Neural Novel Actor: Learning a Generalized Animatable Neural Representation for Human Actors

Qingzhe Gao*, Yiming Wang*, Libin Liu†, Lingjie Liu†, Christian Theobalt, Baoquan Chen†

Abstract—We propose a new method for learning a generalized animatable neural human representation from a sparse set of multi-view imagery of multiple persons. The learned representation can be used to synthesize novel view images of an arbitrary person and further animate them with the user's pose control. While most existing methods can either generalize to new persons or synthesize animations with user control, none of them can achieve both at the same time. We attribute this accomplishment to the employment of a 3D proxy for a shared multi-person human model, and further the warping of the spaces of different poses to a shared canonical pose space, in which we learn a neural field and predict the person- and pose-dependent deformations, as well as appearance with the features extracted from input images. To cope with the complexity of the large variations in body shapes, poses, and clothing deformations, we design our neural human model with disentangled geometry and appearance. Furthermore, we utilize the image features both at the spatial point and on the surface points of the 3D proxy for predicting person- and pose-dependent properties. Experiments show that our method significantly outperforms the state-of-the-arts on both tasks.

Index Terms—Neural Rendering, Neural Radiance Field, Human Synthesis

1 Introduction

Synthesizing high-quality free-viewpoint videos of an arbitrary human character using a sparse set of cameras is crucial for many computer graphics applications, including VR/AR, film production, video games, and telepresence. Many of these applications require user control over human poses in the synthesis. Achieving these with traditional methods is difficult because it needs an expensive capturing setup [4], [11], [12], [15], [19], [55], the production-quality human geometry and appearance models, and manual invention and corrections [10], [13], [57].

Recently, neural human representation and rendering algorithms based on the neural radiance fields (NeRF) [40] have demonstrated the ability to overcome the limitations of the traditional approaches. Some methods [6], [31], [44], [56] can learn an animatable human representation from multi-view imagery in a person-specific setting, but they are not able to generalize to new persons. Other works [7], [27], [39], [75] proposed generalizable radiance fields for humans conditioned on input image features, inspired by the generalized neural representation for static scenes [5], [9], [34], [63], [73]. With the learned representations, they can generate novel views of an arbitrary person from sparse multi-view images without training. However, their representations are not animatable and thus cannot generate images with user's pose control.

In this paper, we address a challenging yet practical problem – training a model capable of rendering unseen individuals in a feed-forward manner using only still images captured from sparse viewpoints as input, as illustrated in Fig. 1. To tackle this task, we introduce a new approach to learning a generalized animatable neural human representation from sparse multi-view input imagery of multiple persons. This representation allows us to generalize to new persons without training and further animate the representation with pose control.

Specifically, our method uses a Skinned Multi-Person Linear (SMPL) model as a 3D proxy and transforms each pose space to a shared canonical pose space. Then a neural radiance field in the canonical space is learned and we estimate person- and pose-dependent geometry and appearance with the features extracted from the input images. To efficiently learn this representation for multiple persons, we disentangle the geometry and appearance in our human model by extracting separate features for geometry and appearance properties to condition the prediction of these properties. Furthermore, we extract the image features at both the spatial points and the surface points of SMPL to better infer the person- and pose-dependent properties.

We evaluate our method on the ZJU-MoCap [45], DeepCap [21] and DynaCap [20] datasets. Our method significantly outperforms the state-of-the-art, MPS-NeRF [17], in this challenging task. To demonstrate the effectiveness of our representation, we also separately evaluate its performance in two key aspects: generalization of novel view synthesis and animation with user-defined pose control. The experiments show that our method outperforms the state-of-the-art on both these tasks. These evaluations provide compelling evidence that our method has the potential to serve as a robust baseline model for future research endeavors in the domain of generalizable 3D human rendering.

In summary, our technical contributions are:

- We present a new method for achieving both the novel view synthesis of arbitrary persons and the animation synthesis

- *Qingzhe Gao is with the Department of Computer Science, Shandong University, China and Peking University, China. E-mail: gaoqingzhe97@gmail.com*
- *Yiming Wang, Libin Liu and Boquan Chen are with SIST & KLMP (MOE), Peking University,China E-mail: {wym12416, libin.liu, baoquan}@pku.edu.cn,*
- *Lingjie Liu and is Christian Theobalt are with the Graphic, Vision & Video group of Max Planck Institute for Informatics in Saarbrücken, Germany. E-mail:{lliu, theobalt}@mpi-inf.mpg.de*

* *The first two authors contributed equally to this work.*
† *These authors contributed senior supervision equally to this work.*

with pose control at the same time.

- We design a new generalized animatable neural human representation with disentangled geometry and appearance, which can be learned efficiently from sparse multi-view imagery of multiple persons.
- We present a new way to predict the person- and pose-dependent properties by taking the features at both the spatial points and the surface points of SMPL into account.

2 Related Work

2.1 Human Performance Capture

There have been multiple studies addressing novel view synthesis of human performance. While many methods based on pre-scanned human models [4], [11], [15], [55] can capture humans in a sparse multi-view setting, pre-scanned human models are often unavailable in most cases. Recent works rely on depth sensors [10], [13], [57] or dense arrays of cameras [12], [19] to achieve high-fidelity reconstruction, but these settings are not easily accessible. By employing neural networks, some methods [37], [38], [65] can compensate for geometric artifacts through the modification of the rendering pipeline. More recently, several works [22], [23], [41], [49], [50], [51], [76] have been able to reconstruct 3D humans from a single image using 3D human geometry priors. However, these methods rely on 3D geometry data and cannot generalize to complex poses that are not present in the training data. In contrast, our method is capable of generalizing to new persons using only sparse multi-view image supervision.

2.2 Neural Representations for human

Neural rendering techniques [28], [32], [53], [58], [65] have enabled neural networks to learn 3D object reconstruction from 2D images. Various 3D representations, such as 3D voxel-grid [29], [35], [54], [69], point clouds [1], [65], textured mesh [30], [33], [58], [67], and multi-plane images [14], [60], [77], have been learned from 2D images through differentiable rendering to enhance novel view synthesis performance. However, achieving higher resolution remains challenging due to memory constraints.

NeRF [40] represents scenes with implicit fields of density and color. To extract more accurate surfaces, some works [42], [62], [71], [72] employ the signed distance function (SDF) to represent geometry in a scene. Building on these representations, numerous studies [8], [16], [43], [45], [48], [56], [64], [66], [68] use neural representation to capture humans. However, optimizing for each novel video is time-consuming.

Generalizable neural representation methods [5], [9], [34], [63], [70], [73] address this issue by employing implicit fields conditioned on image features. Inspired by these works, some studies [7], [27], [39], [75] propose generalizable radiance fields for humans, but they fail to achieve an animatable human model.

Leveraging the Skinned Multi-Person Linear (SMPL) model [36], several works [3], [6], [31], [44], [56] manage to obtain an animatable neural representation for humans. However, they still require subject-specific training. In contrast, our proposed novel deformable neural human representation enables us to acquire an animatable neural human representation from a single multi-view image of a new person without the need for additional training.

MPS-NeRF [17] can also learn an animatable neural human representation from multi-view images of a single frame of the target person. It mainly relies on inverse skinning weights of SMPL [36] to animate the representation, while our method learns an additional residual deformation mapping to compensate for the deformation that cannot be modeled by inverse kinemetric transformation. Furthermore, the representation of MPS-NeRF is based on NeRF [40] , while our representation additionally disentangles geometry and appearance by formulating them as two separate template implicit functions based on NeuS [62].

3 Method

Given a set of multi-view RGB videos of several persons performing various motions, our goal is to learn a generalized animatable neural human representation (Fig. 2) in the training. At inference time, our model enables two tasks: (1) Generalization: given a sparse multi-view (e.g., 3 or 4 views) videos of an unseen person performing arbitrary motions, we can synthesize novel views of the person performing these motions without training. (2) Animation: given a sparse multi-view images of an unseen person in a static pose, we can animate the neural representation of the person to generate novel pose images according to the user's pose control.

Our method uses the Skinned Multi-Person Linear (SMPL) model [36] as a 3D proxy and learns a canonical pose space shared by all the persons and poses. For each 3D point in a posed space, we convert it into this canonical space using the inverse skinning transformations [23] and the non-rigid deformations predicted by a neural network. Then, we learn a neural field [62] in the canonical space to infer Signed Distance Fields (SDF) and color for generating the final images. Our key idea is to extract geometry and appearance features for each 3D point from the sparse input images and use these features to infer the non-rigid residual deformations, SDF and color at the 3D point. To better train our model on multiple persons, we propose two new designs: (1) We disentangle geometry and appearance by formulating them as two separate template implicit functions in the canonical space. (2) We extract the geometry and appearance features at both the 3D point and its nearest SMPL vertices; the rationale is that our model is a person-agnostic model and the 3D proxy (i.e., SMPL model) for different persons is shared. Therefore, taking the properties at both the 3D point and its SMPL surface points into consideration would better infer the distance of the 3D point to the SMPL surface for different persons.

In the following, we first introduce our deformable neural human representation (Sec. 3.1) and then explain how we construct the geometry and appearance features used in such a representation (Sec. 3.2). Based on our neural representation, we can synthesize novel view images for arbitrary human poses (Sec. 3.3).

3.1 Deformable Neural Human Representation

In order to represent different human identities and poses, we adopt SMPL [36] as the base model in our framework. SMPL is a mesh-based human model consisting of a template mesh with $N_v = 6890$ vertices $V \in \mathbb{R}^{N_v \times 3}$ and driven by N_J joints. It deforms the template mesh according to a set of parameters, $\boldsymbol{\rho}$, representing the body shape and pose of a person. This process can be written as

$$\boldsymbol{v}_{\boldsymbol{\rho}} = \mathrm{SMPL}\left(\boldsymbol{v}, \boldsymbol{\rho}, \boldsymbol{w}_v\right), \tag{1}$$

where $\boldsymbol{v} \in V$ represents a surface vertex and $\boldsymbol{w}_v$ is the skinning weight of $\boldsymbol{v}$. We consider a predefined pose $\boldsymbol{\rho}_0$ as the canonical model of our framework. The canonical space is then defined

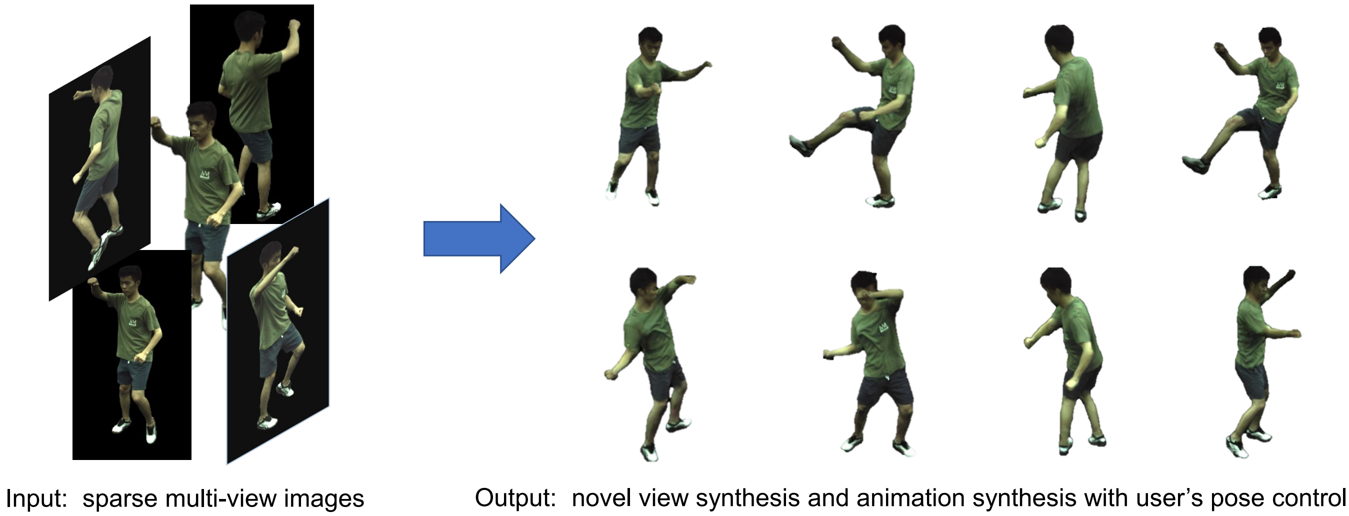
Input: sparse multi-view images
Output: novel view synthesis and animation synthesis with user's pose control

Fig. 1. Our model learns a generalized animatable neural human representation with multi-view RGB videos of several persons performing various motions. At inference time, given sparse multi-view images, our model can directly get novel view synthesis and animation synthesis with user's pose without further optimization.

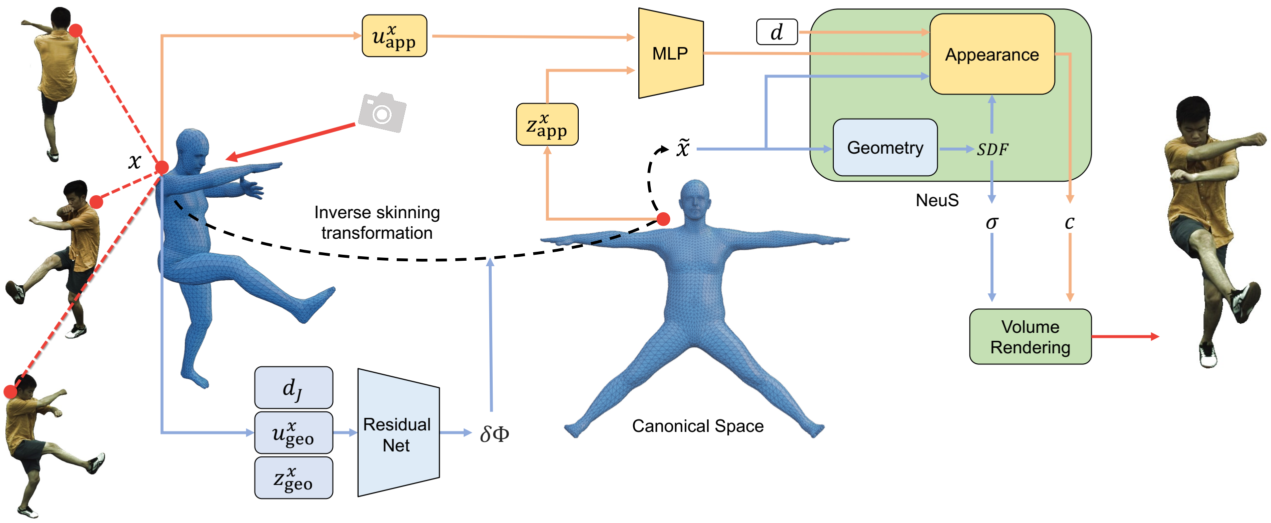

Fig. 2. Overview of our framework. Given a query point $\boldsymbol{x}$ in the posed space, we use inverse skinning transformation of its nearest surface point and a predicted residual deformation $\delta\Phi$ to transform $\boldsymbol{x}$ to the canonical space. The deformed point $\tilde{\boldsymbol{x}}$ is used as input to our geometry network and appearance network. The pose-dependent residual deformation is predicted using geometry features $\boldsymbol{u}^x_{\text{geo}}, \boldsymbol{z}^x_{\text{geo}}$ and the relative displacement $\boldsymbol{d}_J$ between the query point $\boldsymbol{x}$ and every joint. The appearance features $\boldsymbol{u}^x_{\text{app}}, \boldsymbol{z}^x_{\text{app}}$ are used as input to the appearance network. The appearance network also takes the view direction $\boldsymbol{d}$ as input.

as the corresponding 3D space containing this canonical model. For an arbitrary pose $\boldsymbol{\rho}$, we can transform the spacial points in the corresponding posed space into this canonical space using the SMPL surface as guidance. Such transformation links the canonical space to different posed spaces and thus allow those spaces to share the common features defined in the canonical space.

3.2 Geometry and appearance features

Our framework utilizes two latent variables, $\boldsymbol{F}_{\text{geo}}$ and $\boldsymbol{F}_{\text{app}}$, to represent the geometry and appearance features of a person, respectively. These features are extracted from the input images and then used to render new images of the person performing novel poses from novel views. Our framework obtains the geometry and appearance features in a similar manner, so we omit the subscripts in this section for simplicity when it is not confusing.

Both the geometry and appearance features are defined at every spacial point around the human model. The latent variable $\boldsymbol{F}$ defined at a spacial point $\boldsymbol{x}$ consists of two components $\boldsymbol{F} = (\boldsymbol{u}^x, \boldsymbol{z}^x)$, where $\boldsymbol{u}^x$ represents the image features based on pixel alignment, and $\boldsymbol{z}^x$ are surface features computed based on the connectivity of the SMPL mesh.

3.2.1 *Occlusion-aware image features*

Formally, our framework takes a sparse set of multi-view images, $\{I^c\}, c = 1, \ldots, N_C$, captured by N_C calibrated cameras as input, where each image $I^c \in \mathbb{R}^{H \times W \times (3+1)}$ contains the RGB color of every pixel and a foreground mask indicating the pixels belong to the person. We apply CNN [26] to these images and extract a set of feature maps $\boldsymbol{U}^c$ at multiple resolutions as

$$\boldsymbol{U}^c = \mathrm{CNN}([I^c]). \tag{2}$$

Then, for every spacial point $\boldsymbol{x}$, we obtain a set of image features $\{\boldsymbol{u}^c\}$ by projecting $\boldsymbol{x}$ onto each feature map $\boldsymbol{U}^c$. Inspired by [27], we employ a self-attention mechanism to aggregate these image features from different views to compute the image features $\boldsymbol{u}^x$ at $\boldsymbol{x}$ as

$$\boldsymbol{u}^x = \underset{c}{\mathrm{softmax}}\left(\frac{Q(\boldsymbol{u}^c)\cdot K(\boldsymbol{u}^c)}{\sqrt{d}} + B^c\right)\cdot V(\boldsymbol{u}^c), \tag{3}$$

where $Q(\cdot)$, $K(\cdot)$, and $V(\cdot)$ are the learnable query, key, and value embedding functions proposed in the original self-attention mechanism [61], and d is the dimension of the embedding space.

When a spacial point is occluded in an input image I^c, its corresponding image features $\boldsymbol{u}^c$ is not reliable and should not be weighted the same as the features from the other images. We thus employ an occlusion-aware mechanism to ensure this principle. Specifically, when a 3D query point $\boldsymbol{x}$ is occluded by the posed SMPL model in an input image I^c, we subtract a fixed bias B^c from the corresponding attention weights in Equation (3) to explicitly inform the self-attention mechanism of this occlusion. This mechanism is partially inspired by Attention with Linear Biases (ALiBi) [47]. We find that it effectively mitigates the artifacts caused by occlusion in our experiments.

3.2.2 Pose-aware surface features

The mesh structure of the SMPL model provides a strong prior for determining the shape and appearance of human body. To fully utilize such structural cues, we associate surface features to the mesh and diffuse them to the spacial points in the surrounding space.

To build these surface features, we extract the occlusion-aware image features of every vertex of the mesh from the image and employ the Graph Convolution Networks (GCN) to fuse them as suggested by [25], [46], [52]. GCN is a special convolutional network structure that aggregates the information on each individual vertex based on the connectivity of the mesh. We additionally include the displacement between each pair of connected vertices in this operation. Considering that the SMPL mesh is deformed based on pose parameters, this augmentation effectively allows the GCN to encode an implicit pose description into the computation, thus resulting in a set of pose-aware features.

To enhance the generalization ability of GCN, we use a novel local representation instead of global pose parameters (e.g. SMPL's 72 dimension of pose vector) as the input pose information of GCN. Specifically, the pose information is included in the graph edges, by using the direction and length of the posed edges as their features. The edge features are scattered to the node features to conduct message passing for the GCN. This localized pose representation helps the GCN to generate pose-dependent appearance and geometry to reduce reconstruction loss during training and generalize to new poses after training on a dataset with a large variety of poses.

Formally, assuming the input images correspond to pose parameters $\boldsymbol{\rho}_I$, we compute a deformed mesh V_{ρ_I} using SMPL and project each vertex $\boldsymbol{v} \in V_{\rho_I}$ onto the input images, obtaining a set of occlusion-aware image features $\{\boldsymbol{u}^v\}$. Then, we convert these image features using the pose-aware GCN described above and compute surface features

$$\{\boldsymbol{z}_I^v\} = \mathrm{GCN}(\{\boldsymbol{u}^v\}, V_{\rho_I}). \tag{4}$$

When rendering a new pose $\boldsymbol{\rho}$, we transform these input surface features $\{\boldsymbol{z}_I^v\}$ onto the corresponding new SMPL mesh $V_{\boldsymbol{\rho}}$ via another GCN procedure

$$\{\boldsymbol{z}^v\} = \mathrm{GCN}(\{\boldsymbol{z}_I^v\}, V_\rho). \tag{5}$$

The GCNs of Equation (4) and Equation (5) do not share weights. Note that these surface features, $\{\boldsymbol{z}^v\}$, are defined only on the surface vertices $\boldsymbol{v} \in V_\rho$. We then extend them to the surrounding space through a diffusion process. Specifically, for a spatial point $\boldsymbol{x}$ in the vicinity of the deformed mesh V_ρ, we find K nearest vertices $\{\boldsymbol{v}^k\} \subset V_\rho$ on the mesh and take their features $\{\boldsymbol{z}^k\}$. The surface feature $\boldsymbol{z}^x$ of the query point $\boldsymbol{x}$ is then computed as

$$\boldsymbol{z}^x = \sum_k w_k \,\mathrm{MLP}(\boldsymbol{z}^k, \boldsymbol{x} - \boldsymbol{v}^k), \tag{6}$$

where $w_k = (\|\boldsymbol{x}-\boldsymbol{v}^k\| + \varepsilon)^{-1} / \sum_k (\|\boldsymbol{x}-\boldsymbol{v}^k\| + \varepsilon)^{-1}$, and ε is a small scalar used to prevent dividing by zero.

3.2.3 Implementation

For a spacial point $\boldsymbol{x}$, the latent variable $\boldsymbol{F}_{\mathrm{geo}} = (\boldsymbol{u}_{\mathrm{geo}}^x, \boldsymbol{z}_{\mathrm{geo}}^x)$ and $\boldsymbol{F}_{\mathrm{app}} = (\boldsymbol{u}_{\mathrm{app}}^x, \boldsymbol{z}_{\mathrm{app}}^x)$ are computed using the images features and the surface features defined above. To further enforce disentanglement of the features, we let the appearance features $\boldsymbol{F}_{\mathrm{app}}$ be independent of the driving pose $\boldsymbol{\rho}$. This is achieved by computing $\boldsymbol{z}_{\mathrm{app}}^x$ using the canonical pose $\boldsymbol{\rho}_0$ in Equation (5), as depicts in Figure 2.

3.3 Pose-driven Volume Rendering

Using the geometry and appearance features $\boldsymbol{F}_{\mathrm{geo}}$ and $\boldsymbol{F}_{\mathrm{app}}$ extracted from the input images, our framework can render new images from a novel viewpoint given an arbitrary driving pose $\boldsymbol{\rho}$. We employ NeuS [62], an SDF-based differential renderer, to synthesize those images. NeuS predicts the color of each pixel by accumulating the radiance along the camera ray $\boldsymbol{r}$ passing through the pixel. This computation can be discretized using a series of spacial points $\{\boldsymbol{x}_i\}$ sampled along $\boldsymbol{r}$. Specifically, NeuS computes

$$\tilde{C}(\boldsymbol{r}) = \sum_{i=1}^{n} T_i \alpha_i c_i, \tag{7}$$

where $\tilde{C}(\boldsymbol{r})$ is the predicted color, $T_i = \prod_{j=1}^{i-1}(1-\alpha_i)$ is the discrete accumulated transmittance, α_i represents the opacity values defined as

$$\alpha_i = \max\left(\frac{\phi(s_i) - \phi(s_{i+1})}{\phi(s_i)}, 0\right), \tag{8}$$

and $\phi(x) = (1+e^{-kx})^{-1}$ with a learnable scalar k. The color values c_i and the SDF values s_i in the above equations are evaluated at every sample point $\boldsymbol{x}_i$. Our framework handles every sample point in the same way, so we omit the subscript i in the rest of this section for simplicity.

3.3.1 Pose-driven deformation field

Given a driving pose $\boldsymbol{\rho}$, we transform each sample point $\boldsymbol{x}$ into the canonical space and evaluate c and s based on the canonical position $\tilde{\boldsymbol{x}}$. This mechanism is inspired by recent studies [31], [59], which have shown its efficiency in modeling dynamic scenes and human poses. We define the deformation mapping Φ using the inverse skinning transformation [23]. As suggested by [31], an additional residual deformation mapping $\delta\Phi$ is employed to compensate for the deformation that cannot be captured by the inverse skinning, such as the deformation of cloth. The canonical position $\tilde{\boldsymbol{x}}$ of a sample point $\boldsymbol{x}$ is thus computed as

$$\tilde{\boldsymbol{x}} = \Phi(\boldsymbol{x}, \boldsymbol{\rho}) + \delta\Phi(\boldsymbol{x}, \boldsymbol{\rho}). \tag{9}$$

$\Phi(\boldsymbol{x}, \boldsymbol{\rho})$ is the inverse skinning mapping

$$\Phi(\boldsymbol{x}, \boldsymbol{\rho}) = \sum_{j=1}^{N_J} w_j \cdot (R_j(\boldsymbol{x} - \delta\boldsymbol{v}) + \boldsymbol{t}_j), \tag{10}$$

where R_j and $\boldsymbol{t}_j$ represent the rotation and translation that transform joint j from pose $\boldsymbol{\rho}$ back to the canonical pose $\boldsymbol{\rho}_0$, $w_j \in \boldsymbol{w}_v$ is the corresponding skinning weight of the surface point $\boldsymbol{v}$ that is the nearest to $\boldsymbol{x}$, and N_J is the number of joints. Note that we allow the pose parameters to also define the body shape of the target person. A displacement $\delta\boldsymbol{v}$ is leveraged to compensate for the deformation caused by the change of body shape. Specifically,

$$\delta\boldsymbol{v} = \mathrm{SMPL}\left(\boldsymbol{v}, \beta(\boldsymbol{\rho}), \boldsymbol{w}_v\right) - \mathrm{SMPL}\left(\boldsymbol{v}, \beta(\boldsymbol{\rho}_0), \boldsymbol{w}_v\right), \tag{11}$$

where $\beta(\cdot)$ extract the body shape parameters from $\boldsymbol{\rho}$.

The residual deformation $\delta\Phi$ is computed using the geometry features $\boldsymbol{F}^x_{\text{geo}}$ extracted from the input images. We further consider the relative displacement between the query point $\boldsymbol{x}$ and every joint, collectively represented by $\boldsymbol{d}_J$, as an extra cue. The residual displacement is thus computed as

$$\delta\Phi(\boldsymbol{x}, \boldsymbol{\rho}) = \mathrm{MLP}(\boldsymbol{F}^x_{\text{geo}}, \boldsymbol{d}_J). \tag{12}$$

3.3.2 SDF-based volume rendering

After transforming the sample points $\boldsymbol{x}$ into the canonical space $\tilde{\boldsymbol{x}}$ using the deformation field, we compute its SDF value s and color c as

$$s = \mathscr{S}(\tilde{\boldsymbol{x}}) \tag{13}$$

$$c = \mathscr{C}(\tilde{\boldsymbol{x}}, \boldsymbol{F}_{\text{app}}, \boldsymbol{d}, s, \boldsymbol{n}_x). \tag{14}$$

Both $\mathscr{S}$ and $\mathscr{C}$ are implemented as MLPs. The SDF function $\mathscr{S}$ only takes the canonical position of $\boldsymbol{x}$ as input. The color function $\mathscr{C}$ considers the appearance feature $\boldsymbol{F}_{\text{app}}$, the view direction $\boldsymbol{d}$, the SDF value s of the sample point, as well as the normal vector $\boldsymbol{n}_x$ of the implicit surface at the sample point. $\boldsymbol{n}_x$ can be computed as the gradient of the SDF function $\boldsymbol{n}_x = \nabla_x \mathscr{S}(\boldsymbol{x})$. The results of these functions are then used by NeuS to predicts the color of the pixel as described above.

3.4 Training

For every training image, we render m random pixels and sample n spacial points on each generated camera ray. The loss function is then defined as

$$\begin{aligned} \mathscr{L} = \frac{1}{m}\sum_r \underbrace{\|\tilde{C}(r) - C(r)\|_1}_{\text{color loss}} + \lambda_1 \sum_r \underbrace{\mathrm{BCE}(\tilde{M}_r, M_r)}_{\text{mask loss}} \\ + \lambda_2 \frac{1}{nm}\sum_{\boldsymbol{x}} \underbrace{(\|\boldsymbol{n}_{\boldsymbol{x}}\|_2 - 1)^2}_{\text{eikonal term}} + \lambda_3 \underbrace{\mathrm{LPIPS}(\tilde{C}(P), C(P))}_{\text{LPIPS loss}}, \end{aligned} \tag{15}$$

where the color loss measures the difference between the predicted color $\tilde{C}(r)$ and the ground truth $C(r)$, the mask loss matches the predicted mask $\tilde{M}_r = \sum_{i=1}^{n} T_i \alpha_i$ with the foreground mask M_r by computing the binary cross entropy (BCE) between them, and an Eikonal term [18] is adopted to regularize the SDF function. We further employ a perceptual loss, LPIPS [74], to ensure the quality of the synthesized image. This LPIPS loss is computed by rendering random patches P sampled on the target image.

During training, we utilized multi-view inputs with multiple frames to introduce pose variation. In detail, we randomly selected a frame of a character and used its multi-view inputs to train the model. This involved reconstructing an image of another view using any three available views and calculating the loss. During testing, we only use a single frame of multiview image to generate an animatable avatar, which is consistent with our training setting. The process of obtaining image features to render a new pose involves transforming a 3D point from the new pose to the input pose to acquire image features, using the canonical space as a bridge. Besides image features, we also get pose-aware surface features for the new pose using the GCN as described.

4 Experiment

Implementation details. We train our models using the Adam [24] optimizer. We follow a two-stage training regime for a faster convergence. We first pre-train the SDF network using the canonical model mentioned in Sec 3.2 and then train all the networks jointly. The learning rate is first linearly warmed up from 0 to 5×10^{-4} in the first 2k iterations and then is controlled by the cosine decay schedule. In the second stage, we freeze the SDF network after training for 50K iterations. We use the sampling strategy proposed in NeuS [62], and the numbers of the coarse and fine sampling are 32 and 32 respectively. Following Neural Actor [31], we also adopt a geometry-guided ray marching process for the volume rendering. Specifically, we only sample points near the SMPL surface to speed up the rendering process. We first train our models on 4 Nvidia V100 32G GPUs and sample 1024 rays per batch per GPU for 80K iterations without LPIPS loss. Then, we randomly select two additional patches P (size 24×24) per GPU to continue training the model with LPIPS loss for 20,000 iterations. The training takes about $38 + 18$ hours to complete. Because the number of subjects in the ZJU-MoCap is small, we augment the data using color jittering during the training. For the loss weight, we set λ_1, λ_2 both to 0.1, and λ_3 to 1.

Evaluation metric. We measure the quality with two evaluation metrics: Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM). Following previous work, we project the 3D bounding box of the human body onto the image plane to get a 2D mask and only calculate the PSNR and SSIM of the mask area instead of the whole image. For the 3D reconstruction, we only provide the qualitative results, as shown in Fig. 5, because the ground truth is not available.

The goal of our work is to get novel-view and novel pose human synthesis. To the best of our knowledge, only MPS-NeRF [17] can achieve both at the same time. Hence, we conduct two experiments: novel view synthesis generation (Sec 4.1) and pose control animation (Sec 4.2).

4.1 Generalization

In this part, we evaluate our approach on the novel view synthesis generalization task. Given a sparse multi-view (e.g., 3 or 4 views)

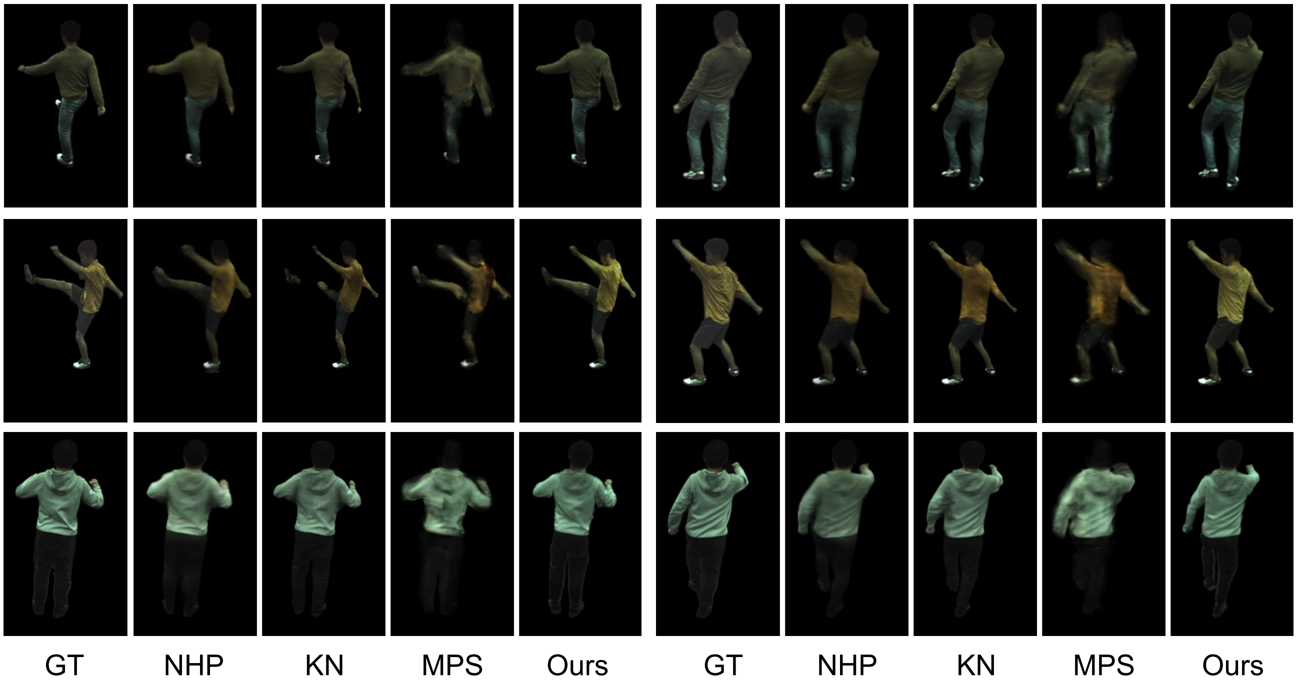

Fig. 3. Qualitative comparison of identity generalization on the ZJU-MoCap [45] dataset. Our method outperforms three baselines, Neural Human Performer (NHP) [27], Keypoint NeRF (KN) [39] and MPS-NeRF (MPS) [17] in terms of synthesized wrinkles and appearance details. All methods are **trained on all the source subjects** and directly **tested on the target subjects** without training.

TABLE 1
Quantitative comparison of the generalization task in the four settings. We evaluate the synthesis quality on two metrics: PSNR and SSIM . We compare with other generalized model, Neural Human Performer (NHP) [27],Keypoint NeRF (KN) [39] and MPS-NeRF (MPS) [17]. Our method achieves significantly better performance in identity and cross-dataset generalization. In other two setting, our method is on par with other methods. Additionally, we provide the results of a person-specific model, Neural Body (NB) [45], in the task of seen poses for seen subjects as an upper-bound baseline. Our method achieves comparable performance with this benchmark.

Setting	Identity generation		Cross-dataset		Pose generation		Seen subjects	
Method	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
NB	-	-	-	-	-	-	28.56	0.943
NHP	24.85	0.908	23.16	0.869	26.19	0.915	26.90	0.927
KN	24.92	0.910	22.31	0.861	**27.64**	**0.933**	**27.91**	**0.938**
MPS	22.99	0.877	22.38	0.842	24.66	0.880	24.84	0.887
Ours	**25.14**	**0.914**	**24.19**	**0.886**	27.36	0.929	27.83	**0.938**

videos of an unseen person performing arbitrary motions, our method can synthesize novel views of the person performing these motions without training. In this setting, we compare our method with MPS-NeRF (MPS) [17], Keypoint NeRF (KN) [39], Neural Human Performer (NHP) [27] and Neural Body (NB) [45]. We also compare our method with Neural Body (NB) [45] which is a person-specific model

ZJU-MoCap dataset consists of 10 human subjects captured from 23 synchronized cameras. Following NHP, we split the dataset into two parts: 7 **source** subjects and 3 **target** subjects. We evaluate our method and the baseline methods in the following four different settings. Note that for all the comparisons except the cross-dataset generalization, the first 300 frames of the source or target subjects are used for training, and the rest frames (unseen poses) are used for testing.

1) Identity generalization. First, we evaluate the generalization to different identities by testing on the **target subjects**. All methods are trained on all the source subjects and directly tested on target subjects. As Tab. 1 and Fig. 3 show, our method gives the best performances quantitatively and qualitatively.

2) Cross-dataset generalization. To further test the generalizability of our method to new datasets, we train all methods on the ZJU-MoCap dataset and directly test on the DeepCap [21] and DynaCap [20] dataset without fine-tuning. As shown in the Tab. 1, our method significantly improves the performance compared to other methods. Even though the training and testing datasets are significantly different in the appearance distribution and the distance from the camera to subject, our method still can achieve impressive results without fine-tuning, as shown in Fig. 4.

3) Pose generalization. In this setting, all the methods are trained on the source subjects and tested on **unseen poses of the same subject**. All methods are trained on all source subjects together. As Tab. 1 shows, our model outperforms MPS, NHP significantly on PSNR and SSMI. Our method also achieve comparable results to KN.

4) Seen poses of seen subjects. To demonstrate the superiority

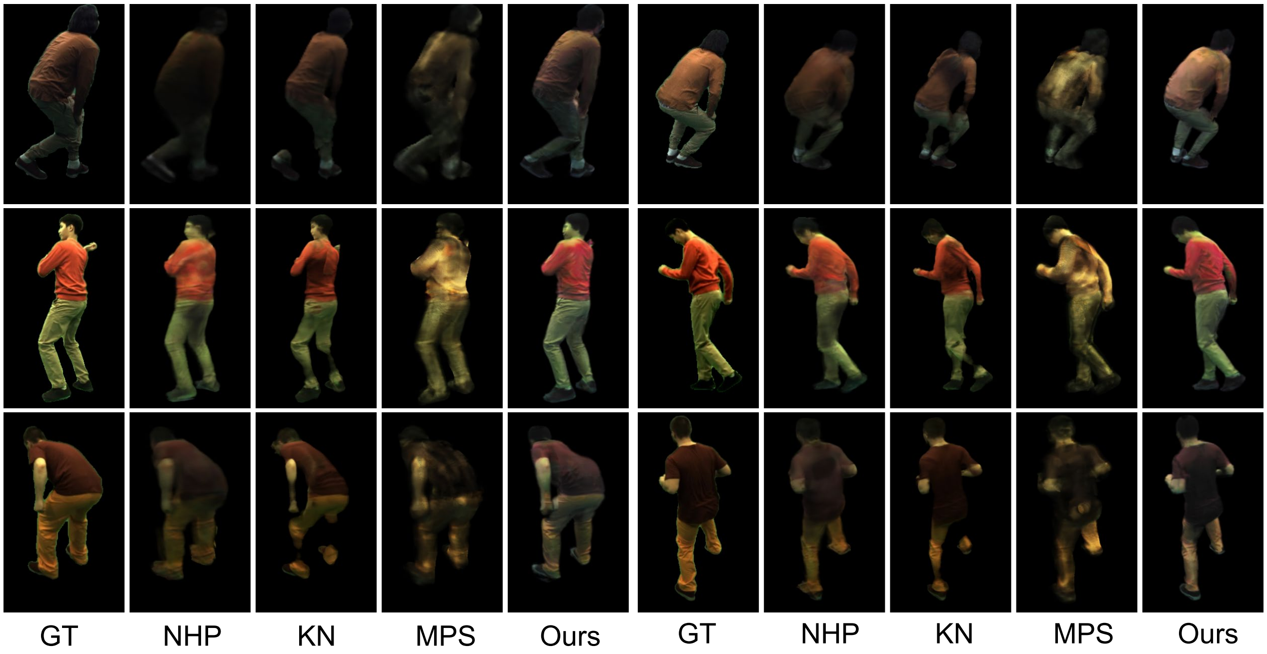

Fig. 4. Qualitative comparison of cross-dataset generalization on novel view synthesis. All methods are trained on the ZJU-MoCap [45] and directly tested on the DeepCap [21] and DynaCap [20]. Our method significantly outperforms other baselines.

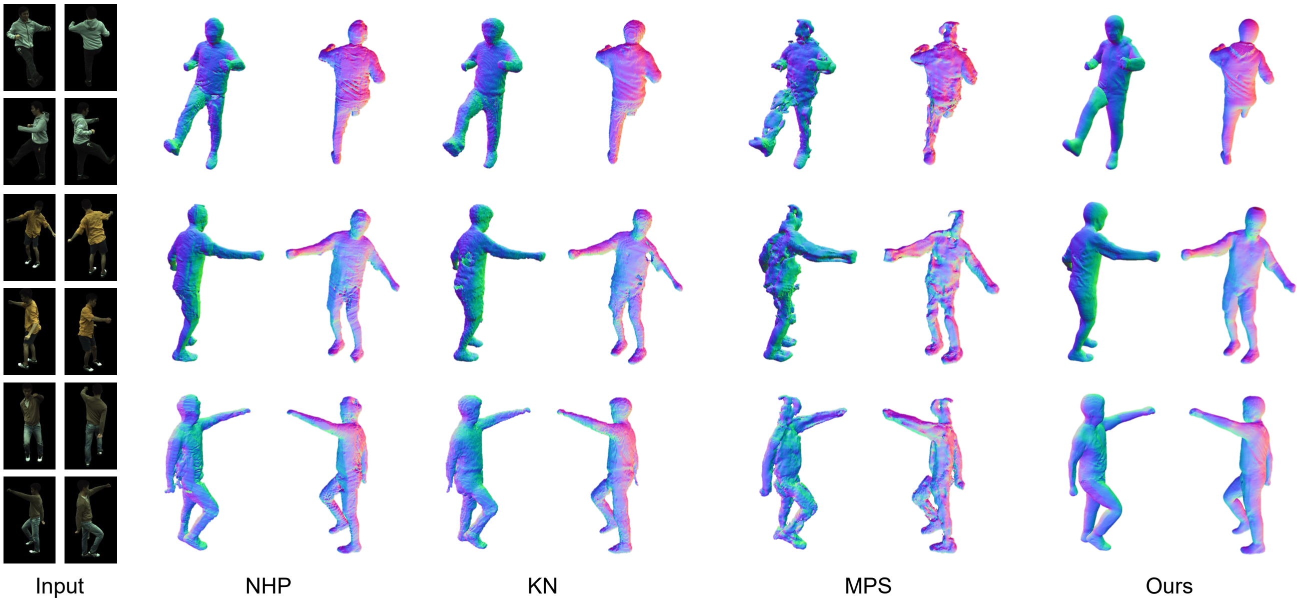

Fig. 5. Visualization of 3D reconstruction results. The meshes are extracted by running the Marching Cubes algorithm on the predicted volume density. Other methods contain more unwanted artifacts compared to our method.

of our model, we evaluate the performance for seen poses of source subjects. All methods except NB are trained on the all source subjects and tested on the seen poses of them. Tab. 1 demonstrates that our method outperforms MPS and NHP, and it is comparable to NB and KN.

Note that KN is sightly better than our method in task 3 and 4, but our method outperform it significantly in task 1 and 2. Moreover, there are artifacts in the results of KN for cross dataset, which do not appear in the results of the ZJU-MoCap dataset. The reason for this is that KN sets a fixed hyper-parameter that controls the impact of each keypoint. The hyper-parameter determines relative spatial encoding for 3D query point and keypoints, which is sensitive to human body shape and pose. As a result, KN can overfit the training dataset, but it does not generalize well for the pose or shape not seen in the training dataset. And KN is not animatable human models.

Our method shares the same goal as MPS-NeRF but differs in our approach. MPS-NeRF mainly relies on the SMPL model to fuse image features for generating novel views, while we associate surface features with the SMPL model and diffuse them onto surrounding spatial points using a specially designed GCN network. This allows us to reconstruct reasonable geometry details, such as

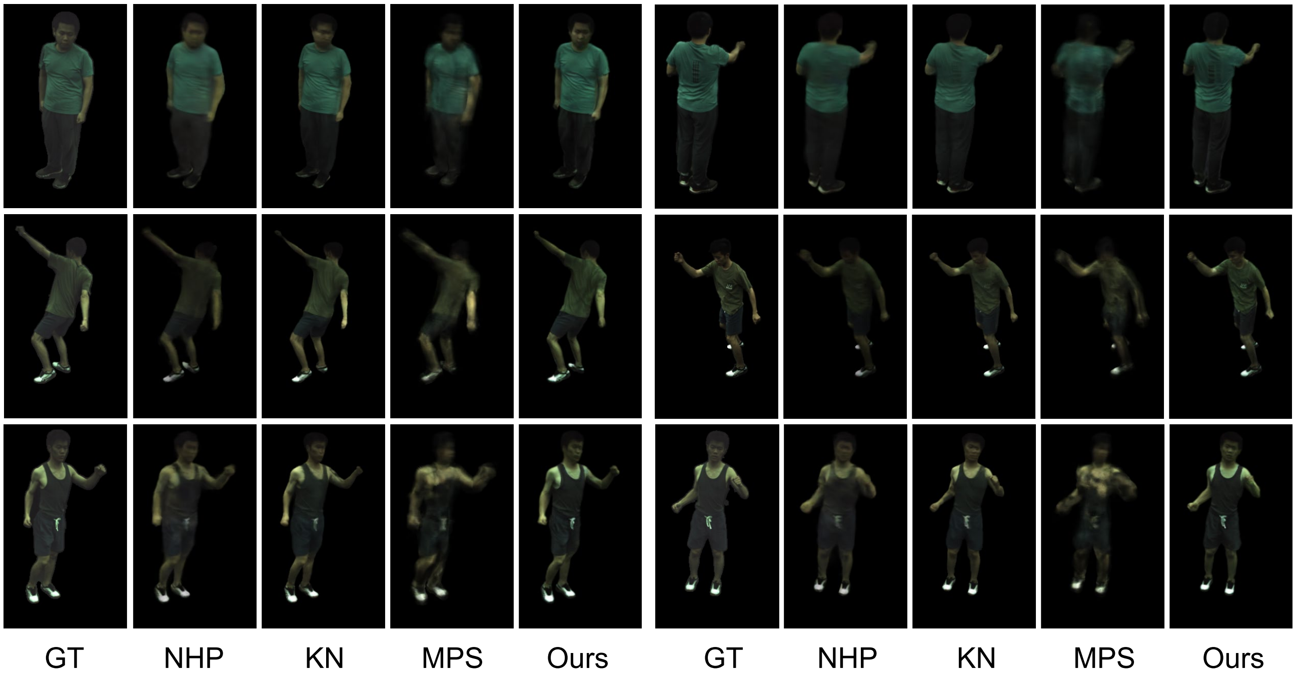

Fig. 6. Qualitative comparison of Pose generalization on the ZJU-MoCap dataset. Our method significantly outperforms three baselines, Neural Human Performer (NHP) [27], Keypoint NeRF (KN) [39] and MPS-NeRF (MPS) [17], in terms of synthesized wrinkles and appearance details.

clothes and hair, even with sparse input images. Additionally, our occlusion-aware self-attention mechanism can effectively handle mesh points that are occluded in the sparse input images. As shown in Tab. 1, these mechanisms allow our method to significantly outperform MPS-NeRF in all tasks.

In conclusion, our method achieves state-of-the-art results on the novel view synthesis generalization task for new character (identity and cross-dataset generalization). And our method is comparable to other methods in the fitted task.

4.2 Animation

The task of animation is that, given sparse multi-view images of an **unseen** person in a static pose, the model needs to generate novel pose images under user's pose control. This task is distinct from the pose generation task discussed in Sec 4.1. In the case of generating images, the task only requires pose input without any image input. On the other hand, pose generation tasks involve using image input to generate novel view images. To evaluate the performance of this task, we compare our method with Neural Body (NB), Animatable Nerf (AN) [44], Neural Actor (NA) [31] and MPS-NeRF (MPS) [17] on the ZJU-MoCap dataset. Neural Human Performer (NHP) [27] and Keypoint NeRF (KN) [39] are not animatable human models and thus cannot generate images with user's pose control. The spilt for the ZJU-MoCap dataset is the same as that described in Sec. 4.1.

MPS-NeRF and our method are the generalized animatable human model for this task, so we train them on all the source subjects. At test time, we directly obtain an animatable model of the target person just from the sparse camera views of one frame of the target person without training. Other person-specific methods are trained in a person-specific manner on the first 300 frames of the target person and tested on the remaining frames of the same person.

Tab. 2 and Fig. 7 demonstrate that our method significantly outperforms MPS-NeRF quantitatively and qualitatively. MPS-NeRF relies on the inverse skinning over the SMPL model to animate the character, while our method employs an additional residual deformation mapping to compensate for the deformation that SMPL cannot model. This mechanism results in a more accurate shape and appearance, as shown in Fig. 7.

Moreover, our method outperforms other baselines even though it is trained and evaluated in a more difficult setting (i.e., unlike other baselines, our method is not overfitted to the target subject during training). Neural Actor (NA) [31] is designed for input videos with dense camera views and requires obtaining textures for each frame of the input video. However, in sparse view settings, acquiring high-quality textures can be challenging due to issues like self-occlusion that reduce effectiveness. Moreover, Animatable NeRF (AN) [44] and Neural Body (NB) [45] do not account for self-occlusion in sparse view and lack specific designs for completing the missing parts during training. In contrast, our method can obtain high-quality avatars from sparse view inputs using carefully designed modules, such as GCN and the occlusion-aware self-attention mechanism. We believe that our design can also enhance per-scene optimization methods. Additionally, our method is trained on multi-person data, which allows for better pose generalization compared to methods trained on single-person data.

To more effectively demonstrate the efficacy of our approach, we assess its performance on a cross-dataset evaluation. Among existing methods, only MPS-NeRF and our method are capable of accomplishing this task. We train all method on the ZJU-MoCap dataset and directly test on the DeepCap [21] and DynaCap [20] dataset. To demonstrate the difficulty of this task, building upon prior research by VideoAvatar [2], we implemented a baseline projection method (denoted as Tex), which projects pixel colors onto the surface of the SMPL model. The SMPL model, learned from scans of unclothed humans, is unable to represent clothing

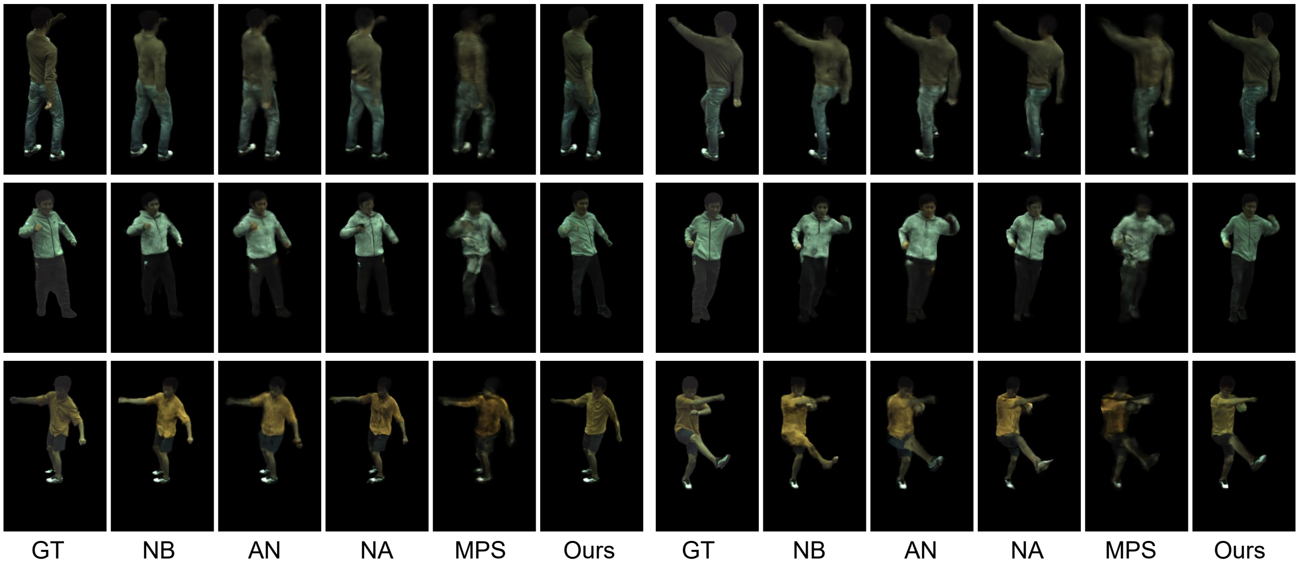

Fig. 7. Qualitative comparison of the Animation task on the ZJU-MoCap dataset. We compared with three person-specific methods, Neural Body (NB) [45], Animatable Nerf (AN) [44], Neural Actor (NA) [31] and the generalized method, MPS-NeRF (MPS) [17]. Note that the person-specific methods are trained with 300 frames of a target person and tested on the same person, while MPS-NeRF and our method is trained on all source subjects and obtains an animatable human model of a target person just from **one** frame of the person. Our method outperforms all baselines even in a disadvantaged setting.

TABLE 2
Quantitative comparison of the animation task. Our method achieves the best performance in two metrics, compared to three person-specific animatable human models, NeuralBody (NB) [45], AnimatbleNerf (AN) [44], and Neural Actor (NA) [31] and MPS-NeRF (MPS) [17]

Method	NB	AN	NA	MPS	Ours
PSNR↑	23.03	22.95	22.50	21.80	**23.18**
SSIM↑	0.880	0.875	0.878	0.858	**0.886**

or other personal surface details. We estimate per-vertex deformations of the SMPL model by rendering it to fit the foreground mask. However, it cannot accurately model geometry, due to the limitations of the SMPL template as a base. In contrast to simple projection, we reconstructed better geometry and used our GCN module to complete the missing parts from a few viewpoints, resulting in better performance, as shown in Fig. 8. Furthermore, the state-of-the-art method MPS-NeRF, which relies exclusively on pixel-aligned features and does not incorporate human prior information or disentangle appearance and geometry, produced inferior results, particularly when applied across different datasets.

We also fine-tune our model for approximately 10 minutes with the same input, using 3 images as input and another image as the target each time. As shown in Fig. 8, fine-tuning can reduce color deviations caused by differences in lighting distribution between the test data and training dataset. Additionally, fine-tuning can result in more accurate geometry.

4.3 Appearance and geometry control

As our method disentangles the geometry and appearance in the human modeling, we can either change the appearance while keeping the geometry fixed or control the body shape of modeled humans by manipulating shape parameters in SMPL. Fig. 9 demonstrates the synthesized images after changing the body shape and exchanging the appearance.

4.4 Ablation Study

We conduct ablation studies using the ZJU-MoCap dataset on both the generalization and animation tasks. The same experiment settings as described in Sec. 4.1 and Sec. 4.2 are used for these two tasks. The results are shown in Table 3. We also provide visual result in Fig. 10 to further demonstrate the significance of different components.

We first evaluate the effect of the GCN used in extracting the surface features. The baseline, w/o GCN, is performed by computing the surface features directly using the image features. In the absence of the GCN module, the model cannot effectively handle occlusion and utilize prior knowledge about the human body as discussed in Sec. 3.2.2, which leads to artifacts both in shape and appearance as showed in Fig. 10. Additionally, the drop in performance shown in Tab. 3 underscores the importance of the GCN module.

We also evaluate the effect of the image features. We compare with: 1) removing the image features from the geometry features (w/o $\boldsymbol{u}_{\text{geo}}$); 2) removing the image features from the appearance features (w/o $\boldsymbol{u}_{\text{app}}$); 3) disabling the occlusion-aware self-attention mechanism (w/o Occ) by letting $B^c = 0$ in Equation (3). As visible in Fig. 10, the predicted results have significant geometric deviations when lacking the geometry feature $\boldsymbol{u}_{\text{geo}}$, mainly reflected in the lack of clothing geometry and errors in the shape of the person. Similarly, when lacking the appearance feature $\boldsymbol{u}_{\text{app}}$, the model's generalization ability is significantly impaired, as evidenced by reconstructing incorrect clothing colors and confusing clothing with limbs. Furthermore, when lacking the occlusion module, the model's ability to process multi-view information is greatly reduced, and it may incorrectly utilize the input information that should not be used, such as reconstructing frontal clothing information in the back. These comparisons show that the image

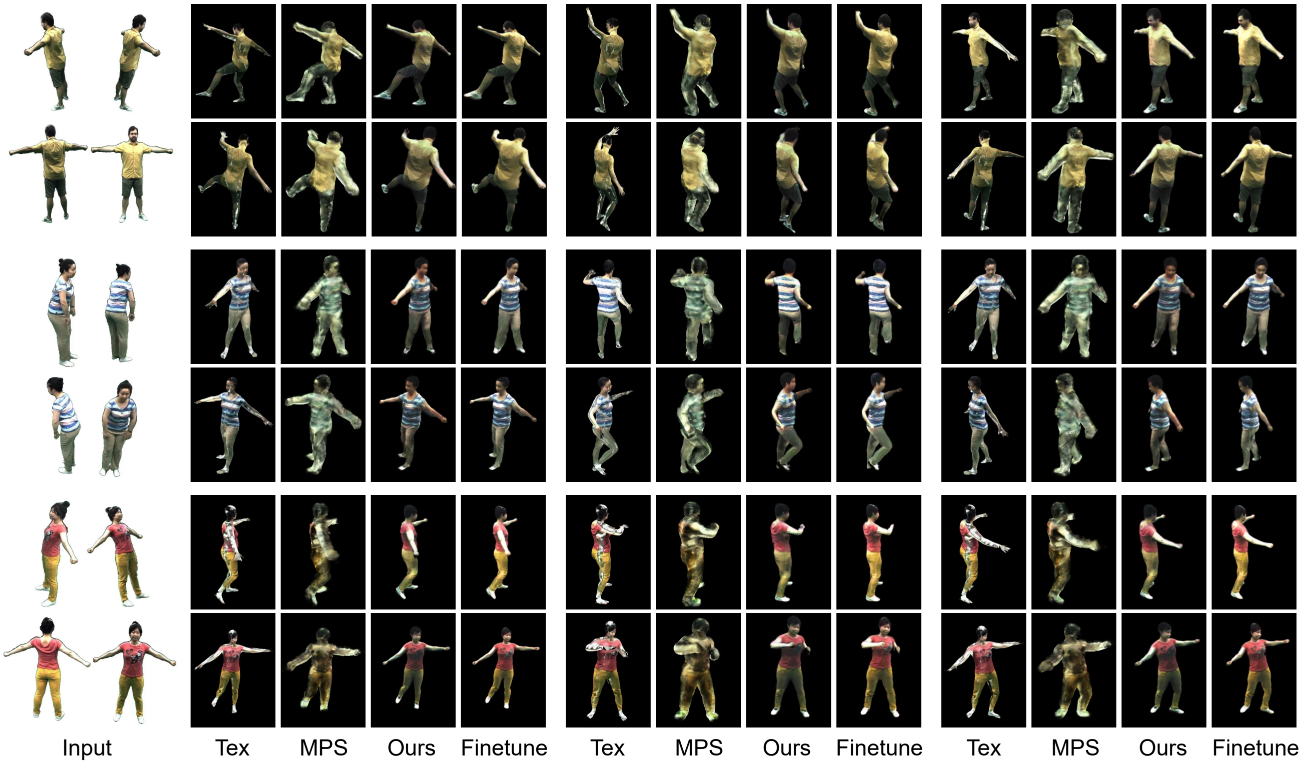

Fig. 8. Qualitative comparison of cross-dataset generalization on animation. Tex: projects pixel colors onto the surface of the SMPL model with per-vertex deformations. Finetune: fine-tune our model for approximately 10 minutes with the same input. MPS-NeRF and our method trained on the ZJU-MoCap [45] and directly tested on the DeepCap [21] and DynaCap [20]. Our method shows significant superiority over other methods, while also yielding superior outcomes with fine-tuning.

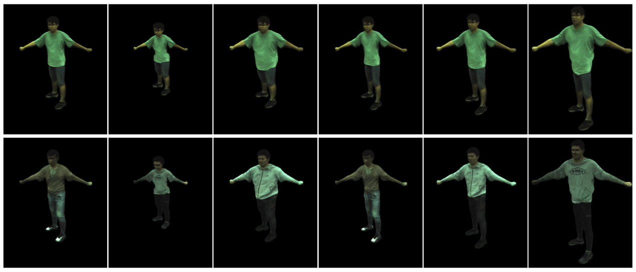

Fig. 9. Results of changing appearance and geometry. We can directly change human shape (top row) and human appearance (bottom row) while keeping other factors fixed.

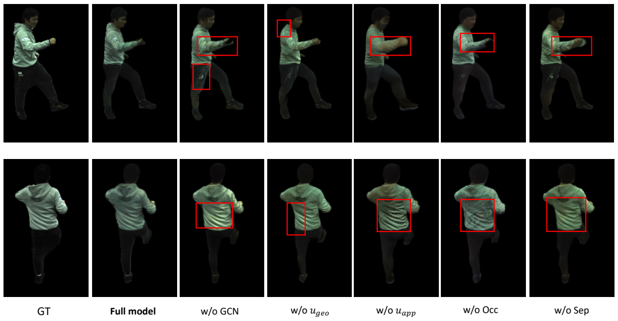

Fig. 10. Visual result of the ablation studies for different components. Occ: the occlusion-aware self-attention mechanism, Sep: separation of geometry and appearance features.

features are critical to quality results, while our occlusion-aware mechanism effectively improves the perceptual realism.

To validate our design of disentangling the geometry and appearance features, we train a model with a single variable for both the appearance and geometry features (w/o Sep). When the disentangling module is missing, the model is prone to confusing geometry and appearance during training, resulting in many artifacts in both geometry and appearance in the predicted results. For example, in Figure 10, an erroneous human body shape is depicted, and clothing is mistakenly reconstructed on the hand. As shown in Tab. 3, our disentangled features achieve better image quality in terms of all of the three metrics.

In table 4, we show the performance of our model in terms of the generalization to novel views and novel animations with different numbers of input views. The performance of our method degrades slightly when given fewer input views.

We also compared our method using NeuS and the original volume rendering algorithm in Nerf. By using the NeuS rendering method, the sampling points are more concentrated on the object surface, which makes the deformation field easier to train. This also allows for more accurate human geometry and fewer appearance artifacts, as shown in the pose driven results in the Fig. 11.

5 Discussion

Despite its success, our framework still has several limitations. First, our method relies on the accuracy of pose tracking, and the low-quality SMPL estimation may result in artifacts. It would be interesting to optimize the SMPL parameters in the framework.

TABLE 3
Ablation study for different components. Occ: the occlusion-aware self-attention mechanism, Sep: separation of geometry and appearance features.

	Unseen subjects		Animation	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
w/o GCN	24.27	0.891	22.81	0.872
w/o $\boldsymbol{u}_{\text{geo}}$	23.79	0.890	22.34	0.875
w/o $\boldsymbol{u}_{\text{app}}$	23.39	0.881	22.38	0.870
w/o Occ	24.10	0.889	22.79	0.874
w/o Sep	23.68	0.891	22.46	0.874
Full model	**25.14**	**0.914**	**23.18**	**0.886**

TABLE 4
Quantitative evaluation of using different numbers of input camera views during inference. The performance of our method degrades slightly when given fewer input views.

view number		1 view	2 views	3 views	4 views
Unseen subjects	PSNR↑	23.28	24.26	24.78	**25.14**
	SSIM↑	0.887	0.900	0.909	**0.914**
Animation	PSNR↑	22.30	22.90	23.08	**23.18**
	SSIM↑	0.875	0.881	0.883	**0.886**

Second, our method can only handle the clothing types that follow the topology of the SMPL model, and it is challenging to model very loose clothes, such as skirts. Finally, our method does not model complex lighting effects and we assume the uniform lighting in the input multi-view videos. When the assumption cannot be met (e.g., the ZJU-MoCap dataset), our model tends to learn the average lighting and produce the results with color shift. We leave them for future work. Moreover, it is noteworthy that our proposed model exhibited exceptional performance despite being trained on a restricted dataset. The potential benefits of augmenting our framework to accommodate larger models and incorporating additional training data warrant further exploration.

We extract image features from both the spatial and surface points of the SMPL model to more effectively infer person- and pose-dependent properties, and our GCN integrates local pose information by using the edges of the posed SMPL as its edge features, enabling the generalization of pose-aware surface features. In contrast, for the novel view synthesis generalization task, the state-of-the-art method, Keypoint NeRF (KN) [39], relies on spatial encoding for 3D query points and keypoints, which is sensitive to human body shape and pose. The person- and pose-dependent properties of our method lead to improved performance in pose-dependent appearance and geometry in the task of cross-dataset and identity generalization. On the other hand, in the task of animation, although we design a residual module to compensate for the deformation caused by changes in body shape and pose, using the extracted pose-aware surface features. However, determining the changes in appearance and geometry with different poses is challenging, given only four static images as input. Consequently, the performance of pose-dependent appearance and geometry in this task is not as strong as in novel view synthesis. As a result, exploring the use of multi-view videos in testing to enhance this capability is another interesting direction.

6 Conclusion

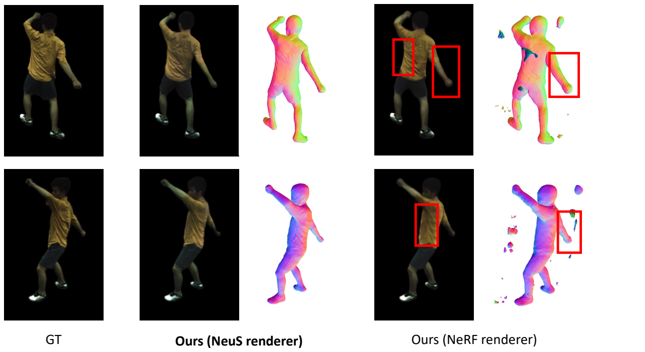

Fig. 11. Comparison of NeuS rendering and NeRF rendering: The NeuS rendering method enables more accurate human geometry and fewer appearance artifacts.

We presented Neural Novel Actor, a new method for learning a generalized animatable neural human representation from a sparse set of multi-view imagery of multiple persons. With the learned representation, we can synthesize novel view images of an arbitrary person using a sparse set of cameras and further synthesize animations with user's pose control. To efficiently learn this representation for multiple persons, we design our proposed human representation with disentangled geometry and appearance. Furthermore, we leverage the features at both the spatial points and the surface points of SMPL to infer pose- and person-dependent geometry and appearance. Extensive experiments demonstrate that our method significantly outperforms the state-of-the-arts on the tasks of novel view synthesis of new persons and the animation synthesis with pose control.

Acknowledgments

This work was supported in part by NSFC Projects of International Cooperation and Exchanges (62161146002). Christian Theobalt was supported by ERC Consolidator Grant 4DReply (770784). Lingjie Liu was supported by Lise Meitner Postdoctoral Fellowship.

References

[1] K.-A. Aliev, A. Sevastopolsky, M. Kolos, D. Ulyanov, and V. Lempitsky. Neural point-based graphics. In European Conference on Computer Vision, pp. 696–712. Springer, 2020.

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8387–8397, 2018.

[3] anonymous. Hdhumans: A hybrid approach for high-fidelity digital humans. 2022.

[4] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. ACM transactions on graphics (TOG), 22(3):569–577, 2003.

[5] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14124–14133, 2021.

[6] J. Chen, Y. Zhang, D. Kang, X. Zhe, L. Bao, X. Jia, and H. Lu. Animatable neural radiance fields from monocular rgb videos. arXiv preprint arXiv:2106.13629, 2021.

[7] M. Chen, J. Zhang, X. Xu, L. Liu, Y. Cai, J. Feng, and S. Yan. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII, pp. 222–239. Springer, 2022.

[8] X. Chen, W. Li, D. Cohen-Or, N. J. Mitra, and B. Chen. Moco-flow: Neural motion consensus flow for dynamic humans in stationary monocular cameras. arXiv preprint arXiv:2106.04477, 2021.

[9] J. Chibane, A. Bansal, V. Lazova, and G. Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7911–7920, 2021.

[10] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. ACM Transactions on Graphics (ToG), 34(4):1–13, 2015.

[11] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In ACM SIGGRAPH 2008 papers, pp. 1–10. ACM New York, NY, USA, 2008.

[12] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pp. 145–156, 2000.

[13] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. ACM Transactions on Graphics (ToG), 35(4):1–13, 2016.

[14] J. Flynn, M. Broxton, P. Debevec, M. DuVall, G. Fyffe, R. Overbeck, N. Snavely, and R. Tucker. Deepview: View synthesis with learned gradient descent. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2367–2376, 2019.

[15] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1746–1753. IEEE, 2009.

[16] C. Gao, Y. Shih, W.-S. Lai, C.-K. Liang, and J.-B. Huang. Portrait neural radiance fields from a single image. arXiv preprint arXiv:2012.05903, 2020.

[17] X. Gao, J. Yang, J. Kim, S. Peng, Z. Liu, and X. Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–12, 2022. doi: 10.1109/TPAMI.2022.3205910

[18] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman. Implicit geometric regularization for learning shapes. arXiv preprint arXiv:2002.10099, 2020.

[19] K. Guo, P. Lincoln, P. Davidson, J. Busch, X. Yu, M. Whalen, G. Harvey, S. Orts-Escolano, R. Pandey, J. Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. ACM Transactions on Graphics (ToG), 38(6):1–19, 2019.

[20] M. Habermann, L. Liu, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. Real-time deep dynamic characters. ACM Transactions on Graphics (TOG), 40(4):1–16, 2021.

[21] M. Habermann, W. Xu, M. Zollhofer, G. Pons-Moll, and C. Theobalt. Deepcap: Monocular human performance capture using weak supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5052–5063, 2020.

[22] T. He, Y. Xu, S. Saito, S. Soatto, and T. Tung. Arch++: Animation-ready clothed human reconstruction revisited. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11046–11056, 2021.

[23] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung. Arch: Animatable reconstruction of clothed humans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3093–3102, 2020.

[24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, eds., 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[25] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. CoRR, abs/1609.02907, 2016.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 2012.

[27] Y. Kwon, D. Kim, D. Ceylan, and H. Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. Advances in Neural Information Processing Systems, 34, 2021.

[28] Y. Kwon, S. Petrangeli, D. Kim, H. Wang, E. Park, V. Swaminathan, and H. Fuchs. Rotationally-temporally consistent novel view synthesis of human performance video. In European Conference on Computer Vision, pp. 387–402. Springer, 2020.

[29] Y. Kwon, S. Petrangeli, D. Kim, H. Wang, E. Park, V. Swaminathan, and H. Fuchs. Rotationally-temporally consistent novel view synthesis of human performance video. In European Conference on Computer Vision, pp. 387–402. Springer, 2020.

[30] Y. Liao, K. Schwarz, L. Mescheder, and A. Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5871–5880, 2020.

[31] L. Liu, M. Habermann, V. Rudnev, K. Sarkar, J. Gu, and C. Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. ACM Transactions on Graphics (TOG), 40(6):1–16, 2021.

[32] L. Liu, W. Xu, M. Habermann, M. Zollhöfer, F. Bernard, H. Kim, W. Wang, and C. Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. arXiv preprint arXiv:2001.04947, 2020.

[33] L. Liu, W. Xu, M. Zollhoefer, H. Kim, F. Bernard, M. Habermann, W. Wang, and C. Theobalt. Neural rendering and reenactment of human actor videos. ACM Transactions on Graphics (TOG), 38(5):1–14, 2019.

[34] Y. Liu, S. Peng, L. Liu, Q. Wang, P. Wang, C. Theobalt, X. Zhou, and W. Wang. Neural rays for occlusion-aware image-based rendering. arXiv preprint arXiv:2107.13421, 2021.

[35] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh. Neural volumes: Learning dynamic renderable volumes from images. ACM Trans. Graph., 38(4):65:1–65:14, July 2019.

[36] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG), 34(6):1–16, 2015.

[37] R. Martin-Brualla, R. Pandey, S. Yang, P. Pidlypenskyi, J. Taylor, J. Valentin, S. Khamis, P. Davidson, A. Tkach, P. Lincoln, et al. Lookingood: Enhancing performance capture with real-time neural re-rendering. arXiv preprint arXiv:1811.05029, 2018.

[38] M. Meshry, D. B. Goldman, S. Khamis, H. Hoppe, R. Pandey, N. Snavely, and R. Martin-Brualla. Neural rerendering in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6878–6887, 2019.

[39] M. Mihajlovic, A. Bansal, M. Zollhoefer, S. Tang, and S. Saito. Keypoint-nerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In European Conference on Computer Vision, pp. 179–197. Springer, 2022.

[40] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In European conference on computer vision, pp. 405–421. Springer, 2020.

[41] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima. Siclope: Silhouette-based clothed people. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4480–4490, 2019.

[42] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2020.

[43] A. Noguchi, X. Sun, S. Lin, and T. Harada. Neural articulated radiance field. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5762–5772, 2021.

[44] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao. Animatable neural radiance fields for modeling dynamic human bodies. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14314–14323, 2021.

[45] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9054–9063, 2021.

[46] T. Pfaff, M. Fortunato, A. Sanchez-Gonzalez, and P. W. Battaglia. Learning mesh-based simulation with graph networks. arXiv preprint arXiv:2010.03409, 2020.

[47] O. Press, N. A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. ArXiv, abs/2108.12409, 2021.

[48] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10318–10327, 2021.

[49] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville. On the spectral bias of neural networks. In International Conference on Machine Learning, pp. 5301–5310. PMLR, 2019.

[50] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human

digitization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2304–2314, 2019.

[51] S. Saito, T. Simon, J. Saragih, and H. Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 84–93, 2020.

[52] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. Battaglia. Learning to simulate complex physics with graph networks. In International Conference on Machine Learning, pp. 8459–8468. PMLR, 2020.

[53] A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov, et al. Textured neural avatars. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2387–2397, 2019.

[54] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2437–2446, 2019.

[55] C. Stoll, J. Gall, E. De Aguiar, S. Thrun, and C. Theobalt. Video-based reconstruction of animatable human characters. ACM Transactions on Graphics (TOG), 29(6):1–10, 2010.

[56] S.-Y. Su, F. Yu, M. Zollhöfer, and H. Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. Advances in Neural Information Processing Systems, 34, 2021.

[57] Z. Su, L. Xu, Z. Zheng, T. Yu, Y. Liu, and L. Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In European Conference on Computer Vision, pp. 246–264. Springer, 2020.

[58] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG), 38(4):1–12, 2019.

[59] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhöfer, C. Lassner, and C. Theobalt. Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12959–12970, 2021.

[60] R. Tucker and N. Snavely. Single-view view synthesis with multiplane images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 551–560, 2020.

[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., Advances in Neural Information Processing Systems, vol. 30, 2017.

[62] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. Advances in Neural Information Processing Systems, 34, 2021.

[63] Q. Wang, Z. Wang, K. Genova, P. P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699, 2021.

[64] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. arXiv preprint arXiv:2201.04127, 2022.

[65] M. Wu, Y. Wang, Q. Hu, and J. Yu. Multi-view neural human rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1682–1691, 2020.

[66] W. Xian, J.-B. Huang, J. Kopf, and C. Kim. Space-time neural irradiance fields for free-viewpoint video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9421–9431, 2021.

[67] F. Xiang, Z. Xu, M. Hasan, Y. Hold-Geoffroy, K. Sunkavalli, and H. Su. Neutex: Neural texture mapping for volumetric neural rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7119–7128, 2021.

[68] T. Xu, Y. Fujita, and E. Matsumoto. Surface-aligned neural radiance fields for controllable 3d human synthesis. arXiv preprint arXiv:2201.01683, 2022.

[69] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. Advances in neural information processing systems, 29, 2016.

[70] G. Yao, H. Wu, Y. Yuan, and K. Zhou. Dd-nerf: Double-diffusion neural radiance field as a generalizable implicit body representation. arXiv preprint arXiv:2112.12390, 2021.

[71] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman. Volume rendering of neural implicit surfaces. In Thirty-Fifth Conference on Neural Information Processing Systems, 2021.

[72] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. Advances in Neural Information Processing Systems, 33, 2020.

[73] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelnerf: Neural radiance fields from one or few images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4578–4587, 2021.

[74] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018.

[75] F. Zhao, W. Yang, J. Zhang, P. Lin, Y. Zhang, J. Yu, and L. Xu. Humannerf: Generalizable neural human radiance field from sparse inputs. arXiv preprint arXiv:2112.02789, 2021.

[76] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. Deephuman: 3d human reconstruction from a single image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7739–7749, 2019.

[77] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely. Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817, 2018.

Qingzhe Gao is a a forth-year Ph.D. student in the School of Computer Science and Technology, Shandong University. His research interests include e 3D reconstruction, image segmentation, and neural rendering.

Yiming Wang is a four-year undergraduate student in the School of Electronics Engineering and Computer Science, Peking University. His research interests include neural scene representations and 3D reconstruction.

Libin Liu is an assistant professor at the School of Intelligence Science and Technology, Peking University. He received his Ph.D. degree in computer science from Tsinghua University. His research interests include character animation, physics-based simulation, motion control, and related areas in machine learning and robotics.

Lingjie Liu is a post-doctoral researcher at the Graphic, Vision & Video group of Max Planck Institute for Informatics in Saarbrücken, Germany. She received her BEng degree from the Huazhong University of Science and Technology in 2014 and PhD degree from the University of Hong Kong in 2019. Her research interests include 3D reconstruction, neural rendering and human performance capture. She has received Hong Kong PhD Fellowship Award (2014) and Lise Meitner Postdoctoral Fellowship Award (2019).

Christian Theobalt is a Professor of Computer Science and the head of the research group “Graphics, Vision, & Video” at the Max-Planck-Institute for Informatics, Saarbruecken, Germany. He is also a professor at Saarland University. His research lies on the boundary between Computer Vision and Computer Graphics. For instance, he works on 4D scene reconstruction, markerless motion and performance capture, machine learning for graphics and vision, and new sensors for 3D acquisition. Christian received several awards, for instance the Otto Hahn Medal of the Max-Planck Society (2007), the EUROGRAPHICS Young Researcher Award (2009), the German Pattern Recognition Award (2012), an ERC Starting Grant (2013), and an ERC Consolidator Grant (2017). In 2015, he was elected one of Germany’s top 40 innovators under 40 by the magazine Capital. He is a co-founder of theCaptury (www.thecaptury.com).

Baoquan Chen is a professor of Peking University with research interests in computer graphics, visualization and computer vision. He has published more than 200 papers and has received the Best Paper Awards for IEEE Visualization 2005 and ACM SIGGRAPH Asia 2022 and honorary mention for ACM SIGGRAPH 2022. He was elected IEEE Fellow in 2020 and was inducted to IEEE Visualization Academy in 2021. He served as conference chairs of ACM SIGGRAPH Asia 2014 and IEEE Visualization 2005, associate editor of ACM Transactions on Graphics and IEEE Transactions on Visualization and Graphics, and many other roles in professional organizations.