DNA: Deformable Neural Articulations Network for Template-free Dynamic 3D Human Reconstruction from Monocular RGB-D Video

Khoa Vo, Trong-Thang Pham, Kashu Yamazaki, Minh Tran, Ngan Le
AICV Lab, University of Arkansas
Fayetteville, Arkansas, USA
{khoavoho, tp030, kyamazak, minht, thile}@uark.edu

Abstract

In this paper, we present a novel Deformable Neural Articulations Network (DNA-Net), which is a template-free learning-based method for dynamic 3D human reconstruction from a single RGB-D sequence. Our proposed DNA-Net includes a Neural Articulation Prediction Network (NAP-Net), which is capable of representing non-rigid motions of a human by learning to predict a set of articulated bones to follow movements of the human in the input sequence. Moreover, DNA-Net also include Signed Distance Field Network (SDF-Net) and Apearance Network (Color-Net), which take advantage of the powerful neural implicit functions in modeling 3D geometries and appearance. Finally, to avoid the reliance on external optical flow estimators to obtain deformation cues like previous related works, we propose a novel training loss, namely Easy-to-Hard Geometric-based, which is a simple strategy that inherits the merits of Chamfer distance to achieve good deformation guidance while still avoiding its limitation of local mismatches sensitivity. DNA-Net is trained end-to-end in a self-supervised manner directly on the input sequence to obtain 3D reconstructions of the input objects. Quantitative results on videos of DeepDeform dataset show that DNA-Net outperforms related state-of-the-art methods with an adequate gaps, qualitative results additionally prove that our method can reconstruct human shapes with high fidelity and details.

1. Introduction

Dynamic 3D human reconstruction is a long-standing problem, which requires us to (i) track human movements at the pixel-level throughout every frame (RGB or RGB-D) of an input stream, and (ii) reconstruct the human 3D shape based on the tracked movements. Resolving this problem is an essential step in understanding the 4D world (i.e., the 3D world within the temporal dimension), and opens a door to a wide range of computer vision applications, such as automatically creating and controlling realistic avatars in AR/VR environments. On one hand, template-based methods (8; 26; 33) have achieved considerable results in 3D human pose tracking, however, their reliance on a pre-designed human model (17) significantly limits their reconstruction capability, making it impossible to fully capture varying human clothes or be applied to cases where the human is partially observed (e.g., videos of the upper body). On the other hand, template-free methods such as DynamicFusion (20) and OcclusionFusion (16) do not require a pre-defined model and can process an arbitrary object, however, their frame-by-frame paradigm accumulates error created during the process, which easily leads to failure for extreme cases where the object movements are too large.

Recently, with the success of neural radiance field networks (18; 31) (NeRF) in stationary scenes reconstruction from multiple RGB images, (24; 25) leverage them to model short RGB videos of dynamic object by training a multi-layer perceptron (MLP) to model a volumetric deformation field. Such MLP warps observed points from each frame to a canonical space, so that corresponding points in every timestamp are projected to the same location in canonical space. Nevertheless, their performance degrades drastically on long videos with excessive motions due to their proneness to local minimas. Alternatively, LASR (34), ViSER (35), and BANMo (36) propose to model deformations from a RGB video by a set of nodes representing articulated bones, with an observation that objects movements are controlled by articulated bones, and points in a region around a bone should move rigidly with that bone. Yet, RGB images create geometric ambiguities and unstable results, which causes artifacts and inaccurate final reconstructed shape. Inspired by the idea of articulated bones for deformation modeling and motivated to resolve the geometric ambiguity, we propose a novel method, namely Deformable Neural Articulations Network (DNA-Net), which is optimized in a self-supervised manner only using the input RGB-D video of

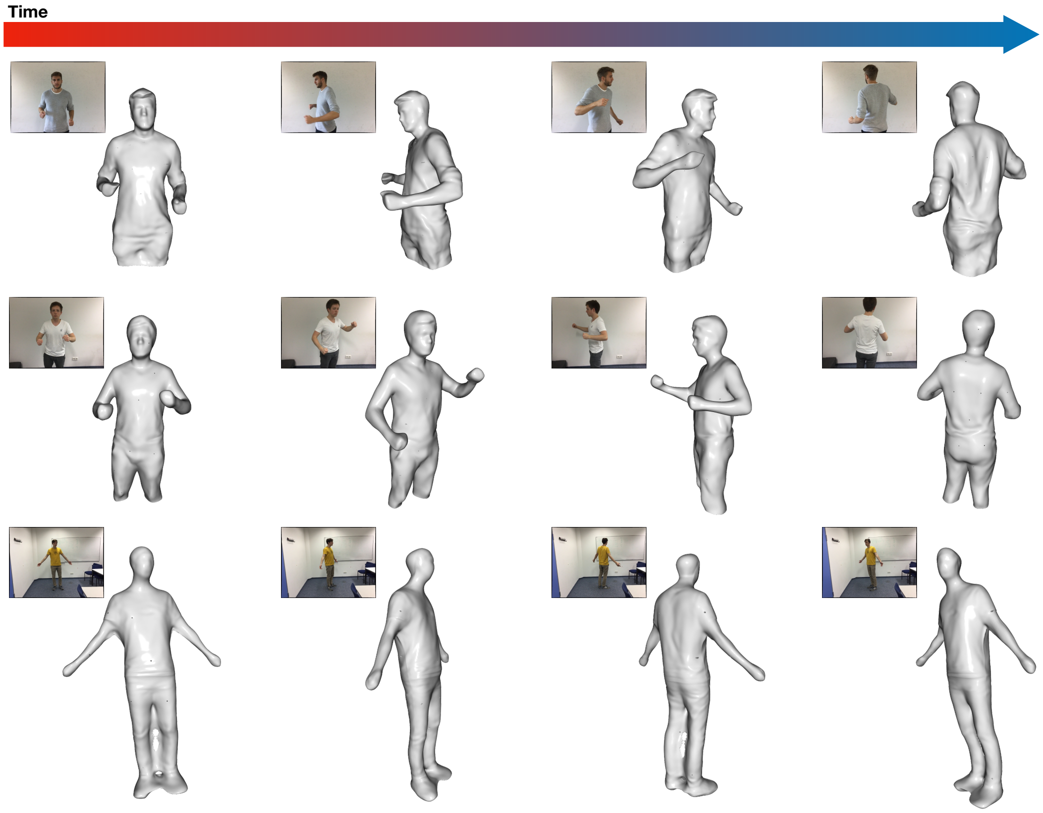

Figure 1. Examples of reconstructed results obtained by our DNA-Net given a RGB-D video. Although being trained in self-supervised manner solely on RGB-D input and without any prior information as well as known human pose, our DNA-Net can achieve high-fidelity shapes of non-rigidly moving humans. Examples illustrate top-to-hip pose (*top row*), top-to-knee pose (*middle row*), and full body pose (*bottom row*).

non-rigidly moving human to densely track every observable point on their surface throughout the video and reconstruct their observable shape.

Specifically, the core component of our proposed DNA-Net is Neural Aticulations Prediction Network (NAP-Net), which reasons on the occupancy grid generated from each depth frame of the input video to predict bones parameters representing articulated pose of the human in the video. From the predicted bones parameters of a frame, we apply linear blend skinning (11) to warp any point in the observing frame of the video to the canonical space. In this space, we optimize neural implicit functions (9; 18) to represent the inferred 3D surface through the signed distance field (SDF) and its corresponding appearance through the color field, which we call as SDF-Net and Color-Net for short. All modules of DNA-Net are optimized jointly only through the color and 3D cues from the input RGB-D stream without the needs of external annotations.

Furthermore, not like other articulation-based methods (34; 35; 36), which rely on an off-the-shelf optical flow estimator to extract spatial correspondence between nearby frames, we propose a self-contained, simple yet effective Easy-to-Hard (E2H) Geometric-based Deformation loss (E2HGD) to obtain strong cues of geometric deformation for the optimization process. In short, given source and target point clouds from two nearby frames, the proposed loss function computes the deviation between transformed source point cloud, which is obtained by NAP-Net, and the target point cloud using forward-only Chamfer distance. As Chamfer distance is a nearest-neighbor-based metric, which may likely give false correspondences if the resolved deformations are not good enough. Therefore, we propose an E2H strategy, which starts with sampling only adjacent frame pairs, then increases the sampled frame distance during optimization process. We show that our easy-to-hard strategy can avoid the drawback of Chamfer distance while inheriting its merits, which helps obtaining high quality deformation cues that are equivalent to optical flow.

Our empirical results show that DNA-Net can recover high-fidelity results of 3D dynamic humans in various scenarios, completely without support of any template keypoints detector as well as guidances from optical flow as in previous state-of-the-art (SOTAs) methods. Illustrations are shown in Fig. 1.

Our key contributions can be summarized as follows:

- A novel Deformable Neural Articulation Network

(DNA-Net) for human non-rigid 3D reconstruction from a single view RGB-D video stream. DNA-Net consists of (i) a Neural Articulations Prediction Network (NAP-Net), which models and tracks movements of the human throughout input video to project observed points into a common canonical space, (ii) SDF-Net models the signed distance field in the canonical space using the projected observations, (iii) Color-Net models colors of every surface points. Both SDF-Net and Color-Net are neural implicit functions, which inherits the powerful representation of neural networks. An overview flowchart of the proposed DNA-Net is shown in Fig. 2.

- We propose an Easy-to-Hard Geometric-based Deformation Loss, which is a self-contained, simple and effective function in guiding our DNA-Net in solving the non-rigid movements of human the video. The proposed loss function helps eliminating the reliance of previous articulation-based methods on an off-the-shelf optical flow estimator.
- We compare our proposed DNA-Net with related methods on various scenarios and experimental results show that our DNA-Net can consistently achieve lower geometric error and better qualitative results.

2. Related Works

2.1. Frame-by-Frame Non-rigid Reconstruction

Frame-by-frame non-rigid reconstruction is a very intuitive approach where 3D deformable motion is tracked and reconstructed in an incremental fashion. As this approach mainly relies on 3D information, it utilizes RGB-D inputs captured by commercial RGB-D cameras such as Realsense, Kinect, Kinect Azure. DynamicFusion (20) is a seminal work in this approach, which aggregates observable object surface through each frame by optimizing a dense volumetric 6D motion field. While DynamicFusion is solely optimized on depth information to resolve depth correspondences, follow-up methods leverage RGB information to improve its robustness and performance, e.g., sparse SIFT correspondences (10), learnt correspondences (2; 3), or optical flow (16). Other methods incorporate regularization over the volumetric truncated signed distance fields (TSDFs) regularization (28; 29). Although these approaches showed steady improvements over time, they still suffer a major limitation of the recursive fashion, which suffers from the accumulative tracking errors, and completely fails if encountering fast motion or extreme noise during the process.

To overcome aforementioned limitations of frame-by-frame methods, we follow more advanced learning-based articulation methods in deformation modeling and human reconstruction from RGB-D videos.

2.2. NeRF-based Dense Motion Fields

To recap, Neural Radiance Field (NeRF) (18) is a fundamental learning-based technique in scene representation that optimizes a scene-specific MLP for novel views synthesis and rendering given a set of images of the scene. NeuS (31) and VolSDF (37) are two NeRF variants that are specifically devised for 3D scene reconstruction by modeling the volume density as a function of the learnt signed distance fields (SDFs).

Many attempts have been made to adopt NeRF into dynamic scenes or objects from a RGB video stream. For example, (24; 25; 27) model non-rigid motions by additional functions that deform every observed point forward or backward in time to its corresponding location, through a canonical space, in which their NeRF-related parts are operated to reconstruct the scene. However, as there is no cue provided for deformation functions, these methods heavily rely on prior camera pose registration using background, hence they easily fail when object motions are too large and can only handle very short videos. Most recently, NDR (5) overcomes those issues by adding depth cues into the set of loss functions beside RGB cue, and externally utilizing Robust ICP algorithm (38) for relative camera pose registration based on the color point cloud of the object at every frame.

Our method shares several similarities with (5), e.g., we both incorporate depth cues as one of the constraints, and we also utilize an external method to pre-estimate camera poses relative to the object at every frame, although our proposed registration method is inspired by convolutional neural networks and reasons on semantically richer embedding features (19). Different from aforementioned works, our DNA-Net models the dynamic motions by articulated bones, which helps the model converge more quickly and more open to applications such as human pose manipulation.

2.3. Articulated Motion Representation

Articulations have a long history of usage in both frame-by-frame and learning-based methods owing to its effectiveness in representing articulated motion of animals and humans. Prior works in articulated motion representation can be categorized into two groups, i.e., template-based and template-free methods.

Template-based 3D reconstruction methods leverage a category-specific parametric template shape (17; 22; 26; 40; 41) as a guidance to recover the 3D shapes of commonly known input objects with single or multiple views, e.g., humans (8; 21; 39), quadrupted animals (1; 33). Although template-based methods have achieved great suc-

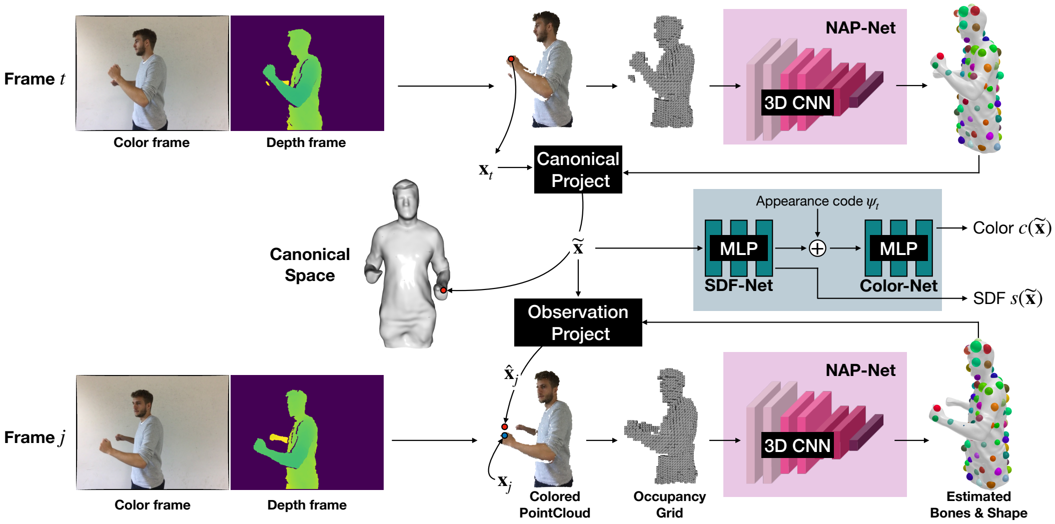

Figure 2. An overview of our proposed DNA-Net consisting of three modules: (i) NAP-Net models and tracks movements of the human to project observed points into a common canonical space; (ii) SDF-Net models the signed distance field in the canonical space using the projected observation; (iii) Color-Net models colors of every surface point.

cess on these object categories, they cannot be adopted well to novel object categories (e.g., chicken, fish, bird), parts of known objects (e.g., upper or lower half human body, human arms or hands), or known objects with complex appearances (e.g., lions, giraffe, elephants, or humans with uncommon garments) without pre-designing proper parametric templates.

In contrast, template-free 3D reconstruction methods can handle arbitrary objects under input-specific self-supervised learning manner. LASR (34), ViSER (35), and BANMo (36) together create a series of template-free articulation-based 3D object reconstruction from RGB videos. These methods train model shape and articulation learnable parameters using reconstruction losses on color frames, and external cues of silhouttes, optical flow that are estimated by off-the-shelf models (13; 30). Regardless of many impressive showcases are reported, these methods do not perform robustly in all cases due to the ambiguity of RGB-only inputs that the optical flow, which is not fully reliable, cannot fully resolve.

Our proposed DNA-Net follows template-free approach and operates over RGB-D videos to be able to reconstruct human shapes in many cases while avoiding depth ambiguities. Moreover, we propose a novel Easy-to-Hard Geometric-based Deformation loss to remove the requirement of external optical flow, which is costly to be estimated but unreliable.

3. Our Method

Our DNA-Net is trained on a specific RGB-D input sequence $\mathcal{F} = \{\mathcal{I}_t, \mathcal{D}_t\}_{t=1}^{N}$ captured by a single sensor, where $\mathcal{I}_t \in [0, 255]^{H\times W\times 3}$ is the t-th color frame and $\mathcal{D}_t \in \mathbb{R}^{H\times W\times 1}$ is the corresponding depth map. Prior to training, we apply a pre-trained video human matting model (15) to every image $\mathcal{I}_t$ to obtain a binary mask $\mathcal{M}_t \in [0, 1]^{H\times W\times 1}$ of the human whom we want to reconstruct the shape of. With the binary mask $\mathcal{M}_t$, we can convert $\mathcal{I}_t$ and $\mathcal{D}_t$ into a point cloud of observed human at frame t, i.e., $\mathcal{O}_t = \{\mathbf{x}_t \in \mathbb{R}^3, \mathbf{c}_t \in \mathbb{R}^3\}$, where $\mathbf{x}_t$ is a point in the point cloud $\mathcal{O}_t$, and $\mathbf{c}_t$ is its corresponding color.

In our DNA-Net, we maintain two spaces including (i) time-invariant canonical space, where all observations at every frame is projected into, for training the SDF-Net and Color-Net to estimate canonical shape of the human, and (ii) time-dependent observation space at every particular frame t, where the canonical shape of the human is projected into, for non-rigid reconstruction of the shape at frame t. An overview of our proposed DNA-Net is illustrated in Fig. 2.

In the following parts of this section, we carefully discuss through each module of our proposed DNA-Net, including root pose initialization (Sec. 3.1), Neural Articulation Prediction network (NAP-Net, Sec. 3.2), SDF and Color networks (SDF-Net & Color-Net, Sec. 3.3).

3.1. Root Pose Initialization

Root poses initialization is an important pre-processing step of estimating per-frame global orientation of the articulation. The initial root pose aims to reduce training time as well as helping our model avoid being stuck in local minima, as suggested by (5; 36). In this paper, we define a root pose of the human in a frame as a coordinate with its center at their waist, z-axis points to their head, y-axis points to their front, and x-axis is parallel with their shoulder.

We pre-train a simple *RootNet* to predict the root pose from an RGB frame. Following (36), *RootNet* would reason on DensePose CSE feature (19) extracted from RGB image, which allows itself to be trained on simple synthesized data settings while provide more robust predictions to various scenarios compared to directly reasoning over RGB data. However, training process of *RootNet* is not simple like in (36) as we do not have direct access to the per-vertex CSE feature of humanoid 3D mesh (19) due to its copyright. Instead, we utilize a 3D humanoid mesh Ω_{ref} from synthetic dataset of DeformingThings4D (14) as a reference, normalized be in a unit scale and centered at the world coordinate. Please note that our DNA-Net do not rely on any humanoid reference mesh, but just utilize the raw root predicted by pre-trained *RootNet*.

To generate a training sample for *RootNet*, we first sample a camera $\mathcal{C}_{gt} = [\mathbf{R}_{gt}|\mathbf{t}_{gt}] \in \mathbb{R}^{3\times 4}$ that is positioned over a unit sphere surrounding Ω_{ref} and always points to its center. Afterwards, we carefully synthesize several light sources according to the sampled camera position and do rendering to obtain a synthesized image of size 224 × 224. The image is encoded using pretrained Densepose CSE (19) to obtain a feature map $\phi_{\text{cse}} \in \mathbb{R}^{224\times 224\times 16}$. ϕ_{cse} is fed to *RootNet* to predict root pose $\mathbf{R}_{\text{root}} \in SE(3)$, which is trained by the following loss function:

$$\mathcal{L}_{\mathit{RootNet}} = \mathcal{L}_2(\mathbf{R}_{\text{root}}, \mathbf{R}_{gt}^T) \tag{1}$$

where L_2 is a mean squared error loss, and the predicted root pose is optimized to become the inversion of the camera view.

RootNet is trained once at the pre-training procedure. It is then employed in all reconstruction experiments. We utilize ResNet-34, which is initially pre-trained on ImageNet (7), as a backbone of *RootNet*. We use Adam optimizer (12) with a batch size of 512 and a learning rate of 10^{-4} to train *RootNet* for 30K iterations. The entire process would take around an hour to finish.

At deployment phase, given a RGB frame $\mathcal{I}_t$, we first resize it to 224×224 and encode it to Densepose CSE feature map $\phi_i \in \mathbb{R}^{224\times 224\times 16}$. We then feed ϕ_i into pre-trained *RootNet* to predict the raw root pose $\mathbf{G}_t^{\text{raw}} \in SE(3)$, which is summarized as follows:

$$\mathbf{G}_t^{\text{raw}} = \mathit{RootNet}\left(\text{DenseposeCSE}(\mathcal{I}_t)\right) \tag{2}$$

3.2. NAP-Net

As stated above, we model non-rigid motion globally through a canonical space. A sampled 3D point $\mathbf{x}_t$ in frame t can be transformed to any other frame j by consecutively applying projection $\mathcal{P}_t^{\text{can}} : \mathbf{x}_t \to \widetilde{\mathbf{x}}$, which transforms $\mathbf{x}_t$ to corresponding 3D location $\widetilde{\mathbf{x}}$ in the canonical space, then observation projection $\mathcal{P}_j^{\text{obs}} : \widetilde{\mathbf{x}} \to \hat{\mathbf{x}}_j$, which transforms canonical point $\widetilde{\mathbf{x}}$ into the observed frame j. We follow (35; 36) by modeling such projections through a set of controlling bones.

Specifically, there are two kinds of controlling bones defined in our work, i.e., (i) a set of fixed bones $\widetilde{\mathcal{B}} = \{\widetilde{\mathbf{P}}^b, \tilde{s}^b\}_{b=1}^B$ in the canonical space, where $\widetilde{\mathbf{P}}^b$ and $\tilde{s}^b$ are canonical position and importance radius of a bone b; (ii) a set of articulated bones $\mathcal{B}_t = \{\mathbf{P}_t^b, \mathbf{R}_t^b\}_{b=1}^B$, at an observed frame t, where $\mathbf{P}_t^b \in \mathbb{R}^{3\times 1}$ is the position of a bone b, and $\mathbf{R}_t^b \in \mathbb{R}^{3\times 3}$ is an rotation of a bone b.

With those two kinds of controlling bones, both canonical projection and observation projections can be carried out through linear blend skinning (11). Given a 3D point $\mathbf{x}_t$ sampled in the observation space of frame t, canonical projection $\mathcal{P}_t^{\text{can}}$ is formulated as follows:

$$W^b(\mathbf{x}_t) = \exp\Big(\frac{||\mathbf{x}_t - \mathbf{P}_t^b||_2^2}{(\tilde{s}^b)^2}\Big) \tag{3}$$

$$\widetilde{\mathbf{x}} = \mathcal{P}_t^{\text{can}}(\mathbf{x}_t) = \sum_{b=1}^{B} W^b(\mathbf{x}_t)\mathbf{G}_t\big[(\mathbf{R}_t^b)^T(\mathbf{x}_t - \mathbf{P}_t^b) + \widetilde{\mathbf{P}}^b\big] \tag{4}$$

where $W^b(\mathbf{x}_t)$ is a weighted Gaussian function that returns weight value of $\mathbf{x}_t$ wrt. bone b of $\mathcal{B}_t$, and G_t is an estimated root pose at frame t that will be discussed later. Likewise, the observation projection to frame j, $\mathcal{P}_j^{\text{obs}}$, is applied onto $\widetilde{\mathbf{x}}$ and expressed as follows:

$$\widetilde{W}^b(\widetilde{\mathbf{x}}) = \exp\Big(\frac{||\widetilde{\mathbf{x}} - \widetilde{\mathbf{P}}^b||_2^2}{(\tilde{s}^b)^2}\Big) \tag{5}$$

$$\mathcal{P}_j^{\text{obs}}(\widetilde{\mathbf{x}}) = \sum_{b=1}^{B} W^b(\widetilde{\mathbf{x}})\mathbf{G}_j\big[(\mathbf{R}_j^b)(\widetilde{\mathbf{x}} - \widetilde{\mathbf{P}}^b) + \mathbf{P}_j^b\big] \tag{6}$$

where G_j is an estimated root pose at frame j, $\widetilde{W}^b(\widetilde{\mathbf{x}})$ is the weighted Gaussian function of $\widetilde{\mathbf{x}}$ wrt. canonical bone b in canonical bones set $\widetilde{\mathcal{B}}$.

Our NAP-Net aims to estimate parameters of controlling bones at every frame of the RGB-D sequence. Inspired by the success of 3D convolutional neural networks (3D CNN) in understanding 3D point clouds (4; 6), our NAP-Net utilizes the architecture of 3D CNN in NeuralGraph (4) to regress controlling bones at every frame of the input sequence. Besides, to eliminate noise and unreliable results from initialized root pose $\mathbf{G}_t^{\text{raw}}$, we also incorporate an offset regressor into NAP-Net to obtain $\mathbf{G}_t$. Bones $\mathcal{B}_t$ and root pose $\mathbf{G}_t$ are realized as following equations:

$$(\mathcal{B}_t, \Delta\mathbf{G}_t) = \textbf{NAP-Net}(\text{Occ}(\mathcal{O}_t)) \tag{7}$$

$$\mathbf{G}_t = \Delta\mathbf{G}_t\mathbf{G}_t^{\text{raw}} \tag{8}$$

where Occ($\cdot$) is a function that voxelizes the input point cloud into an occupancy grid (we use a grid size of 64 × 64 × 64 for all experiments).

3.3. SDF-Net and Color-Net

In the canonical space, SDF-Net and Color-Net are defined as follows:

SDF-Net: We define SDF-Net as a MLP reasoning directly over coordinates of sampled canonical point $\widetilde{\mathbf{x}}$ to predict its signed distance $s(\widetilde{\mathbf{x}}) \in \mathbb{R}$ to the surface:

$$s(\widetilde{\mathbf{x}}), f(\widetilde{\mathbf{x}}) = \textbf{SDF-Net}(\widetilde{\mathbf{x}}) \tag{9}$$

where $f(\widetilde{\mathbf{x}}) \in \mathbb{R}^{256}$ is feature vector that is additionally regressed to provide to Color-Net. The human shape is finally represented by the zero level-set of the signed distance field realized by the optimized SDF-Net:

$$S = \{\widetilde{\mathbf{x}} \in \mathbb{R}^3 | s(\widetilde{\mathbf{x}}) = 0\} \tag{10}$$

Color-Net: Color-Net is also a MLP that predicts RGB color $c(\widetilde{\mathbf{x}}) \in \mathbb{R}^3$ of the sampled canonical point $\widetilde{\mathbf{x}}$ from its position, feature $f(\widetilde{\mathbf{x}})$, and a learnable per-frame appearance code $\Psi_t \in \mathbb{R}^{64}$:

$$c(\widetilde{\mathbf{x}}) = \textbf{Color-Net}(\widetilde{\mathbf{x}}, f, \Psi_t) \tag{11}$$

4. Optimization

Given a sequence of colored point cloud $\mathcal{O}_t$ $(1 \geq t \geq N)$, our objective is to optimize the parameters corresponding to learnable code Ψ_t for each frame, trainable parameters in neural networks of NAP-Net, SDF-Net, Color-Net. Our DNA-Net is optimized end-to-end by a set of losses, which can be clustered into three types, i.e., surface losses, near-surface losses, and E2H geometric-based deformation loss:

$$\mathcal{L} = \underbrace{\Big(\mathcal{L}_{\text{sdf}} + \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{vis}}\Big)}_{\text{surface}} + \underbrace{\Big(\mathcal{L}_{\text{eik}} + \mathcal{L}_{\text{cov}} + \mathcal{L}_{\text{cyc}}\Big)}_{\text{near-surface}} + \mathcal{L}_{\text{deform}} \tag{12}$$

4.1. Surface Losses

Surface losses are meant to penalize sampled points on the observed surface at a frame t, i.e., $\mathbf{x}_t \in \mathcal{O}_t$. Let $\widetilde{\mathbf{x}} = \mathcal{P}_t^{\text{can}}(\mathbf{x}_t)$, SDF loss $\mathcal{L}_{\text{sdf}}$ encourages SDF predictions to vanish for these points (absolute values are close to zero):

$$\mathcal{L}_{\text{sdf}} = \sum_{\mathbf{x}_t} \big|\big|s(\widetilde{\mathbf{x}})\big|\big|_1 \tag{13}$$

Next, color loss $\mathcal{L}_{\text{rgb}}$ adds constraints on reconstructed color of sampled point $\mathbf{x}_t \in \mathcal{O}_t$ to be as close as its corresponding observed color $\mathbf{c}_t$:

$$\mathcal{L}_{\text{rgb}} = \sum_{\mathbf{x}_t} \big|\big|c(\widetilde{\mathbf{x}}) - \mathbf{c}_t\big|\big|_1 \tag{14}$$

Finally, with an intuition that every observed surface point always faces to camera, the visible loss $\mathcal{L}_{\text{vis}}$ is added to maintain this property. Specifically, it encourages the angle between the normal of a surface point and camera view not to be smaller than 90°:

$$\mathcal{L}_{\text{vis}} = \sum_{\mathbf{x}_t} \max\Big(\big\langle \frac{\nabla_{\mathbf{x}_t} s(\widetilde{\mathbf{x}})}{||\nabla_{\mathbf{x}_t} s(\widetilde{\mathbf{x}})||_2}, \frac{\mathbf{v}_{\text{cam}}}{||\mathbf{v}_{\text{cam}}||_2} \big\rangle, 0\Big) \tag{15}$$

where $\nabla_{\mathbf{x}_t} s(\widetilde{\mathbf{x}})$ denotes the gradient of SDF function wrt. $\mathbf{x}_t$, which is also the normal vector at this point. $\mathbf{v}_{\text{cam}}$ is the camera view, which is a constant vector $[0, 0, -1]^T$ at every observed frame, and $\langle \cdot, \cdot \rangle$ denotes cross inner product of two vectors.

4.2. Near-Surface Losses

To generate near-surface points, we splat an isotropic Gaussian, $\mathcal{N}(\mathbf{x}_t, \sigma^2\mathbf{I})$, on every point $\mathbf{x}_t$ that is uniformly sampled from $\mathcal{O}_t$. Denoting $\mathbf{x}'_t = \mathcal{N}(\mathbf{x}_t, \sigma^2\mathbf{I})$, the corresponding canonical point is $\widetilde{\mathbf{x}}' = \mathcal{P}_t^{\text{can}}(\mathbf{x}'_t)$. Additionally, as the depth map $\mathcal{D}_t$ for each frame is also available, we can easily project sampled point $\mathbf{x}'_t$ into the depth map. Comparing depth of $\mathbf{x}'_t$ with recorded depth value at the corresponding pixel would let us know whether $\mathbf{x}'_t$ is outside or inside the surface. Let denote $l'_t \in \{0, 1\}$ is a label of $\mathbf{x}_t$ to denote if it is inside $(l'_t = 1)$ or otherwise, outside $(l'_t = 0)$.

Near-surface losses aim to at constraints on a narrow band around the observed surface. Particularly, Eikonal loss (9) $\mathcal{L}_{\text{eik}}$ plays a role of regularization to encourage gradients $\nabla_{\mathbf{x}'_t} s(\widetilde{\mathbf{x}}')$ to be a unit 2-norm:

$$\mathcal{L}_{\text{eik}} = \sum_{\mathbf{x}'_t} \big(||\nabla_{\mathbf{x}'_t} s(\widetilde{\mathbf{x}}')||_2 - 1\big)^2 \tag{16}$$

Coverage loss $\mathcal{L}_{\text{cov}}$ encourages the SDF value obtained by SDF-Net of sampled points that are inside of the surface to be negative and vice versa:

$$\mathcal{L}_{\text{cov}} = \sum_{\mathbf{x}'_t} CE\big(\text{sigmoid}(100 s(\widetilde{\mathbf{x}}')), l'_t\big) \tag{17}$$

where $CE(\cdot, \cdot)$ is Cross-Entropy loss function. In the above equation, we upscale the predicted SDF value $s(\widetilde{\mathbf{x}})$ by ×100 to reduce the effective area to just a narrow band around object. Afterwards, we re-scale resulted value to a range of $[0, 1]$ by Sigmoid function, to match with the requirements of Cross-Entropy loss.

Finally, cycle loss $\mathcal{L}_{cyc}$ directly regularizes the consistency of NAP-Net through cycle projections. Concretely, it constrains that if we project a point $\mathbf{x}'_t$ in frame t into canonical space by canonical projection $\mathcal{P}_t^{\text{can}}$ and then project the result back to the same frame by observation projection $\mathcal{P}_t^{\text{obs}}$, then the final result should be located at the same position as the original position. Such constraint is formulated as follows:

$$\mathcal{L}_{\text{cyc}} = \sum_{\mathbf{x}'_t} \Big|\Big|\mathcal{P}_t^{\text{obs}}\big(\mathcal{P}_t^{\text{can}}(\mathbf{x}'_t)\big) - \mathbf{x}'_t\Big|\Big|_2^2 \tag{18}$$

4.3. E2H Geometric-based Deformation Loss

We leverage a simple Chamfer distance and propose a novel loss function, namely Easy-to-Hard (E2H) Geometric-based Deformation Loss (E2HGD), which aims at providing high quality deformation cues through the easy-to-hard strategy. We firstly introduce a general geometric-based deformation loss from a pair of frames t and j, then, we will go further into how we conduct an easy-to-hard strategy.

Given a surface point observed from frame t, $\mathbf{x}_t \in \mathcal{O}_t$, we denote its transformed version in frame j as $\hat{\mathbf{x}}_j = \mathcal{P}_j^{\text{obs}}\big(\mathcal{P}_t^{\text{can}}(\mathbf{x}_t)\big)$. Now, the geometric-based deformation loss between frames t and j is designed as follows:

$$\mathcal{L}_{\text{deform}}^{tj} = \sum_{\mathbf{x}_t} \begin{cases} 0, \text{ if } \langle \frac{\nabla_{\hat{\mathbf{x}}_j} s(\mathcal{P}_j^{\text{can}}(\hat{\mathbf{x}}_j))}{||\nabla_{\hat{\mathbf{x}}_j} s(\mathcal{P}_j^{\text{can}}(\hat{\mathbf{x}}_j))||_2}, \frac{\mathbf{v}_{\text{cam}}}{||\mathbf{v}_{\text{cam}}||_2} \rangle < 0. \\ FCD(\hat{\mathbf{x}}, \mathcal{O}_j), \text{ otherwise.} \end{cases} \tag{19}$$

$$FCD(\hat{\mathbf{x}}, \mathcal{O}_j) = \min_{\mathbf{x}_j \in \mathcal{O}_j} ||\hat{\mathbf{x}} - \mathbf{x}_j||_2 \tag{20}$$

where $FCD(\hat{\mathbf{x}}_j, \mathcal{O}_j)$ computes forward-only Chamfer distance (FCD) between transformed point $\hat{\mathbf{x}}_j$ and point cloud $\mathcal{O}_j$ of frame j.

Concisely, the above equation only returns FCD when the normal of the transformed point $\hat{\mathbf{x}}_j$ faces to the camera view at frame j, which means that it belongs to a visible part of the human at j, i.e., $\mathcal{O}_j$. Contrarily, a point with a normal vector not facing to the camera view belongs to the occluded part of the human at j, therefore it does not belong to $\mathcal{O}_j$ and taking its FCD value into account will add noise to the total loss.

As FCD only returns the nearest distance from the query point to the search point cloud, it is unreliable because of its sensitivity to local mismatches (32), especially when the query point is far from the point cloud of interest. This case happens when we sample a pair of t and j frames that are far away but NAP-Net is not yet capable to model accurate deformation. Thus, we propose an simple easy-to-hard strategy, which starts with closest pairs of frames and increases the difficulty of the sampled pairs when our NAP-Net has learned well to model deformations of easy pairs.

Specifically, given that we are at a training iteration i and I is the total number of iterations, we define a linear sampling schedule as follows:

$$d(i) = \max\Big(\big(d_{\text{max}} - 1\big)\frac{i - i_{\text{begin}}}{I - i_{\text{begin}}} + 1, 0\Big) \tag{21}$$

where d_{max} is a pre-defined threshold for maximum sampled distance between two frames, i_{begin} is the iteration where we start increasing sampled distance between two frames. With the above equation, given training sample at frame t, E2HGD loss randomly samples an integer d in a range of $[-d(i), d(i)]$ and compute deformation loss. The procedure is formulated as below:

$$j = t + \text{round}\Big(\mathcal{U}\big(-d(i), d(i)\big)\Big) \tag{22}$$

$$\mathcal{L}_{\text{deform}} = \mathcal{L}_{\text{deform}}^{tj} \tag{23}$$

where round$(\cdot)$ rounds the input value to the nearest integer.

5. Experiments

Implementation Details: Following (31), we initialize weights of SDF-Net to be approximate a unit sphere. Then, our proposed DNA-Net is trained end-to-end using Adam optimizer (12) with a learning rate of 5×10^{-4} and a batch size of 128 frames. In all of our experiments, we train our networks for 60K iterations on a machine equipped with a single NVIDIA RTX A6000 48GB GPU.

For each frame in a mini-batch, we uniformly sample $1,500$ for observed surface points, free-space points, and near surfance points, separately. As suggested in (24), we also applied a coarse-to-fine positional embedding scheme during the training to achieve high-fidelity reconstructed shape. For E2HGD loss, we set maximum sampling frame distance $d_{\text{max}} = 30$, and $i_{\text{begin}} = 10^4$.

Dataset: We mainly evaluate our DNA-Net and compare it with other baselines on the dataset of DeepDeform (3), which is a popular benchmark for non-rigid humans/objects reconstruction. Videos in DeepDeform (3) captured by an Ipad with an attached structure sensor, color frames from the Ipad are calibrated to depth frames of the external sensor. All videos are in RGB-D format, with the resolution of 640 × 480, and captured at 30 frames per second. As our DNA-Net is trained on each specific input video to obtain 3D reconstruction of its subject, we skip the training set of DeepDeform. From the validation and testing sets, there are a total of 20/60 videos having human subjects, in which there are 4 distinct subjects, therefore, we selected 4 videos out of these videos, each for a human subject for our experiments.

5.1. Quantitative Comparisons

Metrics: quantitative comparisons are conducted using the geometric error in centimeters (cm). This metric is computed by comparing 3D distance depth pixels inside foreground mask with the corresponding point from reconstructed mesh. For a video, we compute mean geometric error of all foreground pixels for a sparse set of frames, and report the averaged value of these errors.

Method	Subset / Video name				
	val/seq024	val/seq026	test/seq001	test/seq006	test/seq010
LASR (34)	15.78	7.57	4.84	8.99	4.94
BANMo (36)	8.98	6.92	4.04	6.82	7.22
NDR (5)	**1.35**	*1.15*	*0.71*	*0.74*	*0.83*
Ours	*1.51*	**0.53**	**0.32**	**0.31**	**0.39**

Table 1. Quantitative comparisons on **geometric error** between our proposed DNA-Net and other baselines for videos of DeepDeform dataset (23). We highlight the **best** and the *second-best* performances. All values are in *cm* and the lower is the better.

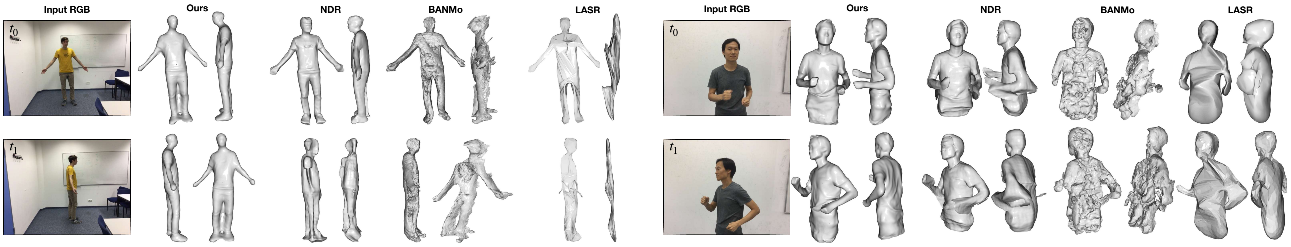

Figure 3. Qualitative results of our DNA-Net, NDR (5), BANMo (36), and LASR (34) on two videos of DeepDeform (3) (val/seq024 on left side and test/seq001 on right side).

We compare our proposed DNA-Net with several baselines that are most related to ours, which are LASR(34), BANMo (36), and NDR (5). Table 1 summarizes the performances of ours and other baselines on every video.

Compared to LASR and BANMo, our proposed method outpeforms with very large margins on all experiments. Meanwhile, DNA-Net consistently surpasses NDR in all experiments and performs competitively on val/seq024. Such performance pattern can be explained because LASR and BANMo does not utilize depth information and suffer from the 2D ambiguities, while NDR and ours utilizes both depth and color cues. However, dense motion fields from NDR shows deficient capability in modeling deformation compared to articulated controlling bones of our DNA-Net.

5.2. Qualitative Comparisons

Qualitative comparisons between our DNA-Net and other baselines are shown in Fig. 3.

We can observe that although NDR (5) quantitatively performs better in video val/seq024 and reconstructs human meshes with a bit higher details, there are still several frames that it gives defective reconstructions Fig. 3(left). On Fig. 3 (right), we can see that our DNA-Net can reconstruct much better details than NDR in the head and arm regions. These observations show that our DNA-Net well preserves the geometrical properties of reconstructed surface and performs robustly throughout the entire video when compared to NDR.

5.3. Robustness of E2HGD Loss

To evaluate the robustness of our proposed E2HGD loss, we conduct an experiment on two videos of DeepDeform (3) where we replace E2HGD loss by optical flow guidance extracted by an external. We follow the training pipeline in (16) to pre-train RAFT (30) (a SOTA method is optical flow estimation) on RGB-D videos for optical flow estimation. Empirically, the provided optical flows are accurate when we set the distance between two frames as 5. As the estimated optical flow maps are in 2D, we align them with the corresponding depth to convert them to the 3D scene flows $\mathcal{F}^t_{\text{flow}} = \{\mathbf{x}_t, \mathbf{x}_{t+5}\}$ at every frame t. The replaced loss function is as follows:

Loss	Subset / Video name	
	test/seq006	test/seq010
Scene Flow	0.28	0.36
E2HGD	0.32	0.39

Table 2. Comaprisons between our proposed E2HGD loss and scene flow loss.

$$\mathcal{L}_{\text{flow}} = \sum_{\mathbf{x}_t, \mathbf{x}_{t+5} \in \mathcal{F}^t_{\text{flow}}} \left|\left| \mathcal{P}^{\text{obs}}_{t+5} \mathcal{P}^{\text{can}}_t (\mathbf{x}_t) - \mathbf{x}_{t+5} \right|\right|^2_2 \quad (24)$$

Comparisons between our E2HGD loss and scene flow loss are shown in Tab. 2. As we can see, the performance improvements with optical flow guidance are not large when compared with our E2HGD loss. Therefore, our E2HGD loss is effective and proves to be a lower-cost alternative to optical flow guidance.

6. Conclusions

We propose a novel method in dynamic 3D human reconstruction, namely Deformable Neural Articulations Network (DNA-Net), which includes a Neural Articulation Prediction Network (NAP-Net) that learns to predict articulated controlling bones to model non-rigid motions of human in the input video. DNA-Net also includes powerful neural implicit functions to model 3D shape and appearance of the human. Both quantitative and qualitative results show that DNA-Net outperforms other SOTA methods. Furthermore, we propose a novel easy-to-hard geometric-based loss, which is proved to be a low-cost alternative to optical flow in providing deformation guidance to train DNA-Net.

Limitation: Although obtaining human reconstructions with high fidelity in most cases, DNA-Net, as a learning-based method, still needs a lot of time to be trained compared to frame-by-frame methods that run in real-time. Therefore, there is still a long way to incorporate DNA-Net into real-world applications.

Acknowledgement: This work is supported by the National Science Foundation (NSF) under Award No OIA-1946391, NSF 1920920, and NSF 2223793.

References

[1] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and smal: Recovering the shape and motion of animals from video. In C.V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 3–19, Cham, 2019. Springer International Publishing. 3

[2] Aljaz Bozic, Pablo Palafox, Michael Zollhöfer, Angela Dai, Justus Thies, and Matthias Nießner. Neural non-rigid tracking. *Advances in Neural Information Processing Systems*, 33:18727–18737, 2020. 3

[3] Aljaz Bozic, Michael Zollhofer, Christian Theobalt, and Matthias Nießner. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7002–7012, 2020. 3, 7, 8

[4] Aljaž Božič, Pablo Palafox, Michael Zollhöfer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1450–1459, 2021. 5

[5] Hongrui Cai, Wanquan Feng, Xuetao Feng, Yan Wang, and Juyong Zhang. Neural surface reconstruction of dynamic scenes with monocular rgb-d camera. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3, 4, 7, 8

[6] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 5

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[8] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20470–20480, June 2022. 1, 3

[9] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3789–3799. PMLR, 13–18 Jul 2020. 2, 6

[10] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 362–379, Cham, 2016. Springer International Publishing. 3

[11] Alec Jacobson, Zhigang Deng, Ladislav Kavan, and J. P. Lewis. Skinning: Real-time shape deformation (full text not available). In *ACM SIGGRAPH 2014 Courses*, SIGGRAPH '14, New York, NY, USA, 2014. Association for Computing Machinery. 2, 5

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5, 7

[13] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[14] Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 4dcomplete: Non-rigid motion estimation beyond the observable surface. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12706–12716, October 2021. 5

[15] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 238–247, January 2022. 4

[16] Wenbin Lin, Chengwei Zheng, Jun-Hai Yong, and Feng Xu. Occlusionfusion: Occlusion-aware motion estimation for real-time dynamic 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1736–1745, 2022. 1, 3, 8

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1, 3

[18] 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 *ECCV*, 2020. 1, 2, 3

[19] Natalia Neverova, David Novotny, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17258–17270. Curran Associates, Inc., 2020. 3, 5

[20] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 1, 3

[21] H. Onizuka, Z. Hayirci, D. Thomas, A. Sugimoto, H. Uchiyama, and R. Taniguchi. Tetratsdf: 3d human reconstruction from a single image with a tetrahedral outer shell. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6010–6019, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. 3

[22] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12695–12705, October 2021. 3

[23] 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. 7

[24] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien

Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 1, 3, 7

[25] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021. 1, 3

[26] 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 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1, 3

[27] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10318–10327, June 2021. 3

[28] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5474–5483, 2017. 3

[29] Miroslava Slavcheva, Maximilian Baust, and Slobodan Ilic. Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2646–2655, 2018. 3

[30] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 402–419, Cham, 2020. Springer International Publishing. 4, 8

[31] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 1, 3, 7

[32] Tong Wu, Liang Pan, Junzhe Zhang, Tai WANG, Ziwei Liu, and Dahua Lin. Balanced chamfer distance as a comprehensive metric for point cloud completion. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 7

[33] Yuefan Wu, Zeyuan Chen, Shaowei Liu, Zhongzheng Ren, and Shenlong Wang. CASA: Category-agnostic skeletal animal reconstruction. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1, 3

[34] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15980–15989, 2021. 1, 2, 4, 7, 8

[35] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. In *NeurIPS*, 2021. 1, 2, 4, 5

[36] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022. 1, 2, 4, 5, 7, 8

[37] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 3

[38] Juyong Zhang, Yuxin Yao, and Bailin Deng. Fast and robust iterative closest point. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 3

[39] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15893–15903, June 2022. 3

[40] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6365–6373, 2017. 3

[41] Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3