MPS-NeRF: Generalizable 3D Human Rendering from Multiview Images

Xiangjun Gao[1*] Jiaolong Yang[2] Jongyoo Kim[2] Sida Peng[3] Zicheng Liu[4] Xin Tong[2]
[1]Beijing Institute of Technology [2]Microsoft Research Asia [3]Zhejiang University [4]Microsoft

Figure 1. Given a sparse set of multiview ***images*** of an ***unseen*** person, our method is able to synthesize novel views of the person and animate it with novel poses. It works in a feed-forward fashion without any optimization.

Abstract

There has been rapid progress recently on 3D human rendering, including novel view synthesis and pose animation, based on the advances of neural radiance fields (NeRF). However, most existing methods focus on person-specific training and their training typically requires multi-view videos. This paper deals with a new challenging task – rendering novel views and novel poses for a person unseen in training, using only multiview images as input. For this task, we propose a simple yet effective method to train a generalizable NeRF with multiview images as conditional input. The key ingredient is a dedicated representation combining a canonical NeRF and a volume deformation scheme. Using a canonical space enables our method to learn shared properties of human and easily generalize to different people. Volume deformation is used to connect the canonical space with input and target images and query image features for radiance and density prediction. We leverage the parametric 3D human model fitted on the input images to derive the deformation, which works quite well in practice when combined with our canonical NeRF. The experiments on both real and synthetic data with the novel view synthesis and pose animation tasks collectively demonstrate the efficacy of our method.

1. Introduction

Free-view human character rendering and animation has numerous applications in avatar creation, telepresence, movie production, among others. While traditional methods [6, 26, 32] use dense multiview camera rigs or depth sensors to accomplish this task, recent neural rendering approaches [30, 36, 37, 46] have shown that free-view rendering and animation can be achieved using sparse color cameras, which could significantly reduce the device setup and capture cost. In particular, promising results have been shown by methods [30, 36, 37] that are based on the neural radiance field (NeRF) [29] representation. With NeRF, the generated multiview images not only have decent quality, but also enjoy high 3D consistency by virtue of an explicit, physics-based rendering process, thus enabling visually pleasing free-view and animated human video creation.

However, due to the high complexity of human motion and appearance, existing methods [30, 36, 37] are typically trained in a person-specific setup, *i.e.*, one model is trained for one single person, to obtain best rendering quality. Such a setup is clearly not scalable, as rendering any new character would require a tedious model training process. Moreover, these methods needs multiview *video* for training in order to handle different human poses. This requirement further restricts the practical usefulness of these methods.

This paper deals with a challenging but more practical problem setup – *training a model that can render unseen persons directly in a feed-forward manner, and using only*

*This work was done when X. Gao was an intern at MSRA.

still images captured at sparse viewpoints as input, as illustrated in Fig. 1. A simple yet effective method is proposed for this task. Our key insight is that human bodies share a similar geometric structure, which can be leveraged to learn a rendering model transferable to new subjects. Prior studies [43, 49] have shown that using images as conditional input, it is possible to train a generic NeRF for common objects, especially for those in a same category. Although these methods focus on rigid objects or static scenes, the spirit of generalization applies to our problem as well.

To leverage the shared human body structure and attain better generalization ability, we learn a radiance field in a canonical space where different people with various poses are well aligned. We use a pose-aligned canonical space defined by a 3D human body parametric model – SMPL [24] in this work – to achieve this. We apply pre-defined volumetric deformations to connect the canonical space with the target space (for the target image) and the observation space (for the input images). The deformation is simply defined by propagating the skinning weights defined on the SMPL human surface to the 3D volume and applying standard linear blend skinning. The input to NeRF is a 3D point in the canonical space and corresponding image features retrieved on the input images, similar to [43, 49].

Our method is named Multi-Person Skinning NeRF, or MPS-NeRF, for the goal to handle generic person and its skinning-based deformation derived by SMPL [24]. It allows for not only novel view synthesis but also plausible novel pose animation, thanks to the animatable nature of our deformation-based NeRF representation and its seamless combination with the parametric human model. We evaluate our method on the Human3.6M dataset [13] and another synthetic dataset created by the body models from the THuman [51] dataset. The results on both novel-view and novel-pose synthesis tasks have demonstrated the effectiveness of our method.

The contributions of this paper can be summarized as follows. First, we explore a challenging task of novel-view and novel-pose rendering for unseen persons given only sparse multiview images. To our knowledge, this task is not handled by any existing method. We propose a novel and simple method for this task, which defines a shared canonical space to facilitate generalizable NeRF training and leverages a parametric human model to derive volume deformation. Albeit its conceptual simplicity, the method works surprisingly well as demonstrated on both real and synthetic datasets. The rendering quality is even on par with or better than previous person-specific models. Our method could serve as a strong baseline model for future works towards generalizable 3D human rendering.

2. Related Work

Neural 3D representations and rendering. Recently, neural rendering based on implicit function has emerged as an effective way of novel view synthesis. In this context, various 3D representations were studied such as occupancy networks [28], SDF [33], Implicit field [5], and SRN [41]. In particular, NeRF [29] achieves photorealistic synthesis by modeling the 3D scene as a continuous 5D function. Though NeRF yields impressive quality, it can only deal with static scenes. Several studies try to handle non-rigid objects by employing deformation to map the observation space to the canonical space [34, 38, 42]. Some works are devoted to generalizing NeRF with images as conditional input [43, 49]. This paper presents a NeRF-based method dedicated to generalizable human rendering.

Human performance capture and 3D reconstruction. 3D human reconstruction has been widely studied in the literature. Earlier works are based on 2D keypoints [9], multi-view consistency [7], or depth maps [3, 45]. Later, parametric 3D human models [14, 24] have been frequently deployed [4, 17, 18, 20, 35], where optimal parameters which deforms the human model to match the input data are fitted. However, the parametric 3D model has the limited capacity, leading to less generalizability to clothed people. Recently, implicit field based methods were proposed to capture detailed surface geometry from one or few views [39, 40]. Yet, the reconstructed geometry is difficult to be deformed into novel poses. ARCH [12] proposes to combine the parametric model with the implicit field to predict rigged 3D clothed humans. The model yields the detailed geometry as well as the pixel-aligned colors. However, these approaches require large-scale ground-truth meshes of clothed persons, while ours is trained on sparse views with coarse SMPL meshes.

Human rendering. A few studies have been proposed for human image synthesis with the NeRF framework. Generally, human-prior information such as a skeleton or a parametric model was adopted to mitigate the large deformation of human bodies. NeuralBody [37] adopts SMPL [24] and uses per-vertex latent code which is used to generate a continuous latent code volume. By deforming the SMPL mesh, any pose and view can be rendered through the proposed framework. In Neural Actor [23] and Ani-NeRF [36], SMPL is deployed to unwrap the 3D space into a canonical pose, where pose-dependent residual deformation is also considered. H-NeRF [47] combines SDF with NeRF for rendering and temporal reconstruction. NHP [19] purposes a generalizable model but they still need sparse videos as input. Though these studies can synthesize plausible images, all of them are person-specific models or conditioned on sparse videos. In contrast, we aim at learning a person-agnostic model generalizable to unseen persons using only still images.

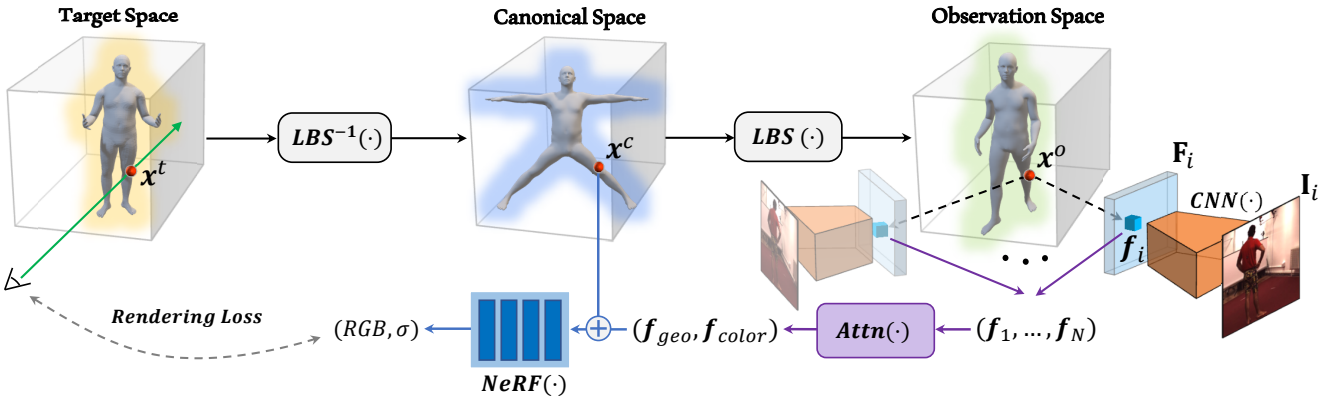

Figure 2. **System Overview.** To render the target image, we first cast rays and sample points in target space. Then, the sampled point $\mathbf{x}^t$ is deformed to canonical space by inverse linear blend skinning algorithm and the deformed point $\mathbf{x}^c$ is used as the input to NeRF. The point $\mathbf{x}^c$ is further deformed to the observation space and projected onto the input images to retrieve the multiview features which are fused by self-attention blocks. The fused features are also used for density and radiance prediction.

3. Method

Ours goal is to train a generic model which can directly synthesize a human image with novel viewpoint and/or pose in a feed-forward fashion using only multiview images as conditional input. For the input multiview images, we assume the calibration parameters and the human region masks are known. We also assume the parameters of a 3D human parametric models fitted to the multiview images are given. In this work, we use SMPL [24] as our parametric model. The SMPL parameters can be obtained by applying existing fitting methods based on keypoints, silhouette, multiview consistency or any other possible means [4, 14, 22, 50].

3.1. Overview

The overall framework of our method is presented in Fig. 2. The input consists of a sparse set of observed multiview images $\{\mathbf{I}_i\}$ associated with their camera parameters $\{\mathbf{v}_i^o\}$ and fitted SMPL model shape and pose parameters $\boldsymbol{\beta}, \mathbf{p}^o$, and the camera and pose parameters $\mathbf{v}^t, \mathbf{p}^t$ of the target view. The output is the human image rendered at $\mathbf{v}^t$ with pose $\mathbf{p}^t$. Note that if $\mathbf{p}^t = \mathbf{p}^o$, the task is known as novel view synthesis; otherwise we synthesize novel poses as in an animation task.

To render the target image, we follow recent works [30, 36, 37] and base our rendering scheme on NeRF [29], which is a compact yet powerful representation for neural rendering. To compute the color of a pixel on the target image, we cast a ray to the 3D space which passes the camera center and the pixel. We then sample points along the ray, predict their densities and colors using a neural network, and accumulate the colors following the volumetric rendering scheme [27, 29].

The key lies in how we compute the density σ and color $\mathbf{c}$ for a point $\mathbf{x}^t$ in the target space. Inspired by [43, 49], we use image features as conditional input to the network for density and color prediction. A straightforward solution would be projecting $\mathbf{x}^t$ onto the input image planes to get features and then combining the features with $\mathbf{x}^t$ as the input to NeRF. However, the complex, non-rigid human motions make the radiance field prediction difficult to learn and generalize to different people with various poses. Besides, this method can only be applied to the novel view synthesis task. For novel pose synthesis, it lacks a way to handle pose change for feature retrieval, nor does it offer a mechanism to drive an animation with desired pose change.

Our method defines a canonical space where 3D human bodies are aligned based on the SMPL model. To predict the density and color, we deform $\mathbf{x}^t$ to this canonical space and use the deformed point as the input to the network instead of $\mathbf{x}^t$. The point is further deformed to the observation space and projected onto the input images to retrieve the multiview features which are also used for radiance prediction.

3.2. Canonical Space and Volume Deformation

The canonical space is a pose-aligned space for different human bodies. Its coordinate system is defined the same as the SMPL model. The pose shown in Fig. 2 is defined as our canonical pose, and we found this using this canonical pose leads to better geometry and rendering results for the leg region compared to the T-shape rest pose in SMPL. The notion of a canonical space has been used in previous deformable NeRF schemes such as [34, 36, 38]. Compared to these methods, our main motivation is to learn a NeRF for different subjects in one shared space to foster generalization, whereas their goal is to model the dynamics of one single scene or object.

We apply two deformation fields to connect the canonical space with the observation space and target space, respectively. The former is used to retrieve image features from the input images for radiance prediction, while the lat-

ter is for rendering the output image. We formulate our deformation as an extended skinning process by propagating the surface skinning weights of SMPL model to the volume. The SMPL model provides a pre-defined skinning weight vector $\boldsymbol{\omega} \in \mathbb{R}^{24}$ for each vertex on the body surface for skinning. Following [2,12,36], for each point in the volume, we assign the skinning weights of its closest body vertex. Note that nothing prevents us from using multiple nearest neighbors on body surface to build a fuzzy association similar to [48]. In practice, we did not observe significant improvements in our experiments and thus opted for the simplest formulation. With the assigned skinning weights, volume deformation can be calculated by the linear blend skinning (LBS) algorithm [21]. Specifically, for a point $\mathbf{x}^c$ in the canonical space, the function $\rho^{c\to o}(\cdot)$ deforming it to the observation space can be written as

$$\mathbf{x}^o = \rho^{c\to o}(\mathbf{x}^c) = LBS\big(\mathbf{x}^c, \boldsymbol{\omega}(\mathbf{x}^c)\big) = \Big(\sum_{j=1}^{24} \omega_j(\mathbf{x}^c)\mathbf{T}_j\Big)\mathbf{x}^c, \tag{1}$$

where $\mathbf{T}_j \in \mathrm{SE}(3), j = 1, \ldots, 24$ are the known rigid transformations of the body joints. Deforming a point $\mathbf{x}^t$ in the target space to the canonical space requires an inverse LBS function:

$$\mathbf{x}^c = \rho^{t\to c}(\mathbf{x}^t) = LBS^{-1}\big(\mathbf{x}^t, \boldsymbol{\omega}(\mathbf{x}^t)\big) = \Big(\sum_{j=1}^{24} \omega_j(\mathbf{x}^t)\mathbf{T}_j\Big)^{-1}\mathbf{x}^t. \tag{2}$$

Discussion. The deformation so-obtained provides an approximation of the true correspondence field when the target and observed poses are different (*i.e.*, $\mathbf{p}^t \neq \mathbf{p}^o$). While it is possible to learn a refined deformation field in some constrained situation (*e.g.*, a residual skinning weight field is learned in the person-specific model of [36]), this task is extremely challenging for the inverse LBS. This is because a point $\mathbf{x}^t$ in the 3D space could be on arbitrary body part for different target poses. In [36], a per-frame latent code is jointly learned to alleviate this issue, and a test-time optimization is further needed to optimize the latent code for a novel pose. However, learning such a deformation network that is generalizable to any unseen person under arbitrary pose is prohibitively difficult. Our finding is that the simple deformation scheme works quite well in our method, where an image-conditioned NeRF is trained with this deformation scheme and will learn to adapt to it to ensure best rendering quality.

3.3. Image-Conditioned Rendering

The input to our canonical NeRF is a point in the canonical space and features extracted from the input images. To get the image features, we first apply a CNN on the images $\{\mathbf{I}_i\}$ to extract feature maps $\{\mathbf{F}_i\}$. For a canonical space point $\mathbf{x}^c$, we deform it to the observation space as $\mathbf{x}^o$ and project it onto the input image planes. A set of features is extracted using bilinear interpolation. We also sample RGB colors on the images and append them to the extracted features to form the final image features $\{\mathbf{f}_i\}$:

$$\mathbf{f}_i = \big[\mathbf{F}_i\big(\Pi(\mathbf{x}^o, \mathbf{v}_i^o)\big), \mathbf{I}_i\big(\Pi(\mathbf{x}^o, \mathbf{v}_i^o)\big)\big], \tag{3}$$

where $\Pi(\cdot)$ denotes the 3D-2D projection function.

Next, we fuse the extracted multiview features for the subsequent radiance field prediction. Human images contain severe self-occlusions, thus feature fusion is not trivial especially under very few views (*e.g.*, three or four) with large view angle differences. In our method, we employ the attention mechanism in Transformers [8, 44] for effective feature fusion. Two self-attention blocks are applied to generate two features, one for geometry prediction and another for color:

$$\mathbf{f}_{geo} = Attn_{geo}\big(\{\mathbf{f}_i\}\big),\ \mathbf{f}_{color} = Attn_{color}\big(\{\mathbf{f}_i\}\big). \tag{4}$$

The fused features, together with the coordinate of canonical space point, are then fed into an MLP network to predict density σ and color $\mathbf{c}$. Our network is adapted from [29] with a few tweaks. Notably, we replace the view direction input with our fused color feature for the final color prediction:

$$\sigma, \mathbf{h} = MLP_1(\gamma(\mathbf{x}^c), \mathbf{f}_{geo}), \tag{5}$$

$$\mathbf{c} = MLP_2(\mathbf{h}, \mathbf{f}_{color}), \tag{6}$$

where MLP_1 and MLP_2 are partial MLP layers with same structures as [29] except for the first layer, and $\gamma(\cdot)$ is the positional encoding function.

Finally, the volumetric rendering procedure [15, 29] for a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ in the target space can be written as:

$$\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\rho^{t\to c}(\mathbf{r}(t)))\mathbf{c}(\rho^{t\to c}(\mathbf{r}(t)))dt, \tag{7}$$

$$T(t) = \exp\Big(-\int_{t_n}^{t} \sigma(\rho^{t\to c}(\mathbf{r}(s)))ds\Big), \tag{8}$$

where the integrals can be estimated by using point samples along the ray.

3.4. Training Loss

Our method contains three submodules that need to be trained: the CNN for image feature encoding, the attention blocks for feature fusion, and the MLP for canonical NeRF. All these networks are trained in an end-to-end manner using images from a collection of people in the training set. After training, our method can be applied on new subjects for novel view and novel pose synthesis. The following loss functions are used for training.

Color loss. Given ground-truth target images, we apply the color loss to supervise the training, defined as:

$$L_{color} = \sum_{\mathbf{r}\in\Omega} \|\mathbf{C}(\mathbf{r}) - \mathbf{C}^*(\mathbf{r})\|_2^2, \tag{9}$$

where Ω is the ray collections for pixels on the target image, and $\mathbf{C}^*(\mathbf{r})$ is the ground-truth color.

Mask loss. We also leverage the mask labels of the ground-truth images to train the radiance fields. The mask loss is defined as:

$$L_{mask} = \sum_{\mathbf{r}\in\Omega} \|M(\mathbf{r}) - M^*(\mathbf{r})\|_2^2, \tag{10}$$

where $M(\mathbf{r})$ is the accumulated volume density and $M^*(\mathbf{r})$ is the binary mask label.

Smoothness loss. We incorporate a normal smoothness prior to encourage smooth geometry. Similar to [31], we enforce the normal vector of a canonical point $\mathbf{x}^c$ and a point sampled in its neighborhood to be close:

$$L_{smooth} = \sum_{\mathbf{x}^c} \|\mathbf{n}(\mathbf{x}^c) - \mathbf{n}(\mathbf{x}^c + \boldsymbol{\epsilon})\|_2^2, \tag{11}$$

where $\boldsymbol{\epsilon}$ is a small perturbation randomly drawn from uniform distribution during training. The normal at a point $\mathbf{x}^c$ is calculated by $\mathbf{n}(\mathbf{x}^c) = \frac{\nabla_{\mathbf{x}^c}\sigma(\mathbf{x}^c)}{\|\nabla_{\mathbf{x}^c}\sigma(\mathbf{x}^c)\|_2}$ where ∇ denotes the spacial gradient which can be obtained by network back-propagation.

Shape loss. To further regularize the learned geometry and avoid overfitting, we add a weak constraint that the learned geometry should not be too far away from the fitting SMPL model. Concretely, we penalize the normal of a point $\mathbf{x}^c$ in our learned geometry and that of its nearest body vertex $\mathbf{y}$ on the SMPL body surface:

$$L_{shape} = \sum_{\mathbf{x}_c\in\Omega} \|\mathbf{n}(\mathbf{x}^c) - \mathbf{n}(\mathbf{y}^c)\|_2^2, \tag{12}$$

where SMPL vertex normal $\mathbf{n}(\mathbf{y}^c)$ is constant and can be pre-computed.

In summary, the overall loss function we use to train our MPS-NeRF can be written as:

$$L = L_{color} + \lambda_1 L_{mask} + \lambda_2 L_{smooth} + \lambda_3 L_{shape}, \tag{13}$$

where λ's are the balancing weights.

4. Experiments

Implementation details. Our method is implemented with PyTorch[1]. We use Adam optimizer [16] with a learning rate of $5e-4$ to train the models. The loss weights are set as λ_1=1.0, λ_2=0.1, and λ_3=0.1 in all the experiments. We use 2 Nvidia Tesla V100 GPUs for training with a batch size of 350 rays on each GPU. Training takes about one day on the datasets we used.

More implementation details can be found in the *suppl. material*.

Datasets. We use two datasets to evaluated our method and compare it with previous methods. The first one is *Human3.6M* [13], which contains video sequences of different human actors captured from 4 synchronized cameras. Following [36], we conduct experiments on 7 subject: S1, S5, S6, S7, S8, S9, and S11. Specifically, we train 7 MPS-NeRF models where each model is trained with 6 subjects and tested on the remaining 1 subject in a leave-one-out cross-validation setup. Three views out of the four are selected as the input to our method, with the remaining view left out for supervision (training set) and result evaluation (test set). The same three input views as in [36] are used. We train and test our method using fitted SMPL parameters and image masks provided by [36] which are obtained using [14] and [10], respectively.

To evaluate our method on a larger people collection, we construct another dataset using textured 3D human meshes of 30 randomly-selected subjects from the *THuman* dataset [51]. Each person has about 30 meshes with different poses, from which 20 are randomly chosen to construct our multi-view, multi-pose image dataset. We randomly split the 30 subjects into a training set with 25 subjects and a test set with the remaining 5 subjects. For each person and each pose, we render 24 images from different camera viewpoints. All the cameras point to body center with their azimuth and elevation angles evenly-sampled in $[0°, 360°]$ and $[0°, 35°]$, respectively. The reconstructed SMPL parameters provided by the THuman dataset [51] are used.

4.1. Results and Comparison with Prior Art

Competing methods. To our knowledge, MPS-NeRF is the first person-agnostic framework for novel-view and novel-pose human synthesis using multiview images. Hence, for reference purpose, we compare our method with recent person-specific models NeuralBody (NB) [37] and Animatable NeRF (AniNeRF) [36]. For these methods, one model is trained and tested on a single subject, so 7 models on Human3.6M and 5 models on THuman are used in our experiments. For the evaluation on Human3.6M, we report the numerical and visual results of NB and AniNeRF provided by the authors. To evaluate the methods on THuman, we train the two methods using the released source codes.

Note that for fair comparison, ***all experimental configurations such as input SMPL parameters, image masks, input views, test image sets, and evaluation protocols are made identical for all the methods***.

[1] All source codes and trained models will be publicly released.

Table 1. Comparison of our method with NB [37], AniNeRF [36] on the Human3.6M dataset. Best and second best results are marked with bold and underline, respectively.

Subject	Novel View Synthesis						Novel Pose Synthesis					
	PSNR			SSIM			PSNR			SSIM		
	NB	AniNeRF	Ours	NB	AniNeRF	Ours	NB	AniNeRF	Ours	NB	AniNeRF	Ours
S1	22.87	22.05	**25.40**	0.897	0.888	**0.926**	**22.11**	21.37	21.87	0.879	0.868	**0.880**
S5	**24.6**	23.27	24.30	**0.917**	0.892	0.908	**23.51**	22.29	21.49	**0.897**	0.875	0.871
S6	22.82	21.13	**23.94**	0.888	0.854	**0.893**	23.52	22.59	**23.63**	0.889	0.884	**0.891**
S7	23.17	22.50	**24.27**	**0.914**	0.890	0.911	**22.33**	22.22	21.88	**0.889**	0.878	0.868
S8	21.72	22.75	**23.66**	0.894	0.898	**0.920**	20.94	**21.78**	21.15	0.876	0.882	**0.888**
S9	24.28	**24.72**	24.55	**0.910**	0.908	0.899	23.04	**23.72**	23.33	0.884	**0.886**	0.875
S11	23.70	24.55	**25.12**	0.896	0.902	**0.913**	23.72	**23.91**	23.53	0.884	0.889	**0.891**
Average	23.31	23.00	**24.46**	0.903	0.890	**0.910**	**22.74**	22.55	22.41	**0.885**	0.880	0.881

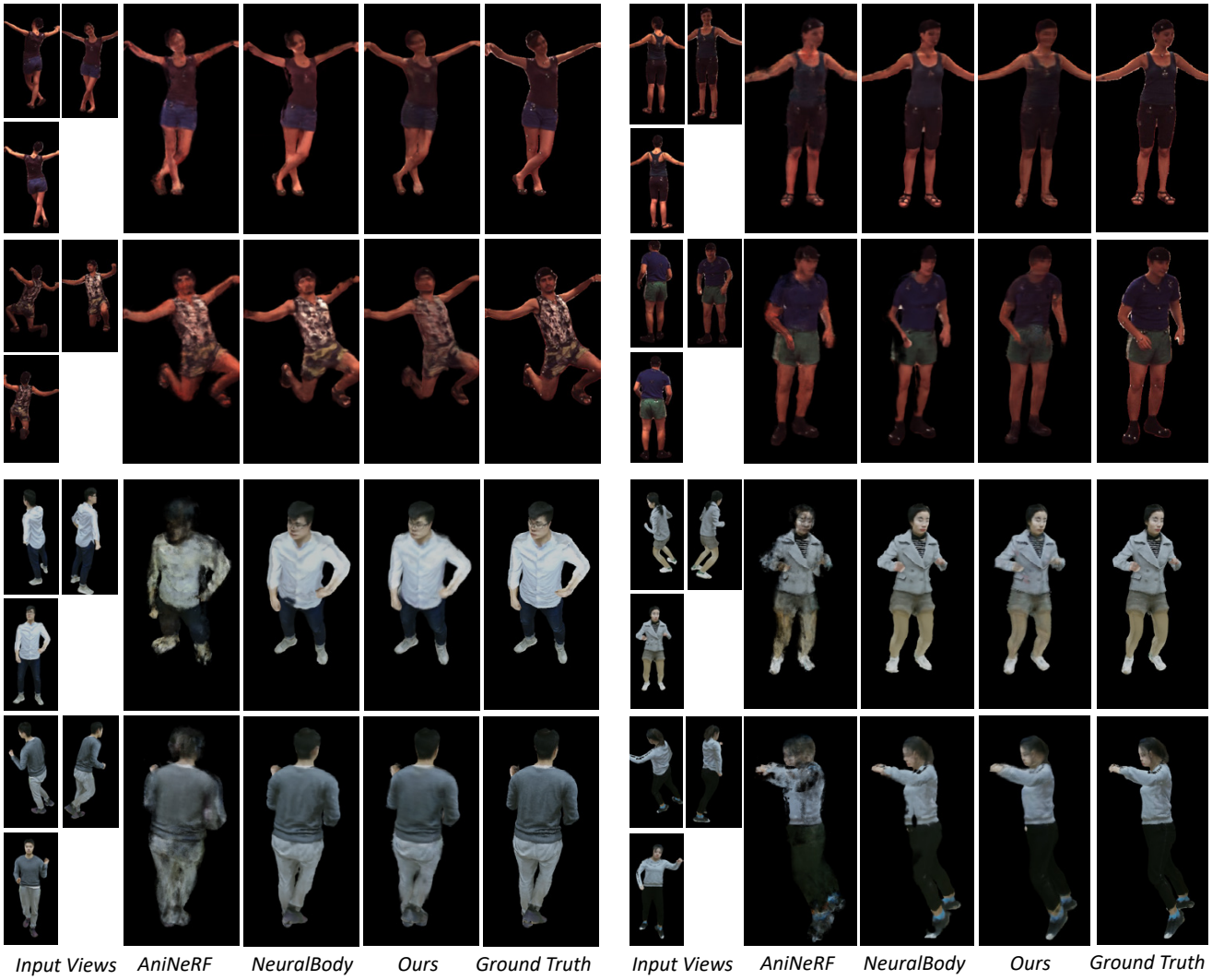

Figure 3. Novel view synthesis results on Human3.6M (top two rows) and THuman (bottom two rows). Note that the three images on the left are only used by our method as input to render these *unseen* subjects at inference time. NeuralBody [37] and AniNeRF [36] are person-specific models which only need camera parameters to render a novel view of these *trained* subjects.

Novel view synthesis results. In our experiments on Human3.6M, all the methods are evaluated on the testing splits of NB and AniNeRF, which contain 49-200 frames for each person. Table 1 presents the quantitative comparison of our

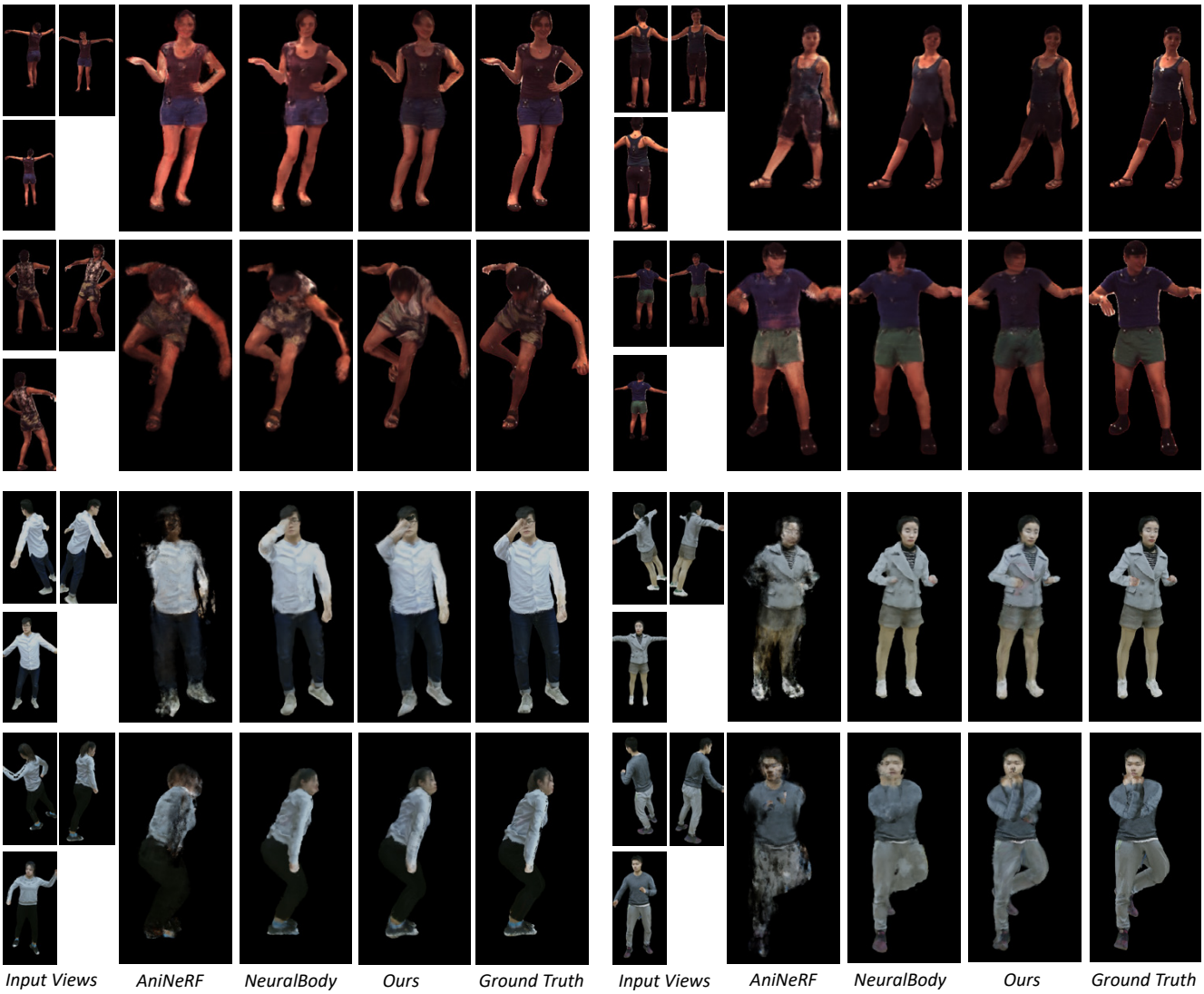

Figure 4. Novel pose synthesis results on the Human3.6M (top two rows) and THuman (bottom two rows) datasets. Note that the three images on the left are only used by our method as input to render these *unseen* subjects at inference time. NeuralBody [37] and AniNeRF [36] are person-specific models which only need camera and pose parameters to render these *trained* subjects.

MPS-NeRF with the other two methods. Although all the testing persons are unseen to MPS-NeRF, it yields better novel view synthesis results than the other two approaches (*e.g.*, 1dB higher in average PSNR). Figure 3 shows the visual result comparison of our method and the others. It can be observed that our method generalizes well to the novel testing persons in the view synthesis task. Compared to NB and AniNeRF, our results exhibit less texture distortions.

To evaluate the methods on THuman, we use images with three evenly-spaced azimuth angles 0°, 120° and 240° as the input for all methods and test on another 8 views. With these 3 views, NB and AniNeRF are trained on all 20 poses of each of the 5 testing person. In contrast, we train a single MPS-NeRF model on 25 training persons each with 20 poses, and use all 24 views as supervision. Table 2 shows the numerical results of different methods on the test set. Again, our method yields better results than both NB and AniNeRF in terms of the average PSNR and SSIM. AniNeRF clearly underperforms in this case. As the pose number for each person is limited and there is no global body rotation, AniNeRF struggles to predict a reasonable novel view image. Figure 3 presents the qualitative results of different methods. Visually inspected, NB and our method are able to generate reasonable results with comparable quality.

Novel pose synthesis results. For the novel pose synthesis task on Human3.6M, we evaluate the results of different methods with the testing view and poses from the test sets of NB and AniNeRF. For our MPS-NeRF, we choose one pose from the training set of NB and AniNeRF to synthesize the novel pose of an unseen testing person. The quanti-

Table 2. Comparison of NB [37], AniNeRF [36], and our method on the THuman dataset.

Method	Novel View		Novel Pose	
	PSNR	SSIM	PSNR	SSIM
NB	24.86	0.929	23.36	0.903
AniNeRF	20.10	0.841	17.25	0.791
Ours	**25.63**	**0.935**	**23.92**	**0.911**

tative results are presented in Table 1. It shows that all three methods performed comparably, with our result marginally worse than NB and AniNeRF (*e.g.*, average PNSR 22.41dB from our method *vs.* 22.74dB and 22.55dB from NB and AniNeRF, respectively). This indicates that our trained model generalizes well to not only unseen person but also various poses. Figure 4 shows some visual results of different methods. While it appears that our results are slightly more blurry than NB and AniNeRF, it generally contains less high-frequency artifacts.

In the novel pose synthesis experiments on THuman, we also use three input views as in the novel view synthesis task. NB and AniNeRF are trained on all the 20 poses of each person. For evaluation, we take another 5 poses from the THuman dataset and render a novel pose test set for each subject. Our method is trained on 25 training subjects each with 20 poses, and tested on the same test set as NB and AniNeRF. The quantitative results presented in Table 2 shows that our method performed best. It is slightly better than NB (average PSNR 23.92dB *vs.* 23.36dB). Again, AniNeRF performed poorly due to limited training poses. Some visual results are presented in Figure. 4.

4.2. Analysis and Ablation study

We further test MPS-NeRF with more input views on the THuman dataset and analyze their impact. We also conduct ablation studies to validate the efficacy of different modules.

Impact of input view number. In this experiment, we train and test our method using different numbers of input views from 3 to 12. Table 3 shows that the performance of our method gradually increases with more input views and gets saturated when the number is greater than 8. One visual example is presented in Fig. 5, which shows that the noise is gradually suppressed and boundary becomes sharper and more accurate with an increasing input view number.

Impact of canonical space To validate the importance of using a canonical space, we remove it in our implementation and retrain our model. Note that in this setup, our method is very similar to the image-conditioned NeRF methods for generic scenes [43, 49]. As shown in Table 4, the performance drops significantly for both the novel-view and novel-pose synthesis tasks, which indicates that our canonical space design is critical for generalizable human rendering. Due to space constrain, the visual comparison is presented in the *suppl. material*.

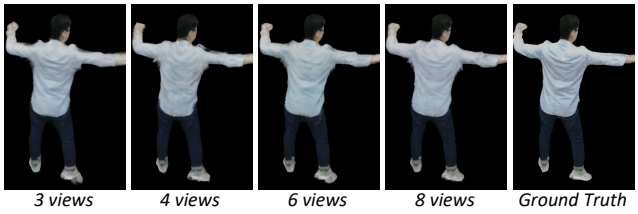

Figure 5. Visual results with different input view numbers.

Table 3. Quantitative results with different input view numbers.

View Number	Novel View		Novel Pose	
	PSNR	SSIM	PSNR	SSIM
3	25.63	0.935	23.92	0.911
4	25.84	0.938	23.95	0.912
6	26.46	0.943	24.25	0.916
8	27.37	0.951	24.45	0.92
12	**27.77**	**0.953**	**24.56**	**0.921**

Table 4. Ablation study on canonical space and attention block

	Novel View		Novel Pose	
	PSNR	SSIM	PSNR	SSIM
w/o canonical	23.34	0.913	21.77	0.888
w/o attention	22.05	0.907	21.18	0.888
Our full model	**25.63**	**0.935**	**23.92**	**0.911**

Impact of attention-based feature fusion. We also verify the efficacy of our attention-based feature fusion scheme by replacing it with a simple feature average pooling strategy. The performance also drops dramatically as shown in Table 4, demonstrating that multiview feature fusion cannot be handle naively and our feature fusion method is substantial for good performance. Due to space constrain, the visual examples are presented in the *suppl. material*.

5. Conclusion

In this paper, we proposed a novel-view and novel-pose human synthesis approach which can be generalizable for unseen persons with sparse multiview images as input. Our key idea is to leverage a canonical-space NeRF and a volume deformation scheme derived by human body parametric model to achieve better generalizability. Our method is simple but works surprisingly well as demonstrated by the extensive experiments. We hope that our method can serve as a strong baseline model for generic 3D human rendering.

Our current deformation scheme cannot handle very loose clothing and garments with complex geometry (*e.g.*, a long coat or opened jacket). Our future work will be devoted to designing a generalizable, learning-based deformation scheme to better handle such cases.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *arXiv preprint arXiv:2103.13415*, 2021. 11

[2] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Advances in Neural Information Processing Systems*, 2020. 4

[3] Federica Bogo, Michael J Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 2300–2308, 2015. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 2, 3

[5] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2

[6] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics*, 34(4):1–13, 2015. 1

[7] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *ACM SIGGRAPH*, pages 1–10. 2008. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4

[9] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1753. IEEE, 2009. 2

[10] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *European Conference on Computer Vision*, pages 770–785, 2018. 5

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 11

[12] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 2, 4

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013. 2, 5, 11

[14] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2, 3, 5

[15] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. In *Annual Conference on Computer Graphics and Interactive Techniques*, volume 18, pages 165–174, 1984. 4

[16] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015. 5

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

[18] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 2

[19] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[20] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 2

[21] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Annual Conference on Computer Graphics and Interactive Techniques*, pages 165–172, 2000. 4

[22] Zhongguo Li, Magnus Oskarsson, and Anders Heyden. 3d human pose and shape estimation through collaborative learning and multi-view model-fitting. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1888–1897, 2021. 3

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

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16, 2015. 2, 3

[25] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 11

[26] Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, et al. Lookingood: enhancing performance capture with real-time neural re-rendering. *ACM Transactions on Graphics*, 37(6):1–14, 2018. 1

[27] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 3

[28] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2

[29] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421, 2020. 1, 2, 3, 4, 11

[30] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *IEEE/CVF International Conference on Computer Vision*, 2021. 1, 3

[31] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *IEEE/CVF International Conference on Computer Vision*, 2021. 5

[32] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Annual Symposium on User Interface Software and Technology*, pages 741–754, 2016. 1

[33] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[34] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2, 3

[35] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[36] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 1, 2, 3, 4, 5, 6, 7, 8, 11

[37] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 1, 2, 3, 5, 6, 7, 8, 11, 12, 13

[38] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2, 3

[39] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

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

[41] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019. 2

[42] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021. 2

[43] Alex Trevithick and Bo Yang. GRF: Learning a general radiance field for 3d representation and rendering. In *IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 2, 3, 8

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 4

[45] Xiaolin Wei, Peizhao Zhang, and Jinxiang Chai. Accurate realtime full-body motion capture using a single depth camera. *ACM Transactions on Graphics*, 31(6):1–12, 2012. 2

[46] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1682–1691, 2020. 1

[47] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[48] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Analyzing clothing layer deformation statistics of 3d human motions. In *European Conference on Computer Vision*, pages 237–253, 2018. 4

[49] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2, 3, 8

[50] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7376–7385, 2020. 3

[51] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019. 2, 5, 11

A. More Network and Implementation Details

We use the first 7 convolution layers of ResNet-34 [11] as our image feature encoder, with the original max-pooling layer removed. The images are downsampled to half resolution before feeding into the encoder CNN. The MLP for canonical NeRF is adapted from [29]. The MLP_1 subnet (Eq. 5 of the main paper) has 8 layers each with 256 feature dimensions. MLP_2 (Eq. 6 of the main paper) has 2 hidden layers with 256 feature dimensions. ReLU activation is used for all hidden layers. Following Mip-NeRF [1], we use a shifted softplus activation: $\log(1+\exp(x-1))$ to produce density σ. For color $\mathbf{c}$, instead of using the sigmoid activation, we use a "widened" sigmoid function $f(x) = (1+2\epsilon)/(1+\exp(-x)) - \epsilon$, with $\epsilon = 0.001$.

To render a pixel on the target image, a ray is cast to the 3D volume and points are sampled along the ray for radiance accumulation. Instead of using the sampling strategy in original NeRF [29] which samples points in the whole volume, we follow [37] to sample points within a 3D bounding box derived based on the SMPL model.

Our learning rate starts from $5e{-}4$ and decays by a factor 0.5 for every 30K interactions, and the maximum iteration number is set to 120K.

B. More Dataset Details

Human3.6M. For the experiments on Human3.6M [13], we use the same frame number setup as in AniNeRF [36] for the novel-view and novel-pose synthesis tests. The details are presented in Table 5. During training, we use three input views (#0, #1, #2) as input, and use all the four view (#0, #1, #2, #3) as supervision. For the novel view synthesis task on an unseen person, we input 3 views (#0, #1, #2) and test on view #3 for each frame. For novel pose synthesis, we choose 3 fixed images from views (#0, #1, #2) as our input, construct the canonical human NeRF, and deform it to all the target novel poses.

Table 5. Number of frames used to evaluate the novel-view and novel-pose synthesis tasks.

	S1	S5	S6	S7	S8	S9	S11
Novel view	150	250	150	300	250	260	200
Novel pose	49	127	83	200	87	133	82

THuman. The 25 training subjects and 5 test subjects we use from THuman [51] are listed in Table 6 and Table 7, respectively.

Evaluation metric As shown in the main paper, PSNR and SSIM metrics are used for quantitative evaluation. Instead of directly calculating PSNR and SSIM for the whole image, we follow AniNeRF [36] and NeuralBody [37] to project the 3D bounding box of a body onto image plane to obtain a 2D mask and only calculate PSNR and SSIM in the masked region.

Table 6. List of the 25 training subjects from THuman.

gyc_20181010_hsc_1_M	gyc_20181010_wyl_1_M
gyx_20181011_lty_1_M	gyx_20181011_lty_2_M
gyx_20181012_yw_2_F	gyx_20181013_dcj_1_F
gyx_20181013_fjj_1_F	gyx_20181013_fyh_1_F
gyx_20181011_scw_1_M	gyx_20181011_scw_2_M
gyx_20181011_wlf_1_M	gyx_20181011_wsc_1_M
gyx_20181013_gfz_2_F	gyx_20181013_gy_1_F
gyx_20181013_hj_1_F	gyx_20181013_jyq_1_F
gyx_20181011_wsc_2_M	gyx_20181011_zcj_1_M
gyx_20181011_zcj_2_M	gyx_20181011_zxh_1_M
gyx_20181011_zxh_2_M	gyx_20181011_zyq_1_F
gyx_20181011_zyq_2_F	gyx_20181012_hl_1_M
gyx_20181012_hl_2_M	

Table 7. List of the 5 test subjects from THuman.

gyx_20181012_sty_1_M	gyx_20181012_xsx_2_M
gyx_20181013_hyd_1_M	gyx_20181012_lw_2_F
gyx_20181013_xyz_1_F	

C. Reconstructed 3D Shapes

Although the NeRF-based representation does the explicitly recover 3D geometry, we can still extract proxy 3D shapes using the density field predicted by the network. Specifically, we first define a 3D human bounding box based on the target pose SMPL model which has a size of $2m \times 2m \times 2m$, and discretize it to a $256 \times 256 \times 256$ voxel grid. Then we deform the voxel centers from the target space to the canonical space, and evaluate their volume densities using our canonical NeRF network. Finally, the density values are binarized and the Marching Cubes algorithm [25] is applied to extract a mesh. Figure 6 presents the extracted proxy 3D shapes for the input images as well as the reposed target shapes using our deformation scheme. Note that we did not apply any post-processing on the extracted 3D shapes such as the Gaussian smoothing used in AniNeRF [36].

D. More Ablation Study Results

Person-specific training. Although our goal is to build a generalizable model which can handle unseen persons, we also evaluate our method in a person-specific training setup for a reference. In these tests, we train and test our MPS-NeRF on one person using the the setup of NB and AniNeRF. As shown in Table 8, the models trained in this way yield better results than our original setting for both the novel-view and novel-pose synthesis tasks. They also outperform NB and AniNerF on these two tasks.

Figure 6. Proxy 3D shapes reconstructed by our method. The meshes are extracted by running the Marching Cubes algorithm on the binarized volume density, and we did not apply any post-processing such as mesh smoothing.

Table 8. More comparisons and ablation studies on the Human3.6M dataset. Ours^{+} denotes our method trained with 6 subjects and tested on the remaining 1 subject in a leave-one-out cross-validation setup. Ours^{ps} denotes our method trained in a person-specific manner similar to NB and AniNeRF.

Subject	Novel View Synthesis								Novel Pose Synthesis							
	PSNR				SSIM				PSNR				SSIM			
	NB	AniNeRF	Ours^{+}	Ours^{ps}	NB	AniNeRF	Ours^{+}	Ours^{ps}	NB	AniNeRF	Ours^{+}	Ours^{ps}	NB	AniNeRF	Ours^{+}	Ours^{ps}
S1	22.87	22.05	25.40	24.24	0.897	0.888	0.926	0.907	22.11	21.37	21.87	22.27	0.879	0.868	0.880	0.879
S5	24.6	23.27	24.30	24.39	0.917	0.892	0.908	0.912	23.51	22.29	21.49	22.25	0.897	0.875	0.871	0.882
S6	22.82	21.13	23.94	24.18	0.888	0.854	0.893	0.894	23.52	22.59	23.63	24.04	0.889	0.884	0.891	0.892
S7	23.17	22.50	24.27	24.35	0.914	0.890	0.911	0.911	22.33	22.22	21.88	22.11	0.889	0.878	0.868	0.877
S8	21.72	22.75	23.66	23.65	0.894	0.898	0.920	0.916	20.94	21.78	21.15	21.20	0.876	0.882	0.888	0.891
S9	24.28	24.72	24.55	25.39	0.910	0.908	0.899	0.909	23.04	23.72	23.33	23.67	0.884	0.886	0.875	0.883
S11	23.70	24.55	25.12	25.47	0.896	0.902	0.913	0.915	23.72	23.91	23.53	24.06	0.884	0.889	0.891	0.895
Average	23.31	23.00	24.46	24.52	0.902	0.890	0.910	0.909	22.74	22.55	22.41	22.80	0.885	0.880	0.881	0.886

Impact of canonical space. The quantitative comparison to evaluate the impact of our canonical space has been presented in the main paper, which shows the performance drops significantly without the canonical space. Here we further show the visual comparison in Figure 7. As can be observed, the results without a canonical space contain more unwanted artifacts. Moreover, the predicted human boundary is clearly erroneous for some cases (*e.g.*, see the arm in the third row).

Impact of attention-based feature fusion The quantitative evaluation in the main paper has shown that our attention-based feature fusion is also critical. Here we further qualitatively compare the results of using our attention-based feature fusion and using naive feature pooling. As shown in Fig. 7, without our attention-based fusion the synthesized images are blurry and suffer from severe color distortions.

E. Results on ZJU-MoCap Dataset

We also tested our method on the ZJU-MoCap dataset [37], where we train on 6 subjects using 4 views as input. Figure 8 shows the results of our method on an unseen testing person. Our method still generates reasonable results for both the novel-view and novel-pose synthesis tasks, despite the person in not the training set. For a reference, we also present the visual results of Neural-

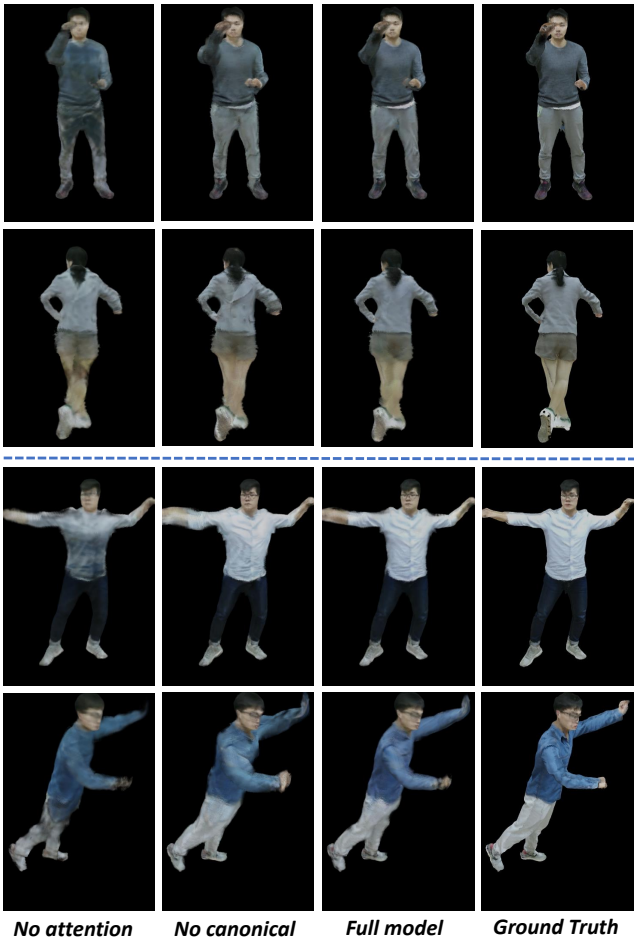

Figure 7. Impact of canonical space and attention-based feature fusion on novel view (top two rows) and novel pose (bottom two rows) synthesis tasks.

Body [37]. Note that in this dataset, the actors make constant body spinning during video shooting, unlike the Human3.6M dataset where the body movements are more natural. This greatly eases the novel-view synthesis learning task for person-specific models. Therefore, NeuralBody can generate remarkable novel view synthesis results. Still, our novel pose synthesis results are comparable to NeuralBody, despite the person is unseen to our model.

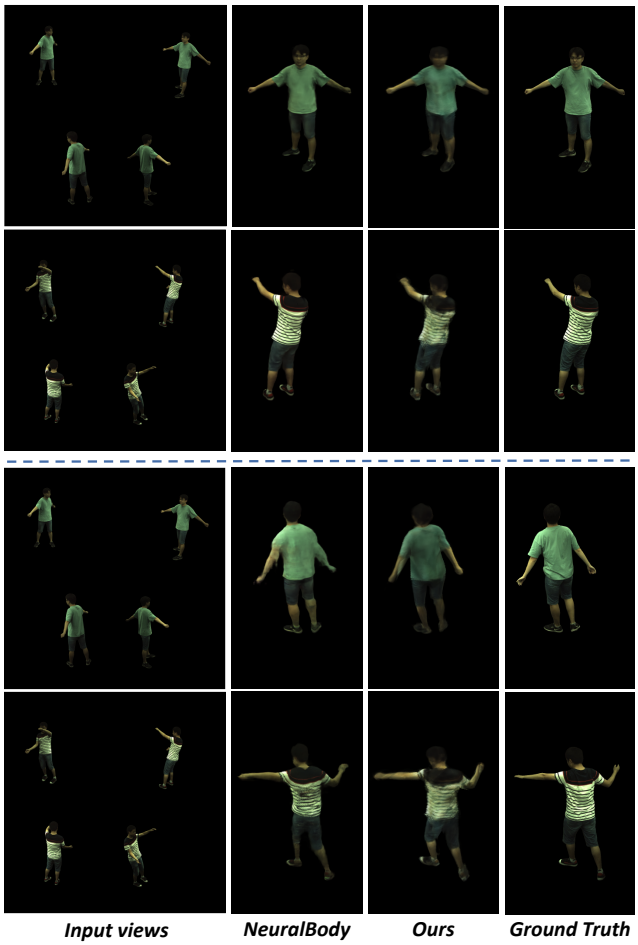

Figure 8. Results of our method tested on one subject from the ZJU-MoCap dataset [37]. Note that in this dataset, the actors make constant body spinning during video shooting, unlike the Human3.6M dataset where the body movements are more natural. This greatly eases the novel-view synthesis learning task for *person-specific* models. Our method still generates reasonable results for both novel-view and novel-pose synthesis despite the person is *unseen* during training.