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

Qingzhe Gao, Yiming Wang, Libin Liu, Lingjie Liu, Christian Theobalt, Fellow, IEEE, and Baoquan Chen, Fellow, IEEE

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.

I. 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], [54], the production-quality human geometry and appearance models, and manual invention and corrections [10], [13], [56].

Recently, neural human representation and rendering algorithms based on the neural radiance fields (NeRF) [39] have demonstrated the ability to overcome the limitations of the traditional approaches. Some methods [6], [30], [43], [55] 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], [38], [74] proposed generalizable radiance fields for humans conditioned on input image features, inspired by the generalized neural representation for static scenes [5], [9], [33], [62], [72]. 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 deformations 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
We present a new way to predict the person- and pose-dependent properties. 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. These evaluations 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 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.

II. RELATED WORK

A. 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], [54] 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], [56] 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 [36], [37], [64] can compensate for geometric artifacts through the modification of the rendering pipeline. More recently, several works [22], [23], [40], [48], [49], [50], [75] 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.

B. Neural Representations for Human

Neural rendering techniques [28], [31], [52], [57], [64] have enabled neural networks to learn 3D object reconstruction from 2D images. Various 3D representations, such as 3D voxel-grid [28], [34], [53], [68], point clouds [1], [64], textured mesh [29], [32], [57], [66], and multi-plane images [14], [59], [76], 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 [39] represents scenes with implicit fields of density and color. To extract more accurate surfaces, some works [41], [61], [70], [71] employ the signed distance function (SDF) to represent geometry in a scene. Building on these representations, numerous studies [8], [16], [42], [44], [47], [55], [63], [65], [67] use neural representation to capture humans. However, optimizing for each novel video is time-consuming.

Generalizable neural representation methods [5], [9], [33], [62], [69], [72] address this issue by employing implicit fields conditioned on image features. Inspired by these works, some studies [7], [27], [38], [74] propose generalizable radiance fields for humans. These methods typically utilize a set of images of a given pose, along with their corresponding pose priors, to...
generate images of new views. However, when applied to novel poses, these methods may encounter difficulties due to the lack of corresponding images. As a result, their models are not truly animatable.

Leveraging the Skinned Multi-Person Linear model (SMPL) [35], several works [3, 6, 30, 43, 55] 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 [35] to animate the representation, while our method learns an additional residual deformation mapping to compensate for the deformation that cannot be modeled by inverse kinematic transformation. Furthermore, the representation of MPS-NeRF is based on NeRF [39], while our representation additionally disentangles geometry and appearance by formulating them as two separate template implicit functions in the canonical space. 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 (Section III-A) and then explain how we construct the geometry and appearance features used in such a representation (Section III-B). Based on our neural representation, we can synthesize novel view images for arbitrary human poses (Section III-C).

### III. 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 [35] 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 [61] 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 (Section III-A) and then explain how we construct the geometry and appearance features used in such a representation (Section III-B). Based on our neural representation, we can synthesize novel view images for arbitrary human poses (Section III-C).

#### A. Deformable Neural Human Representation

In order to represent different human identities and poses, we adopt SMPL [35] 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, \( \rho \), representing the body shape and pose of a person. This process can be written as

\[
v_{\rho} = \text{SMPL}(v, \rho, w_{v}),
\]

where \( v \in V \) represents a surface vertex and \( w_{v} \) is the skinning weight of \( v \). We consider a predefined pose \( \rho_0 \) as the canonical model of our framework. The canonical space is then defined as the corresponding 3D space containing this canonical model. For an arbitrary pose \( \rho \), 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.

### B. Geometry and Appearance Features

Our framework utilizes two latent variables, \( F_{\text{geo}} \) and \( F_{\text{app}} \), 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 \( F \) defined at a spacial point \( x \) consists of two components \( F = (u^x, z^x) \), where \( u^x \) represents the image features based on pixel alignment, and \( z^x \) are surface features computed based on the connectivity of the SMPL mesh.

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 \( U^c \) at multiple resolutions as

\[
U^c = \text{CNN}([I^c]).
\]

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

\[
u^x = \text{softmax}_c \left( \frac{Q(u^c) \cdot K(u^c) + B^c}{\sqrt{d}} \right) \cdot V(u^c),
\]

where \( Q(\cdot), K(\cdot), \) and \( V(\cdot) \) are the learnable query, key, and value embedding functions proposed in the original self-attention mechanism [60], 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 \( u^x \) 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 \( x \) 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 (3) to explicitly inform the self-attention mechanism of this occlusion. This mechanism is partially inspired by Attention with Linear Biases (ALBi) [46]. We find that it effectively mitigates the artifacts caused by occlusion in our experiments.

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, 45, 51]. 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 \( \rho_1 \), we compute a deformed mesh \( V_{\rho_1} \) using SMPL and project each vertex \( v \in V_{\rho_1} \) onto the input images, obtaining a set of occlusion-aware image features \( \{u^c\} \). Then, we convert these image features using the pose-aware GCN described above and compute surface features

\[
\{z^v\} = \text{GCN}(\{u^c\}, V_{\rho_1}).
\]

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

\[
\{z^v\} = \text{GCN}(\{z^v\}, V_{\rho}).
\]

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

$$z^x = \sum_k w_k \text{MLP} \left( z^k, x - v^k \right),$$

(6)

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

### 3) Implementation:

For a spatial point $x$, the latent variable $F_{\text{geo}} = (u^x_{\text{geo}}, z^x_{\text{geo}})$ and $F_{\text{app}} = (u^x_{\text{app}}, z^x_{\text{app}})$ are computed using the images features and the surface features defined above. To further enforce disentanglement of the features, we let the appearance features $F_{\text{app}}$ be independent of the driving pose $\rho$. This is achieved by computing $z^x_{\text{app}}$ using the canonical pose $\rho_0$ in (5), as depicted in Fig. 2.

### C. Pose-Driven Volume Rendering

Using the geometry and appearance features $F_{\text{geo}}$ and $F_{\text{app}}$ extracted from the input images, our framework can render new images from a novel viewpoint given an arbitrary driving pose $\rho$. We employ NeuS [61], an SDF-based differential renderer, to synthesize those images. NeuS predicts the color of each pixel by accumulating the radiance along the camera ray $r$ passing through the pixel. This computation can be discretized using a series of spatial points $\{x_i\}$ sampled along $r$. Specifically, NeuS computes

$$\tilde{C}(r) = \sum_{j=1}^{n} T_i \alpha_i c_i,$$

(7)

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

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

(8)

and $\phi(x) = (1 + e^{-k|x|})^{-1}$. 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 $x_i$. Our framework handles every sample point in the same way, so we omit the subscript $i$ in the rest of this section for simplicity.

1) Pose-Driven Deformation Field: Given a driving pose $\rho$, we transform each sample point $x_i$ into the canonical space and evaluate $c$ and $s$ based on the canonical position $\hat{x}$. This mechanism is inspired by recent studies [30], [58], which have shown its efficiency in modeling dynamic scenes and human poses. We define the deformation mapping $\Phi$ using the inverse skinning transformation [23]. As suggested by [30], 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 $\hat{x}$ of a sample point $x$ is thus computed as

$$\hat{x} = \Phi(x, \rho) + \delta \Phi(x, \rho).$$

(9)

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

$$\Phi(x, \rho) = \sum_{j=1}^{n_j} w_j \cdot (R_j(x - \delta v) + t_j),$$

(10)

where $R_j$ and $t_j$ represent the rotation and translation that transform joint $j$ from pose $\rho$ back to the canonical pose $\rho_0$, $w_j \in w_v$ is the corresponding skinning weight of the surface point $v$ that is the nearest to $x$, 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 $\delta v$ is leveraged to compensate for the deformation caused by the change of body shape. Specifically,

$$\delta v = \text{SMPL}(v, \beta(\rho), w_v) - \text{SMPL}(v, \beta(\rho_0), w_v),$$

(11)

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

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

$$\delta \Phi(x, \rho) = \text{MLP}(F_{\text{geo}}, d_j).$$

(12)

### 2) SDF-Based Volume Rendering:

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

$$s = \mathcal{J}(\hat{x})$$

(13)

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

### D. Training

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

$$L = \frac{1}{nm} \sum_r ||\tilde{C}(r) - C(r)||_1 + \lambda_1 \sum_r \text{BCE}(\tilde{M}_r, M_r)$$

$$+ \lambda_2 \frac{1}{nm} \sum_r (\|n_x\|_2 - 1)^2 + \lambda_3 \text{LPIPS}(\tilde{C}(P), C(P)),$$

(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 $\tilde{M}_r = \sum_{i=1}^{n_j} T_i \alpha_i$ 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 [73], 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.

IV. 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 Section III-B and then train all the networks jointly. The learning rate is first linearly warmed up from 0 to $5 \times 10^{-4}$ in the first 2 k iterations and then is controlled by the cosine decay schedule. In the second stage, we freeze the SDF network after training for 50 K iterations. We use the sampling strategy proposed in NeuS [61], and the numbers of the coarse and fine sampling are 32 and 32 respectively. Following Neural Actor [30], 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 32 G GPUs and sample 1024 rays per batch per GPU for 80 K iterations without LPIPS loss. Then, we randomly select two additional patches $P$ (size $24 \times 24$) 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 $\lambda_1, \lambda_2$ both to 0.1, and $\lambda_3$ 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 (Section IV-A) and pose control animation (Section IV-B).

A. 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) 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) [38], Neural Human Performer (NHP) [27] and Neural Body (NB) [44]. We also compare our method with Neural Body (NB) [44] 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 Table I 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 Table I, 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 Table I shows, our model outperforms MPS, NHP significantly on PSNR and SSIM. Our method also achieve comparable results to KN. The visualizations of the results are shown in Fig. 6.

4) Seen Poses of Seen Subjects: To demonstrate the superiority 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. Table I 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 clothes and hair, even with sparse input images. Additionally, our occlusion-aware self-attention mechanism can effectively handle mesh points that are occluded in
TABLE I

| Setting | Identity generation | Cross-dataset | Pose generation | Seen subjects |
|---------|---------------------|---------------|----------------|---------------|
| Method  | PSNR\(\uparrow\) | SSIM\(\uparrow\) | PSNR\(\uparrow\) | SSIM\(\uparrow\) | PSNR\(\uparrow\) | SSIM\(\uparrow\) |
| NB      | -       | -              | -              | -              | 28.56       | 0.943          |
| NHP     | 24.85   | 0.908          | 23.16          | 0.869          | 26.19       | 0.915          |
| KN      | 24.92   | 0.910          | 22.31          | 0.861          | 27.64       | 0.933          |
| MPS     | 22.99   | 0.877          | 22.38          | 0.842          | 24.66       | 0.880          |
| Ours    | 25.14   | 0.914          | 24.19          | 0.886          | 27.36       | 0.929          |

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) [38] 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) [44], in the task of seen poses for seen subjects as an upper-bound baseline. Our method achieves comparable performance with this benchmark.

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.

**B. 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 Section IV-A. 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) [43], Neural Actor (NA) [30] and MPS-NeRF (MPS) [17] on the ZJU-MoCap dataset. Neural Human Performer (NHP) [27] and Keypoint NeRF (KN) [38] are not animatable human models and thus cannot generate images with user’s pose control. The split for the ZJU-MoCap dataset is the same as that described in Section IV-A.

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.

Table II and Fig. 7 demonstrate that our method significantly outperforms MPS-NeRF quantitatively and qualitatively.
Fig. 4. Qualitative comparison of cross-dataset generalization on novel view synthesis. All methods are trained on the ZJU-MoCap [44] 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.

| 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 |

Our method achieves the best performance in two metrics, compared to three person-specific animatable human models, NeuralBody (NB) [44], animatableNeRF(AN) [43], and neural actor (NA) [30] and MPS-NeRF (MPS) [17].

TABLE II
QUANTITATIVE COMPARISON OF THE ANIMATION TASK

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).

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]...
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) [38] and MPS-NeRF (MPS) [17], in terms of synthesized wrinkles and appearance details.

Fig. 7. Qualitative comparison of the Animation task on the ZJU-MoCap dataset. We compared with three person-specific methods, Neural Body (NB) [44], Animatable Nerf (AN) [43], Neural Actor (NA) [30] 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.

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 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.
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 [44] 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.

C. 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.

D. 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 Sections IV-A and IV-B are used for these two tasks. The results are shown in Table III. We also provide visual result in Fig. 10 to further demonstrate the significance of different components.

Table III

|               | Unseen subjects | Animation |
|---------------|----------------|-----------|
|               | PSNR† | SSIM† | PSNR† | SSIM† |
| w/o GCN       | 24.27  | 0.891 | 22.81 | 0.872 |
| w/o tex       | 23.79  | 0.890 | 22.34 | 0.875 |
| w/o texapp    | 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 |
| NeRF Render   | 24.52  | 0.905 | 22.90 | 0.881 |
| Full model    | 25.14  | 0.914 | 23.18 | 0.886 |

*GCN: The Occlusion-aware self-attention mechanism, Sep: Separation of geometry and appearance features.

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.

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 Section III-B2, which leads to artifacts both in shape and appearance as showed in Fig. 10. Additionally, the drop in performance shown in Table III 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 $u_{geo}$); 2) removing the image features from the appearance features (w/o $u_{app}$); 3) disabling the occlusion-aware self-attention mechanism (w/o Occ) by letting $B^* = 0$ in (3). As visible in Fig. 10, the predicted results have significant geometric deviations when lacking the geometry feature $u_{geo}$, mainly reflected in the lack of clothing geometry and errors in the shape of the person. Similarly, when lacking the appearance feature $u_{app}$, 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 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 Fig. 10, an erroneous human body shape is depicted, and clothing is mistakenly reconstructed on the hand. As shown in Table III, our disentangled features achieve better image quality in terms of all of the metrics. In Table IV, 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. As shown in Table III, NeuS rendering also achieves better image quality in terms of all of the metrics.

Additional visual ablation study results are provided for the cross-dataset animation evaluation. As in Section IV-B, we train all methods on the ZJU-MoCap dataset and directly test them on the DeepCap [21] and DynaCap [20] datasets. As demonstrated in Fig. 12, the modules introduced by our method play a crucial role in this task. With all the proposed modules, our method can achieve satisfactory performance. However, without the GCN or the separation of geometry and appearance features, the performance of our method, especially in terms of generalization and animation capabilities, dropped significantly. The synthesized appearance and geometry of the person may differ noticeably from the input. When the occlusion-aware self-attention mechanism is not employed, the results suffer from poor appearance, showing visual artifacts such as incorrect colors and inaccurate facial features and body structures. The primary reason for this issue is that the model tends to utilize features from all input view images, even when certain views are occluded and should ideally be disregarded. This negatively affects the model’s ability to generate accurate appearances when input views present different occlusion patterns from those in the training data. In contrast, our occlusion-aware self-attention mechanism guides the model to pay greater attention to unoccluded images, thereby ensuring improved image quality.

| 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 |

The performance of our method degrades slightly when given fewer input views.
V. DISCUSSION

A. Limitation of Pose-Dependent Properties

We extract image features from both the spatial and surface points of the SMPL model to more effectively infer person- and pose-dependent properties. Our GCN integrates local pose information by the edges of the posed SMPL as its edge features, enabling the generalization of pose-aware surface features. However, determining the dynamic changes in appearance and geometry with different poses given only four static images as input is still challenging. The performance of our pose-dependent appearance and geometry in this task is not as strong as in novel view synthesis.

Two potential future directions could be pursued to enhance this capability. First, the training phase could make use of a larger dataset to explore pose-dependent information. This might require integrating a generative model to generate pose-driven information and using real or synthesized data to learn temporal consistency. Second, rather than using still images, the testing phase could benefit from directly using multi-view videos. Such an approach would enable the model to learn specific pose-dependent appearance and geometry from the videos.

B. Compared to Per-Scene Optimization

Neural Actor (NA) [30] 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 insufficient coverage and occlusion. Moreover, Animatable NeRF (AN) [43] and Neural Body (NB) [44] 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.

C. Other Limitation and Future Work

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. 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 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.

VI. CONCLUSION

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.

REFERENCES

[1] K.-A. Aliev, A. Sevastopolsky, M. Kolos, D. Ulyanov, and V. Lempitsky, “Neural point-based graphics,” in Proc. Eur. Conf. Comput. Vis., Springer, 2020, pp. 696–712.
[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, “Video based reconstruction of 3D people models,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8387–8397.
[3] M. Habermann, L. Liu, W. Xu, G. Pons-Moll, M. Zollhoefer, and C. Theobalt, “HDHumans: A hybrid approach for high-fidelity digital humans,” in Proc. ACM SIGGRAPH / Eurographics Symp. Comput. Animation, 2023.
[4] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel, “Free-viewpoint video of human actors,” ACM Trans. Graph., vol. 22, no. 3, pp. 569–577, 2003.
[5] A. Chen et al., “MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 14124–14133.
[6] J. Chen et al., “Animatable neural radiance fields from monocular RGB videos,” 2021, arXiv:2106.13929.
[7] M. Chen et al., “Geometry-guided progressive NeRF for generalizable and efficient neural human rendering,” in Proc. 17th Eur. Conf. Comput. Vis., Tel Aviv, Israel, Springer, 2022, pp. 222–239.
[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,” 2021, arXiv:2106.04477.
[9] J. Chibane, A. Bansal, V. Lazova, and G. Pons-Moll, “Stereo radiance fields (SRFs): Learning view synthesis for sparse views of novel scenes,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 7911–7920.
[10] A. Collet et al., “High-quality streamable free-viewpoint video,” ACM Trans. Graph., vol. 34, no. 4, pp. 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 Proc. ACM SIGGRAPH Papers, New York, NY, USA, 2008, pp. 1–10.
[12] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagat, “Acquiring the reflectance field of a human face,” in Proc. 27th Annu. Conf. Conf. Comput. Graph. Interactive Techn., 2000, pp. 145–156.
[13] M. Dou et al., “Fusion4D: Real-time performance capture of challenging scenes,” ACM Trans. Graph., vol. 35, no. 4, pp. 1–13, 2016.
[14] J. Flynn et al., “DeepView: View synthesis with learned gradient descent,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2367–2376.
[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 Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1746–1753.
[16] C. Gao, Y. Shih, W.-S. Lai, C.-K. Liang, and J.-B. Huang, “Portrait neural radiance fields from a single image,” 2020, arXiv:2012.05903.
[17] X. Gao, J. Yang, J. Kim, S. Peng, Z. Liu, and X. Tong, “MPS-NeRF: Generalizable 3D human rendering from multiview images,” IEEE Trans. Pattern Anal. Mach. Intell., to be published, doi: 10.1109/TPAMI.2022.3205910.
[18] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, “Implicit geometric regularization for learning shapes,” 2020, arXiv:2002.10099.

[19] K. Guo et al., “The relightables: Volumetric performance capture of humans with realistic relighting,” ACM Trans. Graph., vol. 38, no. 6, pp. 1–19, 2019.

[20] M. Habermann, L. Liu, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, “Real-time dynamic deep characters,” ACM Trans. Graph., vol. 40, no. 4, pp. 1–16, 2021.

[21] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, “DeepCap: Monocular human performance capture using weak supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5052–5063.

[22] T. He, Y. Xu, S. Saito, S. Soatto, and T. Tung, “ARCH: Animation-ready clothed human reconstruction revisited,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 11046–11056.

[23] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung, “ARCH: Animatable reconstruction of clothed humans,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3093–3102.

[24] D. P. Kingma and M. Welling, “Semi-supervised classification with graph convolutional networks,” 2016, arXiv:1609.02907.

[25] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” 2016.

[26] A. Noguchi, X. Sun, S. Lin, and T. Harada, “Neural articulated radiance field,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 10318–10327.

[27] N. Rahaman et al., “On the spectral bias of neural networks,” in Proc. Int. Conf. Mach. Learn., PMLR, 2019, pp. 5301–5310.

[28] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, “PiPfH: Pixel-aligned implicit function for high-resolution clothed human digitization,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 2304–2314.

[29] S. Saito, T. Simon, J. Saragih, and H. Joo, “PiPfHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 84–93.

[30] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. Battaglia, “Learning to simulate complex physics with graph networks,” in Proc. Int. Conf. Mach. Learn., PMLR, 2020, pp. 8459–8468.

[31] A. Shysheya et al., “Neural avataran,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2387–2397.

[32] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhoefer, “DeepVoxels: Learning persistent 3D feature embeddings,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2437–2446.

[33] C. Stoll, J. Gall, E. De Aguiar, S. Thrun, and C. Theobalt, “Video-based reconstruction of animatable human characters,” ACM Trans. Graph., vol. 29, no. 6, pp. 1–10, 2010.

[34] S.-Y. Su, F. Yu, M. Zollhöfer, and H. Hodini, “A-NeRF: Articulated neural radiance fields for learning human shape, appearance, and pose,” in Proc. Adv. Neural Inf. Process. Syst., 2021, pp. 12278–12291.

[35] 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 Proc. Eur. Conf. Comput. Vis., Springer, 2020, pp. 246–264.

[36] J. Thies, M. Zollhöfer, and M. Nießner, “Deferred neural rendering: Image synthesis using neural textures,” ACM Trans. Graph., vol. 38, no. 4, pp. 1–12, 2019.

[37] 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 Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 12959–12970.

[38] R. Tucker and N. Snavely, “Single-view synthesis with multiplane images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 551–560.

[39] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 6000–6010.

[40] 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,” in Proc. Adv. Neural Inf. Process. Syst., 2021, pp. 27171–27183.

[41] Q. Wang et al., “IBRNet: Learning multi-view image-based rendering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 4690–4699.

[42] C.-Y. Wang, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman, “HumanNeRF: Free-viewpoint rendering of moving people from a monocular video,” 2022, arXiv:2201.04127.

[43] M. Wu, Y. Wang, Q. Hu, and J. Yu, “Multi-view human rendering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 1682–1691.

[44] W. Xian, J.-B. Huang, J. Kopf, and C. Kim, “Space-time neural irradiance fields for free-viewpoint video,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 9281–9291.

[45] F. Xiang, Z. Xu, M. Hasan, Y. Hold-Geoffroy, K. Sunkavalli, and H. Su, “NeuTxt: Neural texture mapping for volumetric neural rendering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 7119–7128.

[46] T. Xu, Y. Fujita, and E. Matsumoto, “Surface-aligned neural radiance fields for controllable 3D human synthesis,” 2022, arXiv:2203.04783.

[47] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee, “Perspective transformer nets: Learning single-view 3D object reconstruction without 3D supervision,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 1704–1712.
[69] G. Yao, H. Wu, Y. Yuan, and K. Zhou, “DD-NeRF: Double-diffusion neural radiance field as a generalizable implicit body representation,” 2021, arXiv:2112.12390.

[70] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, “Volume rendering of neural implicit surfaces,” in Proc. 35th Conf. Neural Inf. Process. Syst., 2021, pp. 4805–4815.

[71] L. Yariv et al., “Multiview neural surface reconstruction by disentangling geometry and appearance,” in Proc. Adv. Neural Inf. Process. Syst., 2020, Art. no. 210.

[72] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, “PixelNeRF: Neural radiance fields from one or few images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 4578–4587.

[73] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 586–595.

[74] F. Zhao et al., “HumanNeRF: Generalizable neural human radiance field from sparse inputs,” 2021, arXiv:2112.02789.

[75] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, “DeepHuman: 3D human reconstruction from a single image,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 7739–7749.

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

Qingzhe Gao is currently working toward the fourth-year PhD degree with the School of Computer Science and Technology, Shandong University. His research interests include 3D reconstruction, image segmentation, and neural rendering.

Yiming Wang is currently working toward the fourth-year undergraduate degree with the School of Electronics Engineering and Computer Science, Peking University. His research interests include neural scene representations and 3D reconstruction.

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

Lingjie Liu received the BEng degree from the Huazhong University of Science and Technology, in 2014, and the PhD degree from the University of Hong Kong, in 2019. She is a post-doctoral researcher with the Graphic, Vision & Video Group of Max Planck Institute for Informatics, Saarbrücken, Germany. 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” with the Max-Planck-Institute for Informatics, Saarbruecken, Germany. He is also a professor with Saarland University. His research lies on the boundary between computer vision and computer graphics. For instance, he works on 4D scene reconstruction, marker-less motion and performance capture, machine learning for graphics and vision, and new sensors for 3D acquisition. He 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 the Captury (www.thecaptury.com).

Baoquan Chen (Fellow, IEEE) 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 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.