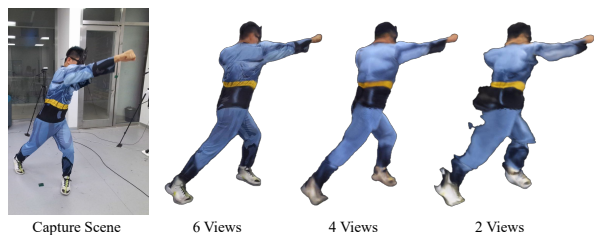

Figure 9. Evaluation of the number of input camera views. Our reconstructed appearance results using two, four and six cameras, respectively.

Figure 10. **Qualitative evaluation on pose generalization.**

facts. We also use results generated with all cameras as the reference to calculate the corresponding PSNR, SSIM, and LPIPS.

Pose Generalization. We further evaluate the pose generalizability of our HumanNeRF, we select 500 frames 'Swing' from videos in our multi-view dataset. We use 400 frames for fine-tuning and test on remaining 100 frames for pose generation. The results are shown in Fig. 10, our HumanNeRF generates visually good results even on the unseen pose and shows good metrics in Tab. 2.

5. Conclusion

We have presented a light weighted method for efficiently generating high-quality novel view synthesis of dynamic humans using only a sparse set of cameras. We leverage the fused image features and pose embedded human deformation module for dynamic human synthesize, and transcend the long-term per-scene optimization scheme of existing approaches. Moreover, our implicit neural appearance blending strategy refines results of volumetric rendering by borrowing fine details from two adjacent views. Experimental results on various datasets demonstrate the effectiveness of our approach in photo-realistic free-view syn-
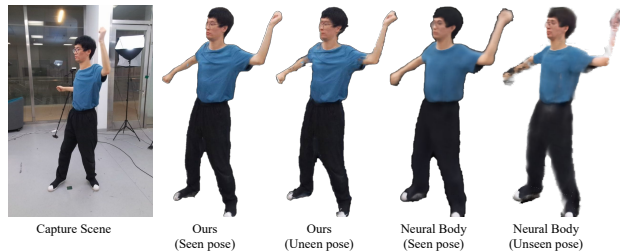
	PSNR↑	SSIM↑	LPIPS ↓	MAE ↓
Ours(Seen)	**36.01**	**0.9897**	**0.0356**	**0.5963**
Ours(Unseen)	34.53	0.9873	0.0386	0.7065
NB(Seen)	32.16	0.9756	0.0626	1.083
NB(Unseen)	27.61	0.9705	0.0640	1.756

Table 2. **Quantitative evaluation on pose generalization**.Results of ours and NB (Neural Body) on the seen pose and unseen pose.

thesis even for challenging human poses and motions. With the ability of efficient generation, we believe that our approach may bring good insights to many critical applications in VR/AR, such as gaming, entertainment, education, immersive telepresence, etc.

6. Acknowledgements

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 696–712. Springer, 2020. 2

[2] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 425–432, 2001. 3

[3] Joel Carranza, Christian Theobalt, Marcus A Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM transactions on graphics (TOG)*, 22(3):569–577, 2003. 1

[4] Gaurav Chaurasia, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM Transactions on Graphics (TOG)*, 32(3):1–12, 2013. 3

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 1, 2

[6] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, 129(10):2846–2864, 2021. 2

[7] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34(4):69, 2015. 1

[8] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 11–20, 1996. 3

[9] Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. Motion2fusion: Real-time volumetric performance capture. *ACM Transactions on Graphics (TOG)*, 36(6):1–16, 2017. 2

[10] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. Fusion4D: Real-time Performance Capture of Challenging Scenes. In *ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques*, 2016. 1, 2

[11] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54, 1996. 3

[12] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 2

[13] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics (TOG)*, 40(4):1–16, 2021. 1, 2

[14] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)*, 38(2):1–17, 2019. 2

[15] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018. 3

[16] Benno Heigl, Reinhard Koch, Marc Pollefeys, Joachim Denzler, and Luc Van Gool. Plenoptic modeling and rendering from image sequences taken by a hand-held camera. In *Mustererkennung 1999*, pages 94–101. Springer, 1999. 3

[17] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 4

[18] Michal Jancosek and Tomás Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR 2011*, pages 3121–3128. IEEE, 2011. 3

[19] Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. Neuralfusion: Neural volumetric rendering under human-object interactions. *arXiv preprint arXiv:2202.12825*, 2022. 2

[20] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984. 4

[21] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[22] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996. 3

[23] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020. 2

[24] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 1, 2, 4

[25] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 38(5):1–14, 2019. 2

[26] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2720–2735, 2013. 2

[27] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4), July 2019. 1, 2, 6

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 3

[29] 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 *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2, 3

[30] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 2

[31] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo-Martin Brualla. Deformable neural radiance fields. *arXiv preprint arXiv:2011.12948*, 2020. 4

[32] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2

[33] 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. 1, 6

[34] Eric Penner and Li Zhang. Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics (TOG)*, 36(6):1–11, 2017. 3

[35] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. 1, 4

[36] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12216–12225, 2021. 2

[37] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2, 3

[38] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 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*, pages 84–93, 2020. 2

[39] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2

[40] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[41] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *2011 International Conference on Computer Vision*, pages 951–958. IEEE, 2011. 2

[42] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[43] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 246–264. Springer, 2020. 2

[44] Zhuo Su, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang. Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular rgbd stream. *arXiv preprint arXiv:2104.14837*, 2021. 1

[45] Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, and Lan Xu. Neuralhumanfvv: Real-time neural volumetric human performance rendering using rgb cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. 2, 6

[46] Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, and Lan Xu. Neuralhumanfvv: Real-time neural volumetric human performance rendering using rgb cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6226–6237, 2021. 3

[47] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, Rohit Pandey, Sean Fanello, Gordon Wetzstein, Jun-Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B. Goldman, and Michael Zollhöfer. State of the Art on Neural Rendering. *Computer Graphics Forum*, 2020. 1

[48] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2

[49] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11708–11718, 2021. 2

[50] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12959–12970, October 2021. 1

[51] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhofer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 4

[52] Twindom dataset. https://https://web.twindom.com/. 5, 6

[53] Liao Wang, Ziyu Wang, Pei Lin, Yuheng Jiang, Xin Suo, Minye Wu, Lan Xu, and Jingyi Yu. ibutter: Neural interactive bullet time generator for human free-viewpoint rendering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4641–4650, 2021. 1

[54] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 1, 2, 6

[55] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1682–1691, 2020. 2

[56] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[57] Lan Xu, Wei Cheng, Kaiwen Guo, Lei Han, Yebin Liu, and Lu Fang. Flyfusion: Realtime dynamic scene reconstruction using a flying depth camera. *IEEE transactions on visualization and computer graphics*, 27(1):68–82, 2019. 2

[58] Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and Lu Fang. Unstructuredfusion: realtime 4d geometry and texture reconstruction using commercial rgbd cameras. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2508–2522, 2019. 1, 2

[59] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4968–4978, 2020. 2

[60] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)*, 37(2):1–15, 2018. 2

[61] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. *Advances in neural information processing systems*, 29, 2016. 2

[62] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1

[63] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7287–7296, 2018. 2

[64] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Trans. Graph.*, 40(4), July 2021. 1, 6

[65] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM transactions on graphics (TOG)*, 23(3):600–608, 2004. 1, 3