Within the Dynamic Context: Inertia-aware 3D Human Modeling with Pose Sequence

Yutong Chen*[1], Yifan Zhan*[1,2], Zhihang Zhong†[1], Wei Wang[1], Xiao Sun†[1], Yu Qiao[1], and Yinqiang Zheng[2]

[1] Shanghai Artificial Intelligence Laboratory
[2] The University of Tokyo

Abstract. Neural rendering techniques have significantly advanced 3D human body modeling. However, previous approaches overlook dynamics induced by factors such as motion inertia, leading to challenges in scenarios where the pose remains static while the appearance changes, such as abrupt stops after spinning. This limitation arises from conditioning on a single pose, which leads to ambiguity in mapping one pose to multiple appearances.

In this study, we elucidate that variations in human appearance depend not only on the current frame's pose condition but also on past pose states. We introduce Dyco, a novel method that utilizes the delta pose sequence to effectively model temporal appearance variations. To mitigate overfitting to the delta pose sequence, we further propose a localized dynamic context encoder to reduce unnecessary inter-body part dependencies. To validate the effectiveness of our approach, we collect a novel dataset named I3D-Human, focused on capturing temporal changes in clothing appearance under similar poses. Dyco significantly outperforms baselines on I3D-Human and achieves comparable results on ZJU-MoCap. Furthermore, our inertia-aware 3D human method can unprecedentedly simulate appearance changes caused by inertia at different velocities. The code, data and model are available at our project website at https://ai4sports.opengvlab.com/Dyco.

Keywords: 3D human · Pose sequence · Neural radiance fields · Inertia

1 Introduction

Pursuing dynamic, lifelike, and high-fidelity human body reconstruction has been a longstanding endeavor, finding versatile applications in AR/VR, filmmaking, gaming, and motion analysis. Previous efforts in dynamic human neural rendering have focused on digitalizing human avatars and modeling human motion [2,4,12,26,27,30,31,37,38,42]. Classical approaches such as SCAPE [4] and SMPL [2] conduct statistical analysis on 3D scans to decompose mesh variation into shape and pose components. However, these methods struggle with lacking geometric details such as clothing depiction. Recent advancements in the neural

* denotes authors with equal contributions. † denotes co-corresponding authors.

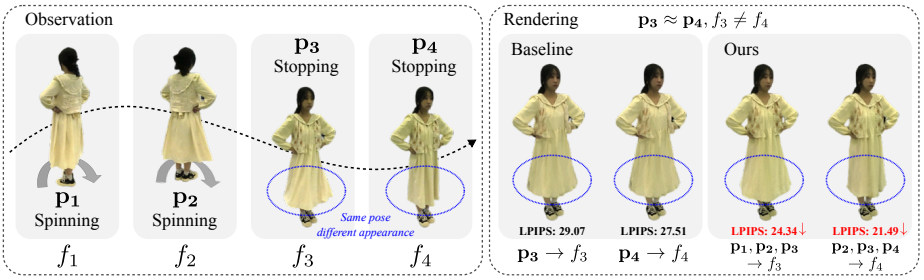


Fig. 1: We emphasize that appearance variations not only depend on different static poses, but can also be induced by inertia, such as the graceful hanging down of the dress drape after a sudden stop in motion. Compared with previous methods solely relying on static poses, we encode past pose trajectory with the pose sequence to accurately capture such dynamic effects. This improves both novel-view renderings and generalization to novel poses.

radiance field [21] have enabled learning more accurate neural human representations from monocular or multi-view videos, typically involving Linear Blend Skinning (LBS) guided skeletal motion module, MLP-based non-rigid motion module, and canonical radiance fields. These methods achieve pose-dependent human motion modeling by learning human radiance fields in canonical space and warping them to observation space conditioned on pre-extracted poses.

Our question is whether pose-based LBS processes are sufficient for modeling 3D human motion. To answer this question, let us consider the scenario captured from a static camera view in Fig. 1, wherein a girl spins ($\boldsymbol{p}_1 \rightarrow \boldsymbol{p}_3$) and then stops ($\boldsymbol{p}_3 \rightarrow \boldsymbol{p}_4$). As she stops, her pose stays mostly the same, but the drape of her dress noticeably falls. Therefore, solely relying on the current single pose is insufficient to model variations in non-rigid human motion and appearance, as it ignores dynamic motion contexts. This limitation causes previous methods to struggle with mapping identical pose inputs to different appearance, resulting in artifacts in the scenario described above.

In this work, we propose a neural avatar model that accounts for dynamic context, dubbed Dyco, which significantly advances dynamic human motion modeling by resolving the ambiguity that arises from relying solely on static poses. Drawing inspiration from dynamic blend shapes in DMPL [2], we encode pose sequences as condition in addition to the current static pose. We compute the changes, or deltas, in pose parameters over the pose trajectory and input the delta sequence into the non-rigid deformation field and canonical volume. Since the increased input dimension poses a risk of overfitting, we design a localized dynamic context encoder to reduce input complexity without sacrificing information. The dynamic context encoder eliminates spurious dependencies between different body parts and efficiently aggregates spatial and temporal information.

In addition to the methodological neglect of dynamic contexts, current datasets fail to accurately capture the associated issues. existing datasets mitigate the issue of appearance variations caused by dynamic contexts by reducing the mag-

nitude of motion (DNA-Rendering Dataset [6]) or wearing tight-fitting garments (ZJU-Mocap Dataset [28], ENeRF-Outdoor Dataset [19] and NHR Dataset [39]). However, real-world human body modeling cannot avoid dealing with large motions and loose clothing. To fully demonstrate the advantages of our inertia-aware 3D human modeling method, we have collected a new dataset of human body data, named I3D-Human. I3D-Human emphasizes significant appearance variations resulting from dynamic contexts, such as skirt swaying and hem dropping, with the aim of facilitating relevant future research. In summary, our contributions are as follows:

1) Dyco, a novel human motion modeling method that utilizes pose sequences, mitigating the ambiguity in appearance learning conditioned solely on poses, thus enhancing the generalization of dynamic contexts to novel poses;
2) a localized dynamic context encoder to alleviate the risk of overfitting associated with the increased representational capacity, thereby contributing to improved model performance;
3) I3D-Human Dataset, a self-collected human dataset capturing richer variations of appearance from dynamic contexts. We achieve best results on our I3D-Human Dataset and demonstrate comparable performance on ZJU-Mocap, compared to state-of-the-art human body modeling methods.

2 Related Work

2.1 Explicit Mesh Representation for Human Avatars

The early success of digital avatars took place in a class of body models [2, 8, 10, 30] that represent the body as a mesh and perform statistical analysis on the database of 3D scans to learn shape and pose components in the mesh variation. Mostly well-known, a family of SMPL(-X)-based human body models [2] has been widely used over the past decade for its flexibility, realism, and ease of use. The key improvement of SMPL over previous models lies in its effective learning of *pose blend shapes* to correct skinning artifacts and pose-dependent deformations. This is achieved by formulating the pose blend shapes as a linear function of the elements of the part rotation matrices. Since then, plenty of work has contributed to estimating the SMPL parameters of subjects from 2D inputs [1, 7, 33]. To capture the momentarily lagging-behind motion of soft body tissues in dynamic human motion, DMPL [2] and Dyna [30] parameterized velocities and acceleration of the body and limbs into their linear models, which brings their animation closer to life. However, these meshed-based body models can only represent naked or tightly-clothed bodies. Though some work [9, 24] extends body mesh models to modeling clothed humans, their mesh-based models of limited resolution make the renderings far from visually realistic.

2.2 Implicit Neural Representation for Human Avatars

In pursuit of photorealism, advanced human avatars exploit neural implicit representation [18, 20, 34, 42]. For instance, a number of methods using neural radiance

field (NeRF [21]), ARAH [37] using neural signed distance function (NeuS [36]), InstantAvatar [12] and FastNeuralBody using InstantNGP [22], and more recent models take advantage of the efficiency of gaussian splatting [14]. Despite variations in their choice of neural representation, these methods typically learn a canonical field to represent body shape and appearance in a rest pose, along with a deformation module to compute correspondence between the rest pose and target pose.

However, while neural 3D representations have demonstrated impressive results in novel-view rendering of static scenes and even some quasi-static dynamic scenes, we find that the rendering results of neural implicit human avatars are not satisfactory enough, exhibiting blurring appearance and failing to manifest realistic dynamic motion, even in the training images. One important reason for the blurriness is that most of the state-of-arts only take the SMPL parameters of the target single pose as input to their neural representation, which will bring dynamic context-induced ambiguity that may influence model fitting.

2.3 Modeling Dynamic Context in Human Avatars

The impact of dynamic context on human appearance was discussed in some previous literature. First, there are physical-based approaches which exploited the material properties of human soft tissues and simulated their dynamic simulation using intricate physical models. For example, Larboulette *et al.* [16] design a mechanical model of a regular limb to simulate the motion of flesh under different accelerations of the bone with the finite element method (FEM). Rohmer *et al.* [32] underscores the importance of secondary animation for liveliness and presents velocity skinning on top of standard LBS to create dynamic effects such as squash and stretch. However, physical-based approach requires manual design for different materials, which is infeasible to apply to full-body animation. Instead of modeling the underlying physics, another line of work attempts to learn the approximation of dynamics from 4D scans in a data-driven manner. DMPL [2] and Dyna [30] learn to predict soft-tissue deformation from the changing of the body pose. However they can only model naked human bodies. DRAPE [9] incorporates motion history into clothing deformation prediction to animate the draping and flowing effect of clothing. However, it requires the clothing template mesh and its proficiency is highly limited to the types of clothing it is trained on. Also, these methods are confined to mesh-based representation and deformation and thus far from high-resolution photo-realistic renderings.

On the other hand, current implicit neural avatars are promising for photo-realisim. However, none of them manage to represent the effect of dynamic context and motion inertia on appearance variation. Our work is the first to make the implicit avatar inertia-aware. One reason for the dynamic effect being overlooked is that current datasets, *e.g.* ZJU-MoCap [28], PeopleSnapshot [3], only focus on tightly-clothed figures with slow and gradual motion. To study the dynamic motion effects that commonly exist in the real scenario, we curate a new dataset of loosely-clothed humans with free and abrupt motion.

3 Preliminaries

3.1 Human NeRF Representation

A HumanNeRF [38] is defined to determine the density σ and color $\boldsymbol{c}$ of any given 3D location $\boldsymbol{x}$, conditioned on a specific pose configuration $\boldsymbol{p} = \{(R_i, \boldsymbol{t}_i) | i \in K\}$ of a person: $F_o(\boldsymbol{x}, \boldsymbol{p}) \rightarrow (\sigma, \boldsymbol{c})$. Typically, the problem is solved in a canonical space, where a standard static Neural Radiance Field (NeRF) [21]

$$F_c(\boldsymbol{x}_c) \xrightarrow{\Theta_{\text{NeRF}}} (\sigma, \boldsymbol{c}), \tag{1}$$

and a point mapping function $T(\boldsymbol{x}, \boldsymbol{p}) \rightarrow \boldsymbol{x}_c$ from observed space back to canonical space are learned. The $\boldsymbol{x}_c$ can be split into a rigid and a non-rigid part, *i.e.*, $\boldsymbol{x}_r$ and $\Delta\boldsymbol{x}$, corresponding to the rigid transformation function T_{R} and non-rigid transformation function T_{NR}, represented as

$$\begin{aligned} \boldsymbol{x}_c &= \boldsymbol{x}_r + \Delta\boldsymbol{x}, & (2)\\ &= T_{\text{R}}(\boldsymbol{x}, \boldsymbol{p}) + T_{\text{NR}}(\boldsymbol{x}_r, \boldsymbol{p}). & (3) \end{aligned}$$

The rigid transformation function T_{R} is formulated as an inverse linear blend skinning process [17], a weighted sum of the rigid transformations defined following

$$T_{\text{R}} = \sum_{i=1}^{K} \omega_o^i(\boldsymbol{x})(R_i\boldsymbol{x} + \boldsymbol{t}_i), \tag{4}$$

where $\omega_o^i(\boldsymbol{x})$ is the corresponding blending weight in observed space and estimated using a set of learnable weights $\omega_c^i(\boldsymbol{x})$ defined in the canonical space as

$$\omega_o^i(\boldsymbol{x}) = \frac{\omega_c^i(\boldsymbol{x})(R_i\boldsymbol{x} + \boldsymbol{t}_i)}{\sum_{i=1}^{K} \omega_c^i(\boldsymbol{x})(R_i\boldsymbol{x} + \boldsymbol{t}_i)}. \tag{5}$$

Using a convolution network that takes a random latent code $\boldsymbol{z}$ as input, a single volume grid $W_c(\boldsymbol{x})$ with $K + 1$ channels are learned to store all the blending weights $W_c(\boldsymbol{x}) = \{\omega_c^i(\boldsymbol{x}) | i \in K\}$: $\text{CNN}_{\Theta_\omega}(\boldsymbol{z}) \rightarrow W_c(\boldsymbol{x})$.

The non-rigid transformation T_{NR} predicts an extra position offset $\Delta\boldsymbol{x}$ caused by different pose configurations by optimizing parameters Θ_{NR},

$$T_{\text{NR}}(\boldsymbol{x}_r, \boldsymbol{p}) \xrightarrow{\Theta_{\text{NR}}} \Delta\boldsymbol{x}. \tag{6}$$

3.2 NeRF and Volume Rendering Revisited

NeRF [21] comprises three primary stages: sampling, volume mapping, and rendering. In the sampling stage, points $\boldsymbol{x} \in \mathbb{R}^3$ are sampled along rays computed from the camera's position. Subsequently, in the volume mapping stage, each 3D point $\boldsymbol{x}$, along with its viewing direction $\boldsymbol{d} \in \mathbb{R}^3$, is queried to determine the volume density σ and color $\boldsymbol{c} = (r, g, b)$ of $\boldsymbol{x}$. Finally, in the rendering process,

the color of each ray is determined using volume rendering techniques [13]. The expected color of the ray $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$ is then computed as follows

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

where t_n and t_f are the near and far bounds, respectively, and

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\boldsymbol{r}(s))ds\right)\,. \tag{8}$$

The ground truth color $C_{\text{gt}}(\boldsymbol{r})$ of each ray serves as supervision for training NeRF, and we define the loss function as follows

$$\mathcal{L}_{\text{MSE}} = \sum_{\boldsymbol{r}\in\mathcal{R}} \|C(\boldsymbol{r}) - C_{\text{gt}}(\boldsymbol{r})\|_2^2\,, \tag{9}$$

where $\mathcal{R}$ is the ray batch. Besides RGB-based loss, previous work [31, 38, 40] shows that utilizing an LPIPS loss [41] $\mathcal{L}_{\text{LPIPS}}$ as an auxiliary supervision can greatly enhance the visual details of the renderings.

4 Method

4.1 Ambiguities under Identical Pose Conditions

Relying solely on rigid transformations, T_R, for modeling details of human motion is not sufficient. In many human NeRF works [18,38], achieving comprehensive motion modeling requires incorporating non-rigid deformations $T_{NR}(\boldsymbol{x}_r, \boldsymbol{p})$ conditioned on human poses $\boldsymbol{p}$. Some studies [23] introduce additional pose conditions for training the canonical volume to capture variations in per-frame appearance. Nevertheless, our analysis suggests that relying exclusively on pose conditions for non-rigid deformation and canonical volume learning still imposes inherent ambiguities. This is exemplified by varying clothing states of a person under identical poses. In instances where an individual wearing loose clothing abruptly halts after rotation, there is a brief period during which the overall pose remains static, yet inertia induces substantial swinging motion at the specific location $\boldsymbol{x}$ on the clothing. Consequently, the non-rigid deformation target $\Delta\boldsymbol{x}$ for point $\boldsymbol{x}$ becomes ambiguous due to this swinging motion, given their shared condition under identical poses. Another type of ambiguity arises from the variations in appearance under identical pose conditions, when two observation points corresponding in canonical space exhibit different local normal characteristics. Even under invariant lighting conditions, differences in local normals can impact shading outcomes, consequently affecting the final appearance.

We emphasize that these two ambiguities are prevalent in everyday scenarios caused by dynamic contexts such as inertia, despite being overlooked by previous methods and datasets focusing on meticulously designed wearing patterns. To alleviate these ambiguities, we introduce pose-sequence information in the following sections to capture the intrinsic details arising from inertia.

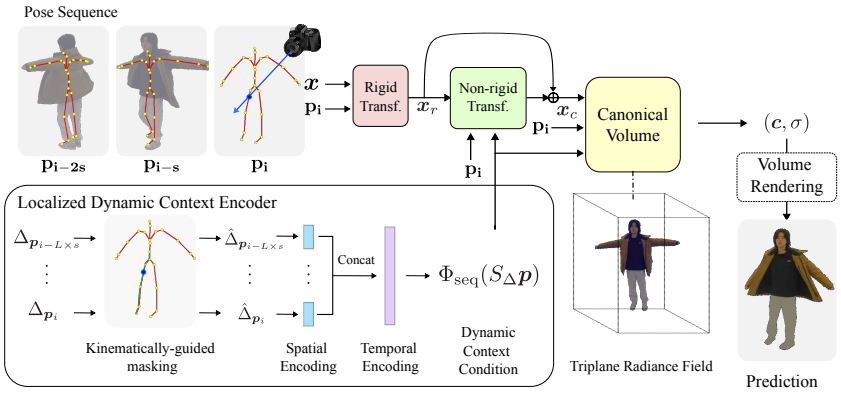


Fig. 2: The overall pipeline of our method. The rigid transformation and non-rigid transformation module deform the coordinate in the pose space into the canonical space, which is then fed into the triplane volume to obtain the color and density in the canonical space. To capture the variation under similar poses within different dynamic contexts, we adopt a localized dynamic context encoder to embed pose sequences as additional conditional inputs into the transformation module and canonical volume.

4.2 Modeling Pose-Sequence Dependent Human Motion

Fig. 2 shows our pipeline. Our core insight is that the inertia properties of the human body cannot be exclusively inferred from the current frame but are encoded within the preceding pose sequence. We introduce two parameters to define such a pose sequence: sequence length L and step s. Formally, a pose-sequence $S_{\boldsymbol{p}}$ with a length L, traced back from the current frame at step s, is defined as

$$S_{\boldsymbol{p}} = \{\boldsymbol{p}_{i-(L-1)s}, \ldots, \boldsymbol{p}_{i-s}, \boldsymbol{p}_i\}, \tag{10}$$

where the sequence length L is the number of frames in the pose-sequence and s is the level of granularity. Both parameters jointly determine the pose sequence's window length and motion details of the pose sequence. However, this pose-sequence design is susceptible to overfitting when used as network input, particularly due to its complexity with increasing sequence length L.

We reassess the inertia-related properties, *e.g.* velocity and acceleration, of the human body and find that they are solely associated with **changes** in pose, independent of absolute pose. Therefore, we further define a delta-pose sequence with length L_d, sequence step s, and delta step s_d as

$$S_{\Delta \boldsymbol{p}} = \{\Delta \boldsymbol{p}_{i-(L_d-1)s}, \Delta \boldsymbol{p}_i, \ldots, \Delta \boldsymbol{p}_{i-s}, \Delta \boldsymbol{p}_i\}, \tag{11}$$

$$\Delta \boldsymbol{p}_i = \delta(\boldsymbol{p}_i, \boldsymbol{p}_{i-s_d}), \tag{12}$$

where δ is a $3K+3$ dimensional vector including the per-joint rotation difference represented in axis-angle form between two poses and the global translation

difference. The proposed $S_{\Delta \boldsymbol{p}}$ reduces input complexity by only encoding pose residuals. The sequence vector is then embedded by a dynamic context encoder Φ_{seq} which will be explained in Sec. 4.3.

As $S_{\Delta \boldsymbol{p}}$ only encompasses temporally local pose difference, we supplement the context condition with the current pose $\boldsymbol{p}_i$. Altogether, we utilize both the current pose condition and the context sequence to guide the learning of the non-rigid deformation and the canonical volume, effectively resolving the appearance ambiguities caused by inertia. For non-rigid deformation, we have

$$T_{NR}(\boldsymbol{x}_r, \boldsymbol{p}_i, \Phi_{\text{seq}}(S_{\Delta \boldsymbol{p}})) \xrightarrow{\Theta_{\text{NR}}} \Delta \boldsymbol{x}\,, \tag{13}$$

and for canonical radiance field, we formulate

$$F_c(\boldsymbol{x}_c, \boldsymbol{p}_i, \Phi_{\text{seq}}(S_{\Delta \boldsymbol{p}})) \xrightarrow{\Theta_{\text{NeRF}}} (\sigma, \boldsymbol{c})\,. \tag{14}$$

As MLP-based canonical representations suffer from slow convergence and require a deep deformation network to model human motion, we employ a triplane-based low-rank model [5] to represent 3D canonical space. This representation allows for rapid convergence while significantly reducing the depth of non-rigid deformation network (2 layers in our implementation), making it possible to achieve fast and high-quality canonical space reconstruction.

4.3 Localized Dynamic Context Encoder

Our designed pose-sequence-related condition enables the model to express variation for the same pose within different inertia states. However, it may also lead the model to learn spurious correlations between the high-dimensional input and the output. To address this issue, we propose a localized dynamic context encoder Φ_{seq} to embed the input context sequence.

Kinematically-guided Spatial Dependency: For a given context sequence $S_{\Delta \boldsymbol{p}}$ and an input coordinate $\boldsymbol{x}$, we begin by processing each $\Delta \boldsymbol{p} \in \mathbb{R}^{3\times K+3}$ independently. We leverage a reasonable physical assumption: the impact of a rotating joint is mainly restricted to the kinematic chains to which it belongs, so the motion and appearance of a point $\boldsymbol{x}$ should mainly depend on the joints within the same kinematic chains. With an input coordinate $\boldsymbol{x}$ and its associated blending weights $W_c(\boldsymbol{x}) \in \mathbb{R}^K$, we first identify its nearest joint $k = \arg\max W_c(\boldsymbol{x})$ to determine the kinematic chains $\boldsymbol{x}$ lies on. Then we gather all joints from the related kinematic chains and apply a binary mask to $\Delta_{\boldsymbol{p}}$ to filter out the joints outside the kinematic chains. We denote the masked conditions as $\hat{\Delta \boldsymbol{p}}$ and the masked sequence as $S_{\hat{\Delta \boldsymbol{p}}}$.

Spatial-Temporal Encoding: For $S_{\hat{\Delta \boldsymbol{p}}}$, we initially employ an MLP to reduce the spatial dimension from $3 \times K + 3$ to 16, resulting in a vector sequence size of $L \times 16$. To integrate temporal information, we flatten the sequence and apply a second MLP to project it to a vector dimension of 32. The resulting 32-dimensional vector is concatenated with the positional embedding of the coordinate and serves as input to the subsequent radiance field.

4.4 I3D-Human Dataset and Dynamic Human Evaluation

Human motion in lightweight wearings and at variable speeds is ubiquitous in daily lives. However, current multi-view avatar benchmarks such as ZJU-MoCap [28], PeopleSnapshot [3] and Human3.6M [11], are collected under controlled speeds and with tight-fitting garments, turning off inertia-induced dynamic effects. This limitation inhibits us from testing the model's ability to capture real-life dynamic effects. To address this gap and advance human avatars toward greater photorealism, we curated a new dataset named the Inertia-aware 3D Human (I3D-Human) dataset. Additionally, we design a new motion-based metric for evaluation purposes.

The I3D-Human Dataset: The dataset focuses on capturing variations in clothing appearance under approximately identical poses. Compared with existing benchmarks, we outfit the subjects in loose clothing such as dresses and light jackets and encourage movements involving acceleration or deceleration, such as sudden stops after spinning, swaying, and flapping sleeves. Our capturing equipment consists of 10 DJI Osmo Action cameras, shooting at a frame rate of 100fps while synchronized with an audio signal. The final processed dataset records 10k frames of sequence from 6 subjects in total. More details about the I3D-Human Dataset can be found in our supplemental materials.

Evaluating Dynamic Motion Effects: Traditional per-frame metrics such as PSNR measure static visual fidelity. Here we introduce a motion-based metric named dynamic motion error (DME) to assess the faithful portrayal of dynamic effects. While we lack ground-truth 3D motion trajectories, we can utilize optical flow to assess motion in pixel space. Formally, we denote the rendered image at the i-th time-step as $\boldsymbol{C}^i$ and its ground truth as $\boldsymbol{C}^i_{\text{gt}}$. We utilize the off-the-shelf optical flow estimator RAFT [35] to calculate the pixel translations between two steps in predictions and ground truths as follows:

$$\boldsymbol{f}^i_{\text{gt}} = \text{RAFT}(\boldsymbol{C}^{i-1}_{\text{gt}}, \boldsymbol{C}^i_{\text{gt}}), \tag{15}$$

$$\boldsymbol{f}^i = \text{RAFT}(\boldsymbol{C}^{i-1}, \boldsymbol{C}^i). \tag{16}$$

Then we compute the end-to-end point error (EPE) between $\boldsymbol{f}^i_{\text{gt}}$ and $\boldsymbol{f}^i$. The final DME for a whole test sequence is the average EPE across all time-steps:

$$\text{DME} = \frac{1}{T}\sum_{i=2}^{T} \left\| \boldsymbol{f}^i_{\text{gt}} - \boldsymbol{f}^i \right\|. \tag{17}$$

5 Experiments

5.1 Datasets

Given our central objective of imbuing neural implicit avatars with inertia-aware dynamic effects, we mainly examine and compare the methods on the I3D-Human dataset. We conduct experiments on 4 sequences from the dataset and

Table 1: Comparison with state-of-the-arts on I3D-Human: The scores are averaged over the four sequences. Dyco outperforms all baselines across all metrics, particularly excelling in the Dynamic Motion Error (DME). LPIPS* = LPIPS $\times 10^3$

Method	Novel View				Novel Pose			
	PSNR↑	SSIM↑	LPIPS* ↓	DME↓	PSNR↑	SSIM↑	LPIPS* ↓	DME↓
NeuralBody [26]	30.33	0.9681	60.89	5.08	28.80	0.9604	67.59	4.63
AniNeRF [27]	29.29	0.9662	61.95	5.73	28.48	0.9628	64.85	4.39
AniSDF [27]	29.20	0.9670	58.94	5.56	28.34	0.9632	62.18	4.48
HumanNeRF [38]	29.53	0.9678	42.04	5.25	28.78	0.9644	45.70	4.39
3DGS-Avatar [31]	30.62	0.9712	39.74	4.93	29.21	0.9658	44.61	4.26
Dyco (Ours)	**31.22**	**0.9738**	**34.54**	**4.52**	**30.12**	**0.9691**	**39.55**	**3.98**

Table 2: Comparison with state-of-the-arts on ZJU-MoCap: The scores are averaged over the 9 sequences. Dyco achieves comparable results with state-of-the-arts, preserving the ability of modeling tightly-clothed humans in slow and gradual motion. LPIPS* = LPIPS $\times 10^3$

Method	Novel View				Novel Pose			
	PSNR↑	SSIM↑	LPIPS* ↓	DME↓	PSNR↑	SSIM↑	LPIPS* ↓	DME↓
NeuralBody [26]	**33.59**	**0.9804**	36.03	**5.02**	29.68	0.9656	49.70	7.22
AniNeRF [27]	32.28	0.9750	47.53	5.44	29.86	0.9667	52.61	6.91
AniSDF [27]	32.16	0.9767	40.63	5.60	29.58	0.9677	45.63	6.96
HumanNeRF [38]	32.13	0.9783	22.30	5.78	29.62	0.9678	31.11	7.43
3DGS-Avatar [31]	32.81	0.9791	23.76	5.47	**30.01**	**0.9699**	**30.20**	**6.98**
Dyco (Ours)	32.80	0.9800	**20.32**	5.20	29.71	0.9681	31.22	7.22

split each sequence into around 600 training frames and 400 test frames. We take 4 cameras for training and the rest for testing. As it remains crucial to maintain performance in the scenario of tightly-clothed subjects in slow and controlled motion, we also compare our method with others on the mostly used ZJU-MoCap [29] benchmark. ZJU-MoCap comprises 9 multi-view RGB sequences. Following the training setup in previous work [25, 27, 29], we use 4 cameras for training and the rest 19 cameras for testing. Each sequence is divided into two consecutive clips, one for training and the other for testing unseen poses.

5.2 Evaluation

We compare the methods on two test splits, namely novel view and novel pose. Novel-view test set evaluates the observed poses in the training frames from novel camera views, assessing whether the model can learn a physically realistic 3D human under the training poses from the 2D RGB observations. Novel-pose test set renders unobserved poses from novel camera views, examining whether the learned 3D avatars can be plausibly animated under unseen poses. For evaluation metrics, we adopt traditional per-frame metrics including PSNR, SSIM, and LPIPS [41]. As explained in Sec. 4.4, we also propose a novel motion-based metric termed DME to measure the accuracy of motion dynamics rendering.

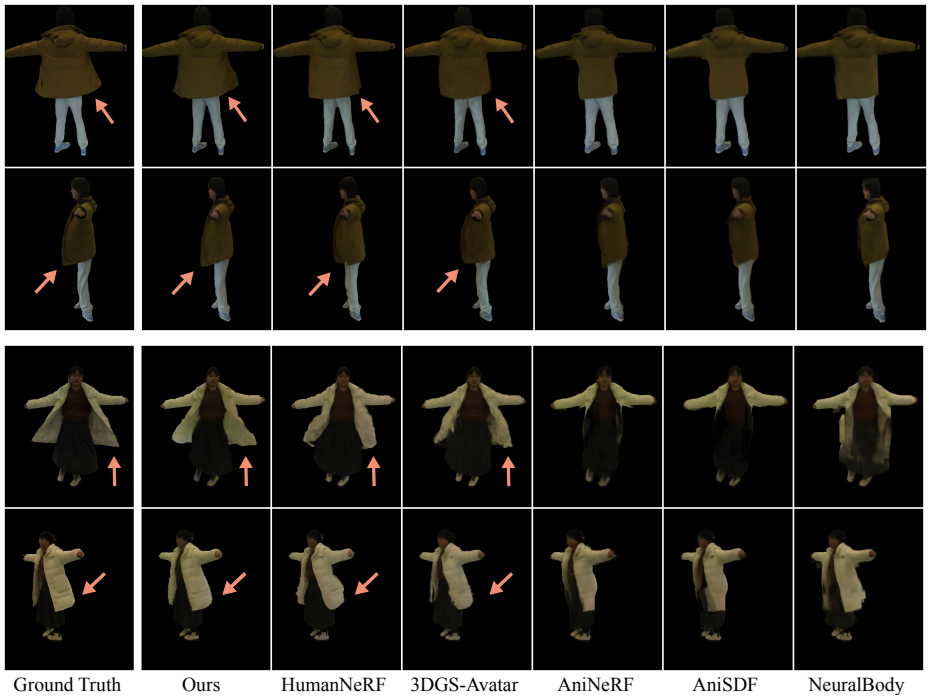


Fig. 3: Qualitative comparison on the novel view of I3D-Human dataset. Our method can render distinct appearance and deformation across various motion contexts. In contrast, HumanNeRF [38] and 3DGS-Avatar [31] struggle to capture the precise details. The other three baselines [26, 27] exhibit noticeable artifacts.

5.3 Comparison with State of the Arts

We compare with state-of-the-art counterparts, including Neural Body [26], AniNeRF [27], AniSDF [27], HumanNeRF [38] and 3DGS-Avatar [31]. The last two models are initially proposed for the monocular training setting so we adapt them to our multi-view setting. Note that the concurrent 3DGS-Avatar harnesses the advanced capabilities of Gaussian splatting representation [15].

I3D-Human: We show the comparisons on I3D-Human in Tab. 1 and Figs. 3 and 4. Our approach not only surpasses state-of-the-arts in terms of standard frame-wise similarity metrics but also achieves significantly lower DME, indicating its effectiveness in modeling inertia-induced dynamics. This improvement remains consistent across both novel view and novel pose test sets, indicating that the learned correlation between dynamic context and appearance variation in training frames can be extended to unseen poses. We will delve deeper into the generalization performance later.

ZJU-MoCap: To show Dyco maintains the capability of modeling motion with minimal inertia influence, we compare it with state-of-the-arts on ZJU-MoCap and report the metrics in Tab. 2. Despite the deliberate minimization of inertia effects in human motion within ZJU-MoCap, our method achieves the best

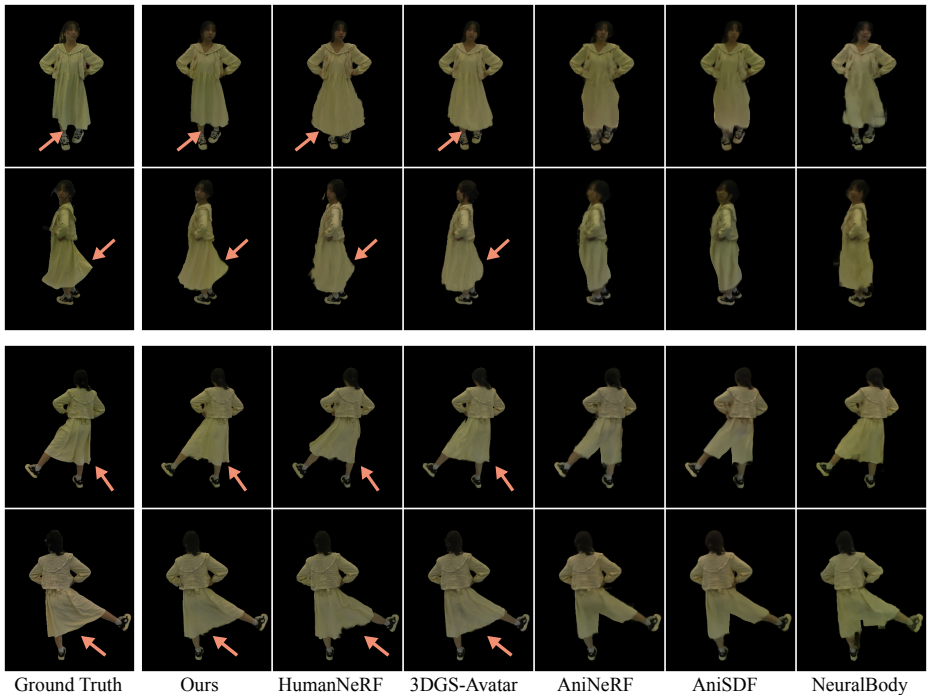


Fig. 4: Qualitative comparison on the novel pose of I3D-Human dataset.

LPIPS scores in the novel-view test set. Even for ZJU-MoCap, in some training frames with similar static poses, variations in cloth wrinkles can still occur due to divergent past motion trajectories. We show that Dyco outperforms other methods in capturing the variation in Fig. 5. In the novel pose test, our method only slightly underperforms HumanNeRF and 3DGS-Avatar. Qualitative comparison can be found in the supplementary material.

5.4 Ablations

Pose condition $\boldsymbol{p}_i$ provides global spatial information for the current frame during training, and its removal results in decreased performance, see Tab. 3.
Delta pose condition $S_{\Delta \boldsymbol{p}}$ incorporates temporal human inertia knowledge, which we believe is the core design for resolving ambiguities related to dynamic context. Discarding this component prevents the network from leveraging inertia information and degrades the performance, see Tab. 3.
Localization in dynamic context encoder reduces the complexity of $S_{\Delta \boldsymbol{p}}$ input by masking redundant dependencies. Without this module, the overfitted network will struggle in generalizing to novel poses, see Tab. 3.
Conditional input to $T_{\mathbf{NR}}$ and F_c: In Sec. 4.1, we analyze two types of dynamic context related ambiguities and propose to input delta pose $S_{\Delta \boldsymbol{p}}$ to both non-rigid transformation and canonical volume. Tab. 4 validates that inputting $S_{\Delta \boldsymbol{p}}$ into both modules helps resolve the ambiguities.

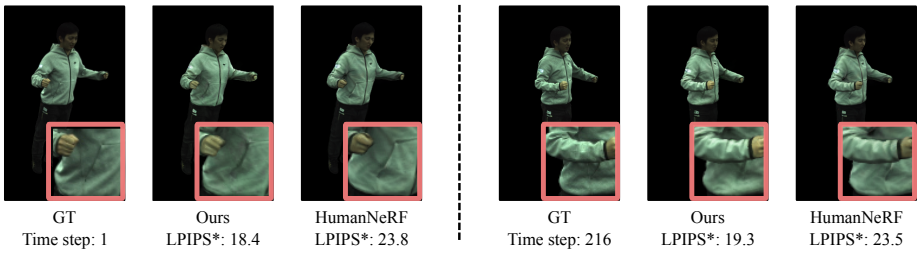


Fig. 5: The ZJU-MoCap dataset also exhibits ambiguous mapping from the static pose to various appearances. At two frames when the subject has similar poses, the clothing wrinkles can differ due to distinct past motion. Our method can reflect the variation, while HumanNeRF [38] generates similar patterns.

Table 3: Ablation on pose condition $\boldsymbol{p}_i$, delta pose condition $S_{\Delta \boldsymbol{p}}$ and localized spatial dependency.

Pose condition	Delta condition	Local	Novel View				Novel Pose			
			PSNR↑	SSIM↑	LPIPS* ↓	DME↓	PSNR↑	SSIM↑	LPIPS* ↓	DME↓
-	-	✓	29.53	0.9676	41.60	5.21	28.75	0.9639	45.28	4.45
✓	-	✓	30.40	0.9708	37.97	4.85	29.16	0.9656	43.67	4.32
-	✓	✓	30.89	0.9722	36.35	4.56	29.92	0.9678	41.85	4.02
✓	✓	-	30.37	0.9703	38.78	**4.42**	29.58	0.9666	42.73	**3.77**
✓	✓	✓	**31.22**	**0.9738**	**34.54**	4.52	**30.12**	**0.9691**	**39.55**	3.98

Table 4: Ablation on delta pose condition $S_{\Delta \boldsymbol{p}}$ in the non-rigid transformation and canonical volume.

$S_{\Delta \boldsymbol{p}}$ in T_{NR}	$S_{\Delta \boldsymbol{p}}$ in F_c	Novel View				Novel Pose			
		PSNR↑	SSIM↑	LPIPS* ↓	DME↓	PSNR↑	SSIM↑	LPIPS* ↓	DME↓
-	-	29.53	0.9676	41.60	5.21	28.75	0.9639	45.28	4.45
✓	-	31.13	0.9729	35.86	5.14	29.99	0.9678	41.52	4.26
✓	✓	**31.22**	**0.9738**	**34.54**	**4.52**	**30.12**	**0.9691**	**39.55**	**3.98**

5.5 Generalization to Novel Dynamic Context

To showcase that Dyco correctly models the correlation between inertia factors and appearance variations, we create novel dynamic context by varying the velocity of the training poses to study how it impacts the rendered results. Please refer to the supplemental video for a clearer visualization.

Varying velocity: We scale $S_{\Delta \boldsymbol{p}}$ with a factor α ranging from 0 to 2, where $\alpha > 1$ speeds up motion and use $\alpha S_{\Delta \boldsymbol{p}}$ as the sequence condition. Fig. 6 demonstrates Dyco's ability to accurately portray varying drape lifting amplitudes with different spinning speeds, with larger α yielding greater amplitudes.

Varying acceleration: We simulate an abrupt stop effect by interrupting poses midway and making subsequent poses identical. The modified pose sequence is then used as the input condition. As shown in Fig. 7, Dyco, when conditioned on

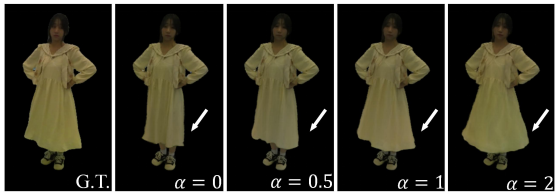


Fig. 6: Appearance variation under varying spinning velocity. We input the scaled delta-sequence $\alpha S_{\Delta \boldsymbol{p}}$ to the trained model to obtain different renderings.

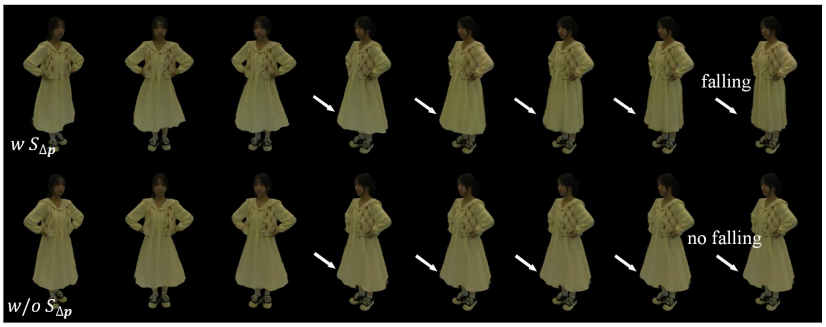


Fig. 7: Dyco correctly models the falling of the drape when the character stops abruptly during spinning, whereas the baseline keeps the drape still.

the novel pose sequence, accurately animates the falling of the hem. In contrast, the model without pose sequence inputs fails to depict this motion.

6 Conclusion

In this work, we introduce Dyco, a novel human motion modeling method that incorporates pose-sequence conditions to address appearance ambiguities resulting from different dynamic contexts. We argue that human appearance is influenced not only by pose conditions but also by past motion trajectories, which can be adequately captured by pose sequences. To mitigate overfitting caused by excessive reliance on delta pose, we design a localized dynamic context encoder. This approach allows us to resolve appearance ambiguities caused by dynamic context and improve the rendering quality of human bodies in loose attire. The I3D-Human dataset addresses previous oversights in datasets and advances research into real-life human motion.

Acknowledgements

This work is supported by the Shanghai Artificial intelligence Laboratory, in part by JSPS KAKENHI Grant Numbers 24K22318, 22H00529, 20H05951, JST-Mirai Program JPMJMI23G1.

References

1. Easymocap - make human motion capture easier. Github (2021), https://github.com/zju3dv/EasyMocap
2. SMPL: A skinned multi-person linear model, author=Loper, Matthew and Mahmood, Naureen and Romero, Javier and Pons-Moll, Gerard and Black, Michael J. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)
3. Alldieck, T., Magnor, M.A., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 8387–8397. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00875, http://openaccess.thecvf.com/content_cvpr_2018/html/Alldieck_Video_Based_Reconstruction_CVPR_2018_paper.html
4. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers, pp. 408–416 (2005)
5. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3d generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 16102–16112. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01565, https://doi.org/10.1109/CVPR52688.2022.01565
6. Cheng, W., Chen, R., Fan, S., Yin, W., Chen, K., Cai, Z., Wang, J., Gao, Y., Yu, Z., Lin, Z., et al.: Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19982–19993 (2023)
7. Dong, J., Fang, Q., Jiang, W., Yang, Y., Bao, H., Zhou, X.: Fast and robust multi-person 3d pose estimation and tracking from multiple views. In: T-PAMI (2021)
8. Freifeld, O., Black, M.J.: Lie bodies: A manifold representation of 3d human shape. In: Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I. Lecture Notes in Computer Science, vol. 7572, pp. 1–14. Springer (2012). https://doi.org/10.1007/978-3-642-33718-5_1, https://doi.org/10.1007/978-3-642-33718-5_1
9. Guan, P., Reiss, L., Hirshberg, D.A., Weiss, A., Black, M.J.: Drape: Dressing any person. ACM Transactions on Graphics (ToG) **31**(4), 1–10 (2012)
10. Hirshberg, D.A., Loper, M., Rachlin, E., Black, M.J.: Coregistration: Simultaneous alignment and modeling of articulated 3d shape. In: Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI. Lecture Notes in Computer Science, vol. 7577, pp. 242–255. Springer (2012). https://doi.org/10.1007/978-3-642-33783-3_18, https://doi.org/10.1007/978-3-642-33783-3_18
11. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014). https://doi.org/10.1109/TPAMI.2013.248, https://doi.org/10.1109/TPAMI.2013.248

12. Jiang, T., Chen, X., Song, J., Hilliges, O.: Instantavatar: Learning avatars from monocular video in 60 seconds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16922–16932 (2023)
13. Kajiya, J.T., Von Herzen, B.P.: Ray Tracing Volume Densities. ACM SIGGRAPH computer graphics **18**(3), 165–174 (1984)
14. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
15. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139:1–139:14 (2023). https://doi.org/10.1145/3592433, https://doi.org/10.1145/3592433
16. Larboulette, C., Cani, M.P., Arnaldi, B.: Dynamic skinning: adding real-time dynamic effects to an existing character animation. In: Proceedings of the 21st Spring Conference on Computer Graphics. p. 87–93. SCCG '05, Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1090122.1090138, https://doi.org/10.1145/1090122.1090138
17. Lewis, J.P., Cordner, M., Fong, N.: Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 811–818 (2023)
18. Li, R., Tanke, J., Vo, M., Zollhofer, M., Gall, J., Kanazawa, A., Lassner, C.: Tava: Template-free animatable volumetric actors (2022)
19. Lin, H., Peng, S., Xu, Z., Yan, Y., Shuai, Q., Bao, H., Zhou, X.: Efficient neural radiance fields for interactive free-viewpoint video. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
20. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: neural free-view synthesis of human actors with pose control. ACM Trans. Graph. **40**(6) (dec 2021). https://doi.org/10.1145/3478513.3480528, https://doi.org/10.1145/3478513.3480528
21. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. Communications of the ACM **65**(1), 99–106 (2021)
22. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
23. Noguchi, A., Sun, X., Lin, S., Harada, T.: Neural articulated radiance field. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5762–5772 (2021)
24. Patel, C., Liao, Z., Pons-Moll, G.: Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2020)
25. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: ICCV (2021)
26. Peng, S., Geng, C., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Zhou, X., Bao, H.: Implicit neural representations with structured latent codes for human body modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
27. Peng, S., Xu, Z., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Bao, H., Zhou, X.: Animatable implicit neural representations for creating realistic avatars from videos. TPAMI (2024)
28. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view

synthesis of dynamic humans. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 9054–9063. Computer Vision Foundation / IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00894, https://openaccess.thecvf.com/content/CVPR2021/html/Peng_Neural_Body_Implicit_Neural_Representations_With_Structured_Latent_Codes_for_CVPR_2021_paper.html

29. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021)
30. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: a model of dynamic human shape in motion. ACM Trans. Graph. **34**(4), 120:1–120:14 (2015). https://doi.org/10.1145/2766993, https://doi.org/10.1145/2766993
31. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting (2024)
32. Rohmer, D., Tarini, M., Kalyanasundaram, N., Moshfeghifar, F., Cani, M.P., Zordan, V.B.: Velocity skinning for real-time stylized skeletal animation. Computer Graphics Forum **40** (2021), https://api.semanticscholar.org/CorpusID:233210320
33. Shuai, Q., Geng, C., Fang, Q., Peng, S., Shen, W., Zhou, X., Bao, H.: Novel view synthesis of human interactions from sparse multi-view videos. In: SIGGRAPH Conference Proceedings (2022)
34. Su, S., Yu, F., Zollhöfer, M., Rhodin, H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 12278–12291 (2021), https://proceedings.neurips.cc/paper/2021/hash/65fc9fb4897a89789352e211ca2d398f-Abstract.html
35. Teed, Z., Deng, J.: RAFT: recurrent all-pairs field transforms for optical flow. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12347, pp. 402–419. Springer (2020). https://doi.org/10.1007/978-3-030-58536-5_24, https://doi.org/10.1007/978-3-030-58536-5_24
36. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021)
37. Wang, S., Schwarz, K., Geiger, A., Tang, S.: Arah: Animatable volume rendering of articulated human sdfs. In: European conference on computer vision. pp. 1–19. Springer (2022)
38. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16210–16220 (June 2022)
39. Wu, M., Wang, Y., Hu, Q., Yu, J.: Multi-view neural human rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1682–1691 (2020)
40. Yu, Z., Cheng, W., Liu, X., Wu, W., Lin, K.Y.: Monohuman: Animatable human neural field from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16943–16953 (2023)

41. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
42. Zheng, Z., Zhao, X., Zhang, H., Liu, B., Liu, Y.: Avatarrex: Real-time expressive full-body avatars. ACM Trans. Graph. **42**(4) (jul 2023). https://doi.org/10.1145/3592101, https://doi.org/10.1145/3592101