






UNIF: United Neural Implicit Functions for Clothed Human Reconstruction and Animation

Shenhan Qian^{1,2*}  Jiale Xu² 
Ziwei Liu³  Liqian Ma^{1†}  Shenghua Gao^{2,4,5} 
{qianshh, xujl1, gaoshh}@shanghaitech.edu.cn
zwliu.hust@gmail.com liqianma.scholar@outlook.com

¹ ZMO AI Inc.

² ShanghaiTech University

³ S-Lab, Nanyang Technological University

⁴ Shanghai Engineering Research Center of Intelligent Vision and Imaging

⁵ Shanghai Engineering Research Center of Energy Efficient and Custom AI IC

Abstract. We propose united implicit functions (UNIF), a part-based method for clothed human reconstruction and animation with raw scans and skeletons as the input. Previous part-based methods for human reconstruction rely on ground-truth part labels from SMPL and thus are limited to minimal-clothed humans. In contrast, our method learns to separate parts from body motions instead of part supervision, thus can be extended to clothed humans and other articulated objects. Our Partition-from-Motion is achieved by a bone-centered initialization, a bone limit loss, and a section normal loss that ensure stable part division even when the training poses are limited. We also present a minimal perimeter loss for SDF to suppress extra surfaces and part overlapping. Another core of our method is an adjacent part seaming algorithm that produces non-rigid deformations to maintain the connection between parts which significantly relieves the part-based artifacts. Under this algorithm, we further propose “Competing Parts”, a method that defines blending weights by the relative position of a point to bones instead of the absolute position, avoiding the generalization problem of neural implicit functions with inverse LBS (linear blend skinning). We demonstrate the effectiveness of our method by clothed human body reconstruction and animation on the CAPE and the ClothSeq datasets. Our code is available at <https://github.com/ShenhanQian/UNIF.git>.

Keywords: clothed human reconstruction, neural implicit functions, shape representation, non-rigid deformation.

1 Introduction

As residents of the 21st century, we are embracing a new life in the virtual world, digitizing everything around us. Recent research interest in human body recon-

* Work conducted during an internship at ZMO AI Inc.

† Corresponding author.

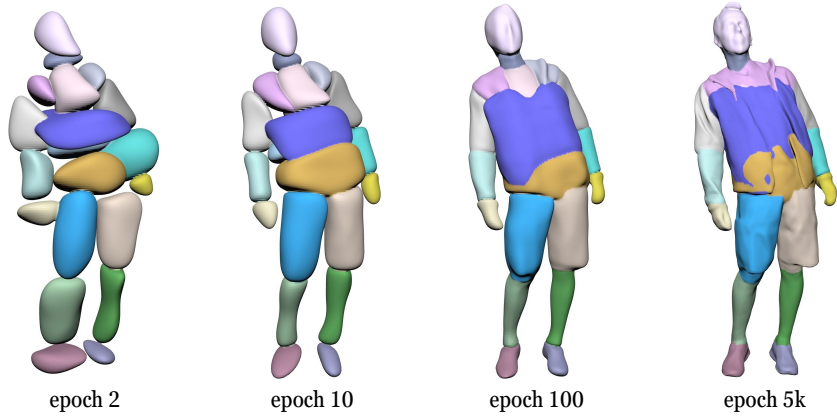


Fig. 1: The evolution of the learned parts of our model.

struction and animation increases dramatically. A popular human body model is SMPL [16], which models minimal-clothed human bodies across genders and figures. Later methods extend SMPL [16] to the clothed human body by overlying vertex offsets [18, 14] or attaching template clothing meshes [25]. However, for complex clothes, the fixed topology of SMPL [16] mesh and the predefined template clothing limit the expressiveness. Recently, the rise of neural implicit representations [5, 19, 24] indicates a higher modeling fidelity and flexibility. These models take in a point position and output an indicator of the geometry such as occupancy and SDF (signed distance function), theoretically supporting an infinitely high resolution. The infinity of resolution is perfect for fidelity but a disaster for skinning since we can no longer store the LBS (linear blend skinning) weights for every point. Although recent methods use another neural implicit function to learn the weights [30, 20, 27, 31], they generalize poorly under unseen poses because the LBS weights of a point vary along with the pose.

Besides learning a whole shape and deforming it with LBS, we can also model an object with separate parts. NASA [6] models the human body with several occupancy networks, each of which is bound to a joint. Therefore, when the skeleton moves, the learned shape is articulated. However, a key limitation of NASA [6] and later part-based methods [1, 15] is that they rely on SMPL’s LBS weights for part division, therefore still limited to minimal-clothed human reconstruction. Another shortage of previous part-based methods is that they model the non-rigid deformation crudely by simply feeding the positions of posed joints or similar pose descriptors into the networks. This results in an overfitted model that produces artifacts under novel poses especially when the training poses are limited.

To push the boundary of part-based methods, we propose UNIF (united neural implicit functions), a method that learns the shape of an object with multiple neural implicit functions. Our method features two novelties: 1) UNIF learns to decouple parts from a whole shape with no need for ground-truth partition labels;

2) UNIF models non-rigid deformations by considering the interaction between parts.

To illustrate the basic idea of automatic part division, let us consider an arm moving relative to the body. For a part-based method, we expect the arm and the body to be modeled by two separate networks. In case they are captured by one network, they will always move as a rigid one, then the model will not be able to reconstruct the same shape when the arm moves. Therefore, when minimizing the surface reconstruction loss with changing poses, we are pushing the networks to converge into separate rigid parts. We call this process *Partition-from-Motion*. However, when the training poses are limited, *e.g.*, the subject in the raw scans never moves the arms, then there will be no driving force to decouple the arm from the body. This is not a big issue for reconstruction but unacceptable for novel-pose animation since the arms can never move. To ensure a good body partition when the training poses are limited, we propose a bone limit loss and a section normal loss that constrain the boundary and the normal of each part by its neighboring joints. These terms significantly enhance the stability of our Partition-from-Motion. Furthermore, motivated by PHASE [12], we derive a minimal perimeter loss on SDF to suppress extra parts and hidden surfaces, which also contributes to a high-quality reconstruction.

Partition-from-Motion helps us separate rigid parts, but this is insufficient because non-rigid deformations are not negligible for human bodies and clothes. We propose an APS (adjacent part seaming) algorithm that deforms points to maintain the connection between parts. APS greatly relieves artifacts such as cracks and exposure of hidden surfaces. Alike other non-rigid deformation algorithms, APS also needs to define the blending weights of a point. Differently, we define blending weights not by the absolute position but by the relative position of a point to each bone and the competition between bones. Such a local definition of blending weights avoids overfitting of absolute positions, generalizing better to unseen poses.

Overall, our contributions can be summarized as:

- We propose united neural implicit functions (UNIF) for clothed human reconstruction and animation from raw scan sequences.
- We decouple rigid parts without partition labels and enhance the robustness with carefully designed initialization and regularization strategies.
- We design an adjacent part seaming (APS) algorithm for non-rigid deformation based on a localized definition of blending weights (Competing Parts).
- We show the effectiveness of our method by clothed human reconstruction and animation on the CAPE [18] and the ClothSeq [31] dataset.

2 Related Work

Our method adopts compound neural implicit functions for human body reconstruction and animation with special attention to part division and non-rigid deformation.

2.1 Neural Implicit Functions

Compared to classic geometry representations such as meshes, point clouds, and voxels that are stored as discrete elements, neural implicit functions [19, 24, 5, 2, 9, 21] are stored with neural networks. They take in the coordinate of a point and output an indicator of geometry, appearance, or other properties. Early methods need dense supervision of occupancy or SDF [19, 24, 5]. Later works make it possible to learn smooth surfaces with sparse supervision [2, 9, 12]. SAL [2] proposes a geometric initialization to realize signed distance learning with unsigned ground-truth data. Benefitting from the Eikonal loss to maintain a valid SDF field, IGR [9] only takes raw scans or triangle soups as the input. Lipman *et al.* [12] unifies SDF and occupancy and proposes a minimal perimeter loss to encourage tight surfaces. Our method follows this line of methods for its lower requirement for the data. It is also possible to model a scene or an object from 2D images without explicitly decoupling the geometry, appearance, and lighting condition [21, 34, 22, 33, 32]. For the usage of compound implicit functions, existing trials mainly lie in template-based shape learning [8, 7].

2.2 Human Body Reconstruction and Animation

As the most popular mesh-based human body model, SMPL [16] and its variations [11, 29, 26] dominate the area of human body reconstruction for its expressiveness and flexibility, supporting innumerable downstream task [14, 25, 2, 9, 28, 27, 13]. Since the new trend of neural implicit functions for shape learning, several papers [6, 20, 1, 15, 4] have attempted to substitute SMPL with an implicit counterpart for higher fidelity and flexibility. Besides the minimal-clothed human body, later works also use neural implicit functions to model clothed humans [30, 31, 23, 28, 27].

For body animation, there exist two types of pose representation - latent vector and skeleton. SAL [2], IGR [9], and NPMs [23] model body poses with a latent space, which is especially useful when no skeleton is available. But they only support interpolation between poses instead of direct animation. As to pose interpolation in the latent space, Atzmon *et al.* [3] regularize the deformation field concerning the latent vector to maintain the as-rigid-as-possible property.

Among the skeleton-based methods, the mainstream practice is to learn a canonical shape and animate it with LBS (linear blend skinning). However, since a neural implicit function lacks point-wise correspondences, a forward and a backward skinning network are introduced [30, 20, 27, 31] to save LBS weights for the bidirectional mapping between a posed shape and the canonical shape. The main limitation here is the poor generalization ability of inverse LBS since the LBS weights vary when the pose changes. SCANimate [30] and LEAP [20] use the cycle consistency to regularize the learned neural skinning weights. In contrast, SNARF [4] only learns the stable forward skinning weights and solves backward skinning by iteratively minimizing the cycle consistency error.

Aside from LBS-based methods, another series of methods model the human body with separate parts. NASA [6] merges the output of a group of occupancy

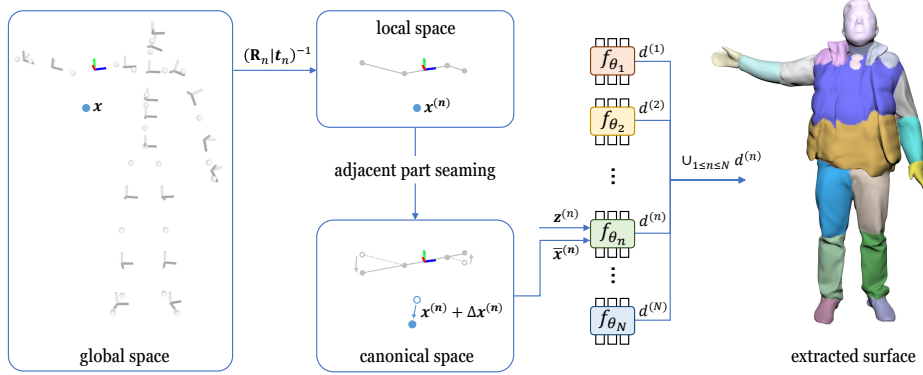


Fig. 2: The pipeline of our method. Given a point \mathbf{x} , we first transform it into the local space of each bone and then apply our adjacent part seaming algorithm to obtain its position $\bar{\mathbf{x}}^{(n)}$ in the canonical space of the n -th bone. Each neural implicit function takes the position $\bar{\mathbf{x}}^{(n)}$ and a pose condition vector $\mathbf{z}^{(n)}$ to predict the SDF value of a part $d^{(n)}$. The final output of our method is the union of all output.

networks, each anchored on a joint of the body. Both LatentHuman [15] and imGHUM [1] train combinational signed distance functions on multi-subject data. LatentHuman [15] pays special attention to relieve part based artifacts, while imGHUM [1] provides additional controllability on hands and expressions. A common feature of the above part-based methods is that they all rely on the LBS weights of SMPL to partite the body. In contrast, we learn part division from body motion. As to the non-rigid deformation, previous part-based methods [6, 15] simply feed skeleton states as an input of networks, leading to limited pose generalization ability.

3 United Neural Implicit Functions

The input of our method is a sequence of point clouds, which captures the shapes of a person in varying poses. For each frame, we fit the body skeleton (*e.g.*, the skeleton of SMPL [16]) represented by the orientations and translations of body joints. Then, we set up local coordinate systems based on the skeleton and learn a neural implicit function in each local space.

We illustrate the pipeline of our method in Fig. 2. For a point \mathbf{x} in the global space, we first transform it to the local space of each bone and get $\mathbf{x}^{(n)}$. Then we deform the point by an offset $\Delta\mathbf{x}^{(n)}$ with an adjacent part seaming (APS) algorithm and get its position $\bar{\mathbf{x}}^{(n)}$ in the canonical space of the n -th bone. Finally, we feed the position $\bar{\mathbf{x}}^{(n)}$ and a pose condition vector $\mathbf{z}^{(n)}$ to each neural implicit function and take the union of their output.

3.1 Shape Representation and Learning

Our united neural implicit functions are based on IGR [9], which adopts a single neural network to model the surface of an object. Given a point cloud $\mathcal{X} = \{\mathbf{x}_i\}_{i \in I} \subset \mathbb{R}^3$ and corresponding surface normals $\mathcal{N} = \{\mathbf{n}_i\}_{i \in I} \subset \mathbb{R}^3$, IGR [9] optimizes the parameters θ of an MLP $f_\theta(\mathbf{x})$ to approximate the signed distance function of the surface behind the point cloud \mathcal{X} with the loss

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{unit}} \mathcal{L}_{\text{unit}}, \quad (1)$$

where

$$\mathcal{L}_{\text{recon}} = \frac{1}{|I|} \sum_{i \in I} (|f_\theta(\mathbf{x}_i)| + \lambda_{\text{normal}} \|\nabla_{\mathbf{x}} f_\theta(\mathbf{x}_i) - \mathbf{n}_i\|_2), \quad (2)$$

$$\mathcal{L}_{\text{unit}} = \mathbb{E}_{\mathbf{x}} (\|\nabla_{\mathbf{x}} f_\theta(\mathbf{x})\|_2 - 1)^2. \quad (3)$$

$\mathcal{L}_{\text{recon}}$ supervises the zero-level set of f to go across \mathcal{X} with the given normals \mathcal{N} . $\mathcal{L}_{\text{unit}}$ encourages the gradient of f to be unit-norm, which is necessary for a signed distance function.

For our UNIF model, we use N ($N = 20$) separate MLPs ($f_{\theta_1}, \dots, f_{\theta_N}$), each learns the SDF of a body part. Given a point \mathbf{x} from the input point cloud \mathcal{X} , the output of UNIF is

$$d = \cup_{1 \leq n \leq N} d^{(n)}, \quad \text{with } d^{(n)} = f_{\theta_n}(\mathbf{x}^{(n)}). \quad (4)$$

\cup is an union operation on the output of all networks. Geometrically, the union of multiple signed distance functions is the minimum of all:

$$d = \min_{1 \leq n \leq N} d^{(n)}. \quad (5)$$

To ease learning and enhance robustness, we use an improved union operation, which is presented in the supplementary material. $\mathbf{x}^{(n)}$ is the local point position for the n -th part with

$$\mathbf{x}^{(n)} = \mathbf{R}_n^T (\mathbf{x} - \mathbf{t}_n), \quad (6)$$

where \mathbf{R}_n and \mathbf{t}_n are the global orientation and translation of the n -th coordinate system.

Finally, the supervision on our UNIF model becomes

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{unit}} \mathcal{L}_{\text{unit}}, \quad (7)$$

where

$$\mathcal{L}_{\text{recon}} = \frac{1}{|I|} \sum_{i \in I} (|d| + \lambda_{\text{normal}} \|\nabla_{\mathbf{x}} d - \mathbf{n}_i\|_2) \quad (\lambda_{\text{normal}} = 0.01), \quad (8)$$

$$\mathcal{L}_{\text{unit}} = \mathbb{E}_{\mathbf{x}} (\|\nabla_{\mathbf{x}} d\|_2 - 1)^2 + \frac{1}{N} \sum_{n=1}^N \mathbb{E}_{\mathbf{x}} \left(\left\| \nabla_{\mathbf{x}} d^{(n)} \right\|_2 - 1 \right)^2. \quad (9)$$

The unit-gradient-norm loss $\mathcal{L}_{\text{unit}}$ has two terms. The first term is applied on the SDF after the union operation (d), and the second term is applied on the output of each part ($d^{(n)}$). Both are necessary according to our experiments.

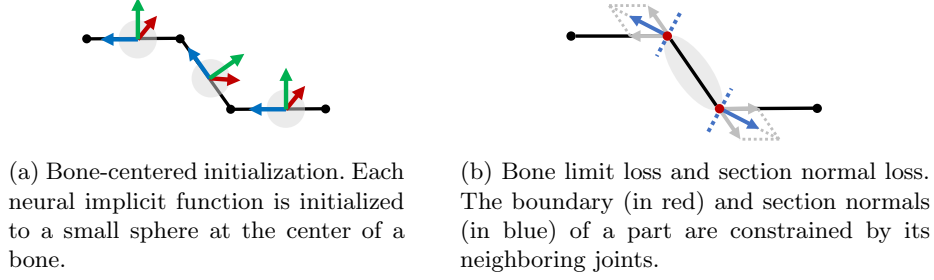


Fig. 3: Improving Partition-from-Motion with skeleton-based priors.

3.2 Partition-from-Motion

Unlike previous methods [6, 1, 15] that use ground-truth partition labels from SMPL [16], we exploit separating parts automatically while learning the entire shape. The key to achieving this is using SDF instead of occupancy because occupancy is constantly zero for locations away from the surface, while SDF provides distance information so that we can determine which part is closer to the query point and then optimize that part to go across the point. This reveals an implicit hypothesis of our method: a point should be assigned to the closest part to it. However, the SDF of a part is randomly initialized and thus may not provide correct distance information at the beginning of training. Therefore, we propose a bone-centered initialization.

Bone-centered initialization. We set up local coordinate systems at the center of bones and use the geometric initialization [2] to turn each part into a small sphere ($r = 0.01$) at the bone center (Fig. 3a). Then, parts are not intersected, and the SDF of a part approximately equals the distance to the bone center. This ensures that most points are assigned to the right part when training begins.

Bone limit loss and section normal loss. With a proper initialization, we can already separate parts, but the quality and stability of body partition highly depend on the variance of training poses. For example, when two parts barely have relative motions in the training set, they are at high risk of overlapping. This leads to artifacts when the model is animated under novel poses. Therefore, we propose a bone limit loss

$$\mathcal{L}_{\text{lim}} = \frac{1}{N \cdot |J^{(n)}|} \sum_{n=1}^N \sum_{j \in J^{(n)}} |d_j^{(n)}|, \quad (10)$$

and a section normal loss

$$\mathcal{L}_{\text{sec}} = \frac{1}{N \cdot |J^{(n)}|} \sum_{n=1}^N \sum_{j \in J^{(n)}} \left\| \nabla_{\mathbf{x}} d_j^{(n)} - \mathbf{n}_j^{(n)} \right\|_2, \quad (11)$$

where $J^{(n)}$ is the n -th bone’s adjacent joints and $|J^{(n)}|$ is the number of its adjacent joints; $d_j^{(n)}$ is the predicted SDF at joint j ; $\mathbf{n}_j^{(n)}$ is the section normal at joint j derived from the angle between adjacent bones. As illustrated by Fig. 3b, these two terms utilize the positions of joints as a prior to limit the range of a part along the axis of its bone and the normal of the sections.

Minimal perimeter loss. In experiments, our method often produces artifacts like extra surfaces, which are due to the insufficiency of the IGR [9] loss. Considering Eq. (1), the reconstruction term $\mathcal{L}_{\text{recon}}$ ensures a zero value at the positions of raw scans and the unit-norm term $\mathcal{L}_{\text{unit}}$ regularize the gradient of the neural field, but neither punish extra surfaces where no scan points lie. Inspired by PHASE [12], we propose a minimal perimeter loss specifically for SDF:

$$\mathcal{L}_{\text{perim}} = \mathbb{E}_{\mathbf{x}} \|\nabla_{\mathbf{x}} \sigma(d)\|^2 + \frac{1}{N} \sum_{n=1}^N \mathbb{E}_{\mathbf{x}} \left\| \nabla_{\mathbf{x}} \sigma(d^{(n)}) \right\|^2, \quad (12)$$

where $\sigma(x) = \frac{1}{1+e^{-\beta x}}$ (we use $\beta = 10$). This minimal perimeter loss $\mathcal{L}_{\text{perim}}$ is applied both globally and locally, similar to $\mathcal{L}_{\text{unit}}$. The global term ensures the tightness of the overall shape, while the local term suppresses the extra surfaces hidden behind the overall shape. We leave further discussion of this loss in the supplementary material.

3.3 Adjacent Part Seaming

Now, we are able to learn separate parts automatically with proper initialization and regularization. But be aware that the entire model is moving as rigid parts, obviously insufficient for either human bodies or clothes. To support non-rigid deformation, previous part-based methods [6, 1, 15] feed a descriptor of joints into networks. We construct a similar descriptor by first transforming the orientation matrices and translation vectors of all joints into each part’s local space, then flattening and concatenating them into a pose condition vector

$$\mathbf{z}^{(n)} = \oplus_{1 \leq j \leq N} (\mathbf{R}_n^T \mathbf{R}_j \oplus \mathbf{R}_n^T (\mathbf{t}_j - \mathbf{t}_n)), \quad (13)$$

where \mathbf{R}_n and \mathbf{t}_n are the orientation and translation of the n -th bone; \mathbf{R}_j and \mathbf{t}_j are the orientation and translation of the j -th joint; N is the number of parts and \oplus refers to vector concatenation. Then, our neural implicit functions become

$$d^{(n)} = f_{\theta_n}(\mathbf{x}^{(n)}, \mathbf{z}^{(n)}), 1 \leq n \leq N. \quad (14)$$

The above pose descriptor does help the network fit the training sequence but generalizes poorly to unseen poses. As demonstrated in recent comparisons [4, 31], part-based models always produce broken parts in unseen poses.

Then, can we make non-rigid deformations of part-based models generalizable to unseen poses? Here is an observation: when two linked parts have a relative rotation (Fig. 4a), some regions are squeezed while others are stretched. We believe that explicitly modeling the phenomenon is the key to relieving part-based artifacts. Therefore, we propose the adjacent part seaming (APS) algorithm.

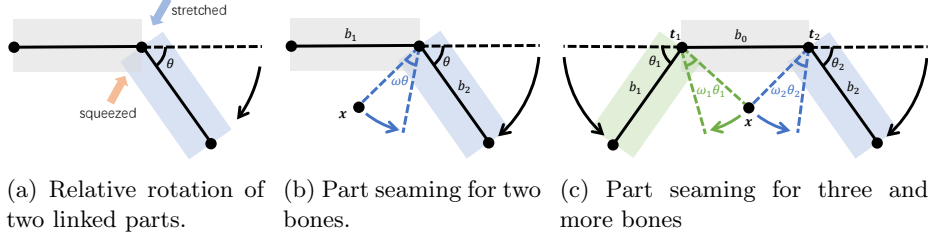


Fig. 4: 2D examples to illustrate the process of part seaming.

Adjacent part seaming by local rotations. Considering a point \mathbf{x} on the bone b_1 (Fig. 4b), the adjacent bone b_2 has rotated for an angle θ from the rest pose. What we pursue is the original position of \mathbf{x} in the rest pose. If the bone b_1 is infinitely rigid, then the point \mathbf{x} on b_1 would not have moved no matter the rotation angle of b_2 . Otherwise, \mathbf{x} should have rotated for an angle of $w\theta$, where $0 < w < 1$ (we assume the blending weight w known at the moment and will discuss it later). Then, we can obtain the original position of \mathbf{x} under the rest pose by

$$\bar{\mathbf{x}} = \mathbf{R}_{w\theta}^T \mathbf{x}, \quad (15)$$

where $\mathbf{R}_{w\theta}$ is the corresponding rotation matrix for $w\theta$. While we use a 2D example for illustration, the same process can be directly generalized to 3D cases with axis-angles.

For the skeleton of SMPL [16], a bone can be connected to up to four neighbors. We then consider the case of three connected bones, which also applies to more connections. As shown in Fig. 4c, when trying to recover a point \mathbf{x} on the bone b_0 to its original position, the point \mathbf{x} is expected to go through two different rotations. Since the axes of the two rotations are not the same, we cannot simply blend the angles. Instead, we blend the offset vectors:

$$\Delta \mathbf{x} = (\mathbf{R}_{w_1\theta_1}^T (\mathbf{x} - \mathbf{t}_1) + \mathbf{t}_1 - \mathbf{x}) + (\mathbf{R}_{w_2\theta_2}^T (\mathbf{x} - \mathbf{t}_2) + \mathbf{t}_2 - \mathbf{x}), \quad (16)$$

where \mathbf{t}_1 and \mathbf{t}_2 are the center points of the two rotations. Then we obtain the original position of \mathbf{x} by

$$\bar{\mathbf{x}} = \mathbf{x} + \Delta \mathbf{x}. \quad (17)$$

Finally, our neural implicit functions are formulated as

$$d^{(n)} = f_{\theta_n}(\bar{\mathbf{x}}^{(n)}, \mathbf{z}^{(n)}), 1 \leq n \leq N, \quad (18)$$

with

$$\bar{\mathbf{x}}^{(n)} = \mathbf{x}^{(n)} + \Delta \mathbf{x}^{(n)} = \mathbf{x}^{(n)} + \sum_{b \in B^{(n)}} (\mathbf{R}_{w_b\theta_b}^T (\mathbf{x}^{(n)} - \mathbf{t}_b) + \mathbf{t}_b - \mathbf{x}^{(n)}), \quad (19)$$

where $B^{(n)}$ is the indices of joints connected to the n -th part.

Taking a step back, you may feel the above APS algorithm is quite like inverse LBS since it cancels deformations by reversing transformations as well. But there is a contradiction for inverse LBS: it needs the LBS weights defined in the canonical space before reverting the deformation; but if we already know where to take the LBS weights in the canonical space, we do not need this inverse deformation. To evade this problem, we should avoid saving blending weights by the absolute position. Therefore, we present “Competing Parts”, a method that defines the blending weights of a point by its relative position to bones so that the blending weights can generalize to arbitrary poses.

Blending weights from “Competing Parts”.

The basic idea here is that the deformation of a point on a part is the result of the interaction between this part and its adjacent parts. We define the tendency of a point to stay static on the part as the rigidity at this point. Then, we can construct a rigidity field for a part with respect to each of its adjacent part. Taking Fig. 5 as an example, we connect the end points of both bones and split the connecting line with point Q by the ratio of bone lengths, then the rigidity of bone b_1 and bone b_2 at the point X are defined as

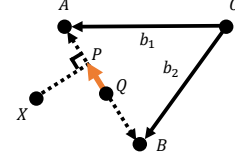


Fig. 5: A 2D example to illustrate the definition of rigidity.

$$r_1 = \exp(\alpha_1 \frac{\vec{QP} \cdot \vec{QA}}{\|\vec{QA}\|^2} + \beta_1), \quad r_2 = \exp(\alpha_2 \frac{\vec{QP} \cdot \vec{QB}}{\|\vec{QB}\|^2} + \beta_2), \quad (20)$$

where P is the projection of X onto the connecting line; α_1 and β_1 are learnable parameters to adjust the rigidity of bone b_1 ; α_2 and β_2 adjust the rigidity of bone b_2 . When the point X moves closer to bone b_1 , its rigidity about bone b_1 increases while its rigidity about bone b_2 decreases.

Based on the defined rigidity, we define the blending weights of a point with respect to bone b_1 and b_2 as

$$w_1 = \frac{r_1}{r_1 + r_2}, \quad w_2 = \frac{r_2}{r_1 + r_2}. \quad (21)$$

Given Eq. (21), $w_1 + w_2 = 1$, which is crucial for perfect part seaming. As an explanation, when two parts undergo a relative rotation for angle θ , their sections will have an angle gap of θ . To maintain the connection, the sum of relative rotations $w_1\theta + w_2\theta$ must equals to θ . Therefore, $w_1 + w_2 = 1$ is required.

3.4 Optimization

The complete supervision of our UNIF model is

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{unit}}\mathcal{L}_{\text{unit}} + \lambda_{\text{lim}}\mathcal{L}_{\text{lim}} + \lambda_{\text{sec}}\mathcal{L}_{\text{sec}} + \lambda_{\text{perim}}\mathcal{L}_{\text{perim}}, \quad (22)$$

where $\lambda_{\text{unit}} = 0.1$, $\lambda_{\text{lim}} = 1.0$, $\lambda_{\text{sec}} = 0.01$, $\lambda_{\text{perim}} = 0.001$.

We follow the same network architecture of IGR [9] except that we lower the largest width of each MLP from 256 to 64 for a comparable number of parameters. For each frame of the raw scan sequence, we sample 5k points as surface points for the reconstruction term $\mathcal{L}_{\text{recon}}$; we also sample 5k points near the surface points with local disturbances ($\sigma_{\text{local}} \sim \mathcal{N}(0, 0.1)$) and another 5k points in the enlarged bounding box ($\sigma_{\text{global}} = 1.5$) of the point cloud for the regularization terms $\mathcal{L}_{\text{unit}}$ and $\mathcal{L}_{\text{perim}}$. We use an NVIDIA A40 GPU for each experiment with 4 scans in a batch. We train our model on each subject for 5k epochs using the Adam optimizer [10] with a learning rate of 1e-3 and scale it down with a coefficient of 0.3 three times every 1k epochs. To extract surfaces from our learned neural implicit functions, we use the Marching Cubes algorithm [17] with the help of MISE [19] under a resolution of 256.

4 Experiments

4.1 Settings

Datasets. We test our method on two datasets with raw scan sequences of clothed humans. The CAPE [18] dataset contains 15 subjects registered by SMPL [16] with additional vertex offsets to model the clothes. Only four of the subjects have their raw scans released. Each of the four subjects has 4 to 6 sequences with 2 clothing types. We learn a model for each clothing type of a subject with one sequence left out for the extrapolation test. The length of each sequence ranges from about 200 to 550. For the training sequences, we use the first frame of every 10 frames for training and the fifth frame of every 10 frames for the interpolation test. The ClothSeq [31] dataset contains three subjects wearing loose clothes, therefore is more challenging. Each subject has one sequence, the length of which ranges from about 500 to 750. We use the first 80 percent frames with a stride of 10 for training, the last 20 percent frames for extrapolation test also with a stride of 10. For the interpolation test, we use frames from the first 80 percent with an offset of 5.

Baselines. NASA [6] is a typical part-based method that learns a group of occupancy networks anchored on joints. SCANimate [30] learns a forward and a backward skinning network with cycle consistency. SNARF [4] conducts backward skinning with iterative root finding to improve generalizing to novel poses.

Metrics. For quantitative evaluation, we sample 100k points from each raw scan and our extracted surface, respectively. We report four metrics during our experiments including the point-to-surface distance (p2s), the recall rate, the Chamfer distance (CD), and the F-score. The point-to-surface distance is computed by the mean distance from a point in the raw scan to its closet point on our extracted surface. The recall rate counts the ratio of points with a point-to-surface distance lower than a threshold (1 mm). The Chamfer distance is the mean of the point-to-surface and the surface-to-point distance, and the F-score is the harmonic mean of recall and precision.

Table 1: Comparison with baselines on the CAPE [18] dataset. In the upper half rows are the results of the extrapolation test, which shows the generalization ability of a model, and the lower half are from the interpolation test, which shows the expressiveness of a model.

seq.	SCANimate				SNARF				NASA				Ours				
	CD↓	F1↑	p2s↓	Rec.↑	CD↓	F1↑	p2s↓	Rec.↑	CD↓	F1↑	p2s↓	Rec.↑	CD↓	F1↑	p2s↓	Rec.↑	
E	0032-SL	10.19	66.16	10.19	67.31	10.71	65.40	10.68	66.96	98.23	15.78	103.35	15.16	8.06	75.09	7.87	75.93
	0032-SS	9.89	66.56	9.45	66.54	15.49	49.58	15.55	48.58	131.73	9.01	74.52	11.81	8.37	72.55	8.18	72.86
	0096-SL	14.53	56.97	16.89	57.25	12.19	63.70	14.35	63.93	92.74	10.69	93.72	10.53	10.40	64.36	10.04	65.54
	0096-SS	11.25	64.84	11.51	65.50	23.57	72.47	24.68	73.42	101.51	14.37	86.82	14.77	8.74	71.57	8.50	72.87
	0159-SL	7.93	75.24	7.49	76.96	29.34	68.65	33.26	67.22	118.10	8.08	153.04	7.47	6.64	82.42	6.28	83.05
	0159-SS	6.52	84.34	6.15	85.71	20.76	78.39	26.82	77.42	85.46	11.73	81.09	12.37	5.91	86.20	5.66	87.61
	3223-SL	8.12	77.95	8.60	78.28	25.29	68.29	30.17	67.20	66.91	21.31	73.49	20.06	6.24	86.77	5.47	88.99
	3223-SS	9.45	75.08	10.93	74.24	13.90	83.83	16.32	84.41	70.15	22.78	67.47	23.04	5.61	87.88	5.31	89.61
I	0032-SL	6.86	85.81	6.80	88.76	4.93	95.51	5.06	97.93	10.00	74.16	10.01	75.38	4.14	95.46	3.72	97.60
	0032-SS	5.70	90.45	5.23	93.39	4.07	96.79	3.99	98.23	10.28	68.45	10.38	69.26	4.17	95.30	3.83	97.01
	0096-SL	8.48	89.50	10.69	91.94	6.48	96.93	8.92	98.07	15.47	61.22	18.34	62.08	4.69	96.05	4.47	98.33
	0096-SS	7.08	82.76	6.47	85.05	4.05	96.41	3.84	97.84	12.73	67.40	11.29	69.12	3.74	97.08	3.42	98.94
	0159-SL	5.18	91.79	4.35	96.01	3.77	96.35	3.18	99.27	11.82	66.37	11.39	69.81	3.39	96.80	2.72	99.91
	0159-SS	4.77	93.75	4.20	96.71	3.42	97.74	3.18	99.19	12.28	65.86	12.04	67.05	2.94	98.00	2.69	99.81
	3223-SL	5.31	93.40	5.26	96.81	5.06	95.70	5.86	95.84	8.17	84.47	7.92	85.61	3.89	96.55	3.07	99.58
	3223-SS	4.89	94.09	4.74	97.15	3.76	97.68	3.88	99.24	7.80	86.61	6.95	87.93	3.09	97.81	2.84	99.68

Table 2: Comparison with baselines on the ClothSeq [31] dataset. In the upper half rows are the results of the extrapolation test, and the lower half are from the interpolation test.

seq.	SCANimate				SNARF				NASA				Ours				
		CD↓	F1↑	p2s↓	Rec.↑	CD↓	F1↑	p2s↓	Rec.↑	CD↓	F1↑	p2s↓	Rec.↑	CD↓	F1↑	p2s↓	Rec.↑
E	JP	14.33	56.25	14.29	58.02	17.72	58.49	21.58	59.16	69.76	16.88	68.24	16.59	13.04	58.60	11.24	62.06
	JS	11.05	61.26	10.85	62.58	13.40	57.48	13.91	57.72	116.14	8.51	97.25	9.61	11.84	65.94	9.00	70.03
	SP	14.32	54.06	14.19	54.80	15.06	60.22	16.47	60.19	65.73	19.93	39.26	21.29	12.10	65.90	9.81	69.46
I	JP	10.05	71.78	7.38	79.40	8.43	80.82	8.67	83.92	22.37	44.95	21.97	45.42	7.87	84.66	5.47	90.01
	JS	8.84	74.84	7.77	78.33	8.81	80.20	7.86	81.80	33.66	31.61	34.35	31.34	8.89	81.48	5.96	86.72
	SP	13.20	57.74	12.52	59.28	11.21	73.04	11.08	74.40	48.69	38.02	33.54	40.06	10.18	75.49	7.42	80.31

4.2 Comparisons

We show quantitative results in Table 1 and Table 2 and qualitative results in Fig. 6. Our method shows clear superiority over NASA [6] (also a part-based method) and outperforms SCANimate [30] and SNARF [4] in most cases.

The extrapolation test is extremely challenging, especially on CAPE [18] because the test poses differ a lot from the limited training poses. Therefore, NASA [6] completely collapses; SCANimate [30] and SNARF [4] produce distortions due to bad neural skinning weights. Our method shows higher robustness under novel poses, benefiting from the proper part division and the generalizable non-rigid deformation modeling.

In the interpolation test, NASA [6] exhibits reasonable results on CAPE [18] but has difficulty in partitioning and reconstructing the subjects in ClothSeq [31]. Our method produces visually comparable results with SCANimate [30] and SNARF [4] on CAPE [18]. However, on the ClothSeq [31] dataset, where clothes are much more complex, our method makes less body distortion or extra surfaces and reconstructs the pose of the subjects more precisely.

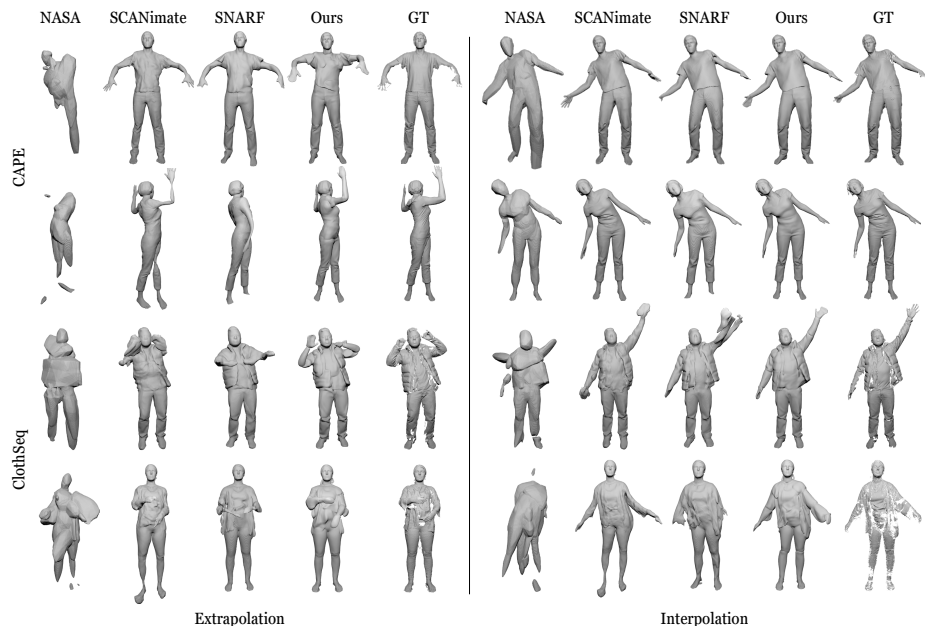


Fig. 6: Qualitative comparison with baselines.

4.3 Visualization and Analysis

Ablation Study. To validate the effectiveness of our main components, we run experiments with each of them disabled and visualize parts in an unseen pose at the early stage of training in Fig. 7. Compared to our full model, dropping the adjacent part seaming algorithm leaves the model almost rigid (*e.g.*, the sections near the knees are exposed and the neck is not completely connected to the body). When disabling the bone limit loss, we lose the restriction on the boundary of a part. Then we see the left foot and leg of the man falsely included in the same part. However, merely using the bone limit loss is not sufficient. If we drop the section normal loss, the model converges to a bad partition where the surface does go across the joint but the main body of the part lies somewhere else. Finally, the minimal perimeter loss is also necessary to suppress extra surfaces such as the one on the right leg.

Limitations. Since the #13 and #14 joints of SMPL [16] are too close to the spine, our method learns a small chest and large shoulders. When shoulders move drastically, our model converges to overlapped parts to reconstruct the shape. Therefore, we can observe inconsistent part division around the chest during animation (Fig. 8). The current framework does not model the dynamics of loose clothes. Seams between parts are still visible due to the generalization problem caused by the pose condition vector.

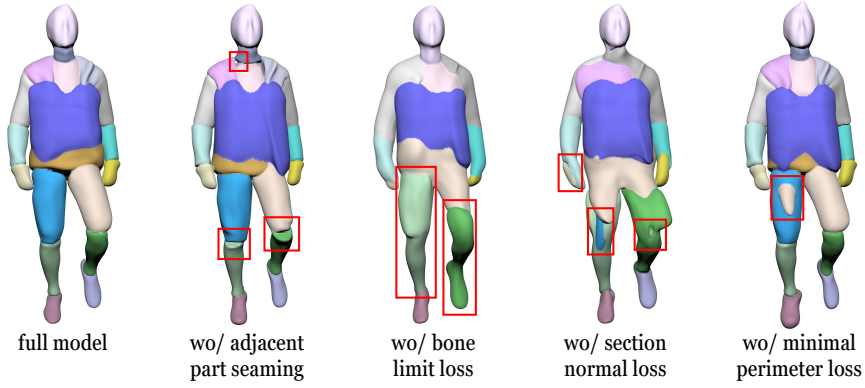


Fig. 7: Ablation study on main components of our method.



Fig. 8: Visualization of learned parts during animation.

5 Conclusions

We present a novel method for clothed human reconstruction and animation. We explore initialization and regularization strategies to learn body parts without ground-truth part labels. Towards a higher generalization ability to novel poses, we propose an adjacent part seaming algorithm to model non-rigid deformations by explicitly modeling the interaction between parts. Experiments on two datasets validate the effectiveness of our method.

Acknowledgments: The work is supported by National Key R&D Program of China (2018AAA0100704), NSFC #61932020, #62172279, Science and Technology Commission of Shanghai Municipality (Grant No. 20ZR1436000), and "huguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission. This work is supported by NTU NAP, MOE AcRF Tier 2 (T2EP20221-0033), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

Bibliography

- [1] Alldieck, T., Xu, H., Sminchisescu, C.: imghum: Implicit generative models of 3d human shape and articulated pose. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5461–5470 (2021)
- [2] Atzmon, M., Lipman, Y.: Sal: Sign agnostic learning of shapes from raw data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2565–2574 (2020)
- [3] Atzmon, M., Novotny, D., Vedaldi, A., Lipman, Y.: Augmenting implicit neural shape representations with explicit deformation fields. arXiv preprint arXiv:2108.08931 (2021)
- [4] Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11594–11604 (2021)
- [5] Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5939–5948 (2019)
- [6] Deng, B., Lewis, J.P., Jeruzalski, T., Pons-Moll, G., Hinton, G., Norouzi, M., Tagliasacchi, A.: Nasa neural articulated shape approximation. In: European Conference on Computer Vision. pp. 612–628. Springer (2020)
- [7] Genova, K., Cole, F., Sud, A., Sarna, A., Funkhouser, T.: Local deep implicit functions for 3d shape. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4857–4866 (2020)
- [8] Genova, K., Cole, F., Vlastic, D., Sarna, A., Freeman, W.T., Funkhouser, T.: Learning shape templates with structured implicit functions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7154–7164 (2019)
- [9] Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. In: ICML (2020)
- [10] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR) (2014)
- [11] Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.* **36**(6), 194–1 (2017)
- [12] Lipman, Y.: Phase transitions, distance functions, and implicit neural representations. arXiv preprint arXiv:2106.07689 (2021)
- [13] 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 Transactions on Graphics (TOG)* **40**(6), 1–16 (2021)
- [14] Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., Gao, S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and

- novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5904–5913 (2019)
- [15] Lombardi, S., Yang, B., Fan, T., Bao, H., Zhang, G., Pollefeys, M., Cui, Z.: Latenthuman: Shape-and-pose disentangled latent representation for human bodies. In: 2021 International Conference on 3D Vision (3DV). pp. 278–288. IEEE (2021)
 - [16] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* **34**(6), 1–16 (2015)
 - [17] Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics* **21**(4), 163–169 (1987)
 - [18] Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3d people in generative clothing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6469–6478 (2020)
 - [19] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4460–4470 (2019)
 - [20] Mihajlovic, M., Zhang, Y., Black, M.J., Tang, S.: Leap: Learning articulated occupancy of people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10461–10471 (2021)
 - [21] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)
 - [22] Oechsle, M., Peng, S., Geiger, A.: Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5589–5599 (2021)
 - [23] Palafox, P., Božič, A., Thies, J., Nießner, M., Dai, A.: Npms: Neural parametric models for 3d deformable shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12695–12705 (2021)
 - [24] Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 165–174 (2019)
 - [25] Patel, C., Liao, Z., Pons-Moll, G.: Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7365–7375 (2020)
 - [26] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019)

- [27] Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14314–14323 (2021)
- [28] Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9054–9063 (2021)
- [29] Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* **36**(6) (Nov 2017)
- [30] Saito, S., Yang, J., Ma, Q., Black, M.J.: Scanimate: Weakly supervised learning of skinned clothed avatar networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2886–2897 (2021)
- [31] Tiwari, G., Sarafianos, N., Tung, T., Pons-Moll, G.: Neural-gif: Neural generalized implicit functions for animating people in clothing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11708–11718 (2021)
- [32] Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS* (2021)
- [33] Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems* **34** (2021)
- [34] Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., Lipman, Y.: Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems* **33**, 2492–2502 (2020)