NeuralHumanFVV: Real-Time Neural Volumetric Human Performance Rendering using RGB Cameras

Xin Suo[1] Yuheng Jiang[1] Pei Lin [1] Yingliang Zhang[2] Minye Wu [1] Kaiwen Guo[3] Lan Xu[1,4]
[1]ShanghaiTech University [2]Dgene [3]Google
[4]Shanghai Engineering Research Center of Intelligent Vision and Imaging

Abstract

4D reconstruction and rendering of human activities is critical for immersive VR/AR experience. Recent advances still fail to recover fine geometry and texture results with the level of detail present in the input images from sparse multi-view RGB cameras. In this paper, we propose NeuralHumanFVV, a real-time neural human performance capture and rendering system to generate both high-quality geometry and photo-realistic texture of human activities in arbitrary novel views. We propose a neural geometry generation scheme with a hierarchical sampling strategy for real-time implicit geometry inference, as well as a novel neural blending scheme to generate high resolution (e.g., 1k) and photo-realistic texture results in the novel views. Furthermore, we adopt neural normal blending to enhance geometry details and formulate our neural geometry and texture rendering into a multi-task learning framework. Extensive experiments demonstrate the effectiveness of our approach to achieve high-quality geometry and photo-realistic free view-point reconstruction for challenging human performances.

1. Introduction

The rise of virtual and augmented reality (VR and AR) to present information in an immersive way has increased the demand of the 4D (3D spatial plus 1D time) content generation. Further reconstructing human activities and providing photo-realistic rendering from a free viewpoint conveniently evolves as a cutting-edge yet bottleneck technique.

Early solutions [27, 28, 58, 9] require pre-scanned templates or two to four orders of magnitude more time than is available for daily usages such as immersive tele-presence. Recently, volumetric approaches have enabled real-time human performance reconstruction and eliminated the reliance of a pre-scanned template model, by leveraging the RGBD sensors and modern GPUs. The high-end solutions [12, 11, 23, 68] rely on multi-view studio setup to achieve high-fidelity reconstruction and rendering in a novel view but are expensive and difficult to be deployed, while the low-end approaches [39, 53, 66, 72, 55] adopt the most handy monocular setup with a temporal fusion pipeline [40] but suffer from inherent self-occlusion constraint. Moreover, these approaches above rely on depth cameras which are not as cheap and ubiquitous as color cameras.

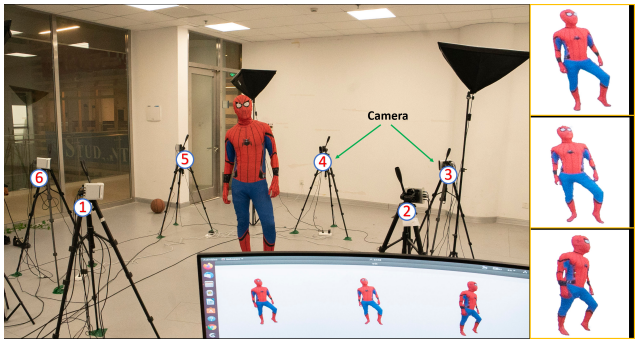

Figure 1. Our NeuralHumanFVV achieves real-time and photo-realistic reconstruction results of human performance in novel views, using only 6 RGB cameras.

The recent learning-based techniques enable robust human attribute reconstruction [35, 48, 73, 29] using only RGB input. In particular, the approaches PIFu [48] and PIFuHD [49] utilize pixel-aligned implicit function to reconstruct clothed humans with fine geometry details, while MonoPort [29] further enables real-time inference in a novel view. However, these methods fail to generate compelling photo-realistic texture due to the reliance of implicit texture representation. On the other hand, neural rendering techniques [32, 7, 64, 38, 25, 46] bring huge potential for photo-realistic novel view synthesis. However, existing solutions rely on per-scene training or are hard to achieve real-time performance due to the heavy network and the complicated 3D representation. Moreover, few researchers explore to combine volumetric geometry modeling and photo-realistic novel view synthesis of human performance in a data-driven manner simultaneously, especially under the light-weight multi-RGB and real-time setting.

In this paper, we attack the above challenges and present

NeuralHumanFVV – a real-time human neural volumetric rendering system using only light-weight and sparse RGB cameras surrounding the performer. As illustrated in Fig. 1, our novel approach generates both high-quality geometry and photo-realistic texture of human activities in arbitrary novel views, whilst still maintaining real-time computation and light-weight setup.

Generating such a human free-viewpoint video by combining volumetric geometry modeling and neural texture synthesis in a data-driven manner is non-trivial. Our key idea is to encode the local fine-detailed geometry and texture information of the adjacent input views into the novel target view, besides utilizing the inherent global information from our multi-view setting. To this end, we first introduce a neural geometry generation scheme to implicitly reason about the underlying geometry in a novel view. With a hierarchical sampling strategy along the camera rays in a coarse-to-fine manner, we achieve real-time detailed geometry inference. Then, based on the geometry proxy above, a novel neural blending scheme is proposed to map the input adjacent images into a photo-realistic texture output in the target view, through efficient occlusion analysis and blending weight learning. A boundary-aware upsampling strategy is further adopted to generate high resolution (e.g., 1k) novel view synthesis result without sacrificing the real-time performance. Finally, we recover the normal information in the target view using the same neural blending strategy, which not only enhances the output fine-grained geometry details but also combines our neural geometry generation and texturing blending into a multi-task learning framework. To summarize, our main contributions include:

- We present a real-time human performance rendering approach, which is the first to reconstruct high quality geometry and photo-realistic texture results in a novel view using sparse multiple RGB cameras, achieving significant superiority to existing state-of-the-arts.
- We propose an efficient neural implicit generation scheme to recover fine geometry details in the novel view via a hierarchical and coarse-to-fine strategy.
- We propose a novel neural blending scheme to provide high-resolution and photo-realistic texture result as well as normal result to further refine the geometry.

2. Related Work

Human Performance Capture. Markerless human performance capture [5, 60] technologies have been widely investigated to generate human free-viewpoint video or geometry reconstruction. The high-end approaches require studio-setup with hundreds of cameras and a controlled imaging environment [54, 31, 22, 9, 23, 14] to produce high quality surface motion and appearance reconstruction. Some recent work only relies on the light-weight and single-view setup [70, 17, 69] and even enables hand-held capture [63, 43, 65] or drone-based capture [67]. However, these methods require the pre-scanned template or naked human model. Only recently, monocular free-form dynamic reconstruction methods [39, 15, 72, 66, 55] with real-time performance have been proposed by combining the volumetric fusion [10] and the nonrigid tracking [56, 27, 74] using RGBD camera. However, these monocular methods still suffer from the inherent self-occlusion constraint and cannot capture the motions in occluded regions. The light-weight multi-view solutions [12, 11, 68] serve as a good compromising settlement between over-demanding hardware setup and high-fidelity reconstruction but still rely on 3 to 8 RGBD streams as input. Comparably, our approach enables real-time high-quality geometry and photo-realistic texture reconstruction in novel views only using 6 RGB cameras surrounding the performer.

Data-Driven Human Modeling. Early human modeling approaches [50, 13] formulate the discriminative performance capture into a regression or classification problem using machine learning techniques. With the advent of deep neural networks, recent approaches obtain various human attributes successfully from only RGB input. Some recent work [6, 35, 24, 16, 18] learns the skeletal pose and even human shape prior by using human parametric models [3, 33]. Various approaches [57, 45, 2, 73] propose to predict human geometry from a single RGB image by utilizing parametric human model as a basic estimation. Several work [20, 42, 36, 48, 49] further reveals the effectiveness of learning the implicit occupancy directly for textured geometry modeling and even real-time inference [29]. Besides, researchers [4, 26] propose to fetch the garment or texture information of the human model. However, these data-driven human modeling methods still fail to recover fine geometry and texture results simultaneously with the level of detail present in the RGB inputs. In contrast, we explore to combine implicit geometry modeling with novel view synthesis in a data driven manner for real-time, high-quality and photo-realistic human performance rendering, achieving significant superiority to previous methods.

Neural Rendering. The recent progress of neural rendering techniques [59, 7, 64, 25] brings huge potential for constructing neural scene representations [51, 32, 52, 38] and photo-realistic novel view blending [37, 19, 61, 46]. For reconstructing neural scenes, various data representations have been explored, such as point-clouds [1, 64], voxels [51, 32] or implicit representations [52, 38, 30]. However, dedicated per-scene training is required in these methods when applying the representation to a new scene. Various methods [19, 46] learn the mapping of features from source images to novel target views to avoid per-scene training, while some recent work [61, 71] further models the

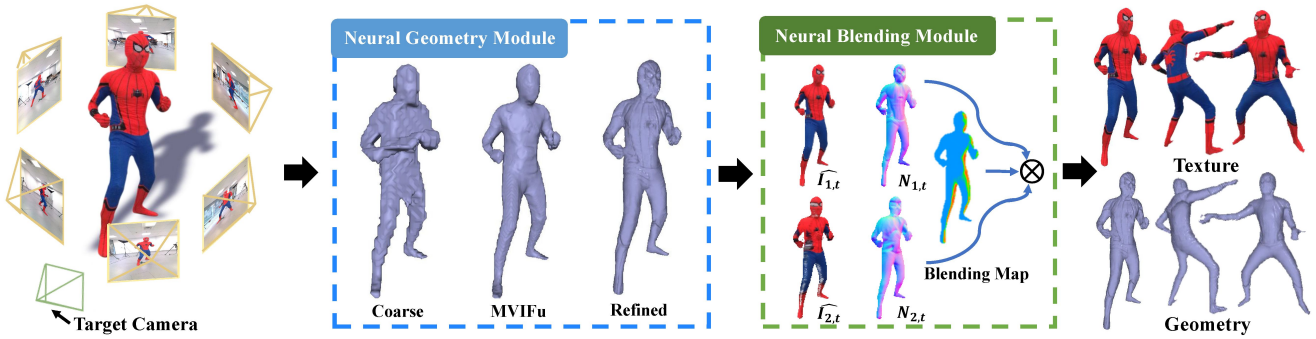

Figure 2. The pipeline of NeuralHumanFVV. Assuming the video input from six RGB cameras surrounding the performer, our approach consists of a neural geometry generation stage (Sec. 4.1) and a neural blending stage (Sec. 4.2) to generate live 4D rendering results.

view-dependent effects. However, these methods rely on heavy networks or complicated 3D proxies which are unsuitable for real-time applications like immersive telepresense. Chen *et al.* [7] propose to predict the output texture using implicit underlying geometry, which enables continues view generation from monocular image. Researchers [25, 21] also utilize such underlying latent geometry for novel view synthesis of human performance in the encoder-decoder manner. However, these approaches suffer from limited representation ability of a single latent code for complex human inferior texture output. Besides, some recent methods [41, 34] combine the neural rendering techniques to provide more visually pleasant results under the traditional RGBD fusion pipeline [12]. Comparably, our method is the first to embrace neural blending into the implicit geometry modeling pipeline under the light-weight multi-RGB and real-time setting, which enables photo-realistic texture and geometry reconstruction in novel views.

3. Overview

The proposed NeuralHumanFVV marries implicit volumetric modeling with neural texture rendering, which generates high-quality geometry and photo-realistic texture of human activities in arbitrary novel views in real-time, and enables various applications like immersive telepresense. Fig. 2 illustrates the high-level components of our system, which takes 6 RGB videos surrounding the performer as input and generates high-quality novel-view synthesis results in challenging scenarios with various poses, clothing types and topology changes as output.

Neural Geometry Generation. We first utilize the inherent geometry prior from our multi-view setting via the shape-from-silhouette [8] technique. Then, we adopt the pixel-aligned implicit function [20, 48, 49] to maintain the complete and continues geometry of the scene. Differently, we further recover the underlying geometry in novel views with a multi-stage hierarchical sampling strategy along the camera rays which enables both real-time detailed geometry inference and the following neural blending stage (Sec. 4.1).

Neural Blending. The core of our pipeline is to encode the local fine-detailed geometry and texture information of the adjacent input views into the novel target view. A novel neural blending scheme is proposed to map the input adjacent images into a photo-realistic texture output in the target view, through efficient occlusion analysis and blending weight learning. A boundary-aware upsampling strategy is further adopted to generate high resolution (e.g., 1k) novel view synthesis result without sacrificing the real-time performance. We also recover the normal information in the target view using the same neural blending strategy, which not only enhances the output fine-grained geometry details but also formulates our neural geometry and texture generation in a multi-task learning framework (Sec. 4.2).

4. NeuralHumanFVV Method

4.1. Neural Geometry Reconstruction

Given the six RGB images input at each frame, we introduce a coarse-to-fine multi-stage neural geometry reconstruction scheme to generate the inherent detailed human geometry in novel views in real-time, as illustrated in Fig. 3.

Coarse Geometry Generation. Firstly, we extract the coarse inherent geometry prior from our multi-view setting. We apply the Shape-from-Silhouette (SfS) [8] algorithm on the human masks segmented off-the-shelf video segmentation method to obtain a coarse human shape.

Accelerated Multi-View Implicit Function. We extend the pixel-aligned implicit function [48, 49] to our multi-view setting. Such multi-view implicit function (MVIFu) maintains the complete and continues geometry of the captured scene, and encodes the human shape priors. Similar to [48], the implicit function f defines the occupancy of every 3D point X in the space, which is formulated as:

$$f(\phi(X), z(X)) = s : s \in [0.0, 1.0],$$
$$\phi(X) = \frac{1}{n}\sum_{i}^{n} F_i(\pi_i(X)), \tag{1}$$

where $\pi_i()$ projects a 3D point into i-th source view; $z(X)$ is the depth value in the camera coordinate space. The projected image feature at the pixel coordinate x is formulated

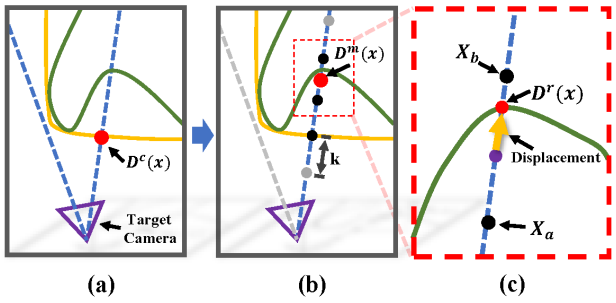

Figure 3. Illustration of our hierarchical and coarse-to-fine strategy in neural geometry generation. Orange curves are the coarse geometry surface recovered by SfS; Green curves are the real geometry. Gray dot line and points are discarded in our hierarchical sampling algorithm. (a) is the result of coarse reconstruction; (b) is the result of MVIFu; (c) is the result after refinement.

as $F_i(x) = g(I_i(x))$, where g denotes a feature extraction network.

Since extracting the whole human geometry is expensive and unnecessary for real-time immersive application, the MVIFu in our pipeline only generates geometry explicitly in the novel view. Thus, we sample evenly spaced 3D points from near to far based on the coarse geometry along each pixel ray with a distance of k in the target view. We select first two adjacent sample points to define the range where the depth value of the ray falls. Specifically, let X_a and X_b be these two points, and s_a, s_b are their occupancy, which satisfy $z(X_a) < z(X_b)$ and $s_a < 0.5, s_b \geq 0.5$. The predicted depth of this pixel x is given by $D^m(x) = \frac{z(X_a)+z(X_b)}{2}$. Moreover, point sampling after X_b can be early terminated. We also prune unnecessary sample points on the background pixel rays outside the coarse geometry generated by SfS algorithm, which inherently contains the whole performer so as to enable real-time reconstruction.

Depth Fine-tuning. The geometry D^m obtained through our accelerated MVIFu is still over smooth because of the depth averaging. In order to recover the geometry details (e.g. clothes wrinkles), we introduce a hierarchical sampling strategy. Specifically, we introduce a depth fine-tuning network h which takes the feature of the midpoint between two selected sample points as input, and outputs the displacement of depth value:

$$h(\phi(\frac{X_a + X_b}{2})) = o : o \in [-1.0, 1.0]. \tag{2}$$

Here, positions on this segment are mapped from -1.0 to 1.0 linearly, and the refined depth value $D^r(x)$ can be composed with the offset o to encode more geometry details:

$$D^r(x) = D^m(x) + k \cdot \frac{o+1}{2}. \tag{3}$$

4.2. Neural Blending

We introduce a neural blending pipeline to encode more local fine-detailed geometry and texture information of the adjacent input views than traditional image-based rendering approaches, so as to produce photo-realistic output in the target view in a data-driven manner, as illustrated in Fig. 4.

Image Warping and Occlusion Analysis. Most of the texture information in a target view can be recovered by its only two adjacent input views in our multi-view setting. Based on this finding, we first generate the depth maps of the target view (D_t^r) and the two input views (D_1^r and D_2^r, respectively) as described in Sec. 4.1. Then, we use D_t^r to warp the input image I_1 and I_2 into the target view, denoted by $I_{1,t}$ and $I_{2,t}$. We also warp source view depth maps into target view and obtain $D_{1,t}^r$ and $D_{2,t}^r$ so as to obtain the occlusion map $O_i = D_{i,t}^r - D_t^r (i = 1, 2)$, which implies the occlusion information.

Texture Blending Network(TBN). $I_{1,t}$ and $I_{2,t}$ may be incorrect due to self-occlusion and inaccurate geometry proxy. Simply blending them will raise strong artifacts. Thus, we introduce a blending network Θ_{TBN}, which utilizes the inherent global information from our multi-view setting, and fuse local fine-detailed geometry and texture information of the adjacent input views with the pixel-wise blending map W, which can be formulated as:

$$W = \Theta_{TBN}(I_{1,t}, O_1, I_{2,t}, O_2). \tag{4}$$

Boundary-Aware Depth Upsampling. For real-time perfomance, depth maps are generated at low resolution (256×256). Aiming to photo-realistic rendering, we need to upsample both the depth map and blending map to 1K resolution. However, naïve upsampling will cause severe zigzag effect near the boundary due to depth inference ambiguity. Thus, we propose a boundary-aware scheme to refine the human boundary area on the depth map. Specifically, we use bilinear interpolation to upsample D_t^r. Then a erosion operation is applied to extract boundary area. Depth values inside boundary area are recalculated by using the pipeline as described in Sec. 4.1 and form $\hat{D}_t^r$ at 1K resolution. Then we warp the original high resolution input images into the target view with $\hat{D}_t^r$ to obtain $\hat{I_{i,t}}$. To this end, our final texture blending result is formulated as:

$$I_r = \hat{W} \cdot \hat{I_{1,t}} + (1.0 - \hat{W}) \cdot \hat{I_{2,t}}, \tag{5}$$

where $\hat{W}$ is the high resolution blending map upsampled by bilinear interpolation directly.

Neural Normal Refinement. We apply networks introduced in [49] on the input RGB images to inference its normal maps. Then, the normal information in the target view is restored via the same neural blending strategy. The blended normal map N_t can further enable the geometry

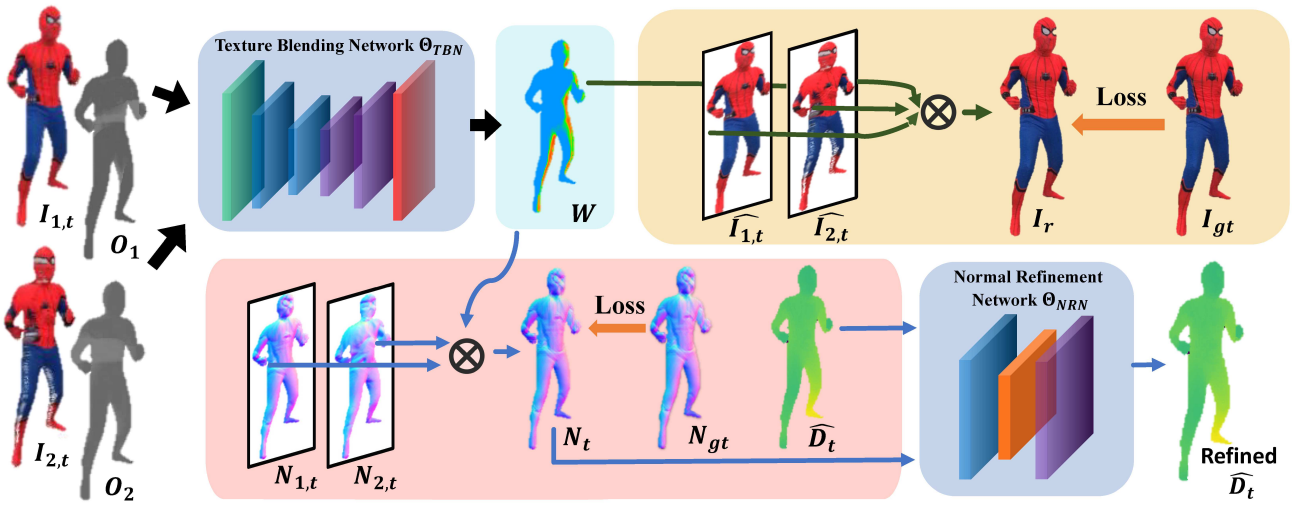

Figure 4. Illustration of our neural blending scheme, which encodes the local fine-detailed geometry and texture information of the adjacent input views into the novel target view.

refinement. Specifically, we introduce a normal refinement network Θ_{NRN} to infer the displacement of the target depth map from N_t and $\hat{D}_t^r$, as illustrated in Fig. 4.

4.3. Data and implementation Details

The key of our NeuralHumanFVV is to train the neural networks in Sec. 4.1 and Sec.4.2 properly, including the feature extractor g, continuous implicit function f, depth fine-tuning network h, as well as our TBN Θ_{TBN} and NRN Θ_{NRN}. Specifically, g is a U-Net [47] and outputs a 64 channels feature maps. f represented by MLPs has the same structure in [48] but the network dimension is further reduced for real-time performance, while h has the same network architecture, only with a different activation function of the last layer replaced by hyperbolic tangent. Besides, both our TBN Θ_{TBN} and NRN Θ_{NRN} adopt the U-Net structure.

We utilize 1820 scans from Twindom [62] and augment the dataset by rigging the 3D model to add more challenging poses, so as to enhance the generation ability of our networks. We fix the six input camera views as a rig surrounding the performer, and sample 180 virtual target views on a sphere. Note that all the 3D models locate on the central regions of the sphere and all the cameras face towards the model. Our training dataset contains the RGB images, normal maps and depth maps for all the views and models.

For the training of g and h in our MVIFu module, we only use the six input camera views in our dataset, and follow the training procedure similar to previous work [48] Then, we train our depth fine-tuning network h using the corresponding pair-wised data provided by the MVIFu.

For the training of our texture blending network Θ_{TBN}, we set out to apply a multi-task learning scheme so as to enable more robust blending weight learning. The training objective is to make both the blended texture and normal map as close as possible to the ground truth, as these two tasks share the same blending map in our Θ_{TBN}. To this end, the loss function includes a appearance term and a normal term with perceptual loss:

$$\begin{aligned}
\mathcal{L}_{rgb} &= \frac{1}{n}\sum_{j}^{n}(\|I_r^j - I_{gt}^j\|_2^2) + \|\varphi(I_r^j) - \varphi(I_{gt}^j)\|_2^2), \\
\mathcal{L}_{norm} &= \frac{1}{n}\sum_{j}^{n}(\|N_t^j - N_{gt}^j\|_2^2) + \|\varphi(N_t^j) - \varphi(N_{gt}^j)\|_2^2), \\
\mathcal{L} &= \lambda \cdot \mathcal{L}_{rgb} + (1.0 - \lambda) \cdot \mathcal{L}_{norm},
\end{aligned} \tag{6}$$

where I_{gt} and N_{gt} are the ground truth RGB images and normal maps; $\varphi(\cdot)$ denotes the output features of the third-layer of pretrained VGG-19.

Our normal refinement network (NRN) Θ_{NRN} need to be adapted to real data, which means we cannot supervise the network training using the same synthetic dataset. Thus, we introduce a self-supervise learning scheme where all the training inputs are collected from the real data generated in our pipeline. The objective is to minimize the loss function:

$$\mathcal{L} = \frac{1}{n}\sum_{j}^{n}\|\nabla(\hat{D_t^{r,j}} + \Theta_{NRN}(N_t^j, \hat{D_t^{r,j}})) - N_t^j\|_2^2 \tag{7}$$

where $\nabla(\cdot)$ is the operator which calculates the normal map from input depth map.

5. Experimental Results

In this section, we evaluate our NeuralHumanFVV method on a variety of challenging scenarios. We run our

Figure 5. The geometry and texture results of our NeuralHumanFVV on several sequences, including "spiderman", "ironman", "undressing", "floral dress", "basketball" and "backpack" from the upper left to lower right.

experiments on a PC with 3.7 GHz Intel i7-8700k CPU 32GB RAM, and Nvidia GeForce RTX3090 GPU. With the live stream data from six RGB cameras, our system generates high-quality geometry and texture results in novel views at 12 fps to enable various interactive immersive applications. The whole pipeline costs approximate 80 ms per frame, where the neural geometry generation takes 64 ms and 16 ms for the neural blending stage. Fig. 5 demonstrates several results of our NeuralHumanFVV, which can generate free-view high quality geometry and texture results simultaneously. Noted that our approach can handle human object interaction scenarios with topology changes, such as playing basketball, carrying bag and removing clothes.

5.1. Comparison

In our real testing data, performers act complex motions with self-occlusion, and dress rich texture clothes, such as floral skirt and plaid shirt. For thorough comparison, we compare our NeuralHumanFVV against the state-of-the-art methods MonoPort [29], Multi-PIFu [48] and Continuous View Control [7] both in geometry and texture. As shown in Fig. 6, our approach achieves significantly better texture and detailed geometry results even when the garment is extraordinarily complex. Even applying per-vertex texture mapping in the geometry from Multi-PIFu [48], image blurs occur much more frequently in contrast to our results.

Method	*MonoPort*	*Multi-PIFu*	*Multi-PIFu**	*CVC*	**Ours**
RGB_1	95.1±4.8	67.5±3.8	43.7±1.9	98.1±21.9	**27.6 ± 1.6**
RGB_6	144.8±6.6	66.4±2.2	56.1±2.1	103.1±27.4	**26.1 ± 1.2**

Table 1. Quantitative comparison of MonoPort [29], Multi-PIFu [48], Continuous View Control [7] and NeuralHumanFVV. Multi-PIFu* denotes per-vertex texture mapping using the geometry from Multi-PIFu as input. RGB_1 and RGB_6 respectively present the MAE in one view and six views.

Then, we make a quantitative comparison on our real testing dataset. The mean absolute error (MAE) is adopted as error metric and we average all MAEs from all images and frames for overall MAE calculation. Since Monoport [29] only takes one image as input, we also evaluate methods with single camera input (RGB_1), compared with using all six cameras (RGB_6). As illustrated in Table. 1, our approach outperforms other methods in all scenarios

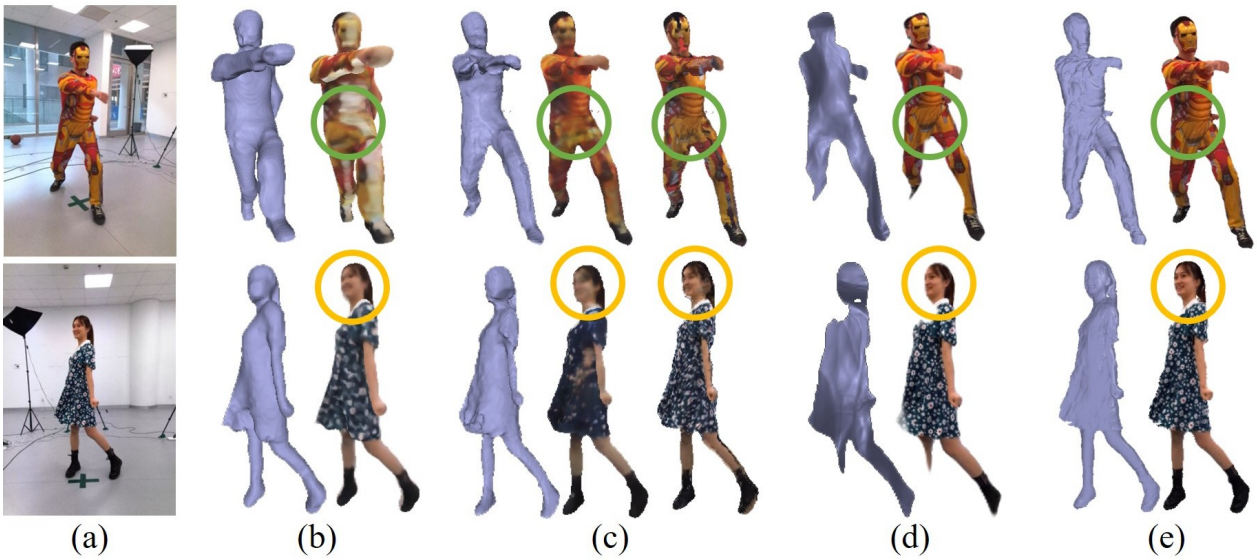

Figure 6. Qualitative comparison. (a) Input images. (b-e) are the geometry and texture results from MonoPort [29], Multi-PIFu [48], Continuous View Control [7] and ours, respectively. Note that the two texture results in (c) corresponds to the implicit texture and per-vertex texture, respectively.

with distinct differences in overall MAE.

Furthermore, we make a comparison on a synthetic dynamic sequence with 600 frames, and generate 90 different target views to evaluate the MAE. The result is shown in Fig. 7. Our method can stay lowest MAE in the entire sequence.

5.2. Ablation Study

Neural Geometry Generation. Here, we evaluate our neural geometry generation scheme. As shown in Fig. 8 (b), the results from SfS [8] only provide coarse geometry priors since only boundary information are utilized. Our scheme without the normal refinement in Fig. 8 (c) can generate mid-level geometry details such as the clothing wrinkles but still suffers from over-smooth results, especially on the face regions. In contrast, our approach with full pipeline in Fig. 8 (d) enables high-quality geometry detail generation almost with the level of details present in the input images.

Neural Texture Blending. We further evaluate our neural texture blending scheme. In Fig. 9, we compare with our variations with different texturing schemes using the same geometry proxy. The per-vertex texturing in Fig. 9 (a) suffers from severe block artifacts, while the offline scheme using the software AGI [44] in Fig. 9 (b) causes inferior results in those regions near the stitching seams. And our neural scheme at low resolution situation in Fig. 9 (c) and the one without the boundary optimization in Fig. 9 (d) suffer from over-smooth texture or coarse boundary, respectively. In contrast, our full neural texture scheme in Fig. 9 (e) enables photo-realistic texture reconstruction in novel views.

For quantitative analysis of the individual components of NeuralHumanFVV, we utilize two different geometries as bases and two different texture methods to make a comparison among these four outputs as shown in as Fig. 10. Fig. 10 (a) is from complete NeuralHumanFVV while Fig. 10 (b) using the geometry from Multi-PIFu, Fig. 10 (c) using the same geometry as Fig. 10 (a) but per-vertex texture mapping, and Fig. 10 (d) is yielded by Multi-PIFu and per-vertex texture mapping. Not only our image outcome which has fewer artifacts but also the per-frame mean error for the three variation of our approach without model completion in Fig. 10 (e) shows the advancement of our NeuralHumanFVV.

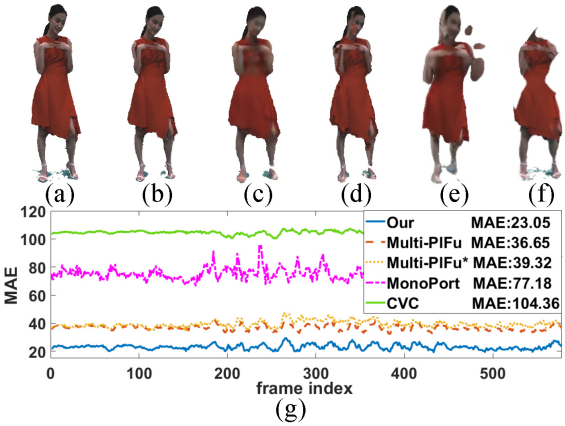

Figure 7. Quantitative and qualitative comparison on synthesis sequence against NeuralHumanFVV, Multi-PIFu [48], MonoPort [29] and Continuous View Control [7].(a) Color image in ground truth; (b) NeuralHumanFVV; (c) Multi-PIFu; (d)Multi-PIFu* (e) MonoPort; (f) Continuous View Control; (g)Error curves. Denote that Multi-PIFu* is the result of per-vertex texture mapping using the geometry from Multi-PIFu.

Camera Number. To evaluate the influence of input views in our multi-view setting, we compare to the variation of our

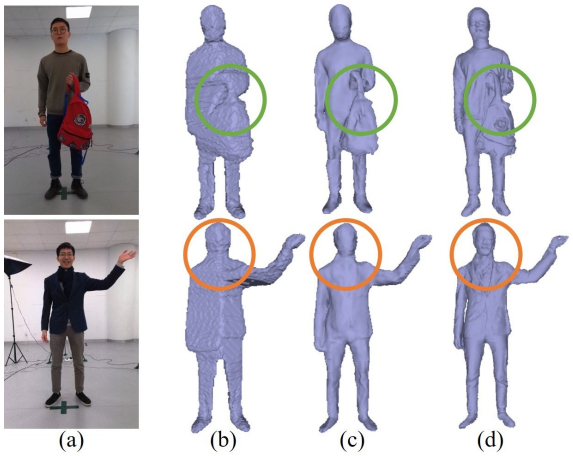

Figure 8. Evaluation of our neural geometry generation. (a) Input images. (b) Geometry from SfS [8]; (c) Geometry without normal refinement; (d) Geometry from NeuralHumanFVV.

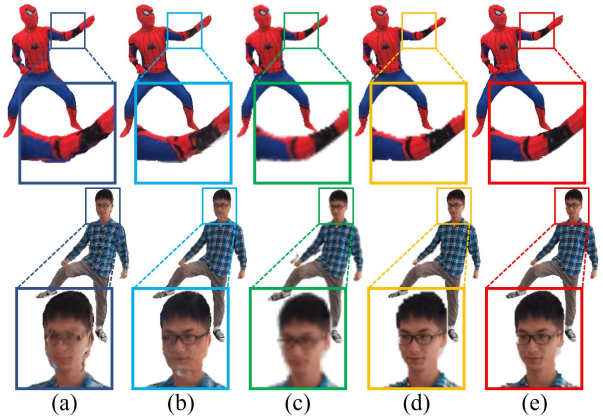

Figure 9. Qualitative evaluation of our neural texture blending scheme. (a) Per-vertex texture mapping; (b) AGI[44]; (c) NeuralHumanFVV at 256 $\times$ 256 resolution ; (d) NeuralHumanFVV without boundary optimization; (e) NeuralHumanFVV.

pipeline using various numbers of input camera views. As shown in Fig. 11, the reconstruction results without enough camera views suffer from severe geometry and blending artifacts and the average error increases significantly as the the camera number decreases. Empirically, the setting with six cameras serve as a good compromising settlement.

5.3. Limitation

As the first trial to enable real-time and photo-realistic neural human performance rendering from only sparse RGB inputs, the proposed NeuralHumanFVV system still own some limitations. First, inaccuracy of segmentation leads to incomplete regions in our final synthesized images. Our texturing results also depend on the input image resolutions. Thin structures like fingers are difficult to reconstruct due to the input with limited resolution. Besides, our approach generates plausible geometry detail from RGB images. But similar to other RGB-based methods, the recovered geomegtry details will be physically inaccurate when the testing images deviate much from the training ones.

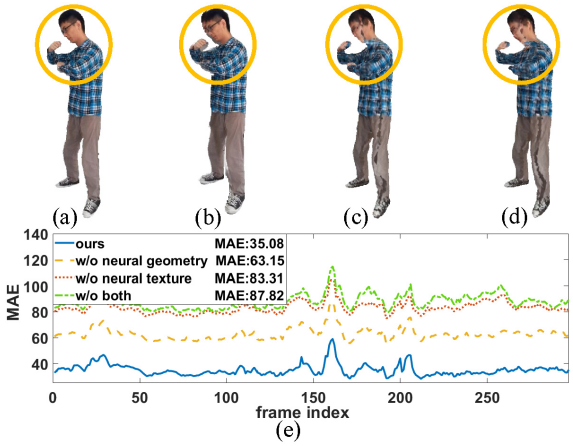

Figure 10. Quantitative evaluation. (a-d) The reconstructed results of ours, w/o neural geometry, w/o neural texture, w/o both.(e) Numerical error curves.

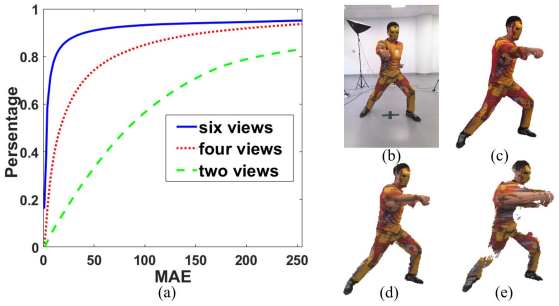

Figure 11. Evaluation of the number input camera views. (a) Cumulative distribution function of the mean absolute error. (b) The reference capture scene. (c, d, e) Our reconstructed texture results using six, four and two cameras, respectively.

6. Conclusion

We have presented a real-time neural performance rendering system to generate high-quality geometry and photo-realistic textures of human activities in novel views only using sparse multiple RGB cameras. Our neural geometry generation benefits inherently from our multi-view setting and enables efficient and implicit reasoning of underlying geometry in a novel view. Our neural blending scheme with occlusion analysis and boundary-aware upsampling further enables to recover high resolution (e.g., 1k) and photo-realistic textures without sacrificing the real-time performance. Our experimental results demonstrate the effectiveness of NeuralHumanFVV for high-quality human performance rendering in challenging scenarios with various poses, clothing types and topology changes. We believe that our approach is a critical step to virtually but realistic teleport human performances, with many potential applications in VR/AR like gaming, entertainment and immersive telepresense.

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. *arXiv preprint arXiv:1906.08240*, 2019.

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, page 408–416, New York, NY, USA, 2005. Association for Computing Machinery.

[4] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019.

[5] Chris Bregler and Jitendra Malik. Tracking people with twists and exponential maps. In *Computer Vision and Pattern Recognition (CVPR)*, 1998.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Xu Chen, Jie Song, and Otmar Hilliges. Monocular neural image based rendering with continuous view control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4089–4099, October 2019.

[8] Kong Man Cheung, Simon Baker, and Takeo Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I, 2003.

[9] 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.

[10] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 303–312, New York, NY, USA, 1996. ACM.

[11] 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 Trans. Graph.*, 36(6):246:1–246:16, Nov. 2017.

[12] 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.

[13] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real time motion capture using a single time-of-flight camera. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 755–762. IEEE, 2010.

[14] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, and et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.*, 38(6), Nov. 2019.

[15] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo and motion reconstruction using a single rgbd camera. *ACM Transactions on Graphics (TOG)*, 2017.

[16] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

[18] Yannan He, Anqi Pang, Xin Chen, Han Liang, Minye Wu, Yuexin Ma, and Lan Xu. Challencap: Monocular 3d capture of challenging human performances using multi-modal references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[19] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Trans. Graph.*, 37(6), Dec. 2018.

[20] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[21] Shi Jin, Ruiynag Liu, Yu Ji, Jinwei Ye, and Jingyi Yu. Learning to dodge a bullet: Concyclic view morphing via deep learning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 230–246, Cham, 2018. Springer International Publishing.

[22] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015.

[23] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[24] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[25] Youngjoong Kwon, Stefano Petrangeli, Dahun Kim, Haoliang Wang, Eunbyung Park, Viswanathan Swaminathan,

and Henry Fuchs. Rotationally-temporally consistent novel view synthesis of human performance video. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 387–402, Cham, 2020. Springer International Publishing.

[26] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *International Conference on 3D Vision (3DV)*, sep 2019.

[27] Hao Li, Bart Adams, Leonidas J Guibas, and Mark Pauly. Robust single-view geometry and motion reconstruction. *ACM Transactions on Graphics (ToG)*, 28(5):1–10, 2009.

[28] Hao Li, Linjie Luo, Daniel Vlasic, Pieter Peers, Jovan Popović, Mark Pauly, and Szymon Rusinkiewicz. Temporally coherent completion of dynamic shapes. *ACM Transactions on Graphics (TOG)*, 31(1):1–11, 2012.

[29] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 49–67, Cham, 2020. Springer International Publishing.

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

[31] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2720–2735, 2013.

[32] 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.

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, Oct. 2015.

[34] Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, and et al. Lookingood: Enhancing performance capture with real-time neural re-rendering. *ACM Trans. Graph.*, 37(6), Dec. 2018.

[35] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4), 2017.

[36] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[37] Moustafa Meshry, Dan B. Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[38] 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 Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 405–421, Cham, 2020. Springer International Publishing.

[39] 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.

[40] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *Proc. of ISMAR*, pages 127–136, 2011.

[41] Rohit Pandey, Anastasia Tkach, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Ricardo Martin-Brualla, Andrea Tagliasacchi, George Papandreou, Philip Davidson, Cem Keskin, Shahram Izadi, and Sean Fanello. Volumetric capture of humans with a single rgbd camera via semi-parametric learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[42] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, June 2019.

[44] Agisoft photoscan professional. http://www.agisoft.com/downloads/installer/, 2019.

[45] Albert Pumarola, Jordi Sanchez-Riera, Gary P. T. Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[46] Gernot Riegler and Vladlen Koltun. Free view synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Cham, 2020. Springer International Publishing.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[48] 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 *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[49] 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 (CVPR)*, June 2020.

[50] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[51] 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 (CVPR)*, June 2019.

[52] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 1121–1132. Curran Associates, Inc., 2019.

[53] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1395, 2017.

[54] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Fast articulated motion tracking using a sums of Gaussians body model. In *International Conference on Computer Vision (ICCV)*, 2011.

[55] 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 Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 246–264, Cham, 2020. Springer International Publishing.

[56] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics (TOG)*, 26(3):80, 2007.

[57] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. A neural network for detailed human depth estimation from a single image. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[58] Jonathan Taylor, Jamie Shotton, Toby Sharp, and Andrew Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 103–110. IEEE, 2012.

[59] 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.

[60] Christian Theobalt, Edilson de Aguiar, Carsten Stoll, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from multi-view video. In *Image and Geometry Processing for 3-D Cinematography*, pages 127–149. Springer, 2010.

[61] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. Image-guided neural object rendering. In *International Conference on Learning Representations*, 2020.

[62] Twindom dataset. `https://https://web.twindom.com/`.

[63] Chenglei Wu, Carsten Stoll, Levi Valgaerts, and Christian Theobalt. On-set performance capture of multiple actors with a stereo camera. *ACM Transactions on Graphics (TOG)*, 32(6):1–11, 2013.

[64] 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 (CVPR)*, June 2020.

[65] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[66] 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.

[67] Lan Xu, Yebin Liu, Wei Cheng, Kaiwen Guo, Guyue Zhou, Qionghai Dai, and Lu Fang. Flycap: Markerless motion capture using multiple autonomous flying cameras. *IEEE Transactions on Visualization and Computer Graphics*, 24(8):2284–2297, Aug 2018.

[68] 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.

[69] 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 (CVPR)*, June 2020.

[70] 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):27:1–27:15, 2018.

[71] Zexiang Xu, Sai Bi, Kalyan Sunkavalli, Sunil Hadap, Hao Su, and Ravi Ramamoorthi. Deep view synthesis from sparse photometric images. *ACM Trans. Graph.*, 38(4), July 2019.

[72] 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. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[73] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a sin-

gle image. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[74] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. Real-time Non-rigid Reconstruction using an RGB-D Camera. *ACM Transactions on Graphics (TOG)*, 33(4):156, 2014.