NDF: Neural Deformable Fields for Dynamic Human Modelling

Ruiqi Zhang and Jie Chen✉

Department of Computer Science, Hong Kong Baptist University
{csrqzhang, chenjie}@comp.hkbu.edu.hk

Abstract. We propose Neural Deformable Fields (NDF), a new representation for dynamic human digitization from a multi-view video. Recent works proposed to represent a dynamic human body with shared canonical neural radiance fields which links to the observation space with deformation fields estimations. However, the learned canonical representation is static and the current design of the deformation fields is not able to represent large movements or detailed geometry changes. In this paper, we propose to learn a neural deformable field wrapped around a fitted parametric body model to represent the dynamic human. The NDF is spatially aligned by the underlying reference surface. A neural network is then learned to map pose to the dynamics of NDF. The proposed NDF representation can synthesize the digitized performer with novel views and novel poses with a detailed and reasonable dynamic appearance. Experiments show that our method significantly outperforms recent human synthesis methods.

Keywords: neural implicit representation, volumetric rendering, novel view synthesis, dynamic motion, human shape and appearance modelling

1 Introduction

Vision-based human performance capture has seen great progress in recent years due to fast development in both hardware and reconstruction algorithms like novel learning-based representation. It enables a wide variety of applications such as tele-presence, sportscast, and mixed reality. The enduring pandemic restricts our travel and public activities, which makes human performance digitization a research topic with great social and economic implications.

Human performance digitization can be roughly divided into human performance capture and human animation. Traditionally, to achieve high-fidelity human performance capture including geometry and texture reconstruction, dense camera rigs [5,8,9] and controlled lighting conditions [2,6] are required. These systems are extremely bulky and expensive, which limits their popularity. Nevertheless, these conventional capture systems could still fail under multi-person scenarios due to severe occlusion, which leads to ambiguity in appearance, pose, and motion sampling. After performance capture, human animation requires skilled artists to manually create a skeleton suitable for the human model and

carefully design skinning weights [11] to achieve realistic animation, which requires countless human labor.

This paper aims to reduce the cost and improve the flexibility of human performance digitization. Many recent works have investigated the potential of neural implicit fields in novel view synthesis. NeRF [20] proposed a neural implicit representation that can be effectively learned from multi-view images. The neural implicit representation is rendered to realistic images from novel views with volume rendering. However, NeRF has a high requirement for the camera numbers and it can only model a static scene which does not apply to multi-view videos of dynamic humans. To extend NeRF to dynamic scenes, an effective idea is to aggregate all observations over different video frames [26,22,24,23,12]. D-NeRF [26] and Nerfies [22] decompose a reconstruction into a canonical neural radiance field and a set of deformation fields that transform points in observation space to canonical space. To further simplify the learning of the deformation fields, Animatable NeRF [24] resorts to a parametric human body model as a strong geometry prior to the deformation fields. However, we claim that the current design of a shared canonical space and deformation fields prevents these methods from learning large movements and detailed geometry changes such as wrinkles of clothes as shown in the experiment.

To solve the above problems, rather than learning shared canonical neural radiance fields from multi-view videos, we use Neural Deformable Fields (NDF) to represent a dynamic scene. Specifically, we unwrap observation space to NDF space using the surface of a parametric body model as reference. NDF space is automatically aligned across frames and we further adopt the skeletal pose as posterior condition to model the dynamic changes. As a result, NDF space is more compact than the original observation space and it can model the dynamic changes caused by different poses. After training, we are able to animate the performer to different views and poses with a high degree of realism.

We evaluate our method on ZJU-MoCap [25] and DynaCap [7] datasets that capture dynamic humans in complex motions with synchronized cameras. The results show that our method can achieve high-fidelity reconstructions, especially for realistic dynamic changes in novel pose synthesis. The code is avaliable at `https://github.com/HKBU-VSComputing/2022_ECCV_NDF`.

In summary, the contributions of this paper are following:

- We propose a compact novel representation called NDF, which can model the dynamic changes caused by different poses.
- The experiment results demonstrate significant improvement on the novel pose synthesis task, especially the detailed and realistic dynamic changes caused by different poses.

2 Related Works

Learning-based Scene Representations. According to the dimensionality of representation, several paradigms have been investigated for 3D content embedding in the context of image-based novel view synthesis. Multiplane image (MPI)

[31,19], voxels [28,16], point cloud [1,3], and neural radiance fields [20,30,18,4,13] have all been under intense research focus recently. MPI learns scene representation in the form of fronto-parallel color and α planes, and novel views are rendered via homography-wraping. Sitzmann et al. [28] proposed to learn a deep-voxel representation by dividing the 3D space into discrete 3D units that embed learned features, which was further replaced with a continuous learnable function [29]. Mildenhall et al. [20] proposed to represent the scene as a neural radiance field (NeRF) by directly mapping a continuous 5D coordinate to the volume density and view-dependent emitted radiance. NeRF has special advantages in that it can represent a continuous scene in arbitrary resolution and it can be effectively learned from multi-view images. Our method follows NeRF to reconstruct scenes from images and further extends it to dynamic scenes.

Neural Implicit Representation for Human. Habermann et al. [7] leverage a 3D scanned person-specific template to learn motion-dependent geometry as well as motion- and view-dependent dynamic textures from multi-view videos. The requirement of a high-quality 3D scanning restricted its use. Several recent works resort to learning a shared representation via deformable functions (in the form of NeRF [26,22,27,21]). Restricted by the design choice of the function, it is difficult for these methods to model relatively large movements efficiently and they show limited generalizability to novel poses. Liu et al. [15] learns a person-specific embedding of the actor's appearance given a monocular video and a textured mesh template of the actor. Neural Body [25] learns neural representations over the same set of latent codes anchored to the deformable human model SMPL [17], and naturally integrate observations across frames. The sparsity allows it to effectively aggregate observations across frames but the result shows it losses details like wrinkles of clothes. Neural Actor [14] learns an unposed implicit human model via inverse linear blend skinning functions (LBS). The model cannot handle surface dynamics and certain geometric information has been lost during the generation of 2D texture maps. Animatable NeRF [24] can animate the performer to novel poses however it requires fine-tuning on the novel pose frames. This would be impossible when applied to a completely novel pose that the performer has never done. Our method does not require fine-tuning and can be directly applied to completely novel poses after training.

3 Proposed Method

Problem Setup. Given a training set of T-frame multi-view video of a dynamic human target over a sparse set of K synchronized and calibrated cameras: $\mathcal{I} = \{I_t^k\}$ $(t = 1 \dots T, k = 1 \dots K)$, our goal is to digitize this performer using the proposed Neural Deformable Field (NDF) representation for both novel-view synthesis (NVS) and novel pose synthesis (NPS). Specifically, in the NVS task, we synthesize free-viewpoint renderings of the performance with novel camera angles. In the NPS task, we synthesize renderings with novel, unseen poses.

We build the NDF representation based on the state-of-the-art volumetric rendering model - Neural Radiance Field (NeRF) [20], which predicts the color $\mathbf{c}$

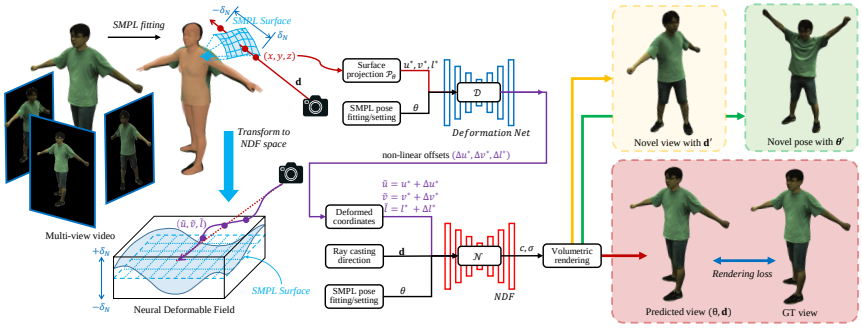


Fig. 1. Overview of proposed method. We query points in observation space, infer their densities and colors in NDF space and adopt volume rendering technique to synthesize images. For a given point $\mathbf{x} = (x, y, z)$ in observation space, we project it to NDF space with surface projection $\mathcal{P}_\theta$ and further adopt deformation net $\mathcal{D}$ to slightly adjust the projection point $\tilde{\mathbf{u}} = (\tilde{u}, \tilde{v}, \tilde{l})$ in NDF space. A radiance field is then learned to predict the color $\mathbf{c}$ and density σ for the point $\tilde{\mathbf{u}}$ in the unwrapped NDF space. The predicted color $\mathbf{c}$ and density σ is then assigned back to the observation-space point $\mathbf{x}$. Finally, volume rendering is used to synthesize an image in the observation space.

and density σ at spatial location $\mathbf{x} \in \mathbb{R}^3$ and view direction $\mathbf{d} \in \mathbb{S}^2$ via a neural network $\mathcal{F}$: $(\mathbf{x}, \mathbf{d}) \mapsto (\mathbf{c}, \sigma)$. Subsequently, volumetric rendering functions are used to render the final pixel color. The differentiable rendering process enables optimization via comparing the output image with ground truth without 3D supervision. However, there are mainly two challenges in this setting. First, in our problem setup, only $K = 4$ cameras are used, which is much less than what is sufficient to train a NeRF network. Second, due to the dynamic property of the human target, directly training a NeRF with all the frames will always cause artifacts and produce a coarse result.

To address these challenges, NDF fits a parametric human body model SMPL to associate 3D points among different video frames and learns a neural implicit field wrapped around and driven by the SMPL surface:

$$\mathcal{N} : (\mathcal{D}(\mathcal{P}_\theta(\mathbf{x})), \mathbf{d}, \theta) \mapsto (\mathbf{c}, \sigma), \tag{1}$$

where $\mathcal{P}_\theta$ is a projection function which projects a point's spatial location $\mathbf{x}$ to NDF space conditioned on the posed SMPL model with parameter θ. $\mathcal{D}$ is a non-linear deformation function which keeps the surface continuity in the projection process. With the spatial alignment reference provided by the SMPL surface, NDF efficiently accumulates visual observations from the multi-view video frames; and given the strong geometry prior, NDF learns a geometry-guided field instead of a volume, which greatly reduces the learning complexity, leading to a much higher modelling efficiency. The details of each module will be introduced in this section.

3.1 SMPL as Projection Reference with Non-linear Deformation

To decrease NeRF's high requirement of camera numbers, a typical solution is to learn a deformation function $\Phi_t(\mathbf{x}) : \mathbb{R}^3 \mapsto \mathbb{R}^3$ to map sample points $\mathbf{x}$ in frame t to a shared canonical space [24] [26]. However, restricted by current design, these methods cannot deal with large movements or detailed geometry changes such as clothes wrinkles. To overcome these drawbacks, we resort to the texture map of SMPL as a reference to align 3D points across different frames and jointly train an integral NeRF model.

SMPL [17] is a skinned vertex-based model, which is defined as a function of shape parameters β, pose parameters θ and a rigid transformation $\mathbf{W}$ using Linear Blending Skinning (LBS). The template model $\bar{\mathbf{T}}$ includes pre-defined 6890 vertices and their connections. With the pose-blend shape $B_P(\theta)$ and shape-blend shape $B_S(\beta)$, the posed mesh $M(\theta, \beta)$ is got from the following equation:

$$M(\theta, \beta) = \mathbf{W}(\bar{\mathbf{T}} + B_S(\beta) + B_P(\theta)). \tag{2}$$

In this paper, we assume the posed mesh is pre-computed from the multi-view video and use the texture map of this mesh to conduct the projection function $\mathcal{P}_\theta$ from observation space to NDF space.

Coordinates Projection. As shown in Figure. 1, a 3D point $\mathbf{x} = (x, y, z)$ is projected to a point $\mathbf{u}^* = (u^*, v^*, l^*)$ in the *unwrapped* Neural Deformable Fields (NDF) space with the projection function $\mathcal{P}_\theta : \mathbf{x} \mapsto \mathbf{u}^*$. $\mathcal{P}_\theta$ first projects the point $\mathbf{x}$ to the closest point $\mathbf{x}' \in \mathbb{R}^3$ on the fitted SMPL surface. $\mathbf{x}'$ has a 2D texel coordinate (u^*, v^*) which is defined over SMPL's texture map and is calculated via:

$$(u^*, v^*, f^*) = \underset{u,v,f}{\arg\min} \|\mathbf{x} - B_{u,v}(\mathcal{V}_{[\mathcal{F}(f)]})\|_2^2, \tag{3}$$

where $f \in \{1 \dots N_F\}$ is the triangle index, $\mathcal{V}_{[\mathcal{F}(f)]}$ is the three vertices of triangle $\mathcal{F}(f)$, $(u, v) : u, v \in [0, 1]$ are the texel coordinates on the texture map and $B_{u,v}(\cdot)$ is the barycentric interpolation function. SMPL is designed for modelling skinned human body and cannot capture surface dynamic changes. To model the dynamic geometry that deviates from the SMPL surface, we extend NDF to 3 dimensions with the euclidean distance l^* between $\mathbf{x}$ and $\mathbf{x}'$ being the third dimension.

Non-linear Deformation. We have projected an observation-space point $\mathbf{x}$ to $\mathbf{u}^*$ in NDF space using the UV coordinate of its nearest point on the SMPL surface as a reference. However, the continuous real surface will become discontinuous after projection. As shown in Figure 2(b), the two yellow points located on the continuous real surface in observation space will be closest to the same vertex on the SMPL surface if they locate in the same intersection of surface normals. After projection, the two yellow points will have the same u^*, v^* but

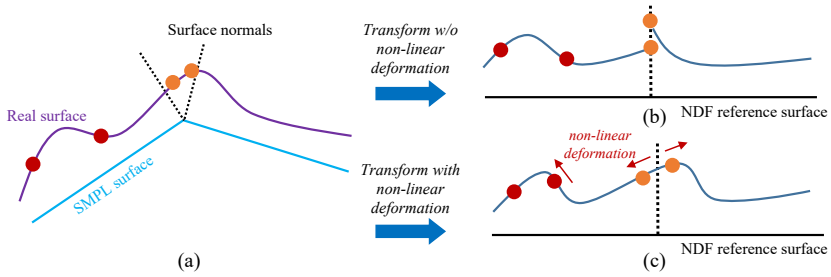


Fig. 2. A simplified 2D demonstration of transformation from the (x, y, z) camera coordinates (a) to the (u, v, l) NDF coordinates with and without non-linear deformation in (c) and (b), respectively.

different l^* in the NDF space. This will cause discontinuity at (u^*, v^*) and hinder the learning of neural radiance fields. To solve this problem, we adopt a deformation net to slightly adjust the projection coordinate. As shown in Figure 2(c), this non-linear deformation can unwrap the surface fragment between the surface normal interval and the continuity of the real surface can be maintained. Formally, the deformed projection location $\tilde{\mathbf{u}} = (\tilde{u}, \tilde{v}, \tilde{l})$ is described as following:

$$\triangle u^*, \triangle v^*, \triangle l^* = \mathcal{D}(\gamma_u(u^*, v^*, l^*), \theta), \tag{4}$$

$$\tilde{u}, \tilde{v}, \tilde{l} = u^* + \triangle u^*, v^* + \triangle v^*, l^* + \triangle l^*, \tag{5}$$

where $\mathcal{D}(\cdot)$ is the deformation net and $\gamma_u(\cdot)$ is the position embedding of $\mathbf{u}^*$. Note that the deformation aims to maintain the surface continuity in projection, but not to align points to a shared canonical space as in D-NeRF [26] and Nerfies [22].

3.2 Neural Deformable Fields

Rendering. For a given 3D spatial location $\mathbf{x}$ along the target camera's tracing ray direction $\mathbf{d}$, a point $\tilde{\mathbf{u}}$ will be found in the NDF space via projection and non-linear deformation as described above. The density for the point $\mathbf{x}$ will be estimated using an MLP M_σ: $\sigma(\mathbf{x}) = M_\sigma(\gamma_u(\tilde{\mathbf{u}}), \theta)$. The color will be estimated with another MLP M_c: $c(\mathbf{x}) = M_c(\gamma_u(\tilde{\mathbf{u}}), \gamma_d(\mathbf{d}), \theta)$, with an additional embedding $\gamma_d(\mathbf{d})$ for viewing direction, which ensures view-dependent effects.

The final image will be rendered via volumetric rendering [10] using numerical quadrature with N consecutive samples $\{x_1, \ldots, x_N\}$ along the tracing ray:

$$I_{out} = \sum_{n=1}^{N} \Big(\prod_{m=1}^{n-1} e^{-\sigma(\mathbf{x_m}) \cdot \delta_m} \Big) \cdot (1 - e^{-\sigma_n \cdot \delta_n}) \cdot c(\mathbf{x_n}). \tag{6}$$

Here $\delta_n = ||\mathbf{x_n} - \mathbf{x_{n-1}}||_2$ denotes the quadrature segment along the ray.

Geometry-guided Sampling Strategy. To further facilitate the learning process of NDF, we use the fitted SMPL as geometry guidance to sample points more effectively and cancel the hierarchical sampling adopted in the original NeRF. Specifically, as shown in Figure 1, we take uniform samples but only accept samples if the projection distance l^* is smaller than a hyper-parameter δ_N.

Remark. NDF representation is lightweight, detailed, and intuitive. As compared with volumetric representations, its underlying geometrical linkage is well-defined by posed SMPL, resulting in reduced dimensionality for geometry reasoning, therefore significantly reducing model complexity and is much easier to train. The feature space of NDF span the whole UV dimension, which records much more details compared with Neural Body [25], where shared canonical features are only located at SMPL vertices. By learning neural radiance fields conditioned on the pose, NDF can recover more intuitive dynamics related to changing pose rather than having to learn how to change query position in the canonical space through a per-frame deformation field like in Neural Actor [14].

3.3 Deformable Fields for Novel Pose Synthesis

Pose-driven NeRF By projecting points from the observation space to the NDF space, we are able to jointly learn a shared neural radiance field across frames. However, this representation would be only capable to capture a static geometry though it can be deformed to different poses. To model the dynamic change of human body geometry, we resort to the skeletal pose of SMPL as the posterior to infer the dynamic changes, i.e. we change the model from simply learning $\mathcal{N} : (\mathcal{D}(\mathcal{P}_\theta(\mathbf{x})), \mathbf{d}) \mapsto (\mathbf{c}, \sigma)$ to learning $\mathcal{N} : (\mathcal{D}(\mathcal{P}_\theta(\mathbf{x})), \mathbf{d}, \theta) \mapsto (\mathbf{c}, \sigma)$, where θ is the pose parameters of SMPL. In SMPL, the pose parameters θ is the axis-angle representation of the relative rotation of part k with respect to its parent in the kinematic tree. Besides being used for changing pose, θ is also used to generate a pose-blend shape that describes the shape deformation caused by different poses. Inspired by this, we infer from pose θ the dynamics of the scene. In practice, we apply an additional feature extractor to extract high-level features of pose parameters which contain significantly more information than the pure pose parameters. The extracted pose features are then concatenated with the position embedding of $\tilde{\mathbf{u}}$ and fed into the following neural networks.

Animation After training, NDF can be generalized to novel views or poses that do not occur in the training data $\mathcal{I}$. Specifically, given a viewing direction $\mathbf{d}$, a shape parameter β and a skeletal pose θ got from a motion capture system or designed by hand, we calculate the mesh vertices through Equation 2. Then we sample points around the SMPL surface and render an image viewing from $\mathbf{d}$ with Equation 6.

Remark. NDF does not need to be fine-tuned on novel pose images compared with Animatable NeRF [25] and can be applied to only sparse cameras compared with Neural Actor [14], where dense cameras are needed to pre-compute a realistic texture map. This animation ability only from sparse cameras would have a wide range of potential applications in VR or the metaverse.

4 Experiment

4.1 Dataset and Metrics

ZJU-MoCap [25] records multi-view videos with 21 synchronous cameras and collects the shape parameters of SMPL as well as the global translation and the SMPL's pose parameters with an off-the-shelf SMPL tracking system [32]. Following [25], we choose 9 sequences and 4 uniformly distributed cameras are used for training and the remaining cameras for testing. The video clips for evaluating novel view synthesis and novel pose synthesis are also the same with [25].
DynaCap. To further evaluate the generalization ability of our method, we select two sequences D1 and D2 from the DynaCap dataset [7]. These two sequences record a performer with over 50 synchronous cameras. We fit neutral SMPL to these cameras using [32] and uniformly select 10 cameras for training and 5 cameras for testing.
Metrics. Following typical protocols [20] and works most related to us [24] [25], we evaluate our method on image synthesis using two metrics: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

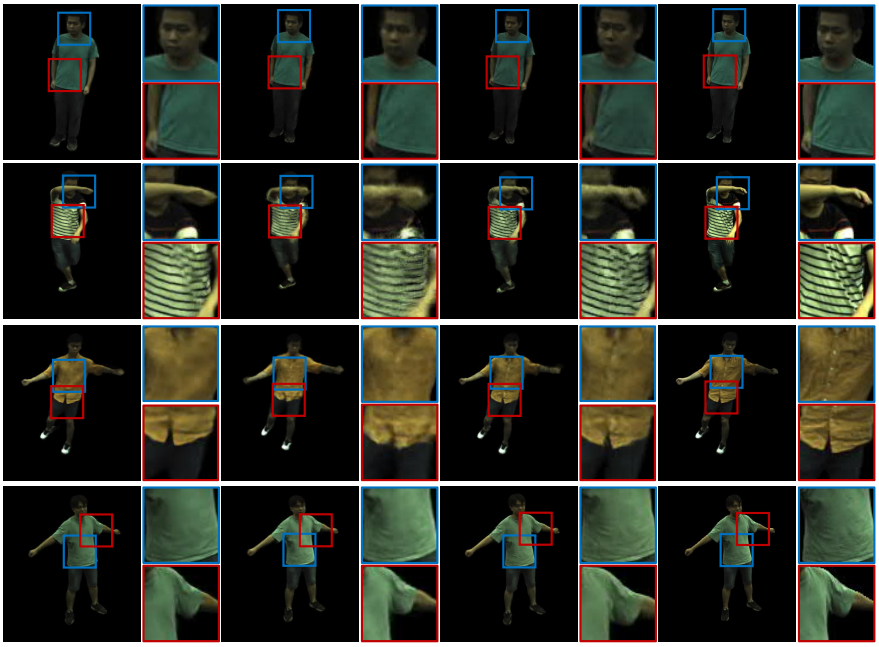
4.2 Performance on NVS and NPS

We compare our method with state-of-the-art view synthesis methods [25,24] that also use SMPL models and can handle dynamic scenes. Neural Body [25] represents the dynamic scene with an implicit field conditioned on a shared set of latent codes anchored on the vertices of SMPL and renders the images using volume rendering. Animatable NeRF [24] predicts the blend weights for each sample point and aggregates observations across frames to a shared canonical representation and further improves on novel pose synthesis by fine-tuning on novel pose images. All methods train a separate network for each scene.

Evaluation on novel view synthesis. Table 1 shows the comparison of our method with Neural Body [25] and Animatable NeRF [24] on ZJU-MoCap dataset. Our method outperforms Animatable NeRF [24] by a margin of 0.49 in terms of the PSNR metric and 0.01 in terms of the SSIM metric. It also performs close to Neural Body. Moreover, our method maintains its superiority when applied to DynaCap dataset as shown in Table 3.

Figure 3 presents the qualitative comparison of our method with [25,24] on the ZJU-MoCap dataset. Both [25] and [24] have difficulty in recovering fine details of the dynamic scene. Neural Body [25] turns to over-smooth the result as shown in the third person and the fourth person of Figure 3. The clothes seam of the third person almost disappears and the small wrinkles on the clothes of the fourth person also disappear. Animatable NeRF [24] shows more artifacts as the blur of the first person's face and the second person's clothes. In contrast, our method can always recover realistic details like the hem of the third person.

Figure 5 further presents the qualitative comparison on the DynaCap dataset. For the first two rows of novel view synthesis, our method can always recover



(a) Neural Body (b) Animatable NeRF (c) NDF (ours) (d) Ground Truth

Fig. 3. Qualitative results of novel view synthesis on the ZJU-MoCap dataset.

realistic details. For the second row, Neural Body [25] losses wrinkles on the back and Animatable NeRF [24] suffers from artifacts. While our method can reproduce high-quality wrinkles on the back.

Evaluation on novel pose synthesis. Table 2 shows the comparison of our method with Neural Body [25] and Animatable NeRF [24] on novel pose synthesis. The result shows that our method outperforms compared method on most of the sequences and performs best for the average metrics. Note that Animatable NeRF [24] needs to be fine-tuned on novel pose images while our method can be directly applied to novel pose synthesis.

The qualitative results are shown in Figure 4. Neural Body [25] learns latent codes for training frames and does not model the dynamic change with respect to poses, thus it always suffers from artifacts when applied to novel pose synthesis. Though fine-tuned on novel pose images, Animatable NeRF [24] has difficulty in modelling large movements and also leads to blur result. Our method is able to recover details such as the hem of clothes for the third person even when applied to novel pose synthesis.

The bottom 2 rows of Figure 5 show the qualitative comparison on the DynaCap dataset. Neural Body [25] fails to recover the face of the second person

Table 1. Results of novel view synthesis on the ZJU-MoCap dataset in terms of PSNR and SSIM (higher is better). "NB" means Neural Body. "AN" means Animatable NeRF. The best and the second best results are highlighted in red and blue, respectively.

	PSNR			SSIM		
	NB [25]	AN [24]	OURS	NB [25]	AN [24]	OURS
313	30.39	29.27	29.84	0.970	0.962	0.969
315	26.53	24.22	25.71	0.954	0.922	0.949
377	27.49	26.63	26.85	0.950	0.941	0.946
386	28.66	26.78	28.21	0.928	0.891	0.923
387	25.52	24.75	24.52	0.922	0.913	0.911
390	27.25	26.19	26.33	0.920	0.915	0.913
392	29.41	27.79	28.40	0.944	0.928	0.937
393	27.41	26.06	26.73	0.934	0.916	0.926
394	28.65	27.53	27.98	0.939	0.925	0.932
average	27.92	26.58	27.17	0.940	0.924	0.934

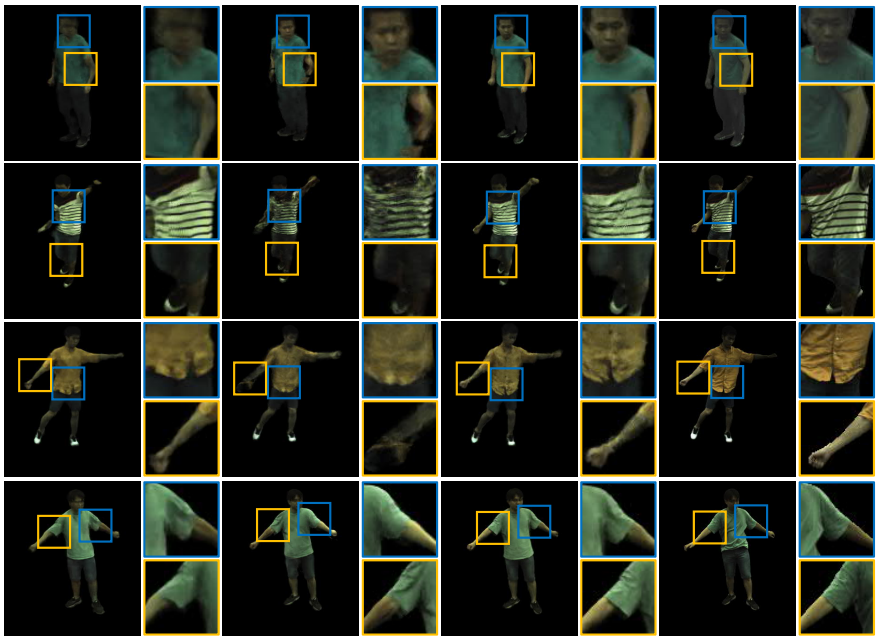
Table 2. Results of novel pose synthesis on the ZJU-MoCap dataset in terms of PSNR and SSIM (higher is better)

	PSNR			SSIM		
	NB [25]	AN [24]	OURS	NB [25]	AN [24]	OURS
313	23.49	23.61	23.29	0.898	0.908	0.903
315	19.38	19.45	19.50	0.847	0.854	0.857
377	23.89	25.03	25.18	0.914	0.927	0.928
386	25.63	25.14	26.33	0.877	0.878	0.893
387	21.75	22.94	22.41	0.865	0.892	0.880
390	23.81	24.51	24.11	0.868	0.889	0.881
392	25.66	24.15	25.62	0.908	0.900	0.914
393	23.30	23.97	24.03	0.891	0.899	0.902
394	23.76	24.29	24.29	0.876	0.893	0.890
average	23.41	23.68	23.86	0.883	0.893	0.894

and Animatable NeRF produces severe artifacts on the face and hands, while our method can produce reliable realistic face and hands for the second person.

4.3 Temporal Consistency

NDF uses pose as condition which changes continuously and smoothly over time, while Neural Body and Animatable NeRF separately learn appearance codes for different frames. This endows NDF with better temporal consistency as can be seen from Figure 6. The red circles point out the flickering part of previous methods while NDF always shows better temporal consistency.



(a) Neural Body (b) Animatable NeRF (c) NDF (ours) (d) Ground Truth

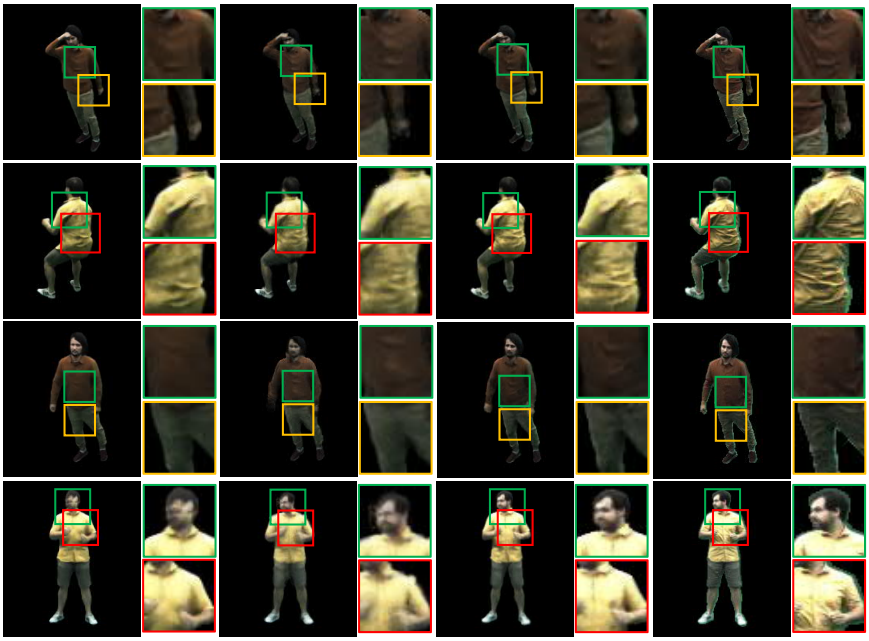
Fig. 4. Qualitative results of novel pose synthesis on the ZJU-MoCap dataset.

4.4 Ablation Study

We conduct ablation studies on one subject (313) of the ZJU-MoCap [25] dataset in terms of the novel view synthesis and novel pose synthesis performance. We test the impact of the surface distance $\tilde{l}$, the impact of using pose as the condition to model dynamic change, the impact of projection from observation space to NDF space, the impact of deformation net, and the reliance of specific reference surface to show the effectiveness of our choice.

Impact of the surface distance $\tilde{l}$ in NDF rendering. To capture the dynamic geometry that cannot be captured by naked SMPL surface, we adopt the distance from a query point to its closest point on SMPL as the third dimension to model the NDF space as a field rather than a naked SMPL surface. To test the impact of this design, we only sample points on the SMPL surface thus the $\tilde{l}$ for projected points are all 0. As shown in the first column of Figure 7 and Table 4, modelling the NDF space as naked SMPL surface causes severe artifacts, especially for clothes that cannot be captured by SMPL surface.

Using pose as condition to model dynamic change. In this experiment, we cancel using pose as the condition and jointly learn a shared canonical NDF



(a) Neural Body (b) Animatable NeRF (c) NDF (ours) (d) Ground Truth

Fig. 5. Qualitative results of novel view synthesis and novel pose synthesis on the DynaCap dataset. Top 2 rows: novel view synthesis. Bottom 2 rows: novel pose synthesis.

for all frames. As shown in the second column of Figure 7, the model cannot handle dynamic changes and produces blur rendering at dynamic regions.

Impact of projection from observation space to NDF space. In this experiment, we directly use the observation-space coordinates (x, y, z) as input to the neural network. The model needs to learn the mapping from pose to the whole 3D volume however it is severely difficult. As shown in the third column of Figure 7, though the model can synthesize novel views of the performer, it totally fails on novel pose synthesis.

Impact of deformation net. The deformation net aims to maintain the surface continuity after projection as claimed in Figure 2. As shown in the fourth column of Figure 7, the face and shoes become slightly noisier and we infer this is because the triangle surfaces of SMPL are small and dense on the face and feet. The result confirms the effectiveness of our design of the deformation net.

Reliance of specific reference surface. NDF does not rely on a specific texture map as the reference surface. To validate this, we replace the default

Table 3. Results of novel view synthesis and novel pose synthesis on the DynaCap dataset in terms of PSNR and SSIM (higher is better).

	PSNR			SSIM		
	NB [25]	AN [24]	OURS	NB [25]	AN [24]	OURS
novel view	23.96	22.99	24.73	0.889	0.872	0.904
novel pose	21.19	20.98	21.42	0.828	0.828	0.841

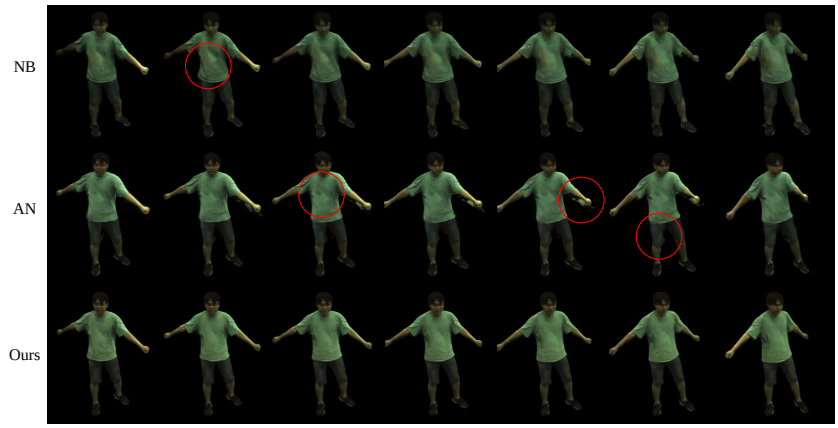


Fig. 6. Qualitative results of continuous frames to show temporal consistency. The red circles point out the flickering part of previous methods.

texture map of SMPL with a self-designed texture map which can be found in the supplementary material. We cut the seam of the SMPL mesh in Blender and unwrap the mesh into one piece in the UV space. As shown in the fifth column of Figure 7 and Table 4, with the 1-piece texture map as reference surface, the face becomes slightly blurred but the whole effect is still robust. This is because the UV region corresponding to the face occurs to be much smaller than in the default texture map of SMPL. The result shows that our method does not rely on a specific texture map and a self-designed texture map can also be used to unwrap points from observation space to NDF space.

5 Limitations and Future Works

Learning neural radiance fields conditioned on pose in NDF space enables us to obtain impressive performances on human digitization. However, our method has a few limitations. 1) Currently our method has a high requirement for the fitting effect of SMPL. Hopefully, in the future, we can integrate the fitting of SMPL in the pipeline and make the fitting and rendering benefit from each other. 2) In more complex scenes, the dynamic content depends both on pose and temporal

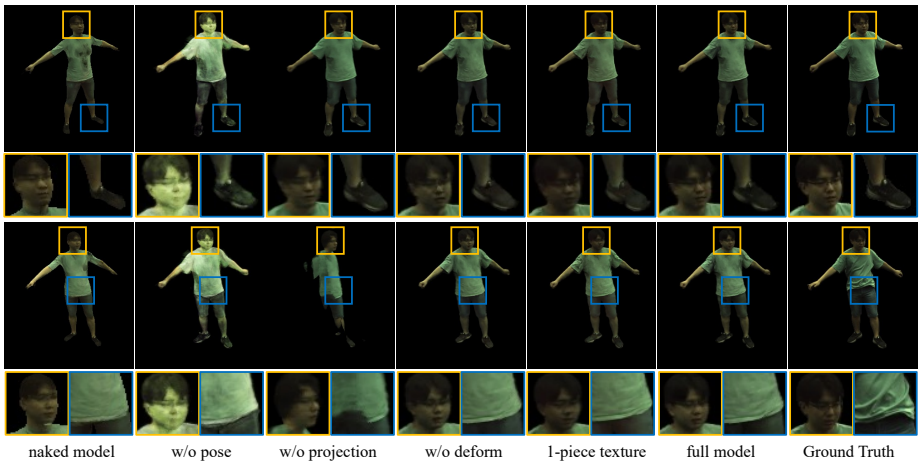


Fig. 7. Qualitative results of ablations. The first row and second row show the visual results for novel view synthesis and novel pose synthesis, respectively.

Table 4. PSNR results of novel view synthesis and novel pose synthesis of ablations(higher is better).

	naked model	w/o pose	w/o projection	w/o deform	1-piece texture	full model
novel view	23.65	21.98	29.73	29.72	29.93	29.75
novel pose	21.71	20.43	18.42	23.35	23.38	23.41

information. A potential solution is to train the model with an auto-regressive way to model the relationship to temporal information.

6 Conclusions

We propose a novel representation of Neural Deformable Fields (NDF) to model dynamic humans. We unwrap observation space to NDF space using a parametric body model as a reference. Then a neural radiance field conditioned on skeletal pose is learned and volume rendering is used to render the pixel color. After training from multi-view videos, our method can synthesize the performer with arbitrary view direction and pose. Extensive experiments on ZJU-MoCap and DynaCap demonstrated that our method outperforms the state-of-the-art in terms of rendering quality and produces faithful pose-dependent appearance changes and wrinkle patterns.

Acknowledgments The research was supported by the Theme-based Research Scheme, Research Grants Council of Hong Kong (T45-205/21-N).

References

1. Aliev, K.A., Sevastopolsky, A., Kolos, M., Ulyanov, D., Lempitsky, V.: Neural point-based graphics. In: European Conferenc on Computer Vision. pp. 696–712 (2020)
2. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Transactions on Graphics **34**(4), 1–13 (2015)
3. Dai, P., Zhang, Y., Li, Z., Liu, S., Zeng, B.: Neural point cloud rendering via multi-plane projection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7830–7839 (2020)
4. Garbin, S.J., Kowalski, M., Johnson, M., Shotton, J., Valentin, J.: Fastnerf: High-fidelity neural rendering at 200fps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14346–14355 (2021)
5. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 43–54 (1996)
6. Guo, K., Lincoln, P., Davidson, P., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., et al.: The relightables: Volumetric performance capture of humans with realistic relighting. ACM Transactions on Graphics **38**(6), 1–19 (2019)
7. Habermann, M., Liu, L., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Real-time deep dynamic characters. ACM Transactions on Graphics **40**(4), 1–16 (2021)
8. Hedman, P., Philip, J., Price, T., Frahm, J.M., Drettakis, G., Brostow, G.: Deep blending for free-viewpoint image-based rendering. ACM Transactions on Graphics **37**(6), 1–15 (2018)
9. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8320–8329 (2018)
10. Kajiya, J.T., Von Herzen, B.P.: Ray tracing volume densities. ACM SIGGRAPH computer graphics **18**(3), 165–174 (1984)
11. Lewis, J.P., Cordner, M., Fong, N.: Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 165–172 (2000)
12. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6498–6508 (2021)
13. Liu, L., Gu, J., Zaw Lin, K., Chua, T.S., Theobalt, C.: Neural sparse voxel fields. Advances in Neural Information Processing Systems **33**, 15651–15663 (2020)
14. 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)
15. Liu, L., Xu, W., Habermann, M., Zollhöfer, M., Bernard, F., Kim, H., Wang, W., Theobalt, C.: Neural human video rendering by learning dynamic textures and rendering-to-video translation. IEEE Transactions on Visualization and Computer Graphics **PP**, 1–1 (May 2020)
16. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. ACM Trans. Graph. **38**(4), 65:1–65:14 (Jul 2019)

17. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM transactions on graphics **34**(6), 1–16 (2015)
18. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7210–7219 (2021)
19. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics **38**(4), 1–14 (2019)
20. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NERF: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421 (2020)
21. 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)
22. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: IEEE International Conference on Computer Vision. pp. 5865–5874 (2021)
23. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. ACM Trans. Graph. **40**(6) (Dec 2021)
24. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14314–14323 (2021)
25. 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. pp. 9054–9063 (2021)
26. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
27. Shao, R., Zhang, H., Zhang, H., Chen, M., Cao, Y.P., Yu, T., Liu, Y.: Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15872–15882 (2022)
28. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhofer, M.: Deepvoxels: Learning persistent 3d feature embeddings. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2437–2446 (2019)
29. Sitzmann, V., Zollhoefer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Advances in Neural Information Processing Systems **32**, 1121–1132 (2019)
30. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
31. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: learning view synthesis using multiplane images. ACM Transactions on Graphics **37**(4), 1–12 (2018)
32. zju3dv: Easymocap. `https://github.com/zju3dv/EasyMocap` (2021)