# iVS-Net: Learning Human View Synthesis from Internet Videos

Junting Dong[1] Qi Fang[2] Tianshuo Yang[1] Qing Shuai [1] Chengyu Qiao[1] Sida Peng[1]†
[1]Zhejiang University [2]NetEase Games AI Lab

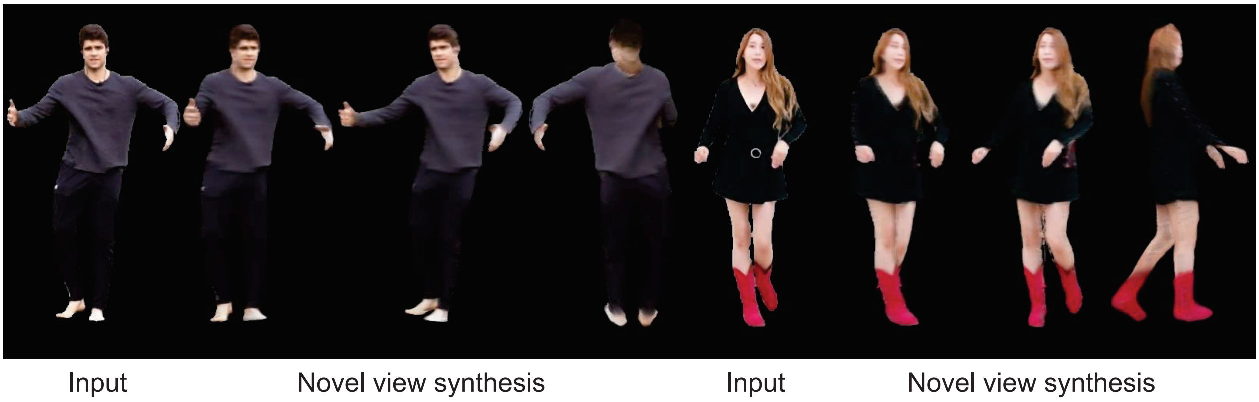

Figure 1: Given a single image of a person as input, our method can generate realistic novel view rendering. For invisible regions such as the back of the subject, we can also obtain plausible results. Instead of limited 3D scans or multi-view images, we only need monocular Internet videos for supervision, which enables it generalizable to in-the-wild images. The code and dataset are available at https://zju3dv.github.io/ivsnet/.

## Abstract

*Recent advances in implicit neural representations make it possible to generate free-viewpoint videos of the human from sparse view images. To avoid the expensive training for each person, previous methods adopt the generalizable human model and demonstrate impressive results. However, these methods usually rely on limited multi-view images typically collected in the studio or commercial high-quality 3D scans for training, which heavily prohibits their generalization capability for in-the-wild images. To solve this problem, we propose a new approach to learn a generalizable human model from a new source of data, i.e., Internet videos. These videos capture various human appearances and poses and record the performers from abundant viewpoints. To exploit the Internet data, we present a video self-supervised pipeline to enforce the local appearance consistency of each body part over different frames of the same video. Once learned, the human model enables realistic novel view synthesis from a single input image. Experiments show that our method can generate high-quality view synthesis on in-the-wild images while only training on monocular videos.*

## 1. Introduction

Generating free-viewpoint videos of a human performer is a core technology in various applications such as AR/VR, telepresence, and gaming. While traditional methods have demonstrated impressive results in free-viewpoint rendering, they typically rely on hundreds of calibrated and synchronized cameras [6, 12] or multiple RGBD sensors [8], which makes them impractical to create free-viewpoint videos for general users.

To make free-viewpoint video creation more accessible, many approaches propose to reconstruct the human model from sparse view or even single view RGB inputs. Given sparse view videos as input (e.g. four views), recent works [37, 36] have achieved photo-realistic novel view synthesis based on the neural radiance field (NeRF) [32] in a per-scene optimization setting. To avoid the expensive per-scene optimization, some works [25, 53, 5] propose generalizable radiance fields conditioned on extracted image features and volumetric features. However, the training of all these methods requires multi-view images usually collected in the studio, which makes them have difficulty in generalizing to in-the-wild images. Instead of using multi-view images, some works [39, 54] achieve remarkable single-image human geometry and appearance reconstruction by super-

†Corresponding author: Sida Peng.

vising the model with commercial high-quality 3D scans. Nonetheless, the limited amount of 3D scans and the domain gap between synthetic and real images make these methods struggle when applied to real-world images. Thus, the key challenge here lies in the generalization ability of the human model.

In this paper, to address this challenge, we follow the single-image reconstruction setting and propose a novel approach to learn a generalizable human model from monocular Internet videos, which supports realistic novel view synthesis. The key observations are as follows: 1) there are lots of human videos on the Internet which contain diverse appearances and actions and abundant viewpoints; 2) due to the articulated structure of the human body, the local appearance of each body part of the same person approximately remains constant. These observations make it possible to learn the human model from Internet videos based on the appearance consistency over different frames of the same video.

To address this new problem, we propose a self-supervised pipeline to learn a generalizable human radiance field from monocular videos. Specifically, we randomly select two frames from the same video called the source frame and paired frame separately. Based on the source frame, we first extract the image feature and utilize the inpainted neural feature map of the parametric human mesh [38] to construct the volumetric feature. Then, these features of sample points are taken as the input of the corresponding human radiance field, which enables rendering novel view images. To supervise the radiance field, we simultaneously leverage the source frame and paired frame for training. The sample points of the paired frame can be transformed into the source frame by linear blend skinning and their density and color can be decoded from the radiance field of the source frame. In addition to reconstruction losses, we further introduce the adversarial loss to encourage the rendering more realistic. Finally, a new dataset consisting of hundreds of Internet videos is created for training.

In summary, this work has the following contributions:

- We introduce a novel approach of novel view synthesis from a single human image, which only utilizes monocular videos for training rather than multi-view images or high-quality 3D scans. A self-supervised pipeline is proposed to learn the human model from monocular inputs.
- We provide a new dataset that consists of more than 600 monocular videos from the Internet totaling more than 120K images, which contain various human and camera viewpoints. For each image, the human mask and SMPL+H parameters are provided.
- We demonstrate that, while only training on monocular videos, our method generates high-quality view synthesis on real-world images.

## 2. Related work

**Novel view synthesis of Human.** Synthesizing novel views of a performer has caused widespread concern. Traditional methods rely on a dense set of cameras [6, 12, 10, 55] or depth sensors placed in the studio [33, 8, 51, 52]. While they have achieved impressive results, it is not applicable in ordinary scenes. A recent series of methods simplify the setting to sparse view RGB cameras. Following the NeRF [31] that represents the scene as a neural radiance field, some works propose to overfit a multi-view human video via per-scene optimization [37, 36, 28, 34]. Conditioning the radiance field on the structured latent codes, Neural Body [37] achieves remarkable results only using four cameras. To improve the generalization ability to novel human poses, [36, 28, 34, 7] propose to define the human model in the canonical space and build the correspondence between canonical space and observation space. To avoid the expensive training for each performer, recent approaches extend the original human model to the generalizable one. [25, 53, 5] combine human prior and image features to reconstruct the generalizable human model. Specifically, they leverage the parametric human model [25, 5] or human skeleton [53] to extract 3D features and then combine them with the pixel-aligned image features as the input of the generalizable human radiance field. However, all these methods require multi-view images as supervision, which makes them struggle when applied to in-the-wild images.

**Single-view human reconstruction.** Reconstructing the 3D human body from a single view is a challenging task. With the help of parametric human models [29, 38, 15], lots of works estimate the model parameters via optimization [4] or neural networks [20, 23, 22]. While they have achieved remarkable progress in 3D human pose estimation, the estimated human shape is quite coarse. To obtain more detailed geometry, some works [47, 13, 46, 14] leverage a prescanned personalized template mesh and deform the mesh by dense non-rigid tracking. Recently, implicit function based methods become popular due to their impressive performance. PiFu based methods [39, 40, 54, 18, 17, 45] represent the human as pixel-aligned implicit functions and can reconstruct detailed human from a single image. For training, they need high-quality 3D scans [3, 2] for 3D supervision. However, both the limited amount of 3D scans and the domain gap between synthesized images and real images limit their generalization ability. Note that the reconstructed appearance of these methods heavily relies on the recovered geometry and the rendering quality will degrade dramatically when the geometry reconstruction is poor.

**Self-supervision in human reconstruction.** Since the 3D ground truth of human model is difficult to collect, self-supervised methods have emerged that directly learn the human model from RGB images without 3D supervision. Here, we focus on the self-supervised methods which only leverage single-view images. For face reconstruction from a single image, [41, 44] leverage the face shape consistency and appearance consistency of the same person across different images to achieve self-supervised training. For the human body reconstruction, [24, 35] also leverages the texture map consistency to learn the network from different frames of the same person. These methods all rely on the parametric human model, which limits the detailed reconstruction. In addition, [42, 19] propose to represent the human model as the depth map and leverage photo consistency or geometry consistency across different frames to supervise the training. While these methods achieve impressive results, they can only reconstruct visible surfaces. Instead, our method reconstructs the complete human model.

# 3. Method

Given a single image, we aim to reconstruct the 3D human model which enables rendering realistic novel view images. To address the problem of poor generalization, our method proposes to learn human view synthesis from abundant Internet videos rather than limited multi-view images or 3D human scans. Figure 2 presents the overview of our approach. We construct the image feature and volumetric feature from the input image (Section 3.1). These two features of sample points are combined as the input of the human radiance field model to predict their density and color, which are further used to synthesize images by volume rendering (Section 3.2). In addition to the supervision from the input image, we also introduce the video self-supervision from another frame of the same video (Section 3.3). Then, we describe the loss functions for training (Section 3.4) and implementation details (Sections 3.5).

For each image, we first adopt the EasyMoCap [1] to estimate SMPL+H [38] parameters and then utilize [26] to generate the human mask. In the following, we describe the details of each component.

## 3.1. Feature construction

Given an input image $\mathbf{I}$, we utilize the 2D CNN to extract the image feature $F_{img}$. For a 3D sample point $\mathbf{x}$, to construct the pixel-aligned image feature $F_{img}(\pi(\mathbf{x}); \mathbf{I})$, we project the point to obtain the image coordinate $\pi(\mathbf{x})$ and bilinearly interpolate the pixelwise features.

In addition to the image feature, we also construct the volumetric feature for each point. Specifically, we unproject the image feature to vertices of the estimated human mesh, resulting in a feature mesh. In practice, the human mesh is rasterized onto the image plane, which produces the projected 2D location and visibility of each vertex. For each visible vertex, we sample the corresponding pixel-aligned image feature at the projected location as the vertex feature. Due to occlusion, the vertex features of invisible vertices are missing, which may damage the rendering quality. To address this, we propose to inpaint the vertex features of those invisible vertices. In particular, we first unwrap the initial incomplete feature mesh to generate the partial UV feature map. Then, this map is processed with a U-Net to produce the complete UV feature map, from which we can obtain the inpainted vertex feature of each vertex. The resulting inpainted feature mesh is further processed with a 3D CNN to obtain the 3D feature volume $F_{vol}$. Based on the feature volume, for each sample point $\mathbf{x}$, we can retrieve the corresponding volumetric feature $F_{vol}(\mathbf{x}; \mathbf{I}, \mathbf{P})$ by trilinealy interpolating, where $\mathbf{P}$ is the human pose.

To obtain the final feature representation $f(\mathbf{x}; \mathbf{I}, \mathbf{P})$ of each sample point, we combine the image feature and the volumetric feature as follows:

$$f(\mathbf{x}; \mathbf{I}, \mathbf{P}) = F_{img}(\pi(\mathbf{x}); \mathbf{I}) + F_{vol}(\mathbf{x}; \mathbf{I}, \mathbf{P}) \tag{1}$$

## 3.2. Human model

Based on the constructed feature of the sample points, we aim to reconstruct the 3D human model, which is represented as the neural radiance field similar to [37, 36]. Specifically, the human geometry and appearance are represented as density fields $F_\sigma$ and color fields $F_{\mathbf{c}}$ given by MLP networks. For a query point $\mathbf{x}$, the human model can be written as follows:

$$\sigma(\mathbf{x}), \mathbf{z}(\mathbf{x}) = F_\sigma(f(\mathbf{x}; \mathbf{I}, \mathbf{P})), \tag{2}$$

$$\mathbf{c}(\mathbf{x}) = F_{\mathbf{c}}(\mathbf{z}(\mathbf{x})), \tag{3}$$

where $\sigma(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ denote the density and color to be decoded. $\mathbf{z}(\mathbf{x})$ denotes the geometry feature. We empirically remove the view direction from the input of the color field.

To learn the human model from images, we leverage volume rendering to synthesize images based on the predicted density and color as follows:

$$\tilde{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma(\mathbf{x}_i)\delta_i))\mathbf{c}(\mathbf{x}_i), \tag{4}$$

$$\text{and} \quad T_i = \exp(-\sum_{j=1}^{i-1} \sigma(\mathbf{x}_j)\delta_j), \tag{5}$$

where $\tilde{C}(\mathbf{r})$ denotes the rendered color of ray $\mathbf{r}$ and $\delta_i = ||\mathbf{x}_{i+1} - \mathbf{x}_i||_2$ denotes the distance between adjacent sampled points. $N$ is set to 64 in all experiments. Based on the rendered images, we can train the human model by comparing them with the observed images.

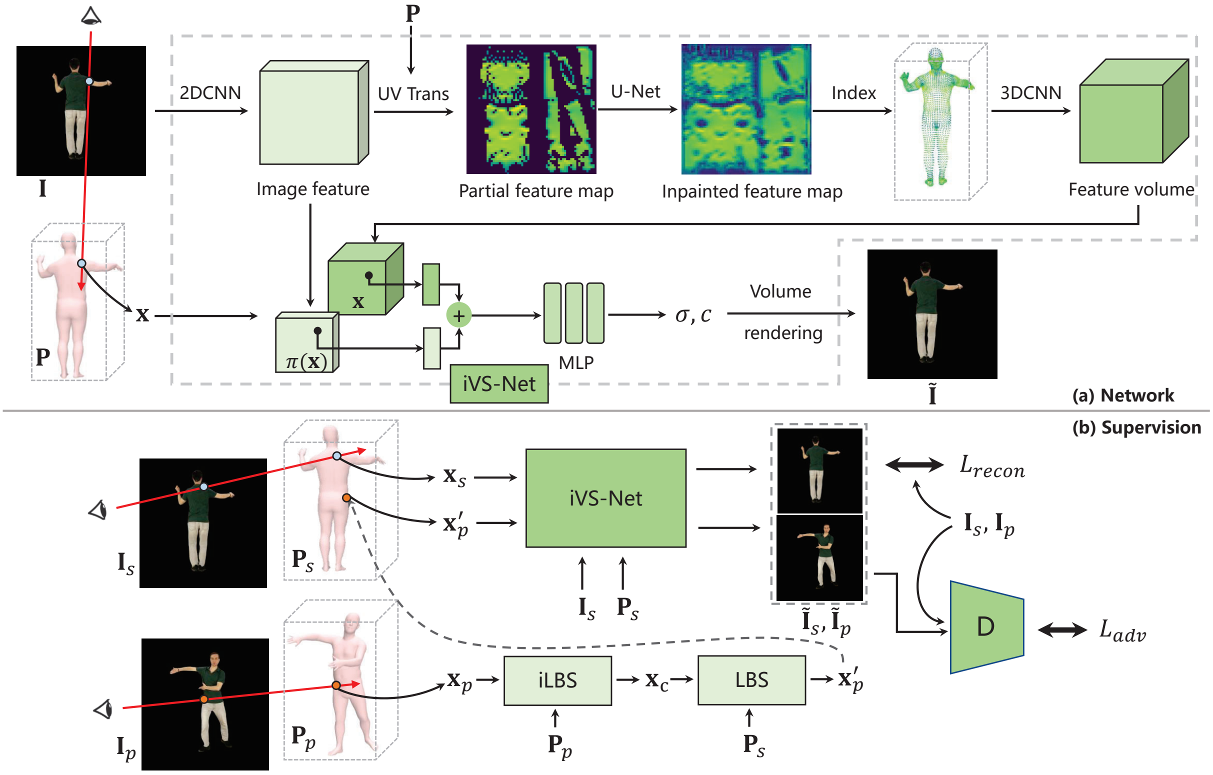

Figure 2: **Overview of the proposed approach.** (a) Given a single image $\mathbf{I}$ and the corresponding human pose $\mathbf{P}$, the proposed iVS-Net aims to reconstruct the 3D human model, represented as the neural radiance field. For a sample point $\mathbf{x}$, its feature representation consists of the pixel-aligned image feature and SMPL+H based volumetric feature. UV feature map inpainting is adopted to improve the volmetric feature. Based on the feature of sample points, the human model predicts the density and color and then synthesizes images by volume rendering. (b) Video self-supervision is proposed to train the iVS-Net with monocular videos. We randomly select two frames from the same video called the source frame $\mathbf{I}_s$ and paired frame $\mathbf{I}_p$, respectively. For the source frame, the iVS-Net reconstructs its human model and synthesizes the corresponding image $\tilde{\mathbf{I}}_s$. For the paired frame, the sample point $\mathbf{x}_p$ is first transformed into the observation space of the source frame and then processed with the human model of the source frame to synthesize image $\tilde{\mathbf{I}}_p$. Finally, the reconstruction loss and adversarial loss are introduced for training.

### 3.3. Video self-supervision

With only monocular input images for supervision, it is ill-posed to learn the neural representation and the resulting human model fails to generalize to novel views as shown in Section 4.3. To address this, we propose to introduce the video self-supervision. Thanks to the articulated structure of the human body, the appearance of each body part approximately remains constant across the video, which makes it possible to apply the appearance consistency constraint over different frames for self-supervision.

Specifically, in addition to the input frame, also known as the source frame $\mathbf{I}_s$, we randomly select another frame from the same video called the paired frame $\mathbf{I}_p$, which is utilized to supervise the reconstructed human radiance field of the source frame. Given a sample point $\mathbf{x}_p$ from the observation space of the paired frame, we first transform it to the canonical space by the inverse linear blend skinning algorithm, which can be written as follows:

$$\mathbf{x}_c = \left( \sum_{m=1}^{M} w^m(\bar{\mathbf{x}}_p)\mathbf{B}(\mathbf{P}_p)^m \right)^{-1} \bar{\mathbf{x}}_p, \tag{6}$$

where $\bar{\mathbf{x}}_p$ is the homogeneous coordinate of $\mathbf{x}_p$, $\mathbf{P}_p$ is the human pose of the paired frame, and $\mathbf{B}(\mathbf{P}_p)^m \in SE(3)$ is the transformation matrix of bone $m$. The blend weight $w^m(\bar{\mathbf{x}}_p)$ of bone $m$ can be obtained by retrieving the blend weight of the closest vertex on the template human mesh. $M$ denotes the number of bones. Then this point $\mathbf{x}_c$ can be transformed into the observation space of the source image using the linear blend skinning algorithm as follows:

$$\mathbf{x}'_p = \left( \sum_{m=1}^{M} w^m(\bar{\mathbf{x}}_c)\mathbf{B}(\mathbf{P}_s)^m \right) \bar{\mathbf{x}}_c, \tag{7}$$

where $\mathbf{P}_s$ is the human pose of the source image. Next, for the sample point $\mathbf{x}'_p$, we can extract the corresponding feature $f(\mathbf{x}'_p; \mathbf{I}_s, \mathbf{P}_s)$ as shown in section 3.1 and predict the density and color. Finally, we can synthesize the image using volume rendering and compare it with the paired frame $\mathbf{I}_p$. Thus, the reconstructed human model of the source image is supervised by the source frame and paired frame simultaneously.

### 3.4. Loss functions

Given the synthesized images of the source frame and paired frame $\tilde{\mathbf{C}} = (\tilde{\mathbf{C}}_{source}, \tilde{\mathbf{C}}_{pair})$ and the corresponding ground truth images $\mathbf{C} = (\mathbf{C}_{source}, \mathbf{C}_{pair})$, we adopt the following reconstruction loss functions for training:

$$L_{recon} = \sum_i (\lambda_{l_1} L_1^i + \lambda_{l_2} L_2^i + \lambda_{lpips} L_{lpips}^i). \tag{8}$$

For $i \in (source, pair)$, we calculate the L1 loss $L_1^i$, L2 loss $L_2^i$, and perceptual loss $L_{lpips}^i$ between the synthesized image $\tilde{\mathbf{C}}_i$ and ground truth image $\mathbf{C}_i$, separately. $(\lambda_{l_1}, \lambda_{l_2}, \lambda_{lpips})$ are the corresponding weight of each loss function.

While the reconstruction loss of self-supervision provides more constraints, the training is still quite under-constrained due to the highly sparse-view supervision. Therefore, we additionally introduce the adversarial loss, which encourages the network to generate realistic images. Specifically, we can regard the human reconstruction model as a genenrator $G$, which generates synthesized images of the specified viewpoints based on the source image. The synthesized images are fed to the discriminator $D$ to determine whether they are real images or not. The discriminator $D$ is parameterized as a CNN with leaky ReLU activation. The non-saturating GAN objective [9] and $R_1$ gradient penalty [30] are adopted for the adversarial training. The adversarial loss function for the generator can be written as follows:

$$min\ L_{adv} = E_{\tilde{\mathbf{C}} \sim p_{\tilde{\mathbf{C}}}}[-log(\sigma(D(\tilde{\mathbf{C}})))], \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $p_{\tilde{\mathbf{C}}}$ denotes the data distribution of synthesized images. The objective for the discriminator can be written as follows:

$$\begin{aligned} min\ L(D) = & E_{\tilde{\mathbf{C}} \sim p_{\tilde{\mathbf{C}}}}[-log(1 - \sigma(D(\tilde{\mathbf{C}})))] \\ & + E_{\mathbf{C} \sim p_{\mathbf{C}}}[-log(\sigma(D(\mathbf{C}))) + \lambda_R \|\nabla D(\mathbf{C})\|^2], \end{aligned} \tag{10}$$

where $\lambda_R = 10$ and $p_{\mathbf{C}}$ denotes the data distribution of real images. For training, we jointly optimize the generator and discriminator.

Finally, the whole loss functions for the human model can be written as follows:

$$L = L_{recon} + \lambda_{adv} L_{adv}, \tag{11}$$

where $\lambda_{adv}$ denotes the weight of the adversarial loss.

### 3.5. Implementation details

**Network architecture and hyper-parameters.** The network architectures of density fields $F_s$ and color fields $F_{\mathbf{c}}$ are almost the same as the original NeRF [31], except that we replace the input of the density field from the positional encoding of the location to the constructed feature $f(\mathbf{x}; \mathbf{I}, \mathbf{P}) \in R^{256}$. We use the ResNet34 [16] to extract the image feature and use the SparseConvNet [11] to extract the volumetric feature. The weights are set as follows: $\lambda_{l_1} = 1, \lambda_{l_2} = 1, \lambda_{lpips} = 0.1, \lambda_{adv} = 0.025$.

**Training details.** Instead of training on random ray samples, we sample a $W \times W$ patch on each image, where $W$ is a random value selected from the range $(72, 90)$. The Discriminator $D$ is trained with 64 × 64 patch images and the batch size is 16. For the training of the human reconstruction model, we adopt the Adam optimizer [21] and the learning rate begins from $5e^{-4}$ and decays exponentially to $5e^{-5}$ during the optimization. For the training of the discriminator $D$, we adopt the RMSprop optimizer [43] and the learning rate is set as $1e^{-4}$. The training of our method takes 24 hours on a single Nvidia V100 GPU.

## 4. Experiments

### 4.1. Datasets and metrics

**Datasets.** Our method leverages monocular human videos for self-supervised learning and the most related dataset is TikTok Dataset [19]. However, the videos of this dataset usually capture little view changing and most videos suffer from severe human truncation, which makes it unsuitable for learning the human view synthesis. Therefore, as shown in Figure 3, we create a new dataset for training and evaluation. Specifically, we manually collect more than 600 videos from the Internet. Each video contains a complete person performing various actions and records the performer from various viewpoints. For each video, we extract about 200 frames, resulting in more than 120K images. For each image, we utilize EasyMoCap[1] to obtain the SMPL+H parameters and [26] to obtain the human mask. We use 90% of the data for training and the rest for qualitative evaluation.

For quantitative evaluation, we utilize ZJU-MoCap dataset [37] that captures complex human actions of 9 subjects using multi-view cameras. We randomly select one camera as input images and surrounding six cameras for the evaluation of novel view synthesis.

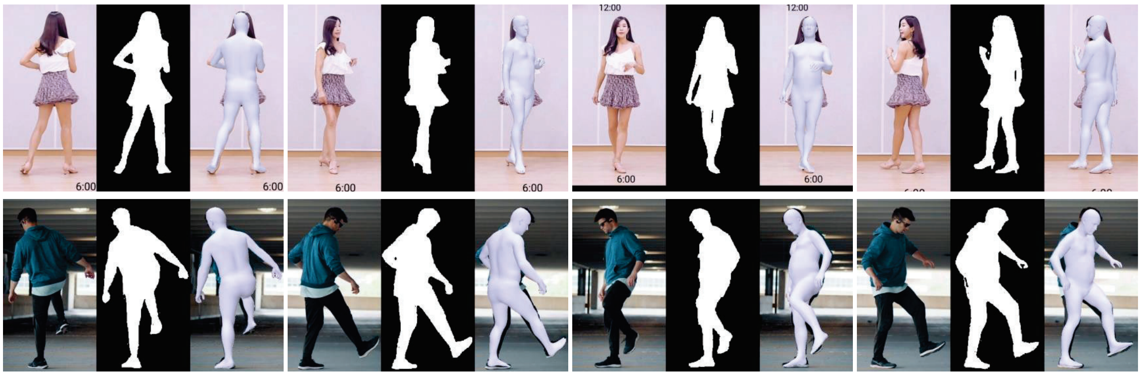

Figure 3: The proposed dataset. We provide a new dataset that consists of more than 600 videos from the Internet. Each video captures a complete human performer doing various actions and records the performer from various viewpoints. This dataset provides more than 120K images along with the corresponding human mask and SMPL+H paramters.

**Metrics.** For evaluating the image synthesis, we adopt the following metrics: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS).

### 4.2. Comparison with the baselines

**Baselines.** We compare with the state-of-the-art generalizable reconstruction methods [50, 25, 39, 40, 54, 45]. We first train the pixelNeRF [50] and NHP [25] on the collected monocular datasets. We also train another NHP [25] model with the multi-view images. Finally, we compare with the 3D scans based model PIFu, PIFuHD, PAMIR, and ICON, whose pre-trained models are publicly available.

**Image synthesis.** We evaluate the image synthesis quality on ZJU-MoCap dataset. Given an input image, we compare the novel view rendering results. The quantitative results are given in Table 1, which shows that our approach outperforms the baseline methods by a large margin. Figure 6 presents the qualitative results on ZJU-MoCap dataset and in-the-wild images. As we can see, the novel view images rendered by our method significantly outperform the counterparts of the previous methods. The pixelNeRF [50] and NHP [25] cannot generate reasonable novel view synthesis when trained on the monocular videos. The reason may be that the supervision from the source images is not sufficient to train the 3D human model. In contrast, our method can generate high-quality novel view images even when the input images are quite different from the training data. For example, the images of ZJU-MoCap are recorded in the studio and the lighting is much darker than the daily videos. The high-quality novel view synthesis on these images demonstrates the generalizable ability of our method. Furthermore, we present the novel view synthesis under large viewpoint changes in Figure 4. The results show that only our method can obtain plausible outcomes for the totally invisible regions such as the back of the subjects.

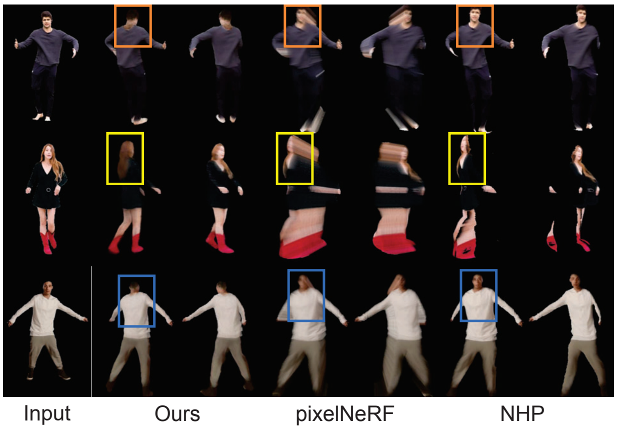

Figure 4: Qualitative comparison of back and side view renderings on the in-the-wild images.

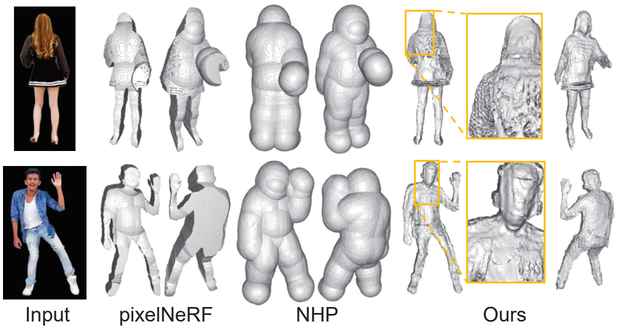

Figure 5: Qualitative results of 3D geometry on in-the-wild images.

**3D reconstruction.** We also compare 3D surface reconstructions of our method and other baselines. We present the qualitative results in Figure 5, which shows that our

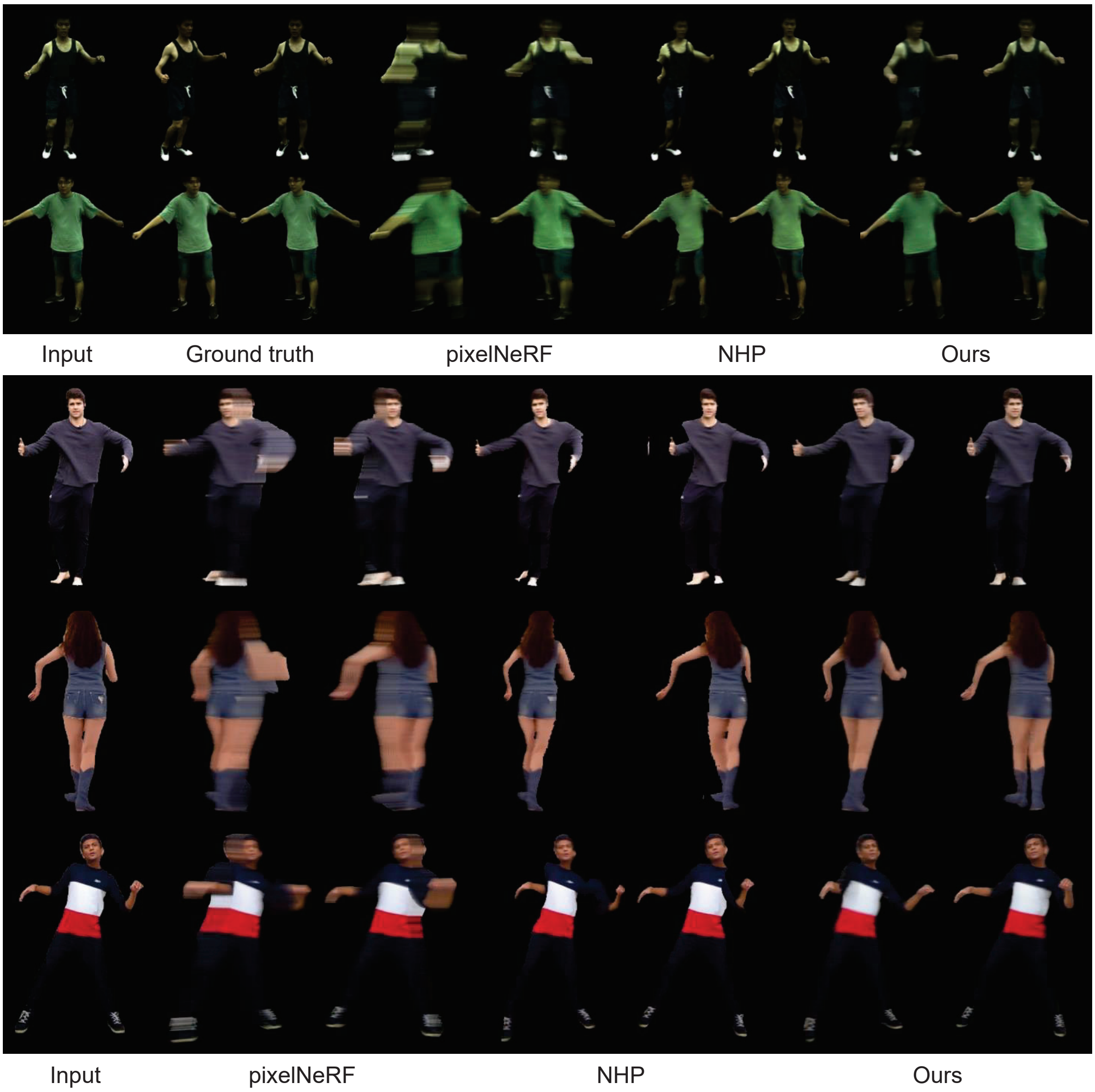

Figure 6: Qualitative results of novel view synthesis on the ZJU-MoCap dataset and in-the-wild images.

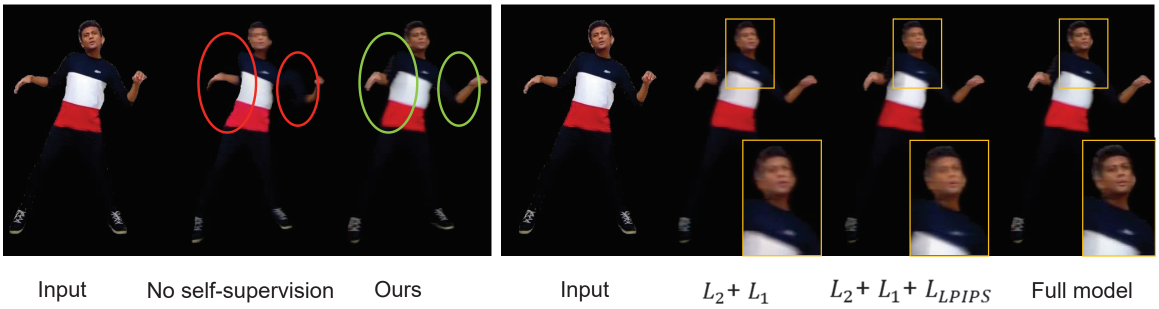

Figure 7: Ablation studies for self-supervision (left) and loss functions (right).

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| pixelNeRF [50] | 18.63 | 0.706 | 0.331 |
| NHP [25] | 17.51 | 0.687 | 0.323 |
| Ours | **24.67** | **0.886** | **0.196** |

Table 1: Results of image synthesis on ZJU-MoCap dataset.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| baseline | 18.71 | 0.710 | 0.349 |
| Ours | **24.08** | **0.892** | **0.192** |

Table 2: Comparison with the baseline trained with multi-view images on ZJU-MoCap dataset.

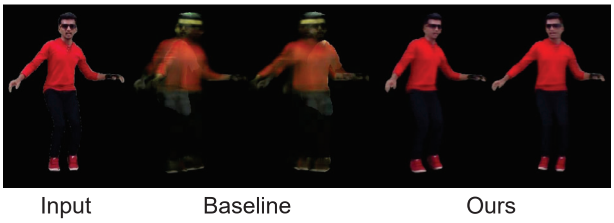

Figure 8: Qualitative comparison with the multi-view images based method on the in-the-wild image.

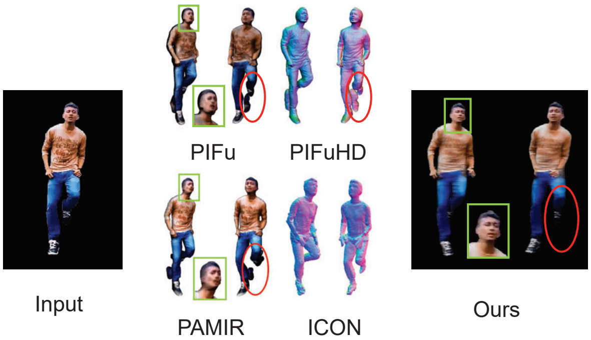

Figure 9: Qualitative comparison with 3D scan based methods on the in-the-wild image. Note that PIFuHD and ICON can only recover human geometry.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| No self-supervision | 17.86 | 0.693 | 0.332 |
| $L_{l2} + L_{l1}$ | 24.62 | 0.881 | 0.251 |
| $L_{l2} + L_{l1} + L_{lpips}$ | **24.72** | 0.885 | 0.230 |
| Ours ($L_{l2}$+$L_{l1}$+$L_{lpips}$+$L_{adv}$) | 24.67 | **0.886** | **0.196** |

Table 3: Ablation studies of image synthesis on ZJU-MoCap dataset.

method significantly outperforms the baselines. The pixelNeRF [50] and NHP [25] cannot reconstruct reasonable human geometry, which may explain their poor novel view rendering. On the contrary, our method can recover more detailed human geometry.

**Comparison with the method supervised by multi-view images.** We conduct the following comparison to demonstrate that the proposed method owns better generalization ability than the ones trained with limited multi-view images. Specifically, given a single image as input, we train another NHP [25] model with multi-view images. We randomly select six subjects from the ZJU-MoCap dataset for training and the remaining three subjects are used for evaluation. The quantitative results are shown in Table 2. Figure 8 shows the qualitative comparison on the in-the-wild image. Note that our method is trained solely on Internet videos. As we can see, our method shows much better quantitative and qualitative results.

**Comparison with the methods supervised by 3D scans.** Additionally, in Figure 9, we qualitatively compare with the PIFu[39] and PIFuHD[40], whose pre-trained models are publicly available. Our method generates more realistic rendering than the PIFu whose rendering heavily relies on the reconstructed geometry. Both the limited amount of 3D scans and the domain gap between synthesized images and real images limit their generalization ability to in-the-wild images, such as the marked red circles.

### 4.3. Ablation studies

**Video self-supervision.** Our method utilizes video self-supervision to provide additional supervision for training the generalizable human model from monocular videos as described in Section 3.3. Here, we compare it to the baseline without video self-supervision, i.e., only using the source images for training. The quantitative results of image synthesis are presented in Table 3. The results show that the proposed video self-supervision significantly improves the quality of synthesized images. We also show the qualitative results in Figure 7(left). As we can see, without the video self-supervision, the baseline method cannot synthesize the complete human under novel views.

**Impact of each loss function.** As described in Section 3.4, we use several loss functions for training: L2 loss, L1 loss, perceptual loss, and adversarial loss. Here, we analyze the effect of each loss function. The quantitative and qualitative results of image synthesis are presented in Table 3 and Figure 8, respectively. The results show that on the basis of L2 loss and L1 loss, adding the perceptual loss and adversarial loss can further improve the image quality of novel view synthesis. As we can see, the adversarial loss is crucial for sharper and more realistic rendering.

### 4.4. Limitations

The proposed method still has the following limitations. First, the reconstructed human geometry still loses much details, such as the face, which degrades the rendering quality. In the future, we will try to replace the density field with the signed distance field like [48] to improve the geometry and rendering. Second, at inference time, the rendering speed of our method is still relatively slow. Recent works [27, 49] leverage the voxel octree to improve the rendering speed and achieve real-time rendering. Combining with this technology is left as future work.

## 5. Summary

In this paper, we propose a new approach to reconstruct the 3D human model from a single image. Different from previous works that require multi-view images or high-quality 3D scans, we only leverage Internet videos for training. We propose a self-supervised pipeline to introduce the local appearance consistency constraint of each body part over different frames of the same video. A new dataset consisting of hundreds of Internet videos is created for training. Extensive experimental results demonstrate the generalization capability of our method for in-the-wild images.

**Acknowledgements.** This work was supported by the Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

## References

[1] Easymocap. https://github.com/zju3dv/EasyMocap/. 3, 5

[2] Mixamo. https://www.mixamo.com/. 2

[3] renderpeople. https://renderpeople.com/. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 2

[5] Hongsuk Choi, Gyeongsik Moon, Matthieu Armando, Vincent Leroy, Kyoung Mu Lee, and Gregory Rogez. Mononhr: Monocular neural human renderer. *arXiv preprint arXiv:2210.00627*, 2022. 1, 2

[6] 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 Trans. Gr.*, 2015. 1, 2

[7] Junting Dong, Qi Fang, Yudong Guo, Sida Peng, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Totalselfscan: Learning full-body avatars from self-portrait videos of faces, hands, and bodies. *Advances in Neural Information Processing Systems*, 35:13654–13667, 2022. 2

[8] 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. Fusion4d: Real-time performance capture of challenging scenes. *ACM Trans. Gr.*, 2016. 1, 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. 2014. 5

[10] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *SIGGRAPH*, 1996. 2

[11] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. pages 9224–9232, 2018. 5

[12] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Trans. Gr.*, 2019. 1, 2

[13] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Trans. Gr.*, 2019. 2

[14] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020. 2

[15] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, 2019. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016. 5

[17] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. pages 11046–11056, 2021. 2

[18] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. pages 3093–3102, 2020. 2

[19] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. pages 12753–12762, 2021. 3, 5

[20] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 5

[22] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2

[23] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2

[24] Jogendra Nath Kundu, Mugalodi Rakesh, Varun Jampani, Rahul Mysore Venkatesh, and R Venkatesh Babu. Appearance consensus driven self-supervised human mesh recovery. pages 794–812, 2020. 3

[25] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. volume 34, pages 24741–24752, 2021. 1, 2, 6, 8

[26] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *CVPR*, 2021. 3, 5

[27] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 33:15651–15663, 2020. 9

[28] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Gr.*, 2021. 2

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Gr.*, 2015. 2

[30] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, pages 3481–3490, 2018. 5

[31] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 5

[32] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1

[33] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015. 2

[34] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, 2021. 2

[35] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *ICCV*, pages 803–812, 2019. 3

[36] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 1, 2, 3

[37] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 1, 2, 3, 5

[38] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Gr.*, 2017. 2, 3

[39] 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 *ICCV*, 2019. 1, 2, 6, 8

[40] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 2, 6, 8

[41] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. pages 7763–7772, 2019. 3

[42] Feitong Tan, Hao Zhu, Zhaopeng Cui, Siyu Zhu, Marc Pollefeys, and Ping Tan. Self-supervised human depth estimation from monocular videos. pages 650–659, 2020. 3

[43] Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012. 5

[44] Yandong Wen, Weiyang Liu, Bhiksha Raj, and Rita Singh. Self-supervised 3d face reconstruction via conditional estimation. pages 13289–13298, 2021. 3

[45] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2, 6

[46] 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 *CVPR*, 2020. 2

[47] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Gr.*, 2018. 2

[48] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. 2021. 9

[49] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *ICCV*, pages 5752–5761, 2021. 9

[50] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. pages 4578–4587, 2021. 6, 8

[51] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021. 2

[52] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *CVPR*, 2018. 2

[53] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. pages 7743–7753, 2022. 1, 2

[54] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3170–3184, 2021. 1, 2, 6

[55] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. Gr.*, 2004. 2