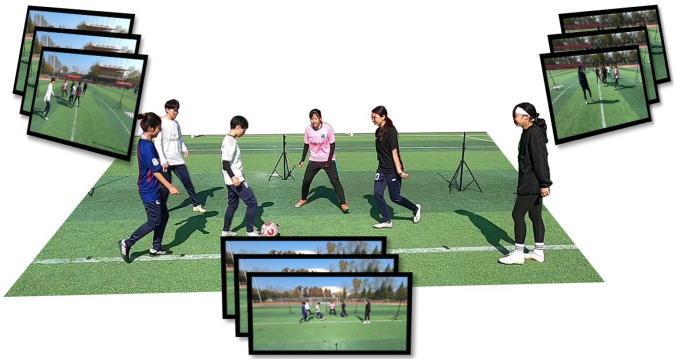
Novel View Synthesis of Human Interactions from Sparse Multi-view Videos

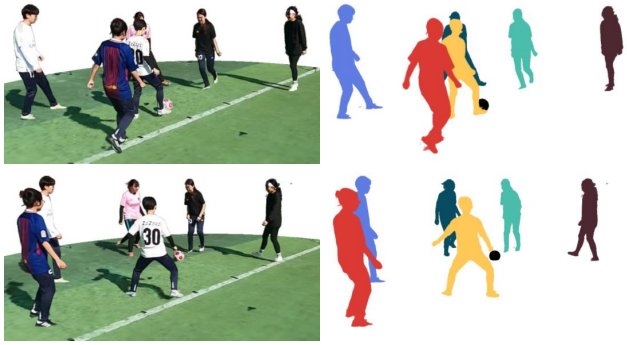
Qing Shuai
Chen Geng
Qi Fang
s_q@zju.edu.cn
gengchen@zju.edu.cn
fangqi19@zju.edu.cn
State Key Laboratory of CAD&CG,
Zhejiang University
China

Sida Peng
Wenhao Shen
pengsida@zju.edu.cn
shenwenhao@zju.edu.cn
State Key Laboratory of CAD&CG,
Zhejiang University
China

Xiaowei Zhou
Hujun Bao*
xwzhou@zju.edu.cn
bao@cad.zju.edu.cn
State Key Laboratory of CAD&CG,
Zhejiang University
China



Input: sparse multi-view videos (eight views in total)



Output: novel view synthesis and instance segmentation

Figure 1: Given sparse multi-view videos of human performers, our approach is able to generate high-fidelity novel views and accurate instance masks even for crowded scenes. Please refer to the supplementary material for the synthesized free-viewpoint video.

ABSTRACT

This paper presents a novel system for generating free-viewpoint videos of multiple human performers from very sparse RGB cameras. The system reconstructs a layered neural representation of the dynamic multi-person scene from multi-view videos with each layer representing a moving instance or static background. Unlike previous work that requires instance segmentation as input, a novel approach is proposed to decompose the multi-person scene into layers and reconstruct neural representations for each layer in a weakly-supervised manner, yielding both high-quality novel view rendering and accurate instance masks. Camera synchronization error is also addressed in the proposed approach. The experiments demonstrate the better view synthesis quality of the proposed system compared to previous ones and the capability of producing an editable free-viewpoint video of a real soccer game using several asynchronous GoPro cameras. The dataset and code are available at https://github.com/zju3dv/EasyMocap.

CCS CONCEPTS

• **Computing methodologies** → **Image-based rendering**.

KEYWORDS

Novel view synthesis, neural rendering, dynamic scene modeling

ACM Reference Format:
Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. 2022. Novel View Synthesis of Human Interactions from Sparse Multi-view Videos. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings), August 7–11, 2022, Vancouver, BC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3528233.3530704

*Corresponding author.

SIGGRAPH '22 Conference Proceedings, August 7–11, 2022, Vancouver, BC, Canada

ACM ISBN 978-1-4503-9337-9/22/08…$15.00
https://doi.org/10.1145/3528233.3530704

1 INTRODUCTION

Synthesizing free-viewpoint videos of human performers is a long-standing problem with wide applicability in immersive viewing experience and telepresence. Unlike static scenes, reconstructing high-fidelity human models especially in crowded scenes for novel view synthesis faces inevitable challenges such as nonrigid motion,

complicated appearance variation and severe occlusion due to close interaction.

Traditional systems either adopt multi-view stereo techniques [Collet et al. 2015; Guo et al. 2019] to reconstruct textured meshes explicitly, or perform image-based rendering via view interpolation [Gortler et al. 1996; Zitnick et al. 2004]. However, all of them require a dense rig of cameras which suffer from high cost and limited mobility. To reduce the complexity of capture systems, some works leverage the commodity depth sensors to build real-time reconstruction systems [Dou et al. 2016; Newcombe et al. 2015; Yu et al. 2021b, 2018], but they are inapplicable in outdoor or large-scale scenes.

Recently, neural scene representation-based techniques [Mildenhall et al. 2020; Niemeyer et al. 2020; Sitzmann et al. 2019] have demonstrated superior performance in modeling scenes and synthesizing photorealistic novel views, given an abundant number of input views. This type of approach has also shown promising results in modeling humans [Lombardi et al. 2019; Wang et al. 2021a; Zhang et al. 2021], while they still require relatively dense multi-view videos as input. To reduce the required number of input views, some recent works [Liu et al. 2021; Peng et al. 2021] propose to utilize human body priors to assist the learning of human representations. For example, NeuralBody first fits an SMPL model [Loper et al. 2015] to the input videos and then learns a set of latent codes that are anchored on the SMPL model to represent geometry and appearance. But these methods are limited to single-person capture.

In this paper, we aim to solve the challenging problem of novel view synthesis of multiple closely interacting human performers from a sparse array of calibrated and roughly synchronized RGB cameras. The background is assumed to be static with simple geometry, e.g., a ground plane. A plausible solution to this problem is to learn a layered scene representation in which each layer represents a human instance or background [Lu et al. 2020; Zhang et al. 2021]. However, this approach generally requires instance segmentation as a preprocessing step and may suffer from inaccuracy of instance masks especially for close interactions, as shown in Fig. 2, which will hinder the subsequent reconstruction and rendering quality. Instead, we propose to learn the layered scene representation and assign the image pixels into different layers simultaneously in a weakly-supervised manner. All layers of radiance fields are jointly learned by directly minimizing the rendering loss, i.e., the difference between the rendered images and input images. Different from [Zhang et al. 2021], we enforce that the color of a pixel (ray) is contributed by only one layer. To this end, we introduce sparsity loss and keypoint loss on the layer logits of each pixel (a one-hot vector that indicates which layer the pixel belongs to). The experiments show that minimizing our loss results in high-quality reconstruction of the layered scene representation and accurate instance masks derived from the layer logits, as demonstrated in Fig. 1.

Two additional challenges exist in practice. The first is the multi-camera synchronization error that causes misalignment between the reconstructed 3D geometry and the images, resulting in blurring and artifacts in rendering. To solve this problem, we propose a novel pose-guided synchronization strategy to compensate for the synchronization error. Another challenge is the moving objects. Accurate modeling of them requires to track their 6DoF poses, which are impractical if the objects are small and fast moving. So we only consider balls in this work and model each ball as a radiance field with constant densities, time-varying colors and translational motion across frames.

Figure 2: Instance masks given by a pretrained segmentation network [Li et al. 2020b] (middle) and our approach (right).

In summary, our main contributions are:

- A novel system for producing editable free-viewpoint video of multiple performers under close interactions from very sparse RGB cameras.
- A new algorithm that is able to decompose the multi-person scene into human instances in a weakly-supervised manner and reconstruct high-quality neural representations for each instance.
- A method to address the camera synchronization error with the guidance of human poses.

2 RELATED WORK

Performance capture. Recently, many works aim to solve human motion capture using RGB cameras. To recover the skeleton motion of human, some previous works propose optimization-based solutions, solving the cross-view matching [Dong et al. 2021; Vo et al. 2020b] or 4D graph parsing [Zhang et al. 2020a], while others use neural networks to regress human poses directly in an end-to-end manner [Iskakov et al. 2019; Tu et al. 2020; Wang et al. 2021b]. To capture volumetric videos of human performers, traditional paradigms require a dense array of cameras [Collet et al. 2015; De Aguiar et al. 2008; Gall et al. 2009; Guo et al. 2019], which are inaccessible for nonprofessionals. Others achieve impressive performance with sparse depth sensors [Dou et al. 2016; Newcombe et al. 2015; Su et al. 2020; Wu et al. 2020; Yu et al. 2021b, 2018], which are however impractical for outdoor or large-scale scenes. For performance capture from monocular videos, some approaches [Habermann et al. 2019, 2020] rely on a person-specific 3D template model and deform it through dense non-rigid tracking. Instead, our work doesn't need a pre-scanned person-specific template model. Template-free methods can produce detailed human meshes beyond capability of the parametric human representation by leveraging image information such as silhouette [Natsume et al. 2019; Zhu et al. 2019]. Some methods attempt to directly regress detailed 3D geometry of human [Huang et al. 2018; Li et al. 2020a; Saito et al. 2019; Suo et al. 2021; Varol et al. 2018; Zheng et al. 2021, 2019] from RGB images but may suffer from limited generalization abilities in practice.

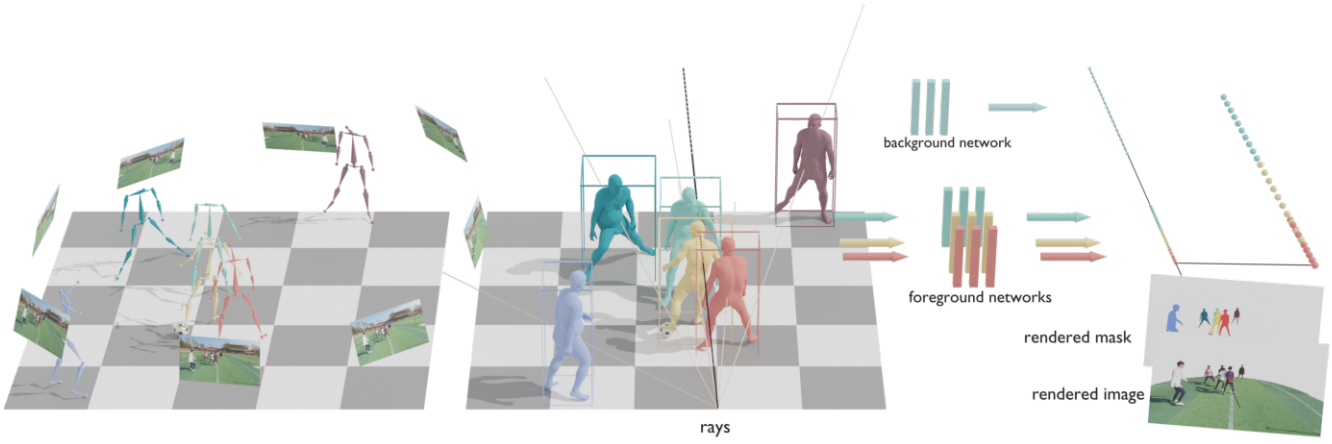


Figure 3: Overview of our method. (a) Given multi-view videos, the skeletal motion of humans and rigid motion of moving objects are first recovered. (b) The SMPL+H models [Romero et al. 2017] are fitted to the estimated skeletons and implicit neural representations of each foreground entity and the background are jointly learned without any mask supervision. (c) The learned neural representations are able to render images and instance masks via volume rendering, which support 360° novel view synthesis and content editing.

Free Viewpoint Rendering. It has been a long-standing problem in the community to render unseen views given a few input views. Most traditional methods use view interpolation, stereo matching, or CNN-based image synthesis to render novel views[Bansal et al. 2020; Levoy and Hanrahan 1996; Shum and Kang 2000]. However, they can only render novel views which are close to input views. Representing a scene with neural implicit functions has been a high-profile research direction recently [Lin et al. 2021; Liu et al. 2020; Mildenhall et al. 2020; Niemeyer et al. 2020; Reiser et al. 2021; Sitzmann et al. 2019; Yu et al. 2021a; Zhang et al. 2020b]. The seminal work NeRF [Mildenhall et al. 2020] models the scene as a neural radiance field and achieves realistic novel view synthesis with volume rendering. Some works also concentrate on object representation[Granskog et al. 2021; Guo et al. 2020; Yang et al. 2021; Yu et al. 2022]. However, these works only focus on static scenes. Several recent works dedicate themselves to handling general dynamic scenes by explicitly modeling motion with rigid transformations [Yuan et al. 2021], nonrigid deformation fields [Gao et al. 2021; Park et al. 2021; Pumarola et al. 2021; Tretschk et al. 2021] and scene graphs [Ost et al. 2021], or learning a time-conditioned dynamic radiance field [Du et al. 2021; Li et al. 2021; Xian et al. 2021]. Compared to these works, we aim to achieve 360° novel view rendering of a scene with close human interactions from sparse input views.

Neural Representations for Humans. Neural modeling of humans has also been deeply investigated. NeuralVolume [Lombardi et al. 2019] is among the first to learn a deep feature volume to represent humans and achieve realistic re-rendering. PIFu [Saito et al. 2019] learns a pixel-aligned implicit function to represent human bodies. NeuralBody [Peng et al. 2021] establishes dense correspondences between video frames by a fitted SMPL model and thus allows for reconstructing NeRF from sparse multi-view videos. MVP [Lombardi et al. 2021] proposes a hybrid representation for efficient rendering of neural avatars. Some other works [Liu et al. 2021; Noguchi et al. 2021] attempt to learn riggable human models from videos for animation. Most of the above works focus on modeling a single human body instead of a dynamic scene including multiple interacting people. ST-NeRF [Zhang et al. 2021] handles the multi-person cases with moderate number of cameras. They propose a layered neural representation to model all entities in the scene and adopt a deformation representation similar to D-NeRF [Pumarola et al. 2021]. Although impressive results have been achieved, human priors are not sufficiently utilized and instance segmentation is required, so the rendering quality degrades in the scenario of sparse input views and close interactions as shown in our experiments.

3 METHODS

Fig. 3 demonstrates the pipeline of our method. Given a dynamic scene with multiple human performers captured by sparse RGB cameras, our goal is to generate editable free-viewpoint videos. We first capture human skeletal motion and object motion (Sec. 3.1), define a layered neural scene representation (Sec. 3.2) and finally learn the representation from input videos (Sec. 3.3).

3.1 Multi-entity motion capture

Data Collection and Camera Calibration. Our capture system only uses a few RGB cameras and thus can be used both indoors and outdoors. The sequence shown in Fig. 1 were recorded with eight portable GoPro cameras at 60 fps. These cameras were calibrated using a calibration board and synchronized by manually selecting key frames.

Detection and Matching. Given synchronized and calibrated multi-view videos, we deal with the humans and objects separately. An off-the-shelf 2D human pose detector [Cao et al. 2017] and an object detector[Redmon et al. 2016] are used to detect 2D human keypoints and object bounding boxes for each frame respectively. Then, we perform matching to find cross-view correspondences of humans

and objects from different views. For humans, we construct the cross-view affinity matrix and solve the multi-view matching problem with an existing algorithm [Dong et al. 2021]. For objects, we simply regard the center of the 2D bounding box as a keypoint and solve the cross-view correspondences similarly. After the matching, the 3D keypoint trajectories of each entity can be recovered via triangulation, which are shown in Fig. 3 (a).

Parametric Model Fitting. As 2D detection results suffer from noises and occlusions in interaction scenes, fitting a parametric model to keypoints in the whole sequence can not only suppress keypoint detection errors by imposing human structural constraints, but also complete missing trajectories.

We adopt SMPL+H [Romero et al. 2017] as the parametric human representation and fit multiple SMPL+H models to the multi-view videos. Let $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ be the disentangled pose parameters and shape parameters of the model. $\boldsymbol{R}$ and $\boldsymbol{T}$ are the global rotation and translation, respectively. $\boldsymbol{M}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{T})$ is a differentiable function that maps the parameters to a mesh. $\mathcal{J}$ is a pre-trained linear regressor used for generating 3D human keypoints from the mesh.

We fit the parametric model to estimated 3D keypoints. The losses of this stage mainly consist of a 3D distance term $\mathcal{L}_{\mathbf{3d}}$ that computes the sum of distances between the detected 3D keypoints and their correspondences on the SMPL-H model, a regularization term $\mathcal{L}_{\mathbf{reg}}$ that constrains the SMPL-H parameters to be in a reasonable range [Pavlakos et al. 2019], and a temporal smoothness term $\mathcal{L}_{\mathbf{t}}$ that penalizes the differences of parameters between consecutive frames. The definition of $\mathcal{L}_{\mathbf{reg}}$ is the L2 regularization of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. $\mathcal{L}_{\mathbf{t}}$ is the L2 regularization of the velocity of parameter $\boldsymbol{\theta}, \boldsymbol{R}, \boldsymbol{T}$. The entire loss function for motion capture $\mathcal{L}_{\mathbf{mocap}}$ is the sum of above losses:

$$\mathcal{L}_{\mathbf{mocap}} = \mathcal{L}_{\mathbf{3d}} + \lambda_{\mathbf{reg}}\mathcal{L}_{\mathbf{reg}} + \lambda_{\mathbf{t}}\mathcal{L}_{\mathbf{t}}, \tag{1}$$

The main loss is the 3d keypoints position loss:

$$\mathcal{L}_{\mathbf{3d}} = \sum \boldsymbol{w}_{\mathbf{3d}} \cdot ||\mathcal{J}\boldsymbol{M}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{T}) - \boldsymbol{J}||_2^2, \tag{2}$$

where $\boldsymbol{J}$ and $\boldsymbol{w}_{\mathbf{3d}}$ are the reconstructed 3D human keypoints and confidence scores for visibility, respectively. The summation operation is performed over all keypoints and frames.

Pose-guided synchronization. In real applications, capturing precisely synchronized videos is usually not easy particularly for in-the-wild scenes where hard synchronization is impractical. Therefore, it is necessary to process asynchronous frames in advance, otherwise the subsequent novel view synthesis may suffer from blurring and artifacts due to the misalignment of multiple views. The first row of Fig. 4 shows this phenomenon. Even when we manually select the nearest frames among views, they are not consistent because of the fast human motion. In order to alleviate this problem, we propose a pose-aware synchronization strategy, which compensates for the synchronization error by minimizing the 2D reprojection loss [Bogo et al. 2016; Pavlakos et al. 2019] of all V views:

$$\min_{[\Delta t_v]} \sum_{v=1}^{V} \mathcal{L}_{2d}(\boldsymbol{\theta} + \Delta t_v \dot{\boldsymbol{\theta}},\ \boldsymbol{\beta},\ \boldsymbol{R} + \Delta t_v \dot{\boldsymbol{R}},\ \boldsymbol{T} + \Delta t_v \dot{\boldsymbol{T}},\ \boldsymbol{W}_v,\ \boldsymbol{w}_v), \tag{3}$$

where Δt_v is the temporal offset of view v to be solved, and $\dot{\boldsymbol{\theta}}$, $\dot{\boldsymbol{R}}$, $\dot{\boldsymbol{T}}$ are the estimated velocities of corresponding variables, which are approximated as the difference between two adjacent frames.

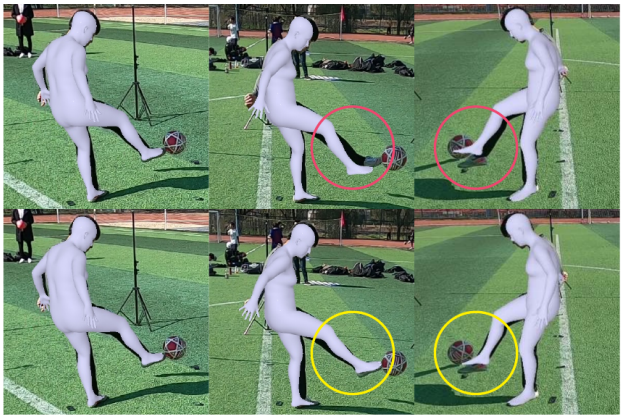


Figure 4: Misalignment caused by asynchronous devices. This figure shows the re-projection of fitting results to different views. The three columns show the reference, lagging and advanced frames, respectively. As the first row shows, even we fit the model to the 3D keypoints, there still exists misalignment in the fast moving regions (highlighted by red circles). The second row shows the fitting results considering the synchronization. The re-projection errors are reduced significantly (highlighted by yellow circlels).

Our method leverages accurate 2D human poses and is able to achieve subframe synchronization accuracy. Finally, the optimized independent temporal offset is applied to each view and more accurate fitting is obtained, as shown in the second row of Fig. 4.

3.2 Layered neural scene representation

Our implicit neural representation follows the NeRF [Mildenhall et al. 2020; Oechsle et al. 2021]. The color of each ray $\mathbf{r}$ is approximated using numerical quadrature. We sample N points in the ray and calculate the occupancy o and the color $\boldsymbol{c}$ by a neural network:

$$\hat{\boldsymbol{C}}(\mathbf{r}) = \sum_{i=1}^{N} o(\boldsymbol{p}_i) \prod_{j<i} (1 - o(\boldsymbol{p}_j)) \boldsymbol{c}(\boldsymbol{p}_i, \mathbf{r}). \tag{4}$$

To model the whole scene including multiple independently moving instances, we adopt a layered scene representation where each layer is a neural radiance field representing a human, object or background, similar to ST-NeRF [Zhang et al. 2021].

Human representation. We follow NeuralBody [Peng et al. 2021] to leverage human motion priors. NeuralBody adopts the parametric human model SMPL [Loper et al. 2015] and defines a set of latent codes on vertices of the SMPL model, followed by a code diffusion process so as to obtain the latent code at any location around the surface. Then a multilayer perceptron (MLP) maps the latent codes to the density and color values at any queried position.

Objects representation. To accurately model general objects, one needs to track the 6DoF motions of objects, which is challenging and beyond the scope of this work. Thus, in this work, we only consider balls and model each ball as a radiance field with time-varying colors and translational motion across frames. The trajectory of the ball is used as the spatial anchor when sampling the points. We convert

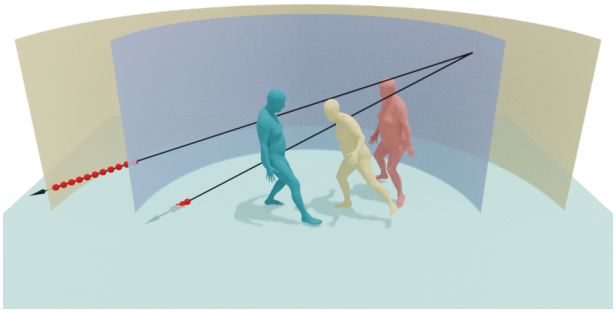


Figure 5: Illustration of the background layer. We simplify the background with the assumption that the actors interact in a limited region. The ground plane of this region is known and the radius of this region can be reasoned from the trajectories of humans. So we just sample points around the ground plane or outside this region (red points).

the points of object m in the frame t to canonical space by the location $\boldsymbol{T}_t^m$. Specifically, given point $\boldsymbol{p}_{it}$, the neural representation of this object $\hat{o}^m, \hat{\boldsymbol{c}}^m$ in canonical space and the temporal latent code of this frame $\boldsymbol{l}_t$, we calculate the density and color by

$$o^m(\boldsymbol{p}_{it}), \boldsymbol{c}^m(\boldsymbol{p}_{it}, \boldsymbol{d}) = \hat{\boldsymbol{o}}^m(\boldsymbol{p}_{it} - \boldsymbol{T}_t^m), \hat{\boldsymbol{c}}^m(\boldsymbol{p}_{it} - \boldsymbol{T}_t^m, \boldsymbol{d}, \boldsymbol{l}_t) \quad (5)$$

Background representation. We also use NeRF [Mildenhall et al. 2020] with temporal latent codes to represent the background and ground. As we only use 8 cameras to cover 360 degrees, compared to 16 cameras with 180 degrees viewing range in ST-NeRF, reconstructing the background from such sparse inputs is ambiguous. To make the background learning better constrained and more efficient, we only sample the points in a limited region, as shown in Fig. 5. There are also shadows in the real environment. We do not explicitly model the dynamic shadows as the lighting conditions are unknown. Instead, we model the changing shadows as a part of dynamic background with the temporal latent codes.

Rendering. Layered rendering [Zhang et al. 2021] is performed to synthesize images. Specifically, for each ray, we first calculate its intersections with each 3D bounding box. If this ray is intersected with entity m, then we uniformly sample N points between the two intersection points $(\boldsymbol{p}_{\text{near}}, \boldsymbol{p}_{\text{far}})$:

$$\boldsymbol{p}_i^m = \boldsymbol{p}_{\text{near}}^m + \frac{i-1}{N}(\boldsymbol{p}_{\text{far}} - \boldsymbol{p}_{\text{near}}), i = 1, 2, \ldots, N. \quad (6)$$

For each point $\boldsymbol{p}_i^m$, it is fed into its neural network and get its occupancy $o^m(\boldsymbol{p}_i^m)$ and color $\boldsymbol{c}^m(\boldsymbol{p}_i^m, \boldsymbol{d})$. Then all the points are merged and sorted by their depth values from near to far. The final color is calculated by Eq. 4.

3.3 Network Training

We optimize the layered scene representation with the rendering loss as previous methods [Mildenhall et al. 2020; Zhang et al. 2021] do:

$$\mathcal{L}_{\mathbf{rgb}} = \sum_{\mathbf{r}\in\mathcal{R}} \left\| \hat{\boldsymbol{C}}(\mathbf{r}) - \boldsymbol{C}(\mathbf{r}) \right\|_2^2, \quad (7)$$

where $\boldsymbol{C}(\mathbf{r})$ and $\hat{\boldsymbol{C}}(\mathbf{r})$ are RGB colors for the ray $\mathbf{r}$ from the ground-truth and volume rendering, respectively. $\mathcal{R}$ represents the set of the sampled rays during training. In the multi-layer rendering, a ray may intersect with multiple layers, which introduces ambiguities to the above optimization problem. To enforce that only one layer contributes to the color of a ray, we propose a sparsity regularization term and a keypoint supervision term, as shown in Fig. 6.

Sparsity regularization. For each ray $\mathbf{r}$, we suppose that $\boldsymbol{\alpha}(\mathbf{r})$ is a vector of layer logits indicating which layer the corresponding pixel belongs to. It is a normalized vector between 0 and 1. We can simply replace the color $\boldsymbol{c}(\boldsymbol{p}_i, \boldsymbol{d})$ in Eq. 4 with the instance label $\boldsymbol{l}(\boldsymbol{p}_i)$ to render the layer logits of ray $\mathbf{r}$:

$$\boldsymbol{\alpha}(\mathbf{r}) = \sum_{i=1}^{N} o(\boldsymbol{p}_i) \prod_{j<i} (1 - o(\boldsymbol{p}_j)) \boldsymbol{l}(\boldsymbol{p}_i), \quad (8)$$

where $\boldsymbol{l}(\boldsymbol{p}_i)$ is the one-hot vector indicating which object the scene point $\boldsymbol{p}_i$ belongs to.

If only one layer contributes to the color of a ray, it means that its layer logits are close to 0 or 1. Therefore, we introduce an entropy loss to regularize the layer logits:

$$\mathcal{L}_{\mathbf{ent}} = -\sum_{\mathbf{r}\in\mathcal{R}} \sum_{m=1}^{|\boldsymbol{\alpha}(\mathbf{r})|} \boldsymbol{\alpha}_m \log(\boldsymbol{\alpha}_m). \quad (9)$$

Keypoint supervision. If the accurate instance segmentation is known, the decomposition can be achieved by making the rendered mask close to the instance segmentation. However, the instance segmentation given by the pretrained model or rendered from SMPL is inaccurate. Thus we additionally supervise the layer logits with human keypoints detected in the images by minimizing:

$$\mathcal{L}_{\mathbf{kpt}} = -\sum_{\mathbf{r}\in\mathcal{R}} \boldsymbol{w}(\mathbf{r}) \| \boldsymbol{\alpha}(\mathbf{r}) - \hat{\boldsymbol{\alpha}}(\mathbf{r}) \|, \quad (10)$$

where $\boldsymbol{w}(\mathbf{r})$ is the confidence score indicating whether the pixel location traversed by the ray $\mathbf{r}$ corresponds to a labeled keypoint and $\hat{\boldsymbol{\alpha}}$ is the instance label of the corresponding keypoint. $\boldsymbol{w}(\mathbf{r})$ equals 0 for locations without keypoint annotations. As the keypoints are very sparse, we also supervise the pixels on the lines that connect two keypoints on each limb.

4 EXPERIMENTS

4.1 Datasets and metrics

Datasets. Since existing human datasets do not contain enough activities involving close interactions, we created a multi-view dataset called *Multi-Human* Dataset for evaluating our approach. This dataset includes 4 dynamic scenes with different human interactions performed. All sequences have more than 300 frames captured by 22 synchronized cameras. We select 8 uniformly distributed cameras for training and 4 cameras for test. ST-NeRF [Zhang et al. 2021] provides two sequences with 16 cameras covering a view range up to 180 degrees. The occlusion in this dataset is relatively low. To evaluate the performance of the proposed approach on in-the-wild data, we created a dataset called *soccer*, which captured a real soccer game using 8 static GoPro cameras at 60 fps. The details and results for this dataset can be found in our supplementary material.

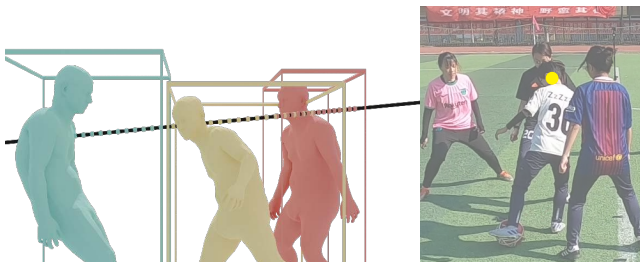


Figure 6: Illustration of the sparsity regularization and keypoint supervision. For a sample ray (black line in the left image and yellow point in the right image), it intersects with three 3D bounding boxes. We sample points and calculate the occupancy for all the three bounding boxes, respectively. If the occupancy values in more that one bounding boxes are nonzero, the rendered layer logits will not be close to 0 or 1 and penalized by the sparsity regularization. If the occupancy values in the red bounding box are large, the rendered instance label will also be red and penalized by the keypoint supervision.

Metrics. Following standard practice in NeRF [Mildenhall et al. 2020], we evaluate our method with two metrics for novel view synthesis: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). For instance segmentation, we adopt the intersection over union (IoU) between the inferred mask and the ground-truth mask.

4.2 Baselines

Two baseline methods are compared in our experiments. 1) The original NeuralBody[Peng et al. 2021] only models a single person. We extend it to multi-person cases. Specifically, for each image, we first perform instance segmentation by an off-the-shelf method [Li et al. 2020b] and assign each instance mask to the fitted SMPL model. Then, a separate NeuralBody model is learned for each person. Specifically, we only sample rays from the visible region of the corresponding person and render each person separately. When rendering the image from a novel view, we first generate the color and depth of each person and then compose the colors according to the depth values. 2) To compare with ST-NeRF [Zhang et al. 2021], we implement the training procedure following their paper since their training code is not released. Note that before the training, they performed scene parsing by utilizing the patch-based multi-view stereo technique [Luo et al. 2019] to generate coarse geometry, which is not feasible in our sparse view setting. Therefore, we use our reconstructed SMPL models to provide the tracked bounding boxes and use the instance segmentation given by [Li et al. 2020b] as the label maps.

4.3 Results

Implementation Details. We adopt the Adam optimizer to train our models. All the networks are trained together from scratch. The learning rate starts from 5e-4 and decays exponentially in iterations. We train our models on four NVIDIA GeForce RTX 3090 GPUs. The training on a multi-view sequence with eight views and 200 frames takes around 30 hours to converge. During inference, it takes around 20s to render a 1920×1080 image and around 6s to render a 960×540 image on a single GPU.

Table 1: Results of novel view synthesis on our Multi-Human dataset. The numbers in brackets are the numbers of people and objects in the scene, respectively. *NB* indicates *NeuralBody* [Peng et al. 2021] and *ST* indicates *ST-NeRF* [Zhang et al. 2021]. The numbers on the left / right of the slash indicate the results without / with masks to remove the background. Note that NeuralBody cannot be evaluated without masks as it cannot model the background of the scene.

	PSNR ↑			SSIM ↑		
Activities	NB	ST	Ours	NB	ST	Ours
Boxing (2p,0o)	- / 27.53	24.33 / 27.89	**25.45 / 30.12**	- / 0.96	0.92 / 0.96	**0.94 / 0.97**
Basketball (2p,1o)	- / 25.85	20.11 / 21.73	**22.82 / 27.76**	- / 0.95	0.86 / 0.91	**0.90 / 0.96**
Handstand (3p,0o)	- / 26.46	25.38 / 30.80	**27.35 / 33.10**	- / 0.95	0.92 / 0.98	**0.95 / 0.99**
Juggling (4p,3o)	- / 27.05	18.80 / 21.73	**26.34 / 30.51**	- / 0.94	0.83 / 0.91	**0.94 / 0.97**
Average	- / 26.72	22.15 / 25.54	**25.50 / 30.37**	- / 0.95	0.88 / 0.94	**0.93 / 0.97**

Results on the Multi-Human dataset. To evaluate the performance on novel view synthesis of multiple entities, we compare our method with NeuralBody [Peng et al. 2021] and ST-NeRF [Zhang et al. 2021] on the Multi-Human dataset. Because NeuralBody is not able to synthesize novel views of the background, we perform experiments that remove the background using masks generated by existing methods[Li et al. 2020b]. Experiments with the background are also conducted. The quantitative results are shown in Table 1 and the qualitative results are shown in Fig. 7. Benefiting from the prior knowledge from SMPL [Loper et al. 2015] model, both our approach and NeuralBody can model human performers vividly. However, NeuralBody suffers from artifacts in crowded scenes with many human interactions due to the lack of layered modeling. On the other hand, ST-NeRF can generate appealing results in simple scenes, but cannot handle more complicated cases where many performers and objects are involved. By proposing a novel layered model with prior knowledge of human bodies, our approach is more robust to complicated cases and outperforms all other methods.

Results on the ST-NeRF dataset. We also perform the quantitative evaluation on the ST-NeRF dataset [Zhang et al. 2021] with different input views for our method. The results are shown in Table 2. The results of ST-NeRF are obtained by their released model. ST-NeRF trained its model with 16 cameras and evaluated on the same cameras, while we train our model with different numbers of cameras and also evaluate on all 16 cameras. The results reveal that we achieve comparable results if using half the number of cameras for training compared to ST-NeRF and significantly surpass it if using all 16 cameras for training.

Results on the soccer dataset. We present the qualitative comparison of our method and other baseline methods on the outdoor *soccer* dataset in Fig.8. We train the three models with all 8 views and 200 frames. The results of NeuralBody have many artifacts due to the wrong instance segmentations. The results of ST-NeRF show that they cannot decompose the moving humans and the background correctly. Our method can render more realistic images and more accurate segmentations, even in the highly occluded regions.

Figure 7: Qualitative results on the *Multi-Human* dataset. *NB* means *NeuralBody* [Peng et al. 2021] and *ST* means *ST-NeRF* [Zhang et al. 2021]. We evaluate these methods both on training views and testing views, with 4 different actions. The blue boxes indicate training views and the red boxes indicate testing views. NeuralBody successfully reconstructs human performers but has artifacts due to inaccurate masks. ST-NeRF fails in some challenging cases as they do not leverage any prior knowledge of human body. Our approach produces photo-realistic results and surpasses the baselines. Please refer to the supplementary video for more results.

Figure 8: Comparison on the *soccer* dataset, where *SCHP* means a pretrained segmentation network [Li et al. 2020b], *NB* means *NeuralBody* [Peng et al. 2021] and *ST* means *ST-NeRF* [Zhang et al. 2021]. For NeuralBody, we render and compose all people. For ST-NeRF and ours, we remove the background layer for comparison. For Both NeuralBody and ST-NeRF fail because of the inaccurate instance segmentation given by SCHP. Our approach can render more realistic images and more accurate masks.

4.4 Ablation studies

We perform the ablation studies on our Multi-Human dataset to validate the design choices in our system.

Effect of the synchronization. To demonstrate the effect of pose-guided synchronization, we evaluate this component on 'ballet' sequence. Our model is trained given the SMPL parameters with and without pose-guided synchronization. The results of a typical

Table 2: Results of novel view syntheis and segmentation on ST-NeRF dataset. View synthesis (PSNR, SSIM) and instance segmentation (IoU) are evaluated. The numbers in brackets indicate the numbers of views used for training. The results of ST-NeRF [Zhang et al. 2021] are obtained by their released model, which are slightly different from those in the original paper.

	Walking			Taekwondo		
	PSNR ↑	SSIM ↑	IoU ↑	PSNR ↑	SSIM ↑	IoU ↑
ST-NeRF (16v)	29.05	0.93	-	29.30	0.95	-
SCHP	-	-	**96.62**	-	-	**96.80**
Ours (4v)	27.25	0.94	94.04	26.35	0.89	95.69
Ours (8v)	31.43	0.97	96.31	30.13	0.95	96.04
Ours (16v)	**34.60**	**0.97**	96.32	**35.83**	**0.97**	96.35

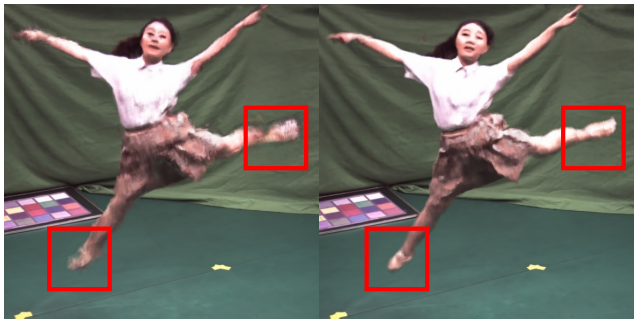


Figure 9: Effect of synchronization error. Left: blurring and artifacts exist due to the synchronization error. Right: the pose-guided synchronization corrects the error.

frame with fast motion are provided in Fig. 9. Our pose-guided synchronization can correct the error and render novel views with fewer artifacts.

The qualitative validation of this component on pose reconstruction is provided in Fig. 4. We conduct an experiment with synthetic data to quantitatively evaluate the synchronization module. Specifically, we downsample a sequence from the public ZJUMoCap dataset [Peng et al. 2021] by different factors and randomly generate misalignment among views. If we set the factor to 2, our method can reduce the mean synchronization error from 0.23 frame to 0.03 frame. If we set the factor to 5, we can reduce the error from 0.41 frame to 0.04 frame.

Effect of the keypoints supervision. We train our model on the 'juggle' sequence with 8 views and 100 frames, which contains 4 moving people and 3 moving balls. Humans can be seen from few views. We train our model with and without $\mathcal{L}_{\mathbf{kpt}}$. Quantitatively, the PSNR/SSIM of the test views are increased from 22.78/0.87 to 26.26/0.92 with the keypoint supervision. Without the keypoint supervision, the training is less constrained and possibly converges to trivial solutions. For example, as shown in Fig. 10, the person who holds the red ball is modeled as part of the background layer. This solution generates more artifacts in the novel view.

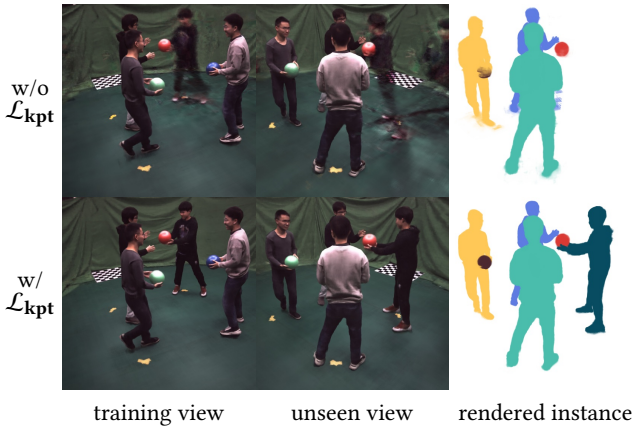


Figure 10: Effect of keypoint supervision. The keypoint supervision provides a good initialization for instance label learning and prevent the convergence to local minima.

5 CONCLUSION

In this paper, we presented a novel approach for novel view synthesis of a scene consisting of multiple human performers with close interactions from a sparse set of calibrated views. The key innovation is a new algorithm that is able to decompose the dynamic scene into independently moving instances and reconstruct a neural representation for each instance. As far as we know, this is the first system that enables high-quality and editable free-viewpoint video synthesis of multiple performers under close interactions from few RGB cameras, which is also applicable for in-the-wild scenes.

Limitations and future work: Currently, the proposed approach is limited to the setting of multiple human performers, only balls as objects, a simple background and a calibrated camera array. As future work, the system can be enhanced in several ways to handle more general settings. First, recovering the human interaction from moving cameras or even a monocular video can be further investigated. Previous works on self-calibration [Huang et al. 2021; Vo et al. 2020a] can be used to remove the need of calibrated camera rig. Second, more general objects can be handled by tracking the 6DoF poses with object pose trackers. Third, if offline scanning of the background is available, the rendering quality of the background can be further improved.

ACKNOWLEDGMENTS

The authors would like to acknowledge support from NSFC (No. 62172364).

REFERENCES

Aayush Bansal, Minh Vo, Yaser Sheikh, Deva Ramanan, and Srinivasa Narasimhan. 2020. 4D Visualization of Dynamic Events From Unconstrained Multi-View Videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5365–5374. https://doi.org/10.1109/CVPR42600.2020.00541

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *Computer Vision – ECCV 2016 (Lecture Notes in Computer Science)*. Springer International Publishing.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*. 7291–7299.

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality streamable free-viewpoint video. *ACM TOG* (2015).

Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. 2008. Performance capture from sparse multi-view video. In *SIGGRAPH*. 1–10.

Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Hujun Bao, and Xiaowei Zhou. 2021. Fast and Robust Multi-Person 3D Pose Estimation and Tracking from Multiple Views. *IEEE TPAMI* (2021).

Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. 2016. Fusion4D: Real-Time Performance Capture of Challenging Scenes. *ACM TOG* 35, 4 (2016), 1–13.

Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. 2021. Neural Radiance Flow for 4D View Synthesis and Video Processing. In *ICCV*. 14324–14334.

Juergen Gall, Carsten Stoll, Edilson de Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. 2009. Motion capture using joint skeleton tracking and surface estimation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 1746–1753. https://doi.org/10.1109/CVPR.2009.5206755

Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. 2021. Dynamic View Synthesis from Dynamic Monocular Video. In *ICCV*.

Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. 1996. The lumigraph. In *SIGGRAPH*.

Jonathan Granskog, Till N Schnabel, Fabrice Rousselle, and Jan Novák. 2021. Neural scene graph rendering. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–11.

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

Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas Funkhouser. 2020. Object-centric neural scene rendering. *arXiv preprint arXiv:2012.08503* (2020).

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

Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. 2020. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*. 5052–5063.

Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. 2021. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *3DV*. 710–720.

Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. 2018. Deep volumetric video from very sparse multi-view performance capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 336–354.

Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. 2019. Learnable triangulation of human pose. In *ICCV*. 7718–7727.

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

Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. 2020b. Self-Correction for Human Parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). https://doi.org/10.1109/TPAMI.2020.3048039

Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. 2020a. Monocular real-time volumetric performance capture. In *European Conference on Computer Vision*. Springer, 49–67.

Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. 2021. Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes. In *CVPR*. 6498–6508.

Haotong Lin, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. 2021. Efficient Neural Radiance Fields with Learned Depth-Guided Sampling. *arXiv preprint arXiv:2112.01517* (2021).

Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. 2020. Neural Sparse Voxel Fields. In *NeurIPS*.

Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. 2021. Neural Actor: Neural Free-view Synthesis of Human Actors with Pose Control. *ACM TOG* (2021).

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM TOG* 38, 4 (2019), 1–14.

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of Volumetric Primitives for Efficient Neural Rendering. *ACM Trans. Graph.* 40, 4, Article 59 (jul 2021), 13 pages. https://doi.org/10.1145/3450626.3459863

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM TOG* 34, 6 (2015), 1–16.

Erika Lu, Forrester Cole, Tali Dekel, Weidi Xie, Andrew Zisserman, David Salesin, William T Freeman, and Michael Rubinstein. 2020. Layered neural rendering for retiming people in video. *ACM TOG* 39, 6 (2020), 1–14.

Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. 2019. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *ICCV*. 10452–10461.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*. Springer International Publishing, Cham, 405–421.

Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. 2019. SiCloPe: Silhouette-Based Clothed People. In *CVPR*. 4480–4490.

Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. 2015. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*. 343–352.

Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. 2020. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*. 3504–3515.

Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. 2021. Neural Articulated Radiance Field. In *ICCV*.

Michael Oechsle, Songyou Peng, and Andreas Geiger. 2021. UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction. In *ICCV*.

Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. 2021. Neural Scene Graphs for Dynamic Scenes. In *CVPR*. 2856–2865.

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable Neural Radiance Fields. In *ICCV*. 5865–5874.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*. 10975–10985.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*. 9054–9063.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *CVPR*. 10318–10327.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *CVPR*. 779–788.

Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. 2021. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *ICCV*. 14335–14345.

Javier Romero, Dimitrios Tzionas, and Michael J Black. 2017. Embodied hands: modeling and capturing hands and bodies together. *ACM TOG* 36, 6 (2017), 1–17.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*. 2304–2314.

Harry Shum and Sing Bing Kang. 2000. Review of image-based rendering techniques. In *Visual Communications and Image Processing 2000*, Vol. 4067. SPIE, 2–13.

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. 2019. Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations. In *NeurIPS*, Vol. 32. 1121–1132.

Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. 2020. RobustFusion: Human Volumetric Capture with Data-Driven Visual Cues Using a RGBD Camera. In *ECCV*. 246–264.

Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, and Lan Xu. 2021. NeuralHumanFVV: Real-Time Neural Volumetric Human Performance Rendering using RGB Cameras. In *CVPR*. 6226–6237.

Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. 2021. Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video. In *ICCV*. 12959–12970.

Hanyue Tu, Chunyu Wang, and Wenjun Zeng. 2020. VoxelPose: Towards Multi-Camera 3D Human Pose Estimation in Wild Environment. In *ECCV*. 197–212.

Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. 2018. Bodynet: Volumetric inference of 3d human body shapes. In *ECCV*. 20–36.

Minh Vo, Ersin Yumer, Kalyan Sunkavalli, Sunil Hadap, Yaser Sheikh, and Srinivasa G Narasimhan. 2020b. Self-supervised multi-view person association and its applications. *IEEE transactions on pattern analysis and machine intelligence* 43, 8 (2020), 2794–2808.

Minh Phuoc Vo, Yaser A Sheikh, and Srinivasa G Narasimhan. 2020a. Spatiotemporal Bundle Adjustment for Dynamic 3D Human Reconstruction in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

Liao Wang, Ziyu Wang, Pei Lin, Yuheng Jiang, Xin Suo, Minye Wu, Lan Xu, and Jingyi Yu. 2021a. iButter: Neural Interactive Bullet Time Generator for Human Free-viewpoint Rendering. In *ACM MM*. 4641–4650.

Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. 2021b. Direct Multi-view Multi-person 3D Human Pose Estimation. *NeurIPS* 34 (2021).

Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. 2020. Multi-View Neural Human Rendering. In *CVPR*. 1682–1691.

Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. 2021. Space-time Neural Irradiance Fields for Free-Viewpoint Video. In *CVPR*. 9421–9431.

Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. 2021. Learning object-compositional neural radiance

field for editable scene rendering. In *ICCV*. 13779–13788.
Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021a. PlenOctrees for Real-time Rendering of Neural Radiance Fields. In *ICCV*.
Hong-Xing Yu, Leonidas J. Guibas, and Jiajun Wu. 2022. Unsupervised Discovery of Object Radiance Fields. In *International Conference on Learning Representations*.
Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021b. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *CVPR*. 5746–5756.
Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. 2018. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *CVPR*. 7287–7296.
Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. 2021. STaR: Self-supervised Tracking and Reconstruction of Rigid Objects in Motion with Neural Rendering. In *CVPR*. 13144–13152.
Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. 2021. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)* 40, 4, 1–18.
Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020b. NeRF++: Analyzing and Improving Neural Radiance Fields. *arXiv:2010.07492* (2020).
Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 2020a. 4D association graph for realtime multi-person motion capture using multiple video cameras. In *CVPR*. 1324–1333.
Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. 2021. DeepMultiCap: Performance Capture of Multiple Characters Using Sparse Multiview Cameras. In *ICCV*.
Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. 2019. DeepHuman: 3D Human Reconstruction From a Single Image. In *ICCV*. 7739–7749.
Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. 2019. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*. 4491–4500.
C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. 2004. High-quality video view interpolation using a layered representation. *ACM TOG* (2004).