MetaAvatar: Learning Animatable Clothed Human Models from Few Depth Images

Shaofei Wang[1]
shaofei.wang@inf.ethz.ch

Marko Mihajlovic[1]
marko.mihajlovic@inf.ethz.ch

Qianli Ma[1,2]
qianli.ma@tue.mpg.de

Andreas Geiger[2,3]
a.geiger@uni-tuebingen.de

Siyu Tang[1]
siyu.tang@inf.ethz.ch

[1]ETH Zürich [2]Max Planck Institute for Intelligent Systems, Tübingen [3]University of Tübingen

Abstract

In this paper, we aim to create generalizable and controllable neural signed distance fields (SDFs) that represent clothed humans from monocular depth observations. Recent advances in deep learning, especially neural implicit representations, have enabled human shape reconstruction and controllable avatar generation from different sensor inputs. However, to generate realistic cloth deformations from novel input poses, watertight meshes or dense full-body scans are usually needed as inputs. Furthermore, due to the difficulty of effectively modeling pose-dependent cloth deformations for diverse body shapes and cloth types, existing approaches resort to per-subject/cloth-type optimization from scratch, which is computationally expensive. In contrast, we propose an approach that can quickly generate realistic clothed human avatars, represented as controllable neural SDFs, given only monocular depth images. We achieve this by using meta-learning to learn an initialization of a hypernetwork that predicts the parameters of neural SDFs. The hypernetwork is conditioned on human poses and represents a clothed neural avatar that deforms non-rigidly according to the input poses. Meanwhile, it is meta-learned to effectively incorporate priors of diverse body shapes and cloth types and thus can be much faster to fine-tune, compared to models trained from scratch. We qualitatively and quantitatively show that our approach outperforms state-of-the-art approaches that require complete meshes as inputs while our approach requires only depth frames as inputs and runs orders of magnitudes faster. Furthermore, we demonstrate that our meta-learned hypernetwork is very robust, being the first to generate avatars with realistic dynamic cloth deformations given as few as 8 monocular depth frames.

1 Introduction

Representing clothed humans as neural implicit functions is a rising research topic in the computer vision community. Earlier works in this direction address geometric reconstruction of clothed humans from static monocular images [32, 33, 59, 60], RGBD videos [34, 35, 67, 74, 76] or sparse point clouds [11] as direct extensions of neural implicit functions for rigid objects [10, 42, 43, 48]. More recent works advocate to learn shapes in a canonical pose [7, 26, 71] in order to not only handle

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

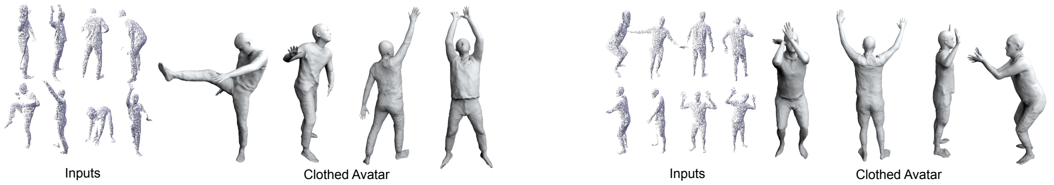

Figure 1: Given as few as 8 monocular depth images and their SMPL fittings, our meta-learned model yields a controllable neural SDF in 2 minutes which synthesizes realistic cloth deformations for unseen body poses. Here we show results of two different subjects wearing different clothes.

reconstruction, but also build controllable neural avatars from sensor inputs. However, these works do not model pose-dependent cloth deformation, limiting their realism.

On the other hand, traditional parametric human body models [38, 47, 50, 73] can represent pose-dependent soft tissue deformations of minimally-clothed human bodies. Several recent methods [12, 45] proposed to learn neural implicit functions to approximate such parametric models from watertight meshes. However, they cannot be straightforwardly extended to model clothed humans. SCANimate [61] proposed to learn canonicalized dynamic neural Signed Distance Fields (SDFs) controlled by human pose inputs and trained with Implicit Geometric Regularization (IGR [20]), thus circumventing the requirement of watertight meshes. However, SCANimate works only on dense full-body scans with accurate surface normals and further requires expensive per-subject/cloth-type training. These factors limit the applicability of SCANimate for building personalized human avatars from commodity RGBD sensors.

Contrary to all the aforementioned works, we propose to use meta-learning to effectively incorporate priors of dynamic neural SDFs of clothed humans, thus enabling fast fine-tuning (few minutes) for generating new avatars given only a few monocular depth images of unseen clothed humans as inputs. More specifically, we build upon recently proposed ideas of meta-learned initialization for implicit representations [63, 68] to enable fast fine-tuning. Similar to [63], we represent a specific category of objects (in our case, clothed human bodies in the canonical pose) with a neural implicit function and use meta-learning algorithms such as [15, 46] to learn a meta-model. However, unlike [63, 68], where the implicit functions are designed for static reconstruction, we target the generation of dynamic neural SDFs that are *controllable* by user-specified body poses. We observe that directly conditioning neural implicit functions (represented as a multi-layer perceptron) on body poses lacks the expressiveness to capture high-frequency details of diverse cloth types, and hence propose to meta-learn a hypernetwork [24] that predicts the parameters of the neural implicit function. Overall, the proposed approach, which we name *MetaAvatar*, yields controllable neural SDFs with dynamic surfaces in minutes via fast fine-tuning, given only a few depth observations of an unseen clothed human and the underlying SMPL [38] fittings (Fig. 1) as inputs. Code and data are public at https://neuralbodies.github.io/metavatar/.

2 Related Work

Our approach lies at the intersection of clothed human body modeling, neural implicit representations, and meta-learning. We review related works in the following.

Clothed Human Body Modeling: Earlier works for clothed human body modeling utilize parametric human body models [5, 25, 27, 38, 47, 50, 73] combined with deformation layers [2, 3, 7, 8] to model cloth deformations. However, these approaches cannot model fine clothing details due to their fixed topology, and they cannot handle pose-dependent cloth deformations. Mesh-based approaches that handle articulated deformations of clothes either require accurate surface registration [30, 40, 75, 79] or synthetic data [21, 23, 49] for training. Such requirement for data can be freed by using neural implicit surfaces [9, 61, 69]. For example, SCANimate [61] proposed a weakly supervised approach to learn dynamic clothed human body models from 3D full-body scans which only requires minimally-clothed body registration. However, its training process usually takes one day for each subject/cloth-type combination and requires accurate surface normal information extracted from dense scans. Recent point-based clothed human models [39, 41, 77] can also be learned from unregistered data. Like our method, concurrent work [41] also models pose-dependent shapes across different subjects/cloth-types, but it requires full-body scans for training. In contrast, our approach enables learning of clothed body models in minutes from as few as 8 depth images.

Neural Implicit Representations: Neural implicit representations [10, 42, 43, 48, 51] have been used to tackle both image-based [26, 32, 33, 53, 55, 59, 60, 82] and point cloud-based [7, 11] clothed human reconstruction. Among these works, ARCH [26] was the first one to represent clothed human bodies as a neural implicit function in a canonical pose. However, ARCH does not handle pose-dependent cloth deformations. Most recently, SCANimate [61] proposed to condition neural implicit functions on joint-rotation vectors (in the form of unit quaternions), such that the canonicalized shapes of the neural avatars change according to the joint angles of the human body, thus representing pose-dependent cloth deformations. However, diverse and complex cloth deformations make it hard to learn a unified prior from different body shapes and cloth types, thus SCANimate resorts to per-subject/cloth-type training which is computationally expensive.

Meta-Learning: Meta-learning is typically used to address few-shot learning, where a few training examples of a new task are given, and the model is required to learn from these examples to achieve good performance on the new task [1, 13, 14, 16–18, 22, 28, 29, 54, 56, 58, 62, 66, 70, 72, 78]. We focus on optimization-based meta-learning, where Model-Agnostic Meta Learning (MAML [15]), Reptile [46] and related alternatives are typically used to learn such models [4, 6, 19, 31, 36, 57]. In general, this line of algorithms tries to learn a "meta-model" that can be updated quickly from new observations with only few gradient steps. Recently, meta-learning has been used to learn a universal initialization of implicit representations for static neural SDFs [63] and radiance fields [68]. MetaSDF [63] demonstrates that only a few gradient update steps are needed to achieve comparable or better results than slower auto-decoder-based approaches [48]. However, [63, 68] only meta-learn static representations, whereas we are interested in dynamic representations conditioned on human body poses. To our best knowledge, we are the first to meta-learn the hypernetwork to generate the parameters of neural SDF networks.

3 Fundamentals

We start by briefly reviewing the linear blend skinning (LBS) method [38] and the recent implicit skinning networks [45, 61] that learn to predict skinning weights of cloth surfaces in a weakly supervised manner. Using the learned implicit skinning networks allows us to canonicalize meshes or depth observations of clothed humans, given only minimally-clothed human body model registrations to the meshes. Canonicalization of meshes or points is a necessary step as the dynamic neural SDFs introduced in Section 4 are modeled in canonical space.

3.1 Linear Blend Skinning

Linear blend skinning (LBS) is a commonly used technique to deform parametric human body models [5, 25, 38, 47, 50, 73] according to user-specified rigid bone transformations. Given a set of N points in a canonical space, $\hat{\mathbf{X}} = \{\hat{\mathbf{x}}^{(i)}\}_{i=1}^{N}$, LBS takes a set of rigid bone transformations (in our case we use 23 local transformations plus one global transformation, assuming an underlying SMPL model) $\{\mathbf{B}_b\}_{b=1}^{24}$ as inputs, each $\mathbf{B}_b$ being a 4×4 rotation-translation matrix. For a 3D point $\hat{\mathbf{x}}^{(i)} \in \hat{\mathbf{X}}$ [1], a skinning weight vector is a probability simplex $\mathbf{w}^{(i)} \in [0,1]^{24}$, s.t. $\sum_{b=1}^{24} \mathbf{w}_b^{(i)} = 1$, that defines the affinity of the point $\hat{\mathbf{x}}^{(i)}$ to each of the bone transformations $\{\mathbf{B}_b\}_{b=1}^{24}$. The set of transformed points $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ of the clothed human is related to $\hat{\mathbf{X}}$ via:

$$\mathbf{x}^{(i)} = LBS\left(\hat{\mathbf{x}}^{(i)}, \{\mathbf{B}_b\}, \mathbf{w}^{(i)}\right) = \left(\sum_{b=1}^{24} \mathbf{w}_b^{(i)} \mathbf{B}_b\right) \hat{\mathbf{x}}^{(i)}, \quad \forall i = 1, \ldots, N \tag{1}$$

$$\hat{\mathbf{x}}^{(i)} = LBS^{-1}\left(\mathbf{x}^{(i)}, \{\mathbf{B}_b\}, \mathbf{w}^{(i)}\right) = \left(\sum_{b=1}^{24} \mathbf{w}_b^{(i)} \mathbf{B}_b\right)^{-1} \mathbf{x}^{(i)}, \quad \forall i = 1, \ldots, N \tag{2}$$

[1] with slight abuse of notation, we also use $\hat{\mathbf{x}}$ to represent points in homogeneous coordinates when necessary.

where Eq. (1) is referred to as the LBS function and Eq. (2) is referred to as the inverse-LBS function. The process of applying Eq. (1) to all points in $\hat{\mathbf{X}}$ is often referred to as *forward skinning* while the process of applying Eq. (2) is referred to as *inverse skinning*.

3.2 Implicit Skinning Networks

Recent articulated implicit representations [45, 61] have proposed to learn functions that predict the forward/inverse skinning weights for arbitrary points in $\mathbb{R}^3$. We follow this approach, but take advantage of a convolutional point-cloud encoder [52] for improved generalization. Formally, we define the implicit forward and inverse skinning networks as $h_{\text{fwd}}(\cdot,\cdot) : (\mathbb{R}^{3\times K}, \mathbb{R}^3) \mapsto \mathbb{R}^{24}$ and $h_{\text{inv}}(\cdot,\cdot) : (\mathbb{R}^{3\times K}, \mathbb{R}^3) \mapsto \mathbb{R}^{24}$, respectively. Both networks take as input a point cloud with K points and a query point for which they predict skinning weights. Therefore, we can analogously re-define Eq. (1, 2) respectively as:

$$\mathbf{x}^{(i)} = \left(\sum_{b=1}^{24} h_{\text{fwd}}(\hat{\mathbf{X}}, \hat{\mathbf{x}}^{(i)})_b \mathbf{B}_b \right) \hat{\mathbf{x}}^{(i)}, \quad \forall i = 1, \ldots, N \tag{3}$$

$$\hat{\mathbf{x}}^{(i)} = \left(\sum_{b=1}^{24} h_{\text{inv}}(\mathbf{X}, \mathbf{x}^{(i)})_b \mathbf{B}_b \right)^{-1} \mathbf{x}^{(i)}, \quad \forall i = 1, \ldots, N \tag{4}$$

Training the Skinning Network: We follow the setting of SCANimate [61], where a dataset of observed point clouds $\{\mathbf{X}\}$ and their underlying SMPL registration are known. For a sample $\mathbf{X}$ in the dataset, we first define the re-projected points $\bar{\mathbf{X}} = \{\bar{\mathbf{x}}\}_{i=1}^{N}$ as $\mathbf{X}$ mapped to canonical space via Eq. (4) and then mapped back to transformed space via Eq. (3). We then define the training loss:

$$\mathcal{L}(\mathbf{X}) = \lambda_r \mathcal{L}_r + \lambda_s \mathcal{L}_s + \lambda_{skin} \mathcal{L}_{skin} \,, \tag{5}$$

where $\mathcal{L}_r$ represents a re-projection loss that penalizes the L2 distance between an input point $\mathbf{x}$ and the re-projected point $\bar{\mathbf{x}}$, $\mathcal{L}_s$ represents L1 distances between the predicted forward skinning weights and inverse skinning weights, and $\mathcal{L}_{skin}$ represents the L1 distances between the predicted (forward and inverse) skinning weights and the barycentrically interpolated skinning weights $\mathbf{w}^{(i)}$ on the registered SMPL shape that is closest to point $\mathbf{x}^{(i)}$; please refer to the Supp. Mat. for hyperparameters and details.

We train two skinning network types, the first one takes a partial point cloud extracted from a depth image as input and performs the *inverse skinning*, while the second one takes a full point cloud sampled from iso-surface points generated from the dynamic neural SDF in the canonical space and performs *forward skinning*.

Canonicalization: We use the learned inverse skinning network to canonicalize complete or partial point clouds $\{\hat{\mathbf{X}}\}$ via Eq. (4) which are further used to learn the canonicalized dynamic neural SDFs.

4 MetaAvatar

Our approach meta-learns a unified clothing deformation prior from the training set that consists of different subjects wearing different clothes. This meta-learned model is further efficiently fine-tuned to produce a dynamic neural SDF from an arbitrary amount of fine-tuning data of unseen subjects. In extreme cases, MetaAvatar requires as few as 8 depth frames and takes only 2 minutes for fine-tuning to yield a subject/cloth-type-specific dynamic neural SDF (Fig. 1).

We assume that each subject/cloth-type combination in the training set has a set of registered bone transformations and canonicalized points, denoted as $\{\{\mathbf{B}_b\}_{b=1}^{24}, \hat{\mathbf{X}}\}$. Points in $\hat{\mathbf{X}}$ are normalized to the range $[-1,1]^3$ according to their corresponding registered SMPL shape. With slight abuse of notation, we also define $\mathbf{X}$ as all possible points in $[-1,1]^3$. Our goal is to meta-learn a hypernetwork [24, 65] which takes $\{\mathbf{B}_b\}_{b=1}^{24}$ ($\{\mathbf{B}_b\}$ for shorthand) as inputs and predicts *parameters of the neural SDFs* in the canonical space. Denoting the hypernetwork as $g_\psi(\{\mathbf{B}_b\})$ and the predicted neural SDF as $f_\phi(\mathbf{x})|_{\phi=g_\psi(\{\mathbf{B}_b\})}$, we use the following IGR [20] loss to supervise the learning of g:

$$\begin{aligned}\mathcal{L}_{\text{IGR}}(f_\phi(\hat{\mathbf{X}})|_{\phi=g_\psi(\{\mathbf{B}_b\})}) = \sum_{\mathbf{x}\in\hat{\mathbf{X}}} \lambda_{sdf}\left|f_\phi(\mathbf{x})|_{\phi=g_\psi(\{\mathbf{B}_b\})}\right| &+ \lambda_{\mathbf{n}}\left(1-\langle\mathbf{n}(\mathbf{x}), \nabla_{\mathbf{x}} f_\phi(\mathbf{x})|_{\phi=g_\psi(\{\mathbf{B}_b\})}\rangle\right) \\ &+ \lambda_E\left|\|\nabla_{\mathbf{x}} f_\phi(\mathbf{x})|_{\phi=g_\psi(\{\mathbf{B}_b\})}\|_2 - 1\right| \qquad \text{(on-surface loss)} \\ + \sum_{\mathbf{x}\sim\mathbf{X}\backslash\hat{\mathbf{X}}} \lambda_O \exp\left(-\alpha\cdot\left|f_\phi(\mathbf{x})|_{\phi=g_\psi(\{\mathbf{B}_b\})}\right|\right)& \\ &+ \lambda_E\left|\|\nabla_{\mathbf{x}} f_\phi(\mathbf{x})|_{\phi=g_\psi(\{\mathbf{B}_b\})}\|_2 - 1\right| \qquad \text{(off-surface loss)}\end{aligned} \tag{6}$$

where $\mathbf{n}_\mathbf{x}$ is the surface normal of point $\mathbf{x}$. We assume that this information, along with the ground-truth correspondences from transformed space to canonical space is available when learning the meta-model on the training set, but do not require this for fine-tuning g on unseen subjects.

In practice, we found that directly learning the hypernetwork g_ψ via Eq. (6) does not converge, and thus we decompose the meta-learning of g_ψ into two steps. First, we learn a meta-SDF [63] (without conditioning on $\{\mathbf{B}_b\}$, Sec. 4.1), and then we meta-learn a hypernetwork that takes $\{\mathbf{B}_b\}$ as input and predicts the *residuals* to the parameters of the previously learned meta-SDF (Sec. 4.2).

4.1 Meta-learned Initialization of Static Neural SDFs

To effectively learn a statistical prior of clothed human bodies, we ignore the input bone transformations $\{\mathbf{B}_b\}$ and meta-learn the static neural SDF $f_\phi(\mathbf{x}) : [-1,1]^3 \mapsto \mathbb{R}$, parameterized by ϕ, from all canonicalized points of subjects *with different genders, body shapes, cloth types, and poses*. Furthermore, for faster and more stable convergence, the neural SDF f_ϕ function additionally leverages the periodic activation functions [64].

The full meta-learning algorithm for the static neural SDFs is described in Alg. 1.

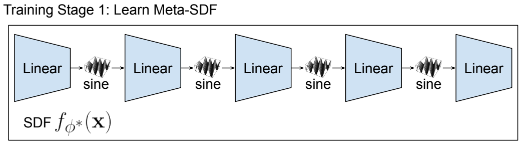
Training Stage 1: Learn Meta-SDF
Linear
sine
Linear
sine
Linear
sine
Linear
sine
Linear
SDF $f_{\phi^*}(\mathbf{x})$

Figure 2: **Overview of the meta-SDF network.** We use a 5-layer SIREN [63] network with 256 neurons for each layer.

Algorithm 1 Meta-learning SDF with Reptile [46]

```
Initialize: meta-network parameters $\phi$, meta learning rate $\beta$, inner learning rate $\alpha$, max training iteration $N$, inner-loop iteration $m$, batch size $M$
for $i = 1, \dots, N$ do
    Sample a batch of $M$ training samples $\{\hat{\mathbf{X}}^{(j)}\}_{j=1}^{M}$
    for $j = 1, \dots, M$ do
        $\phi_0^{(j)} = \phi$
        for $k = 1, \dots, m$ do
            $\phi_k^{(j)} = \phi_{k-1}^{(j)} - \alpha\nabla_\phi \mathcal{L}_{\text{IGR}}(f_\phi(\hat{\mathbf{X}}^{(j)})|_{\phi=\phi_{k-1}^{(j)}})$
        end for
    end for
    $\phi \leftarrow \phi + \beta\frac{1}{M}\sum_{j=1}^{M}(\phi_m^{(j)} - \phi)$
end for
```

4.2 Meta-learned Initialization of HyperNetwork for Dynamic Neural SDFs

The meta-learned static neural SDF explained in the previous section can efficiently adapt to new observations, however it is not controllable by user-specified bone transformations $\{\mathbf{B}_b\}$. Therefore, to enable non-rigid pose-dependent cloth deformations, we further meta-learn a hypernetwork [24] to predict *residuals* to the learned parameters of the meta-SDF in Alg. 1.

The key motivation for meta-learning the hypernetwork is to build an effective unified prior for articulated clothed humans, which enables the recovery of the non-rigid clothing deformations at test time via the efficient fine-tuning process from several depth images of unseen subjects.

Algorithm 2 Meta-learning hypernetwork with Modified Reptile

```
    Initialize: meta-hypernetwork parameters $\psi$, pre-trained meta-SDF parameters $\phi^*$, meta learning rate $\beta$, inner learning rate $\alpha$, max training iteration $N$, inner-loop iteration $m$.
 1: for $i = 1, \ldots, N$ do
 2:     $\psi_0 = \psi$
 3:     Randomly choose a subject/cloth-type combination $n$
 4:     Uniformly sample $M \sim \{1, \ldots, D^{(n)}\}$ where $D^{(n)}$ is the number of datapoints of subject/cloth-type combination $n$
 5:     Sample $M$ datapoints from subject/cloth-type combination $n$, denoting these datapoints as $\mathcal{S} = \{\{\mathbf{B}_b\}^{(j)}, \hat{\mathbf{X}}^{(j)}\}_{j=1}^{M}$
 6:     for $k = 1, \ldots, m$ do
 7:         $\mathcal{L} = \frac{1}{M} \sum_{(\{\mathbf{B}_b\}, \hat{\mathbf{X}}) \in \mathcal{S}} \mathcal{L}_{\text{IGR}}(f_\phi(\hat{\mathbf{X}})|_{\phi = g_{\psi_{k-1}}(\{\mathbf{B}_b\}) + \phi^*})$
 8:         $\psi_k = \psi_{k-1} - \alpha \nabla_{\psi_{k-1}} \mathcal{L}$
 9:     end for
10:     $\psi \leftarrow \psi + \beta(\psi_m - \psi_0)$
11: end for
```

Denoting the meta-SDF learned by Alg. 1 as ϕ^* and our hypernetwork as $g_\psi(\{\mathbf{B}_b\})$, we implement Alg. 2. This algorithm differs from the original Reptile [46] algorithm in that it tries to optimize the inner-loop on arbitrary amount of data. Note that for brevity the loss in the inner-loop (line 7-line 8) is computed over the whole batch $\mathcal{S}$, whereas in practice we used stochastic gradient descent (SGD) with fixed mini-batch size over $\mathcal{S}$ since $\mathcal{S}$ can contain hundreds of samples; SGD is used with the mini-batch size of 12 for the inner-loop.

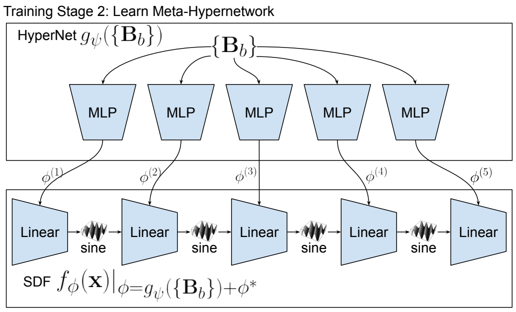
Training Stage 2: Learn Meta-Hypernetwork
HyperNet $g_\psi(\{\mathbf{B}_b\})$
$\{\mathbf{B}_b\}$
MLP
MLP
MLP
MLP
MLP
$\phi^{(1)}$
$\phi^{(2)}$
$\phi^{(3)}$
$\phi^{(4)}$
$\phi^{(5)}$
Linear
sine
Linear
sine
Linear
sine
Linear
sine
Linear
SDF $f_\phi(\mathbf{x})|_{\phi=g_\psi(\{\mathbf{B}_b\})+\phi^*}$

Figure 3: **Overview of the meta-hypernetwork.** It predicts residuals to ϕ^* which is learned in Sec. 4.1

Inference: At test-time, we are given a small fine-tuning set $\{\{\mathbf{B}_b\}^{fine,(j)}, \hat{\mathbf{X}}^{fine,(j)}\}_{j=1}^{M}$ and the validation set $\{\{\mathbf{B}_b\}^{val,(j)}\}_{j=1}^{K}$. The fine-tuning set is used to optimize the hypernetwork parameters ψ ($m = 256$ SGD epochs) that are then used to generate neural SDFs from bone transformations available in the validation set. The overall inference pipeline including the inverse and the forward LBS stages is shown in Fig. 4.

Bone Transformation Encoding: We found that a small hierarchical MLP proposed in LEAP [45] for encoding bone transformations works slightly better than the encoding of unit quaternions used in SCANimate [61]. Thus, we employ the hierarchical MLP encoder to encode $\{\mathbf{B}_b\}$ for g unless specified otherwise; we ablate different encoding types in the experiment section.

5 Experiments

We validate the proposed MetaAvatar model for learning meta-models and controllable dynamic neural SDFs of clothed humans by first comparing our MetaAvatar to the established approaches [12, 45, 61]. Then, we ablate the modeling choices for the proposed controllable neural SDFs. And lastly, we demonstrate MetaAvatar's capability to tackle the challenging task of learning animatable clothed human models from reduced data, to the point that only 8 depth images are available as input.

Datasets: We use the CAPE dataset [40] as the major test bed for our experiments. This dataset consists of 148584 pairs of clothed meshes, capturing 15 human subjects wearing different clothes

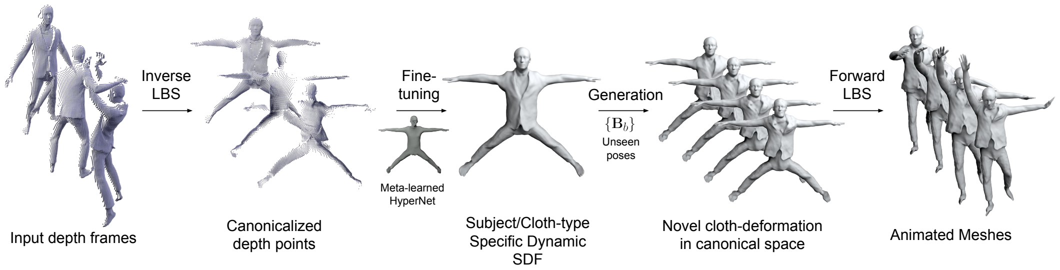

Figure 4: **Overview of our inference pipeline**. The inverse LBS net (Sec. 3.2) takes a small set of input depth frames together with their underlying SMPL registrations to canonicalize the depth points; then the meta-learned hypernetwork (Sec. 4.2) is fine-tuned to represent the instance specific dynamic SDF; given novel poses, the updated hypernetwork generates pose-dependent cloth-deformations in canonical space, and the animated meshes are obtained via the forward LBS network (Sec. 3.2).

while performing different actions. We use 10 subjects for meta-learning, which we denote as the training set. We use four unseen subjects (00122, 00134, 00215, 03375)[2] for fine-tuning and validation; for each of these four subjects, the corresponding action sequences are split into fine-tuning set and validation set. The fine-tuning set is used for fine-tuning the MetaAvatar models, it is also used to evaluate pose interpolation task. The validation set is used for evaluating novel pose extrapolation. Among the four unseen subjects, two of them (00122, 00215) perform actions that are present in the training set for the meta-learning; we randomly split actions of these two subjects with 70% fine-tuning and 30% validation. Subject 00134 and 03375 perform two trials of actions unseen in the training set for meta-learning. We use the first trial as the fine-tuning set and the second trial as the validation set. Subject 03375 also has one cloth type (blazer) that is unseen during meta-learning.

Baselines: We use NASA [12], LEAP [45], and SCANimate [61] as our baselines. NASA and SCANimate cannot handle multi-subject-cloth with a single model so we train per-subject/cloth-type models from scratch for each of them on the fine-tuning set. LEAP is a generalizable neural-implicit human body model that has shown to work on minimally-clothed bodies. We extend LEAP by adding a one-hot encoding to represent different cloth types (similarly to [40]) and train it jointly on the full training and the fine-tuning set.

As for the input format, we use depth frames rendered from CAPE meshes for our MetaAvatar. To render the depth frames, we fixed the camera and rotate the CAPE meshes around the y-axis (in SMPL space) at different angles with an interval of 45 degrees; note that for each mesh we only render it on one angle, simulating a monocular camera taking a round-view of a moving person. For the baselines, we use watertight meshes and provide the occupancy [42] loss to supervise the training of NASA and LEAP, while sampling surface points and normals on watertight meshes to provide the IGR loss supervision for SCANimate. Note that our model is at great disadvantage, as for fine-tuning we only use discrete monocular depth observations without accurate surface normal information.

Tasks and Evaluation: Our goal is to generate realistic clothing deformations from arbitrary input human pose parameters. To systematically understand the generalization capability of the MetaAvatar representation, we validate the baselines and MetaAvatar on two tasks, pose interpolation and extrapolation. For interpolation, we sample every 10th frame on the fine-tuning set for training/fine-tuning, and sample every 5th frame (excluding the training frames) also on the fine-tuning set for validation. For extrapolation, we sample every 5th frame on the fine-tuning set for training, and sample every 5th frame on the validation set for validation.

Interpolation is evaluated using three metrics: point-based ground-truth-to-mesh distance ($D_p \downarrow$, in cm), face-based ground-truth-to-mesh distance ($D_f \downarrow$, in cm), and point-based ground-truth-to-mesh normal consistency ($NC \uparrow$, in range $[-1, 1]$). For computing these interpolation metrics we ignore non-clothed body parts such as hands, feet, head, and neck. For extrapolation, we note that cloth-deformation are often stochastic; in such a case, predicting overly smooth surfaces can result in lower distances and higher normal consistency. Thus, we also conduct a large-scale perceptual study using

[2] We ignore subject 00159 because it has significantly less data compared to other subjects.

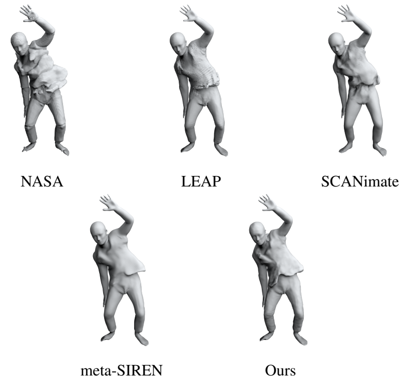

Figure 5: Qualitative comparison on extrapolation results with blazer outfit. NASA shows consistent blocky artifacts. LEAP predicts overly smooth surfaces missing the tails of the blazer outfit. SCANimate does not generalize as this specific pose has not been seen during training. Directly meta-learning a SIREN [64] network that conditions on input poses produces a smooth surface that does not capture the blazer tails well.

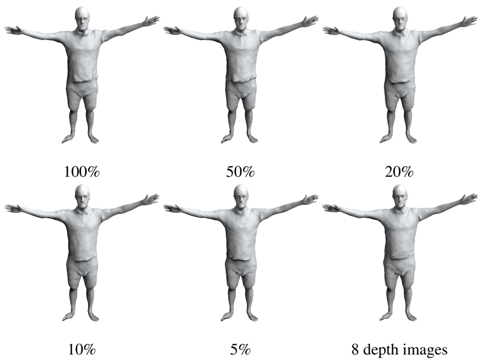

Figure 6: Qualitative comparison on extrapolation results when reducing fine-tuning data on subject 00215 wearing poloshirt. The caption indicates the amount of fine-tuning data used to fine-tune the meta-hypernetwork on this unseen subject. Our meta-learned model captures for this unseen subject the sliding effect of the poloshirt at this pose in which the person raising arms, even fine-tuned with just 8 depth images.

Amazon Mechanical Turk, and report the perceptual scores (PS ↑) which reflects the percentage of users who favor the outputs of baselines over MetaAvatar. Details about user study design can be found in the Supp. Mat.

5.1 Evaluation Against Baselines

In this section, we report results on both interpolation and extrapolation tasks against various baselines described above. For NASA and SCANimate, we train one model for each subject/cloth-type combination on the fine-tuning set. For them it usually takes several thousand of epochs to converge for each subject/cloth-type combination, which roughly equals to 10-24 hours of training. For LEAP, we train a single model on both the training and the fine-tuning set using two days. For the MetaAvatar, we meta-learn a single model on the training set, and for each subject/cloth-type combination we fine-tune the model for 256 epochs to produce subject/cloth-type specific models. The exact fine-tuning time ranges from 40 minutes to 3 hours depending on the amount of available data since we are running a fixed number of epochs; see the Supp. Mat. for detailed runtime comparison on each subject/cloth-type combination. Note that MetaAvatar uses *partial* depth observations while the other baselines are trained on *complete* meshes. The results are reported in Table 1.

Importantly, models of NASA and SCANimate are over-fitted to each subject/cloth-type combination as they cannot straightforwardly leverage prior knowledge from multiple training subjects. LEAP is trained on all training and fine-tuning data with input encodings to distinguish different body shapes and cloth types, but it fails to capture high-frequency details of clothes, often predicting smooth surfaces (Fig. 5); this is evidenced by its lower perceptual scores (PS) compared to SCANimate and our MetaAvatar. In contrast to these baselines, MetaAvatar successfully captures a unified clothing deformation prior of diverse body shapes and cloth types, which generalizes well to unseen body shapes (00122, 00215), unseen poses (00134, 03375), and unseen cloth types (03375 with blazer outfit); although we did not outperform LEAP on the interpolation task for subject 00134 and 03375, we note that 1) our method uses only 2.5D input for fine-tuning, while LEAP has access to ground-truth canonical meshes during training; 2) subject 00134 and 03375 comprise much more missing frames than subject 00122 and 00215, resulting in higher stochasticity and thus predicting smooth surfaces (such as LEAP) may yield better performance; this is also evidenced by LEAP's much lower perceptual scores on subject 00134 and 03375, although obtaining the best performance for pose interpolation. We encourage the readers to watch the side-by-side comparison videos available on our project page: https://neuralbodies.github.io/metavatar/.

		3D Input			2.5D Input
		NASA	LEAP	SCANimate	Ours
Subj 00122, 00215					
Ex.	PS ↑	0.078	0.314	0.333	**0.5**
Int.	$D_p \downarrow$	0.484	0.454	0.586	**0.450**
	$D_f \downarrow$	0.327	0.293	0.489	**0.273**
	$NC \uparrow$	0.752	0.807	0.793	**0.821**
Subj 00134, 03375					
Ex.	PS ↑	0.182	0.224	0.481	**0.5**
Int.	$D_p \downarrow$	0.595	**0.483**	0.629	0.518
	$D_f \downarrow$	0.469	**0.340**	0.542	0.367
	$NC \uparrow$	0.693	**0.780**	0.755	0.773
Averge per-model training/fine-tuning time (hours)					
		>10	-	>10	1.60

Table 1: **Comparison to baselines.** D_p, D_f and NC are reported for interpolation (Int.) while PS is reported for extrapolation (Ex.). Note that MetaAvatar is fine-tuned on *depth images* while all other baselines are trained on *complete meshes*. The training/fine-tuning times are just rough estimates, as ours does not include the time for meta-learning, while many factors, including varying training schedules, disk-IOs and hardware setups, can affect the final speed.

		MLP	PosEnc	SIREN	Hyper Quat	Hyper BoneEnc
Subj 00122, 00215						
Int.	$D_p \downarrow$	3.278	1.806	0.472	**0.460**	0.461
	$D_f \downarrow$	2.201	0.998	0.301	0.288	**0.288**
	$NC \uparrow$	-0.279	-0.045	0.815	0.818	**0.820**
Subj 00134, 03375						
Int.	$D_p \downarrow$	3.320	1.498	0.532	0.526	**0.523**
	$D_f \downarrow$	2.190	0.772	0.385	0.378	**0.374**
	$NC \uparrow$	-0.300	-0.099	**0.773**	0.772	0.772

Table 2: **Ablation for different architectures on the interpolation task.** Hyper-Quat is our model that takes the relative joint-rotations (in the form of unit quaternions) as inputs. Hyper-BoneEnc is our full model with hierarchical bone encoding MLP of LEAP [45]. Models in the table are fine-tuned for 128 epochs.

5.2 Ablation Study on Model Architectures

We further ablate model architecture choices for MetaAvatar. We compare against (1) a plain MLP that takes the concatenation of the relative joint-rotations (in the form of unit quaternions) and query points as input (MLP), (2) a MLP that takes the concatenation of the relative joint-rotations and the positional encodings of query point coordinates as input (PosEnc), and (3) a SIREN network that takes the concatenation of the relative joint-rotations and query points as input (SIREN). The evaluation task is interpolation; results are reported in Table 2. For the baselines (MLP, PosEnc and SIREN), we directly use Alg. 2 to meta-learn the corresponding models with $\phi^* = 0$. For MLP and PosEnc, the corresponding models fail to produce reasonable shapes. For SIREN, it produces unnaturally smooth surfaces which cannot capture fine clothing details such as wrinkles (Fig. 5).

5.3 Few-shot learning of MetaAvatar

Fine-tune data (%)		100	50	20	10	5	<1
Subj 00122, 00215							
Ex.	PS ↑	0.5	0.471	0.509	0.473	0.373	**0.510**
Int.	$D_p \downarrow$	-	**0.450**	0.480	0.512	0.543	0.592
	$D_f \downarrow$	-	**0.273**	0.310	0.353	0.391	0.450
	$NC \uparrow$	-	**0.821**	0.808	0.795	0.785	0.768
Subj 00134, 03375							
Ex.	PS ↑	**0.5**	0.476	0.424	0.463	0.439	0.387
Int.	$D_p \downarrow$	-	**0.518**	0.545	0.576	0.603	0.619
	$D_f \downarrow$	-	**0.367**	0.400	0.438	0.471	0.489
	$NC \uparrow$	-	**0.773**	0.762	0.753	0.745	0.737
Average per-model training/fine-tuning time (hours)							
		1.60	0.8	0.32	0.16	0.08	0.02

Table 3: **Ablation for few-shot learning.** We report performance of MetaAvatar on reduced amount of fine-tuning data. Fine-tuning time scales linearly with the amount of data, since we run for a fixed number of epochs.

In this section, we evaluate the few-shot learning capabilities of MetaAvatar. As shown in Table 3, we reduce the amount of data on the fine-tuning set, and report the performance of models fine-tuned on reduced amount of data. Note that with $<1\%$ data, we require only one frame from each action sequence available for a subject/cloth-type combination, this roughly equals to 8-20 depth frames depending on the amount of data for that subject/cloth-type combination. For interpolation, the performance drops because the stochastic nature of cloth deformation becomes dominant when the amount of fine-tuning data decreases. On the other hand, the perceptual scores (PS) are better than NASA and LEAP even with **<1%** data in the form of **partial depth observations**, and better than or comparable to SCANimate in most cases. The qualitative comparison on extrapolation results of reduced fine-tuning data is shown in Fig. 6. Please see the Supp. Mat. for more qualitative results on few-shot learning, including results on depth from raw scans, results on real depth images and comparison with pre-trained SCANimate model.

6 Conclusion

We introduced MetaAvatar, a meta-learned hypernetwork that represents controllable dynamic neural SDFs applicable for generating clothed human avatars. Compared to existing methods, MetaAvatar learns from less data (temporally discrete monocular depth frames) and requires less time to represent novel unseen clothed humans. We demonstrated that the meta-learned deformation prior is robust and can be used to effectively generate realistic clothed human avatars in 2 minutes from as few as 8 depth observations.

MetaAvatar is compatible with automatic registration methods [8, 71], human motion models [80, 81] and rendering primitives [37, 44] that could jointly enable an efficient end-to-end photo-realistic digitization of humans from commodity RGBD sensors, which has broad applicability in movies, games, and telepresence applications. However, this digitization may raise privacy concerns that need to be addressed carefully before deploying the introduced technology.

7 Acknowledgment

Siyu Tang acknowledges funding by the Swiss National Science Foundation under project 200021_204840. Andreas Geiger was supported by the ERC Starting Grant LEGO-3D (850533) and DFG EXC number 2064/1 - project number 390727645. We thank Jinlong Yang for providing results of SCANimate on the CAPE dataset. We thank Yebin Liu for sharing POSEFusion [35] data and results during the rebuttal period. We also thank Yan Zhang, Siwei Zhang and Korrawe Karunratanakul for proof reading the paper.

References

[1] Ferran Alet, Tomas Lozano-Perez, and Leslie P. Kaelbling. Modular meta-learning. In *Proc. of The 2nd Conference on Robot Learning*, 2018. 3

[2] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *Proc. of CVPR*, 2019. 2

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proc. of CVPR*, 2018. 2

[4] Marcin Andrychowicz, Misha Denil, Sergio Gómez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *Proc. of NeurIPS*, 2016. 3

[5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. *ACM Transasctions Graphics*, 24, 2005. 2, 3

[6] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. In *Proc. of ICLR*, 2019. 3

[7] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *Proc. of ECCV*, 2020. 1, 2, 3

[8] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Proc. of NeurIPS*, 2020. 2, 10

[9] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proc. of ICCV*, 2021. 2

[10] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proc. of CVPR*, 2019. 1, 3

[11] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proc. of CVPR*, 2020. 1, 3

[12] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa: Neural articulated shape approximation. In *Proc. of ECCV*, 2020. 2, 6, 7

[13] Nanqing Dong and Eric P. Xing. Few-shot semantic segmentation with prototype learning. In *Proc. of BMVC*, 2018. 3

[14] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. 3

[15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. of ICML*, 2017. 2, 3

[16] Marta Garnelo, Dan Rosenbaum, Chris J. Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J. Rezende, and S. M. Ali Eslami. Conditional neural processes. In *Proc. of ICML*, 2018. 3

[17] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural processes. In *ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

[18] Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E. Turner. Meta-learning probabilistic inference for prediction. In *Proc. of ICLR*, 2019. 3

[19] Edward Grefenstette, Brandon Amos, Denis Yarats, Phu Mon Htut, Artem Molchanov, Franziska Meier, Douwe Kiela, Kyunghyun Cho, and Soumith Chintala. Generalized inner loop meta-learning, 2019. 3

[20] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proc. of ICML*, 2020. 2, 5

[21] Peng Guan, Loretta Reiss, David A. Hirshberg, Er Weiss, and Michael J. Black. Drape: Dressing any person. *ACM Transasctions Graphics*, 31(4), 2012. 2

[22] Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan, and Jose M. F. Moura. Few-shot human motion prediction via meta-learning. In *Proc. of ECCV*, 2018. 3

[23] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *Proc. of ICCV*, 2019. 2

[24] David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks. In *Proc. of ICLR*, 2017. 2, 4, 5

[25] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A Statistical Model of Human Pose and Body Shape. *Computer Graphics Forum*, 28:337–346, 2009. 2, 3

[26] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable Reconstruction of Clothed Humans. In *Proc. of CVPR*, 2020. 1, 3

[27] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proc. of CVPR*, 2018. 2

[28] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proc. of ICCV*, 2019. 3

[29] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *Proc. of ICLR*, 2019. 3

[30] Zorah Lähner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proc. of ECCV*, 2018. 2

[31] Ke Li and Jitendra Malik. Learning to optimize. In *Proc. of ICLR*, 2017. 3

[32] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *Proc. of ECCV*, 2020. 1, 3

[33] Zhongguo Li, Magnus Oskarsson, and Anders Heyden. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Proc. of NeurIPS*, 2020. 1, 3

[34] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. Robust 3d self-portraits in seconds. In *Proc. of CVPR*, 2020. 1

[35] Zhe Li, Tao Yu, Zerong Zheng, Kaiwen Guo, and Yebin Liu. Posefusion: Pose-guided selective fusion for single-view human volumetric capture. In *Proc. of CVPR*, 2021. 1, 10

[36] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning, 2017. 3

[37] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)*, 2021. 10

[38] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transasctions Graphics*, 34(6), 2015. 2, 3

[39] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *Proc. of CVPR*, 2021. 2

[40] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3D people in generative clothing. In *Proc. of CVPR*, 2020. 2, 6, 7

[41] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Proc. of ICCV*, 2021. 2

[42] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. of CVPR*, 2019. 1, 3, 7

[43] Mateusz Michalkiewicz, Jhony K. Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Proc. of ICCV*, 2019. 1, 3

[44] Marko Mihajlovic, Silvan Weder, Marc Pollefeys, and Martin R Oswald. DeepSurfels: Learning online appearance fusion. In *Proc. of CVPR*, 2021. 10

[45] Marko Mihajlovic, Yan Zhang, Michael J. Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proc. of CVPR*, 2021. 2, 3, 4, 6, 7, 9

[46] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms, 2018. 2, 3, 5, 6

[47] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. Star: Sparse trained articulated human body regressor. In *Proc. of ECCV*, 2020. 2, 3

[48] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. of CVPR*, 2019. 1, 3

[49] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proc. of CVPR*, 2020. 2

[50] 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 *Proc. of CVPR*, 2019. 2, 3

[51] Songyou Peng, Chiyu "Max" Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, and Andreas Geiger. Shape as points: A differentiable poisson solver. In *Proc. of NeurIPS*, 2021. 3

[52] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proc. of ECCV*, 2020. 4

[53] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proc. of CVPR*, 2021. 3

[54] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M. Hospedales, and Tao Xiang. Incremental few-shot object detection. In *Proc. of CVPR*, 2020. 3

[55] Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. Anr-articulated neural rendering for virtual avatars. In *Proc. of CVPR*, 2021. 3

[56] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei A. Efros, and Sergey Levine. Few-shot segmentation propagation with guided networks. In *Proc. of ICML*, 2019. 3

[57] Sachin Ravi and Hugo Larochelle. Optimization as a model for fewshot learning. In *Proc. of ICLR*, 2017. 3

[58] Scott Reed, Yutian Chen, Thomas Paine, Aäron van den Oord, S. M. Ali Eslami, Danilo Rezende, Oriol Vinyals, and Nando de Freitas. Few-shot autoregressive density estimation: Towards learning to learn distributions. In *Proc. of ICLR*, 2018. 3

[59] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. of ICCV*, 2019. 1, 3

[60] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proc. of CVPR*, 2020. 1, 3

[61] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proc. of CVPR*, 2021. 2, 3, 4, 6, 7

[62] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *Proc. of BMVC*, 2017. 3

[63] Vincent Sitzmann, Eric R. Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. In *Proc. of NeurIPS*, 2020. 2, 3, 5

[64] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. of NeurIPS*, 2020. 5, 8

[65] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Proc. of NeurIPS*, 2019. 4

[66] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Proc. of NeurIPS*, 2017. 3

[67] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *Proc. of ECCV*, 2020. 1

[68] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P. Srinivasan, Jonathan T. Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proc. of CVPR*, 2021. 2, 3

[69] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-GIF: Neural generalized implicit functions for animating people in clothing. In *Proc. of ICCV*, 2021. 2

[70] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proc. of ICCV*, 2019. 3

[71] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 10

[72] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Proc. of NeurIPS*, 2019. 3

[73] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proc. of CVPR*, 2020. 2, 3

[74] Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and Lu Fang. Unstructuredfusion: Real-time 4d geometry and texture reconstruction using commercial rgbd cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42, 2020. 1

[75] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Analyzing clothing layer deformation statistics of 3d human motions. In *Proc. of ECCV*, 2018. 2

[76] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proc. of CVPR*, 2018. 1

[77] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. Point-based modeling of human clothing. In *Proc. of ICCV*, 2021. 2

[78] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proc. of ICCV*, 2019. 3

[79] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proc. of CVPR*, 2017. 2

[80] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *Proc. of ICCV*, 2021. 10

[81] Yan Zhang, Michael J. Black, and Siyu Tang. We are more than our joints: Predicting how 3D bodies move. In *Proc. of CVPR*, 2021. 10

[82] Z. Zheng, T. Yu, Y. Liu, and Q. Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 3