DANBO: Disentangled Articulated Neural Body Representations via Graph Neural Networks

Shih-Yang Su[1] Timur Bagautdinov[2] Helge Rhodin[1]

[1] University of British Columbia
[2] Reality Labs Research

Fig. 1: DANBO enables learning volumetric body models from scratch, only requiring a single video as input, yet enable driving by unseen poses (inset) that are out of the training distribution, showing better robustness than existing surface-free approaches. **Real faces are blurred for anonymity.**

Abstract. Deep learning greatly improved the realism of animatable human models by learning geometry and appearance from collections of 3D scans, template meshes, and multi-view imagery. High-resolution models enable photo-realistic avatars but at the cost of requiring studio settings not available to end users. Our goal is to create avatars directly from raw images without relying on expensive studio setups and surface tracking. While a few such approaches exist, those have limited generalization capabilities and are prone to learning spurious (chance) correlations between irrelevant body parts, resulting in implausible deformations and missing body parts on unseen poses. We introduce a three-stage method that induces two inductive biases to better disentangled pose-dependent deformation. First, we model correlations of body parts explicitly with a graph neural network. Second, to further reduce the effect of chance correlations, we introduce localized per-bone features that use a factorized volumetric representation and a new aggregation function. We demonstrate that our model produces realistic body shapes under challenging unseen poses and shows high-quality image synthesis. Our proposed representation strikes a better trade-off between model capacity, expressiveness, and robustness than competing methods. Project website: https://lemonatsu.github.io/danbo.

Keywords: 3D computer vision, body models, monocular, neural fields, deformation

1 Introduction

Animating real-life objects in the digital world is a long-pursued goal in computer vision and graphics, and recent advances already enable 3D free-viewpoint video, animation, and human performance retargeting [17,41,53]. Nevertheless, animating high-definition virtual avatars with user-specific appearance and dynamic motion still remains a challenge: human body and clothing deformation are inherently complex, unique, and modeling their intricate effects require dedicated approaches. Recent solutions achieve astonishing results [10,29,44,47,48] when grounding on 3D data capture in designated studio settings, e.g., with multi-camera capture systems and controlled illumination—inaccessible to the general public for building personalized models.

Less restrictive are methods relying on parametric body models [31] that learn plausible body shape, pose, and deformation from a collection of 3D scans. These methods can thereby adapt to a wide range of body shapes [4,11,36], in particular when using neural approaches to model details as a dynamic corrective [6,8,14]. Even though subject-specific details such as clothing can be learned, it remains difficult to generalize to shapes vastly different from the original scans. Moreover, the most widely used body models have restrictive commercial licenses [31] and 3D scan datasets to train these afresh are expensive.

Our goal is to learn a high-quality model with subject-specific details directly from images. Recent approaches in this class [35,46] use a neural radiance field (NeRF) that is attached to a human skeleton initialized with an off-the-shelf 3D human pose estimator. Similar to the original NeRF, the shape and appearance are modeled implicitly with a neural network that takes as input a query location and outputs density and radiance, and is only supervised with images through differentiable volume rendering. However, unlike the original that models static scenes, articulated NeRFs model time-varying body shape deformations by conditioning on per-frame body pose and representing each frame with the same underlying body model albeit in a different state. This results in an animatable full-body model that is trained directly from videos and can then can be driven with novel skeleton poses.

Not using an explicit surface poses a major difficulty as surface-based solutions exploit surface points to anchor neural features locally as vertex attributes [41], and leverage skinning weights to associate points on or close to the surface to nearby body parts [29,40]. In absence of such constraints, A-NeRF [46] uses an overparametrization by encoding 3D position relative to all body parts. Thereby dependencies between a point and body parts are learned implicitly. By contrast, NARF [35] explicitly predicts probabilities for the association of 3D points to body parts, similar to NASA [15]. However, this probability predictor is conditioned on the entire skeleton pose and is itself prone to poor generalization. Therefore, both approaches rely on large training datasets and generalization to unseen poses is limited—in particular because unrelated body parts remain weakly entangled.

In this paper, we introduce *Disentangled Articulated Neural BOdy* (DANBO), a surface-free approach that explicitly disentangles independent body parts for

learning a generalizable personalized human model from unlabelled videos. It extends the established articulated NeRF-based body modeling with two additional stages, a body part-specific volumetric encoding that exploits human skeleton structure as a prior using Graph Neural Networks (GNN) [23], and a new aggregation module. Both designs are tailored for learning from few examples and optimized to be parameter efficient. Our main contributions are the following:

- A surface-free human body model with better texture detail and improved generalization when animated.
- GNN-based encoding that disentangles features from different body parts and relies on factorized per-bone volumes for efficiency.
- A part-based feature aggregation strategy that improves on and is informed by a detailed evaluation of existing aggregation functions.

We demonstrate that our proposed DANBO results in a generalizable neural body model, with quality comparable to surface-based approaches.

2 Related Work

We start our review with general-purpose neural fields and then turn to human body modeling with a focus on animatable neural representations.

Neural fields. Neural fields [34,37,45] have attracted recent attention due to their ability to capture astonishing levels of detail. Instead of relying on explicit geometry representations such as triangle meshes or voxel grids, these methods represent the scene implicitly - as a continuous function - that maps every point in the 3D space to quantities of interest such as radiance, density, or signed distance. This approach was popularized with Neural Radiance Fields (NeRF) [34] demonstrating impressive results on reconstructing static 3D scene presentation directly from calibrated multi-view images. Various subsequent works focused on improving performance on static scenes in terms of generalization [59], level of detail [5,38], camera self-calibration [27,57], and resource efficiency [28,58]. Most relevant are deformable models that capture non-static scenes with deformation fields [16,17,43,49,53]. However, general deformation fields are unsuitable for animation and no one demonstrated that they can generalize to strongly articulated motion in monocular video.

Template-based body models. The highest level of detail can be attained with specialized performance capture systems, e.g., with dozens of cameras or a laser scanner [19]. The resulting template mesh can then be textured and deformed for capturing high-quality human performances, even from a single video [61]. Neural approaches further enable learning pose-dependent appearance and geometries [3], predict vertex displacements [20] or local primitive volumes [30] for creating fine-grained local geometry and appearance including cloth wrinkles

and freckles. The most recent ones use neural fields to learn implicit body models with the mesh providing strong supervision signals [2,10,44,51,54]. However, template creation is limited to expensive controlled studio environments, often entails manual cleaning, and high-quality ground truth annotations.

Parametric Human Body Models learn common shape and pose properties from a large corpus of 3D scans [4,11,31,36]. For classical approaches the result are factorized parameters for controlling pose, shape [11,31,36,55] and even clothing [47] that can fit to a new subject. Most prevalent is the SMPL [11,31] mesh model with a linear shape space and pose-dependent deformation. However, most existing models have restrictive commercial licenses and modeling person-specific details from images requires additional reconstruction steps.

Personalized Body Models. Learning personalized body models given only videos of a single actor is particularly challenging. Most existing approaches start from estimating a parametric surface model such as SMPL and extend it to learn specifics. For instance, one can anchor neural features spatially by associating each SMPL vertex with a learnable latent feature, and then either diffuse vertex features to the 3D space [26,41] or project the 3D query point to the SMPL surface for feature retrieval. Incorrect shape estimates and missing details can then be corrected by a subsequent neural rendering step. As texturing improves classical approaches, neural texture mapping provides additional rendering quality [29]. Another line of work makes use of SMPL blend skinning weights as initialization for learning deformation fields [40]. The deformation field maps 3D points from 3D world space to canonical space, which enables learning a canonical neural body field that predicts the radiance and density for rendering as for the classical NeRF on static scenes. While the skinning weights in SMPL provide an initialization, [40] showed that fine-tuning the deformation fields via self-supervision helps rendering unseen poses. However, relying on body models imposes the previously discussed limitations.

There are few methods that target learning neural body fields from images without relying on an explicit surface model. Closest to our method are NARF [35] and A-NeRF [46], that learn articulated body models directly from image sequences, leveraging 3D body pose estimates produced by off-the-shelf approaches [25,24]. These methods encode 3D query points with respect to each bone on the posed skeleton, and either explicitly predict blending weights [35] to select the parts of influence or rely on a neural network to learn the assignment implicitly by feeding it a large stack of all relative positions [46]. However, lacking a prior for part assignments leads to spurious dependencies between irrelevant body parts when the training poses are scarce and have low diversity [3,44]. Our approach follows the same surface-free setting but improves upon these by introducing body part disentangled representations and a new aggregation function that achieves better rendering quality and improved generalization on novel body poses. A concurrent work COAP [33] shares a similar part-disentangle concept, but differs significantly. COAP models part geometries separately from 3D scans, whereas DANBO leverages the skeleton structure as a prior to fuse information

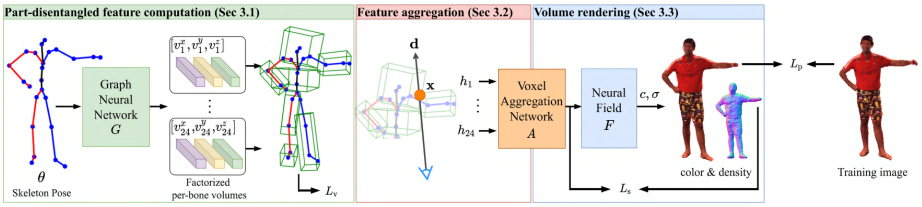

Fig. 2: Overview. The final image is generated via volume rendering by sampling points $\mathbf{x}$ along the ray $\mathbf{d}$ as in the original NeRF. Different is the conditioning on pose. First, pose features are encoded locally to every bone of a skeleton with a graph neural network using factorized volumes to increase efficiency (green boxes). Second, these disentangled features are queried and aggregated with learned weights (red module). Finally, the body shape and appearance are predicted via density and radiance fields σ and c (blue module).

from neighboring body parts, and learns both appearance and body geometry from images without 3D supervisions.

3 Method

Our goal is to learn an animatable avatar from a collections of N images $[\mathbf{I}_k]_{k=1}^N$ and the corresponding body pose in terms of skeleton joint angles $[\theta_k]_{k=1}^N$ that can stem from an off-the-shelf estimator [25,24], without using laser scans or surface tracking. We represent the human body as a neural field that moves with the input body pose. The neural field maps each 3D location to color and density to generate a free-viewpoint synthetic image via volume rendering. See Figure 2 for an overview. Our approach consists of three stages that are trained end-to-end. The first stage predicts a localized volumetric representation for each body part with a Graph Neural Network (GNN). GNN has a limited receptive field and encodes only locally relevant pose information—which naturally leads to better disentangling between body parts in the absence of surface priors. This stage is independent of the query locations and is thus executed only once per frame. Additional performance is gained by using a factorized volume and encouraging the volume bounds to be compact. The second stage retrieves a feature code for each query point by sampling volume features for all body parts that enclose the point and then aggregating the relevant ones using a separate network that predicts blend weights. Finally, the third stage maps the resulting per-query feature code to the density and radiance at that location, followed by the volume rendering as in the original NeRF.

3.1 Stage I: Part-disentangled Feature Computation

Given a pose $\theta = [\omega_1, \omega_2, \cdots, \omega_{24}]$, where $\omega_i \in \mathbb{R}^6$ [62] defines the rotation of the bone $i = 1, 2, \cdots, 24$, we represent the body part attached to each bone i with

a coarse volume V (green boxes in Fig. 2), predicted by a neural network G,

$$[V_1, V_2, \cdots, V_{24}] = G(\theta). \tag{1}$$

We design G as a GNN operating on the skeleton graph with nodes initialized to the corresponding joint angles in θ. In practice, we use two graph convolutional layers followed by per-node 2-layer MLPs. Because the human skeleton is irregular, we learn individual MLP weights for every node. See the supplemental for additional details on the graph network.

Factorized volume. A straight-forward way to represent a volume is via a dense voxel grid, which has cubic complexity with respect to its resolution. Instead, we propose to factorize each volume $V_i = (v_i^x, v_i^y, v_i^z)$ as one vector $v_i \in \mathbf{R}^{H \times M}$ for each 3D axis, where H is the voxel feature channel, and is M the volume resolution. This is similar to [42] doing a factorization into 2D planes.

Figure 3 shows how to retrieve a feature for a given 3D point $\hat{\mathbf{x}}_i$ from the volume by projecting to each axis and interpolating,

$$h_i^x = v_i^x \left[s_i^x \hat{\mathbf{x}}_i(x) \right] \in \mathbf{R}^H, \tag{2}$$

where s_i^x is a learnable scaling factor to control the volume size along the x-axis, and $v_i^x[\cdot]$ returns the interpolated feature when the projected and scaled coordinate falls in $[-1, 1]$, and $\mathbf{0}$ otherwise. The extraction for y and z axes follows the same procedure.

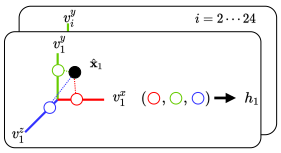

Fig. 3: We retrieve the voxel feature by projecting $\hat{\mathbf{x}}$ to the three axes and linearly interpolating the feature neighboring the projected location.

The GNN attaches one factorized volume to every bone i and is computed only once for every pose. In Section 4.4, we show that the factorized volumes compare favorably against full 3D volumes on short video sequences with sparse or single views, while having 2x lower parameter counts.

3.2 Stage II: Global Feature Aggregation

Given a query location $\mathbf{x} \in \mathbf{R}^3$ in global coordinates, the corresponding voxel feature can be retrieved by first mapping the 3D points to the bone-relative space of i via the world-to-bone coordinates transformation $T(\omega_i)$,

$$\begin{bmatrix} \hat{\mathbf{x}}_i \\ 1 \end{bmatrix} = T(\omega_i) \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}, \tag{3}$$

and then retrieving the factorized features with equation Eq. 2. However, multiple volumes can overlap.

Windowed bounds. To facilitate learning volume dimensions s_x that adapt to the body shape and to mitigate seam artifacts, we apply a window function

$$w_i = \exp(-\alpha(\hat{\mathbf{x}}_i(x)^\beta + \hat{\mathbf{x}}_i(y)^\beta + \hat{\mathbf{x}}_i(z)^\beta)) \tag{4}$$

that attenuates the feature value $h_i = w_i\,[h_i^x, h_i^y, h_i^z]$ for $\hat{\mathbf{x}}_i$ towards the boundary of the volume, with $\alpha = 2$ and $\beta = 6$ similar to [30]. Still, multiple volumes will overlap near joints and when body parts are in contact. Moreover, the overlap changes with varying skeleton pose, demanding for an explicit aggregation step.

Voxel Aggregation Network. Since an $\mathbf{x}$ that is close to the body falls into multiple volumes, we employ a voxel aggregation network A to decide which per-bone voxel features to pass on to the downstream neural field for rendering. We explore several strategies, and conduct ablation studies on each of the options. Our aggregation network A consists of a graph layer followed by per-node 2-layer MLPs with a small network width (32 per layer). We predict the weight p_i for the feature retrieved from bone i and compute the aggregated features via

$$p_i = A_i(h_i), \text{ and aggregated feature } \hat{h} = \sum_{i=1}^{24} p_i h_i. \tag{5}$$

Below, we discuss the three strategies for computing the aggregation weights.

Concatenate. Simply concatenating all features lets the network disentangle individual factors, which is prone to overfitting as no domain knowledge is used.

Softmax-OOB. Instead of simply using Softmax to obtain sparse and normalized weights as in [15,35], we can make use of our volume representation to remove the influence of irrelevant bones

$$p_i = \frac{(1 - o_i)\exp(a_i)}{\sum_{j=1}^{24}(1 - o_i)\exp(a_j)}, \tag{6}$$

where o_i is the out-of-bound (OOB) indicator which equals to 1 when $\hat{\mathbf{x}}_i$ is not inside of V_i. The potential caveat is that $\hat{h}$ is still susceptible to features from irrelevant volumes. For instance, Figure 4 shows that Softmax-OOB produces artifacts when the hand gets close to the chest.

Soft-softmax Due to the design of A, the output logit a_i of bone i is only dependent on itself. We can leverage this design to obtain the weight for each V_i independently and normalize their range to $[0, 1]$ with a sigmoid function,

$$p_i = (1 - o_i) \cdot S(a_i), \text{ where } S = \frac{1}{1 + \exp(-a_i)}. \tag{7}$$

SoftmaxOOB Soft-softmax

Fig. 4: The influence of aggregation strategies.

To nevertheless ensure that aggregated features are in the same range irrespectively of the number of contributors, we introduce a *soft-softmax* constraint

$$L_{\mathrm{s}} = \sum_{\mathbf{x}} \left(\sum_{i=1}^{24} (1 - o_i) p_i - l_{\mathbf{x}} \right)^2, \tag{8}$$

that acts as a soft normalization factor opposed to the hard normalization in softmax. By setting $l_{\mathbf{x}} = 1$ if $T_{\mathbf{x}} \cdot \sigma_{\mathbf{x}} > 0$ and 0 otherwise, the loss enforces the sum of weights of the activated bones to be close to 1 when the downstream neural field has positive density prediction σ (e.g., when $\mathbf{x}$ belong to the human body), and 0 otherwise. The results is a compromise between an unweighted sum and softmax that attained the best generalization in our experiments. A representative improvement on softmax is shown in Figure 4-right.

3.3 Stage III: Neural Field and Volume Rendering

The aggregated features $\hat{h}$ contain the coarse, pose-dependent body features at location $\mathbf{x}$. To obtain high-quality human body, we learn a neural field F to predict the refined radiance c and density σ for $\mathbf{x}$

$$(c, \sigma) = F(\hat{h}, \mathbf{d}), \tag{9}$$

where $\mathbf{d} \in \mathbf{R}^2$ is the view direction. We can then render the image of the human subject by volume rendering as in the original NeRF [34],

$$\hat{\mathbf{I}}(u, v; \theta) = \sum_{q=1}^{Q} T_q (1 - \exp(-\sigma_q \delta_q)) c_q, \quad T_q = \exp\left(- \sum_{j=1}^{q-1} \sigma_j \delta_j \right). \tag{10}$$

Given the pose θ, the predicted image color at the 2D pixel location $\hat{\mathbf{I}}(u, v; \theta)$ is computed by integrating the predicted color c_q of the Q 3D samples along $\mathbf{d}$. δ_q is the distance between neighboring samples, and T_q represents the accumulated transmittance at sample q.

3.4 Training

Our model is directly supervised with ground truth images via photometric loss

$$L_{\mathrm{p}} = \sum_{(u,v) \in \mathbf{I}} |\hat{\mathbf{I}}(u, v; \theta) - \mathbf{I}(u, v; \theta)|. \tag{11}$$

We use L1 loss to avoid overfitting to appearance changes that cannot be explained by pose deformation alone. To prevent the per-bone volumes from growing too large and taking over other volumes, we employ a volume loss on the scaling factors as in [30]

$$L_{\mathrm{v}} = \sum_{i=1}^{24} (s^x \cdot s^y \cdot s^z). \tag{12}$$

For Soft-softmax in Section 3.2, we further regularize the output weights via the self-supervised loss L_{s}.

To summarize, the training objective of our approach is

$$L = L_{\mathrm{p}} + \lambda_{\mathrm{v}} L_{\mathrm{v}} + \lambda_{\mathrm{s}} L_{\mathrm{s}}. \tag{13}$$

We set both λ_{v} and λ_{s} to 0.001 for all experiments. See the supplemental for more implementation details.

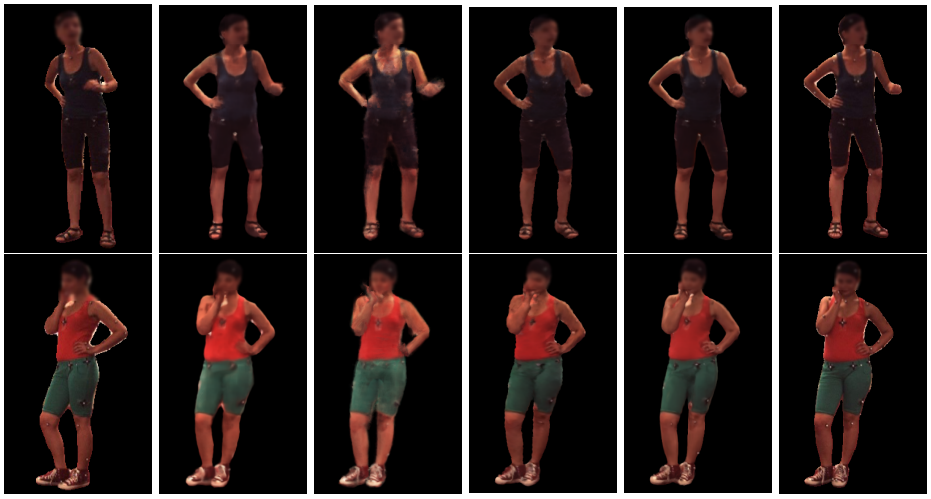

Fig. 5: **Novel-view synthesis results on Human3.6M [21].** DANBO renders more complete limbs and clearer facial features than the baselines.

4 Experiments

In the following, we evaluate the improvements upon the most recent surface-free neural body model A-NeRF [46], and compare against recent model-based solutions NeuralBody [41] and Anim-NeRF [40]. An ablation study further quantifies the improvement of using the proposed aggregation function, local GNN features, and factorized volumes over simpler and more complex [30] alternatives, including the effects on model capacity. The supplemental materials provide additional quantitative and qualitative results, including videos of retargeting applications.

Metrics and protocols. Our goal is to analyze the quality of synthesizing novel views and separately the rendering of previously unseen poses. We quantify improvements by PSNR, SSIM[52], and perceptual metrics KID [7,39] and LPIPS [60] that are resilient to slight misalignments. All scores are computed over frames withheld from training: (1) Novel view synthesis is evaluated on multi-view datasets by learning the body model from a subset of cameras with the remaining ones used as the test set, i.e. rendering the same pose from the unseen view, and (2) novel pose synthesis quality is measured by training on the first part of a video and testing on the latter frames, given their corresponding 3D pose as input. This assumes that only the pose changes as the person moves. Hence, view-dependent illumination changes in (1) but stays similar in (2).

As the background image is not our focus, we report scores on tight bounding boxes either provided by the dataset or computed from the 3D poses.

Datasets. We compare our DANBO using the established benchmarks for neural bodies, covering indoor and outdoor, and single and multi-view settings:

Table 1: **Novel-view synthesis comparisons on Human3.6M [21]**. The disentangled feature enables DANBO to achieve better novel view synthesis.

	NeuralBody [41]				Anim-NeRF [40]				A-NeRF [46]				**DANBO (Ours)**			
	PSNR ↑	SSIM ↑	KID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	KID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	KID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	KID ↓	LPIPS ↓
S1	22.88	0.897	0.048	0.153	22.74	0.896	0.106	0.156	23.93	0.912	0.042	0.153	**23.95**	**0.916**	**0.033**	**0.148**
S5	24.61	0.917	0.033	0.146	23.40	0.895	0.087	0.151	24.67	0.919	0.036	0.147	**24.86**	**0.924**	**0.029**	**0.142**
S6	22.83	0.888	0.050	0.146	22.85	0.871	0.113	0.151	23.78	0.887	0.051	0.164	**24.54**	**0.903**	**0.035**	**0.143**
S7	23.17	0.915	0.043	0.134	21.97	0.891	0.054	0.140	24.40	0.917	**0.025**	0.139	**24.45**	**0.920**	0.028	**0.131**
S8	21.72	0.894	0.071	0.177	22.82	0.900	0.095	0.178	22.70	0.907	0.086	0.196	**23.36**	**0.917**	**0.068**	**0.173**
S9	24.29	0.911	**0.035**	0.141	24.86	0.911	0.057	0.145	25.58	0.916	0.039	0.150	**26.15**	**0.925**	0.040	**0.137**
S11	23.70	0.896	0.080	0.155	24.76	0.907	0.077	0.158	24.38	0.905	**0.057**	0.164	**25.58**	**0.917**	0.060	**0.153**
Avg	23.31	0.903	0.051	0.150	23.34	0.896	0.084	0.154	24.21	0.909	0.048	0.159	**24.70**	**0.917**	**0.042**	**0.146**

Fig. 6: **Unseen pose synthesis on Human3.6M [21] test split**. Our disentangled representation enables DANBO to generate plausible geometry and deformation for held-out testing poses, and achieve better visual quality than both surface-free and surface-based baselines. Note that, unlike Anim-NeRF [40], we do not require test-time finetuning for unseen poses.

- **Human3.6M**[3] [21,22]: We follow the same evaluation protocol as in Anim-NeRF [40], with a total of 7 subjects for evaluation. The foreground maps are computed using [18].
- **MonoPerfCap** [56] features multiple outdoor sequences, recorded using a single high-resolution camera. We use the same two sequences and setting as in A-NeRF [46]: Weipeng_outdoor and Nadia_outdoor with 1151 and 1635 images, respectively, of resolution 1080×1920. Human and camera pose is estimated by SPIN [25] and refined with [46]. Foreground masks are obtained by DeepLabv3 [9].

We further include the challenging motion such as dancing and gymnastic poses from Mixamo [1] and Surreal+CMU-Mocap dataset [50,12] for motion retargeting (detailed in the supplemental). In total, we evaluate on 9 different subjects and 11 different sequences.

[3] Meta did not have access to the Human3.6M dataset.

Fig. 7: **Motion retargeting on Mixamo [1] and Surreal [12,50] dataset** with body models trained on various subjects. DANBO shows better robustness and generalization than the surface-free approach A-NeRF.

4.1 Novel View Synthesis

View synthesis of poses seen during training is simpler as the interplay between body parts is observable. Hence, our explicit disentanglement of body parts is less crucial but still beneficial. Compared to the baselines, higher detail is present and body shape is better preserved, such as visible at facial features and arm contours in Figure 5. Anim-NeRF shows slightly distorted arms and cloud-like artifacts, potentially caused by incorrectly estimated deformation fields. Table 1 verifies these improvements on the test set of Anim-NeRF [40].

Fig. 8: **DANBO better preserves body geometry, showing a less noisy surface than A-NeRF.** We extract the isosurface using Marching cubes [32] with voxel resolution 256. See the supplemental for more results.

Fig. 9: **Ablation study on Human3.6M [21] test split** novel pose (top row) and novel view (bottom row). Our proposed designs together achieve better results with less distortion on the body parts, particularly in the limbs and face.

Table 2: **Novel-pose synthesis comparisons on Human3.6M [21] (row 1-8) and MonoPerfCap [56] (row 9-11)** . Our part-disentangled design enables DANBO to generalize better to unseen poses with superior perceptual qualities.

	NeuralBody [41]				Anim-NeRF [40]				A-NeRF [46]				**DANBO (Ours)**			
	PSNR ↑	SSIM ↑	KID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	KID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	KID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	KID ↓	LPIPS ↓
S1	22.10	0.878	0.110	0.140	21.37	0.868	0.163	0.141	22.67	0.883	0.178	0.143	**23.03**	**0.895**	**0.081**	**0.135**
S5	23.52	0.897	**0.039**	0.151	22.29	0.875	0.123	0.155	22.96	0.888	0.081	0.157	**23.66**	**0.903**	0.049	**0.147**
S6	23.42	0.892	0.095	0.165	22.59	0.884	0.131	0.172	22.77	0.869	0.169	0.198	**24.57**	**0.906**	**0.052**	**0.158**
S7	22.59	0.893	0.046	0.140	22.22	0.878	0.066	0.143	22.80	0.880	0.059	0.152	**23.08**	**0.897**	**0.036**	**0.136**
S8	20.94	0.876	0.137	0.173	21.78	0.882	0.107	0.172	21.95	0.886	0.142	0.203	**22.60**	**0.904**	**0.092**	**0.167**
S9	23.05	0.885	0.043	0.141	23.73	0.886	0.068	0.141	24.16	0.889	0.074	0.152	**24.79**	**0.904**	**0.042**	**0.136**
S11	23.72	0.884	0.060	0.148	23.92	0.889	0.087	0.149	23.40	0.880	0.079	0.164	**24.57**	**0.901**	**0.040**	**0.144**
Avg	22.76	0.886	0.076	0.151	22.56	0.880	0.106	0.153	22.96	0.882	0.112	0.167	**23.76**	**0.902**	**0.056**	**0.146**
Nadia	-	-	-	-	-	-	-	-	**24.88**	**0.931**	0.048	0.115	24.44	0.921	**0.026**	**0.111**
Weipeng	-	-	-	-	-	-	-	-	**22.45**	**0.893**	0.039	0.125	22.07	0.885	**0.024**	**0.117**
Avg	-	-	-	-	-	-	-	-	**23.67**	**0.912**	0.044	0.120	23.25	0.903	**0.025**	**0.114**

4.2 Unseen Pose Synthesis and Animation

Rendering of unseen poses tests how well the learned pose-dependent deformations generalize. Figure 6 shows how differences are most prominent on limbs and faces. DANBO achieves better rendering quality and retains more consistent geometric details, generalizing well to both held out poses and out-of-distribution

Table 3: Ablation on each of the proposed modules.

Method variant	PSNR ↑	SSIM ↑
Ours w/o aggregation	17.08	0.627
Ours w/o volume	24.24	0.892
Ours w/o GNN	23.87	0.896
Ours full	**24.38**	0.**899**

Table 4: Ablation on different aggregation methods.

Aggregation methods	PSNR ↑	SSIM ↑
Softmax	23.80	0.896
Softmax-OOB	24.00	0.897
Sum-OOB	23.22	0.890
Sigmoid-OOB	23.75	0.896
Soft-softmax (Ours)	**24.38**	**0.899**

Table 5: Ablation study of different coarse volumes.

Volume type	PSNR ↑	SSIM ↑
3D Volume (SlabConv) [30]	24.17	0.892
Factorized Volume (Ours)	**24.38**	**0.899**

poses (see Figure 7). Table 2 reports the quantitative results. DANBO consistently outperforms other baselines on Human3.6M. On MonoPerCap, the high-frequency details generated by DANBO yield lower PSNR and SSIM scores, as they penalize slightly misaligned details more than the overly smoothed results by A-NeRF. The perceptual metrics properly capture DANNO's significant quality improvement by 43% on KID and 5% on LPIPS. We attribute the boost in generalization and visual quality to the improved localization via graph neural networks as well as the Soft-softmax that outperforms the default softmax baseline as used in [15,35]. The ablation study below provides further insights.

Manual animation and driving of virtual models, e.g., in VR, requires such pose synthesis, for which Figure 7 provides animation examples. Note that no quantitative evaluation is possible here as no ground truth reference image is available in this mode. Note also that the more difficult outdoor sequences are trained from a monocular video, a setting supported only by few existing approaches. The qualitative examples validate that DANBO achieves better rendering quality on most subjects, and the poses generated by DANBO are sharper, with more consistent body parts, and suffer from less floating artifacts.

4.3 Geometry Comparisons

To further validate that DANBO improves the body shape reconstruction, we visualize the learned body geometry of DANBO and A-NeRF on unseen poses of the Human3.6M [21] dataset in Figure 8. DANBO captures better body shapes and per-part geometry despite also being surface-free. A-NeRF suffers from missing and shrinking body parts, and predicts noisy density near the body surface. Besides the improved completeness, DANBO shows a smoother surface, which we attribute to our coarse per-bone volumetric representation.

4.4 Ablation Study

We conduct the ablation study on S9 of Human3.6M using the same splits as before. To speed up iteration cycles, we reduce the training iterations to 100k, and use every other pose in the training set from the default configuration. We furthermore decreased the factorized volume resolution to $M = 10$. Figure 9 shows results on both novel pose and novel view for all variants. *Proposed Modules.* In Table 3, we report how each of our proposed modules contributes to the final performance. For Ours w/o learned aggregation, we simply concatenate all the retrieved voxel features as inputs to the NeRF network, which is

similar to A-NeRF but using GNN features. This leads to poor generalization, and the w/o aggregation model predicts many floating artifacts in empty space. For Ours w/o volume, the GNN predicts per-bone feature vector instead of factorized volumes. In this variant, the aggregation network takes as input $\hat{\mathbf{x}}_i$ to predict per-bone weights. The feature to neural field F is the aggregated GNN feature and local coordinates. While the w/o volume variant achieves comparable results, the model suffers from overfitting, and produces distorted results on joints. In sum, both our aggregation network and per-bone volume designs provide useful inductive biases for learning robust and expressive models.

Aggregation Strategy. In Table 4, we show the evaluation results with different aggregation methods in Section 3.2. Note that the Softmax variant is equivalent to NARF [35] with our GNN backbone. Strategy with out-of-bound handling shows better robustness to unseen poses, with our Soft-softmax aggregation works better than Softmax-OOB, and the unweighted variant SUM-OOB being the worst.

Choice of Volume Representation. In Table 5, we show the results of using both our factorized volumes, and full 3D volume predicted using SlabConv [30]. We observe that SlabConv, while capturing finer texture details as the model is more expressive, is prone to noises in the empty space. We conclude that more views and pose data are required for using SlabConv as the volume representation.

5 Limitations and Discussion

Similar to other neural field-based approaches, the computation time for DANBO remains the limiting factor for real-time applications. While DANBO offers better generalization to unseen poses, we show in Figure 10 that in extreme cases it sometimes mixes the parts around joints together leading to deformation and blur. Handling such cases remains an open problem as also the surface-based method Anim-NeRF produces candy wrap artifacts around the elbow. It is also worth noting that DANBO is a person-specific model that needs to be trained individually for each person, which is desirable so long as sufficient training time and data is available.

GT Anim-NeRF Ours

Fig. 10: Unseen local poses create artifacts around the joints.

6 Conclusion

We presented a surface-free approach for learning an animatable human body model from video. This is practical as it applies to monocular recordings, alleviates the restrictions of template or parametric models, and works in indoor and outdoor conditions. Our contributions on encoding pose locally with a GNN, factorized volumes, and a soft aggregation function improve upon existing models in the same class and even rival recent surface-based solutions.

Acknowledgements. Shih-Yang Su and Helge Rhodin were supported by Compute Canada, Advanced Research Computing at UBC [13], and NSERC DC.

References

1. Adobe: Mixamo. https://www.mixamo.com/ (2020)
2. Alldieck, T., Xu, H., Sminchisescu, C.: imghum: Implicit generative models of 3d human shape and articulated pose. In: ICCV (2021)
3. Bagautdinov, T., Wu, C., Simon, T., Prada, F., Shiratori, T., Wei, S.E., Xu, W., Sheikh, Y., Saragih, J.: Driving-signal aware full-body avatars. ACM TOG (Proc. SIGGRAPH) (2021)
4. Balan, A.O., Sigal, L., Black, M.J., Davis, J.E., Haussecker, H.W.: Detailed human shape and pose from images. In: CVPR. pp. 1–8 (2007)
5. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. ICCV (2021)
6. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3d human reconstruction. In: ECCV (2020)
7. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: ICLR (2018)
8. Burov, A., Nießner, M., Thies, J.: Dynamic surface function networks for clothed human bodies. In: ICCV (2021)
9. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
10. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: ICCV (2021)
11. Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: ECCV (2020), https://expose.is.tue.mpg.de
12. CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu
13. Computing, U.A.R.: Ubc arc sockeye (2019). https://doi.org/doi: 10.14288/SOCKEYE.
14. Corona, E., Pumarola, A., Alenyà, G., Pons-Moll, G., Moreno-Noguer, F.: Smplicit: Topology-aware generative model for clothed people. In: CVPR (2021)
15. Deng, B., Lewis, J., Jeruzalski, T., Pons-Moll, G., Hinton, G., Norouzi, M., Tagliasacchi, A.: Nasa: neural articulated shape approximation. arXiv preprint arXiv:1912.03207 (2019)
16. Gafni, G., Thies, J., Zollhöfer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: CVPR (2021)
17. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: ICCV (2021)
18. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: ECCV (2018)
19. 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 TOG (Proc. SIGGRAPH) (2019)
20. Habermann, M., Liu, L., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Real-time deep dynamic characters. ACM TOG (Proc. SIGGRAPH) (2021)
21. Ionescu, C., Carreira, J., Sminchisescu, C.: Iterated Second-Order Label Sensitive Pooling for 3D Human Pose Estimation. In: CVPR (2014)

22. Ionescu, C., Li, F., Sminchisescu, C.: Latent Structured Models for Human Pose Estimation. In: ICCV (2011)
23. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
24. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: CVPR (2020)
25. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)
26. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. NeurIPS (2021)
27. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. arXiv preprint arXiv:2104.06405 (2021)
28. Lindell, D.B., Martel, J.N., Wetzstein, G.: Autoint: Automatic integration for fast neural volume rendering. In: CVPR (2021)
29. 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 TOG (Proc. SIGGRAPH Asia) (2021)
30. Lombardi, S., Simon, T., Schwartz, G., Zollhoefer, M., Sheikh, Y., Saragih, J.: Mixture of volumetric primitives for efficient neural rendering. ACM TOG (Proc. SIGGRAPH) (2021)
31. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM TOG (Proc. SIGGRAPH) **34**(6), 1–16 (2015)
32. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. ACM TOG (Proc. SIGGRAPH) (1987)
33. Mihajlovic, M., Saito, S., Bansal, A., Zollhoefer, M., Tang, S.: COAP: Compositional articulated occupancy of people. In: CVPR (Jun 2022)
34. 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: ECCV (2020)
35. Noguchi, A., Sun, X., Lin, S., Harada, T.: Neural articulated radiance field. In: ICCV (2021)
36. Osman, A.A.A., Bolkart, T., Black, M.J.: STAR: A sparse trained articulated human body regressor. In: ECCV (2020)
37. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: CVPR (2019)
38. 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 TOG (Proc. SIGGRAPH) (2021)
39. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: CVPR (2022)
40. 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)
41. 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)
42. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional occupancy networks. In: ECCV. Springer (2020)
43. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural Radiance Fields for Dynamic Scenes. In: CVPR (2020)
44. Saito, S., Yang, J., Ma, Q., Black, M.J.: SCANimate: Weakly supervised learning of skinned clothed avatar networks. In: CVPR (2021)

45. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. NeurIPS **33** (2020)
46. Su, S.Y., Yu, F., Zollhöfer, M., Rhodin, H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In: NeurIPS (2021)
47. Tiwari, G., Bhatnagar, B.L., Tung, T., Pons-Moll, G.: Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In: ECCV (2020)
48. Tiwari, G., Sarafianos, N., Tung, T., Pons-Moll, G.: Neural-gif: Neural generalized implicit functions for animating people in clothing. In: ICCV (2021)
49. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: ICCV (2021)
50. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: CVPR (2017)
51. Wang, S., Mihajlovic, M., Ma, Q., Geiger, A., Tang, S.: Metaavatar: Learning animatable clothed human models from few depth images. In: NeurIPS (2021)
52. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP (2004)
53. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: CVPR (2021)
54. Xu, H., Alldieck, T., Sminchisescu, C.: H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. NeurIPS (2021)
55. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In: CVPR (2020)
56. Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Mehta, D., Seidel, H.P., Theobalt, C.: Monoperfcap: Human performance capture from monocular video. TOG **37**(2), 27 (2018)
57. Yen-Chen, L., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.Y.: inerf: Inverting neural radiance fields for pose estimation. In: IROS (2020)
58. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: PlenOctrees for real-time rendering of neural radiance fields. In: ICCV (2021)
59. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural radiance fields from one or few images. In: CVPR (2021)
60. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
61. Zhou, X., Zhu, M., Pavlakos, G., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. PAMI (2018)
62. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR. pp. 5745–5753 (2019)