SPAMs: Structured Implicit Parametric Models

Pablo Palafox\textsuperscript{1}* Nikolaos Sarafianos\textsuperscript{2} Tony Tung\textsuperscript{2} Angela Dai\textsuperscript{1}
\textsuperscript{1}Technical University of Munich \textsuperscript{2}Meta Reality Labs Research, Sausalito, USA

Abstract

Parametric 3D models have formed a fundamental role in modeling deformable objects, such as human bodies, faces, and hands; however, the construction of such parametric models requires significant manual intervention and domain expertise. Recently, neural implicit 3D representations have shown great expressibility in capturing 3D shape geometry. We observe that deformable object motion is often semantically structured, and thus propose to learn Structured-implicit PARametric Models (SPAMs) as a deformable object representation that structurally decomposes non-rigid object motion into part-based disentangled representations of shape and pose, with each being represented by deep implicit functions. This enables a structured characterization of object movement, with part decomposition characterizing a lower-dimensional space in which we can establish coarse motion correspondence. In particular, we can leverage the part decompositions at test time to fit to new depth sequences of unobserved shapes, by establishing part correspondences between the input observation and our learned part spaces; this guides a robust joint optimization between the shape and pose of all parts, even under dramatic motion sequences. Experiments demonstrate that our part-aware shape and pose understanding lead to state-of-the-art performance in reconstruction and tracking of depth sequences of complex deforming object motion.

1. Introduction

Understanding non-rigidly deforming shapes is essential for perception, as we live in a 4D world where humans, animals, and other 3D objects move in a non-rigid fashion. Dynamic tracking and reconstruction remains a notable challenge, and while significant advances have been made in 4D reconstruction and tracking, they often require complex multi-view setups [5] or build on a domain-specific,
fixed-topology template [1,4,35,36]. In the latter case, para-
metric 3D models in particular have made notable impact in
modeling domain-specific deformable 3D objects, such as
for human bodies [2,20,23], faces [22,34], hands [40], and
animals [51]. However, such parametric 3D models require
a complex construction process involving domain-specific
knowledge and manual efforts, while remaining limited in
expressability of local shape details.

Recently, advances in learned continuous implicit repre-
sentations for modeling 3D shapes have shown impressive
representation power for capturing effective static 3D shape
geometry at relatively high resolutions [7, 8, 17, 28, 29, 33,
39]. Such approaches have also been extended to represent
4D reconstruction of dynamic objects by efficiently disen-
tangling learned implicit spaces representing shape and dy-
namic movement [31, 32]. This has proven to be a very
promising direction, but these approaches characterize ob-
jects as a whole, whereas we observe that the 4D motion of
an object typically maintains a strong structured correlation
on a lower-level part basis.

Thus, we propose Structured-implicit PArametric Mod-
els (SPAMs), which learn a structured, part-based, disentan-
gled representation of deformable 3D objects. Given a set
of observations of various shape identities in different poses
(including a canonical pose) with coarse part annotations,
we learn part-based latent spaces characterizing each part’s
geometry and motion. Note that we do not require com-
prehensive surface correspondence throughout the dataset, nor
complex domain-specific knowledge (e.g., skeleton, kine-
matic chain). We leverage continuous implicit function
representations for each part’s geometry, represented as a
signed distance field in its canonical space, and pose, repre-
sented as a local deformation relative to the canonical space.

At test time, we traverse the learned latent part spaces
to fit to new depth sequences. Crucially, our part-based
representation allows leveraging predicted part segmenta-
tion of the new observation to establish global correspon-
dences with our part-based latent representations. By estab-
lishing correspondence through our part priors, we can
robustly track sequences with significant motion changes by
discovering high-level part correspondence and leveraging
it to guide our joint optimization over part-based shape and
pose. Experiments on non-rigid tracking and reconstruc-
tion of single-camera depth sequences of humans from the
RenderPeople dataset [12] show that the part-aware reason-
ing of our SPAMs can outperform the state of the art by
an order of magnitude on reconstruction (Chamfer distance)
and by 43% on tracking (3D End-Point-Error). In summary,
we present the following contributions:

- We learn a part-based disentanglement of shape and pose,
capturing local characteristics of deformable 3D objects
in latent spaces representing shape and pose of each part.
- Our learned, optimizable spaces enable part-based rea-
soning to guide joint optimization over parts to fit unseen
test sequences. By establishing high-level part correspon-
dence between a new observation and our learned part
spaces, we can robustly guide a joint optimization over
part geometry and pose, resulting in a more globally con-
sistent non-rigid reconstruction and tracking.

2. Related Work

Parametric and Neural Parametric Models. Parametric
body models [23, 48, 50] enable the representation of the
human body variations with a limited number of param-
eters (i.e. a low-dimension descriptor). One of the most
influential works in this line of research is SMPL [23], a
parametric model widely-used to describe the body shape
and pose with deformation blend shapes learned from a
dataset of diverse 3D body scans. Parametric body models
have enabled research work for modeling soft tissue [36]
and clothing [1,25,45]. However, parametric body mod-
els [23,48] employ deformations based on vertex-based
skinning models which have limited resolution and cannot
represent non-linear surface deformations of clothed bod-
ies (e.g., wrinkles). To alleviate these limitations, Neural
Parametric Models [32] learn to disentangle 4D dynamics
into latent-space representations of shape and pose using
implicit functions and can fit to new observations by opti-
mizing over the pose and shape codes. However, they
treat the human body as a single entity, which results in
somewhat unrealistic motions. Genova et al. [16,17] ad-
dressed this problem by introducing LDIF, a 3D represen-
tation that implicitly describes a shape as the sum of lo-
cal 3D functions. LDIF outputs a structured decomposi-
tion into shape elements by encoding 3D points within each
shape using PointNet [15,37,38] and hence enable shape
encoding such as human bodies in local regions arranged in
a global structure. The 3D decompositions of LDIF tend to-
wards temporal consistency, but as tracking is not explicitly
considered, surface tracking tends to become inconsistent in
more challenging motion scenarios. In contrast, we learn a
semantically-driven part decomposition which guides joint
part-based shape and pose optimization that results in ro-
bust, consistent tracking over dynamic sequences.

Continuous Implicit Deformable Representations. Implic-
tic-based and generative representations for humans or
clothing has been an active topic of research for the past
few years [9,10,30,33,42,43,47,49]. Several recent works
have focused on learning identity-specific implicit repre-
sentations to animate clothed people [11,13,19,30,41,46].
Neural-GIF [46] introduced a framework to animate
clothed people from scans as a function of pose directly,
without the need of registration. SCANimate [41] and
LEAP [30] learn a pose or shape representation of the
human body surface. While promising, these methods learn
subject or outfit-specific models, and thus lack general characterization of deformable objects. Given a posed but unclothed body model, POP [27] represents a clothed human with a set of points in order to create pose-dependent clothing animations. New animations can then be created from an unseen scan of a person in clothing. Bhatnagar et al. [3] proposed IP-Net, a method to combine learned implicit functions and traditional parametric models to produce controllable models of humans. IP-Net predicts correspondences to SMPL and leverages a double-layered surface to represent inner and outer surfaces to better represent a clothed body. Our proposed approach also builds on the expressability of learned implicit representations, and we propose to learn from a dataset without strong requirements on surface correspondence annotations to construct a general, semantically-driven decomposition that provides strong high-level guidance for robust pose tracking under challenging motion scenarios.

3. Structured Implicit Parametric Models

We introduce Structured-implicit PArametric Models (SPAMs), a learned approach to build part-aware, implicit, parametric 3D models from a dataset of posed identities of a given object class. We do not require the dataset to have surface correspondence between each instance nor annotations for physical domain-specific properties such as the skeleton or the kinematic chain. Our SPAMs structurally decompose non-rigid object motion into part-based disentangled implicit spaces representing each part’s shape and pose.

SPAMs consist of (1) learned latent spaces that characterize part geometries in the canonical shape space, (2) a part decoder conditional on the part shapes that guides a part-based structural partitioning of the canonical space, and (3) structured, latent pose spaces conditional on both part shape and pose. At test time, SPAMs not only allow for the joint optimization over the learned, local spaces of shape and pose to fit to a new observation, but, crucially, our part-based formulation enables establishing high-level correspondence between predicted parts and the part-based latent representations. This part correspondence enables robust global optimization over the part shape and pose spaces, resulting in robust, consistent non-rigid reconstruction and tracking of unseen sequences of deforming objects.

3.1. Overview

Given a dataset of shape identities from the same class category in different poses, with coarse part segmentation in the canonical space, our goal is to learn a part-based parametric model that structurally disentangles shape and pose. We leverage this structured part decomposition to fit to unseen depth sequences of new identities, where our part-based representation provides a lower-dimensional space to establish semantic part correspondences that provide strong guidance for our part-based shape and pose fitting under
complex motions.

Our SPAMs characterize disentangled shape and pose as sets of local, part-based shape spaces over shape codes \( \{s_q\} \), and their corresponding local pose spaces over \( \{p_q\} \). To obtain the structured partitioning of these local spaces, we first learn a part decoder that predicts a segmentation of the canonical space into a set of local parts (Sec. 3.2). Given the learned space partitioning, we can then construct the structured shape space as a set of local implicit geometric representations (Sec. 3.3). We then build the structured pose space as a set of local implicit deformation fields that characterize the motion of each local shape (Sec. 3.4). Finally, to fit to new depth sequences of unobserved identities at test time, we leverage part correspondence to the set of shape and pose spaces to obtain a robust, joint optimization that accurately represents the observed shape (Sec. 3.5).

### 3.2. Part Decoder

In order to establish a structural space partitioning, we learn a part decoder that segments the canonical space and guides the learning of the part-based shape and pose latent spaces. The part decoder \( f_{\theta_0} \) is characterized by an MLP that predicts part confidences for a query point \( x \in \mathbb{R}^3 \) in canonical space. We implement our part decoder in an auto-decoder fashion [33] and condition on the concatenation of \( D_s \)-dimensional local latent shape codes \( s_q \) (Sec. 3.3). Let \( Q \) denote the number of parts, leading to a \( QD_s \)-dimensional concatenation of local shape codes. Formally, we have:

\[
f_{\theta_0} : \mathbb{R}^{QD_s} \times \mathbb{R}^3 \rightarrow \mathbb{R}^Q, \quad ([s_q], x) \mapsto f_{\theta_0} ([s_q], x) = \tilde{w}. \tag{1}
\]

where \( [s_q] \) is the concatenation of all \( Q \) local shape codes \( \{s_q\} \). We train \( f_{\theta_0} \) with a binary cross entropy loss between predicted and ground-truth part labels as one-hot vectors.

We consider six coarse parts: head, torso, right arm, right leg, left arm and left leg (see Fig. 1-center). A point \( x \in \mathbb{R}^3 \) in canonical space can belong to any of these part classes, where the \( Q \)-dimensional \( \tilde{w} \) denotes the likelihoods of \( x \) belonging to each respective part. In particular, we allow points near the boundary between two parts to belong to both parts at the same time, by indicating both parts to be one in the ground truth part vector. This allows regularization of boundary regions to ensure smooth transitions between implicit functions for shape and pose latent spaces.

Note that \( f_{\theta_0} \) predicts part probabilities for all points \( x \in \mathbb{R}^3 \) in canonical space, not only surface points. This strategy provided more accurate part predictions for off-surface locations during our test-time optimization (Sec. 3.5).

Groundtruth for supervising the part decoder is obtained as follows: as will be shown in Sec. 3.3, to train our shape latent spaces we sample dense surface points with normals in addition to off-surface points randomly sampled in \([-1, 1]^3\).

We re-use these same points to train our part decoder, with part labels for each surface point obtained directly from the original scan or mesh and by nearest neighbor query for off-surface points.

### 3.3. Structured Shape Space

Our multi-part shape space is learned by a dictionary of \( Q \) local MLPs, each learning to represent a local shape part in its canonical pose, characterizing its geometry as the zero iso-surface decision boundary of a signed distance field. The structural decomposition of the \( Q \) local MLPs is guided by the part decomposition predicted by the part decoder \( f_{\theta_0} \). To extract the complete shape from our structured shape space, we query all \( Q \) shape MLPs for every point in a 3D grid, average out their SDF contributions based on predicted part confidence for the given query point, and finally use Marching Cubes [24] to extract a mesh.

Each shape MLP is trained in auto-decoder fashion [33], similar to the part decoder. We directly optimize over a latent code \( s_q \), which is particular to the shape latent space of its assigned part \( q \). Each part \( q \) of a canonically-posed shape identity \( i \) in the training set is then encoded in a \( D_s \)-dimensional latent shape code \( s_q^i \). In turn, each shape MLP \( f_{\theta_q} \) learns to map an input point \( x \in \mathbb{R}^3 \) in the canonical space, conditioned on the local shape code \( s_q^i \), to a predicted SDF value \( \tilde{d}_q \):

\[
f_{\theta_q} : \mathbb{R}^{D_s} \times \mathbb{R}^3 \rightarrow \mathbb{R}, \quad (s_q^i, x) \mapsto f_{\theta_q} (s_q^i, x) = \tilde{d}_q. \tag{2}
\]

As train data typically do not contain watertight meshes, we train directly on oriented point clouds sampled from the (potentially incomplete) train meshes, following [18]. This is accomplished by solving for an Eikonal boundary value problem that constrains the norm of spatial gradients of the SDF to be 1 almost everywhere [18, 44]. In practice, we use the Eikonal loss as presented in [44]. As training data, we only require the surface points with associated normals for each identity, and randomly sampled coordinates in the unit cube. For every identity in a batch, we sample \( N_s \) points, half of which are surface points subsampled from the given identity, and the remaining half as coordinates randomly...
sampled in space. Then for every part \( q \), we train the corresponding shape MLP by minimizing the following reconstruction energy over the \( S \) canonically-posed shape identities of the dataset with respect to the local shape codes \( \{ s_q^i \}_{i=1}^{S} \) and the set of shape MLP weights \( \theta_q^i \):

\[
\text{arg} \min_{\theta_q^i} \sum_{q=1}^{Q} \sum_{i=1}^{S} \left( \sum_{k=1}^{N} w_q^{i,k} \mathcal{L}_s(f_{\theta_q^i}(s_q^i, x_q^i), d_q^i) + \mathcal{L}_f \right).
\]

(3)

Here \( \mathcal{L}_{\xi} \) is the Eikonal loss version proposed in SIREN [44], enforcing that (1) SDF predictions for surface points are 0, (2) groundtruth surface normals match the estimated normals (computed as the spatial gradient of the SDF function at a given position), (3) the norm of the SDF gradient is 1 almost everywhere and (4) off-surface points do not have SDF values close to 0. We refer to [44] for a more detailed explanation. Additionally, we enforce a zero-mean multivariate-Gaussian distribution with spherical covariance \( \sigma_s \) over the latent shape codes, as was proposed in [33]:

\[
\mathcal{L}_{\tau} = \left\| \frac{s_q^i}{\sigma_s^2} \right\|_2^2.
\]

Importantly, the contribution to the loss of each point is weighted by its predicted part confidence \( w_q^{i,k} \) from \( f_{\theta_q^i} \), enabling each local shape MLP to focus on its respective local region. That is, if for point \( x \), \( w_q^{i,k} \) for a class \( q \) is close to 1, then \( x \) likely belongs to \( q \). Then gradients from the contribution of point \( x \) will back-propagate with high weight to the shape MLP of part \( q \), \( \theta_q^i \). On the contrary, if \( w_q^{i,k} \) is near 0, then \( x \) likely does not belong to \( q \) and \( f_{\theta_q^i} \) will not learn to generate geometry at \( x \).

**Implementation details.** Each shape latent space is implemented as a 6-layer SIREN [44] MLP with a hidden dimension of 256 and a frequency \( \omega = 30 \). We use 128-dimensional shape latent codes for each part (\( D_s = 128 \)). Similar to our part decoder, we employ FiLM conditioning [6] to condition the SDF predictions on the local latent shape code, with the mapping network implemented as a 4-layer LeakyReLU MLP with 128 units per layer. We use the Adam optimizer and learning rates of \( 1 \times 10^{-4} \) and \( 1 \times 10^{-3} \) for the local shape MLPs \( f_{\theta_q^i} \) and the local shape codes \( \{ s_q^i \}_{i=1}^{S} \), respectively. The latent shape codes are initialized randomly from \( \mathcal{N}(0, 0.01^2) \).

3.4. Structured Pose Space

Similar to our structured shape latent space, our structured pose space is learned by a dictionary of \( Q \) local MLPs, each optimized to represent a local deformation field that maps query points \( x \) in the canonical space of an identity \( i \) to a deformed space \( j \), by predicting a flow vector \( \Delta \tilde{x}_q \).

This prediction is conditional on both a \( D_p \)-dimensional latent pose code \( p_q^j \) as well as on the latent shape code \( s_q^i \) of the corresponding part \( q \), as pose deformations change with respect to shape. Deformation fields are only defined for a thin layer around the shape surface, since flow vectors become less informative when further away from the surface.

Formally, we have:

\[
f_{\theta_p^j} : \mathbb{R}^{D_p} \times \mathbb{R}^{D_p} \times \mathbb{R} \to \mathbb{R}^3, \\
(s_q^i, p_q^j, x_q^i) \mapsto f_{\theta_p^j}(s_q^i, p_q^j, x_q^i) = \Delta \tilde{x}_q.
\]

Our local pose MLPs are trained with up to \( P \) deformed instances of an identity; note that this does not require different identities to appear in the same pose.

To train these local pose spaces, we sample dense, near-surface correspondences between the canonical and posed frame, which amounts to sampling the canonical and posed raw meshes at the same barycentric coordinates, similar to [32]. Learning one of the local pose spaces \( \theta_p^j \) amounts to minimizing the following energy term over all \( P \) deformation fields with respect to the initial (and local) pose codes \( \{ p_q^j \}_{j=1}^{P} \) and pose MLP weights \( \theta_p^j \):

\[
\text{arg} \min_{\theta_p^j} \sum_{j=1}^{P} \left( \sum_{i=1}^{Q} \sum_{q=1}^{N_p} w_q^{i,k} \mathcal{L}_p + \mathcal{L}_{\tau} \right),
\]

(4)

where \( i(j) \) is a mapping from the index \( j \) of a posed shape to the corresponding index \( i \) of its canonical shape and \( \mathcal{L}_p \) is an \( \ell_2 \) loss between predicted and ground truth flow vectors:

\[
\mathcal{L}_p := \mathcal{L}_p(f_{\theta_p^j}(s_q^i, p_q^j, x_q^i), \Delta x_{ij}^k).
\]

Similar to the structured local shape spaces, we encourage corresponding structural decomposition of the local part spaces by leveraging the predicted part confidences \( w_q^{i,k} \) from \( f_{\theta_q^i} \). We also enforce an analogous zero-mean multivariate-Gaussian distribution over the latent pose codes with \( \mathcal{L}_{\tau} \). While learning these pose latent spaces, we do not optimize over the local shape codes.

**Implementation details.** We implement our structured pose latent spaces as 4-layer SIREN MLPs with hidden dimension equal to 256, \( \omega = 15 \), and set \( D_p = 64 \) as the dimension of our local pose codes. We employ the same training scheme as used for training the shape spaces.

3.5. Test-time Optimization

Our structured latent representations of shape and pose can be optimized over at test time to fit a SPAM to accurately reconstruct and track an input sequence of \( L \) depth maps. This is achieved by solving for the set of \( Q \) local latent shape codes, denoted as \( \{ s \} \), and the \( Q \times L \) latent pose codes \( \{ \{ p_q \} \}_{j=1}^{L} \), that best explain the input sequence.

Each depth map in the sequence is interpreted as a \( 256^3 \)-SDF grid of its back-projected values. A volumetric mask \( M_o \) is also extracted to mask out regions that are further than 0.02 (in normalized units) from the observed surface. We additionally predict part labels for every point.
Figure 3. Our part-aware $\mathcal{L}_{icp}$ can establish high-level correspondences between input points (left) and posed estimates (right), even under difficult motions with poor initial pose estimates (e.g., right arm in green), thus guiding optimization to accurate tracking.

in the input depth map using PointNet++ [38], which we pre-train to predict part labels for the $Q$ parts.

Prior to optimization, we initialize our local shape and pose codes. To initialize each local shape code $s_q$, we use the average optimized train code for each part $q$. To initialize local pose codes, we leverage a learned pose encoder that maps the input depth map to a latent code. While we find that our structured part representations of shape and pose can robustly track from a random pose initialization of all codes from $\mathcal{N}(0, 0.01)^2$, we can obtain improved pose tracking by a learned encoder initialization (c.f. Sec. 4).

Given the initial shape code estimates, we can extract an initial canonical shape by querying our structured shape MLPs on a 3D grid, and extracting the iso-surface with an initial canonical shape by querying our structured shape tracking by a learned encoder initialization (c.f. Sec. 4).

To quantitatively evaluate model fitting to depth sequences, we consider the canonical poses from the current state of the shape codes, re-sample a new set of $N_t = 500k$ points around the mesh and keep those within a distance $\epsilon_{icp}$ from the implicitly represented surface. We use our part decoder (Sec. 3.2) to estimate part labels for these canonical points, and then warp them into a posed frame $j$ using our pose decoders. Then for every point in the input depth map, and given its predicted part $q$ (obtained by PointNet++), we find its nearest neighbor in the warped set of points belonging to $q$ based on $f_{\theta_q}$ to establish correspondences. Crucially, this provides robust correspondences in challenging motion scenarios where pose initialization may be notably misaligned, as demonstrated in Sec. 4.

4. Experiments

We evaluate SPAMs on the task of model fitting to depth sequences (Sec. 4.1), and analyze the effect of our structured, part-based fitting in Sec. 4.2.

Datasets. We train and evaluate on the public RenderPeople dataset [12], which contains real-world 3D scans of people in clothing, post-processed to be minimally noise-free 3D meshes (i.e., removing holes, self-intersection). We train on 338 identities, each rescaled to a common scale in the unit bounding box as a simple data pre-processing step. To learn the structured pose space, we collected a set of 2,446 3D human motion sequences covering a wide variety of action categories of daily activities and sports, and animated canonically-posed identities. From this set of posed, clothed people, we randomly sample scan-motion pairs and obtain 40K randomly posed instances, without requiring seeing multiple identities in the same pose. We evaluate our method on six unseen test identities performing various dancing moves, comprising to a total of 540 test frames organized in 90-frame sequences per identity.

Evaluation metrics. To quantitatively evaluate model fitting to depth sequences, we measure reconstruction quality as well as tracking performance. To measure reconstruction quality, we follow the evaluation protocol of Occu-
the End-Point Error is evaluated as in prior non-rigid tracking works \cite{5, 32} via neighbors in the ground truth mesh. Tracking performance predicted mesh with the normals of the corresponding nearest the mean absolute dot product of the normals of the pre-
tination.

Normal Consistency to any potentially mismatched surface reconstruc-
tion randomly sampled points on the surfaces, giving distance character-
tization between the prediction and ground truth with 100k randomly sampled points in the unit bounding box. 

\[ \text{Chamfer-}\ell_2 (\text{C-}\ell_2) \] measures the bi-directional distance between the prediction and ground truth with 100k randomly sampled points on the surfaces, giving distance characterization to any potentially mismatched surface reconstruction. Normal Consistency (NC) measures surface quality as the mean absolute dot product of the normals of the predicted mesh with the normals of the corresponding nearest neighbors in the ground truth mesh. Tracking performance is evaluated as in prior non-rigid tracking works \cite{5, 32} via the End-Point Error (EPE) as the average \ell_2-distance between predicted and ground truth deformations.

### 4.1. Model Fitting to Monocular Depth Sequences

We evaluate our SPAMs model fitting to new monocular depth sequences in comparison with state of the art on sequences rendered from our RenderPeople \cite{12}-constructed dataset, as well as on CAPE \cite{26} (see supplemental). We compare with the state-of-the-art Neural Parametric Models (NPMs) \cite{32} and IP-Net \cite{3}. We train NPMs on our Renderpeople training split. Since NPMs require watertight meshes for training to determine inside/outside, and inside/outside queries on our train data tend to be unreliable in often-articulated regions such as hands, we adapt NPMs as NPMs* which incorporates the Eikonal loss, SIREN activations, and FiLM conditioning of our approach. For IP-Net \cite{3}, we use a model checkpoint provided by the authors, which was also used to evaluate on RenderPeople.

Table 1 shows a comparison with NPMs* and IP-Net on fitting to monocular depth sequences. Our structural, part-driven representation of shape and pose results in notably improved reconstruction and tracking performance. SPAM achieves higher IoU and normal consistency, as well as significantly reduced Chamfer distance and 3D End-Point-Error, indicating more globally consistent tracking and reconstruction leveraging part-based fitting. Qualitative comparisons are depicted in Figure 4. Under significant, complex motion in the input sequence, our SPAM maintains robust tracking and consistent geometry, while NPM* fails to capture more dramatic motions (e.g. in the arms).

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU ↑</th>
<th>C-\ell_2 ↓</th>
<th>NC ↑</th>
<th>EPE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP-Net</td>
<td>0.729</td>
<td>0.00053</td>
<td>0.837</td>
<td>0.168</td>
</tr>
<tr>
<td>NPMs*</td>
<td>0.755</td>
<td>0.00163</td>
<td>0.856</td>
<td>0.053</td>
</tr>
<tr>
<td>Ours</td>
<td>0.785</td>
<td>0.00032</td>
<td>0.883</td>
<td>0.034</td>
</tr>
</tbody>
</table>

Table 1. Comparison with state-of-the-art NPMs* on the test set of our RenderPeople dataset. Our part-aware decomposition enables notably improved reconstruction and tracking.

pancyNets \cite{28}, computing Intersection over Union, Chamfer distance, and normal consistency on a per-frame basis. Intersection over Union (IoU) measures overlap between predicted and ground truth meshes, computed over \(10^6\) randomly sampled points in the unit bounding box. Chamfer-\ell_2 (C-\ell_2) measures the bi-directional distance between the input sequence, our SPAM maintains robust tracking and consistent geometry, while NPM* fails to capture more dramatic motions (e.g. in the arms).

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU ↑</th>
<th>C-\ell_2 ↓</th>
<th>NC ↑</th>
<th>EPE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPMs* (w/o PE)</td>
<td>0.269</td>
<td>0.01397</td>
<td>0.658</td>
<td>0.182</td>
</tr>
<tr>
<td>Ours (w/o PE, w/o PGM)</td>
<td>0.671</td>
<td>0.00070</td>
<td>0.836</td>
<td>0.052</td>
</tr>
<tr>
<td>Ours (w/o PE)</td>
<td>0.681</td>
<td>0.00065</td>
<td>0.839</td>
<td>0.052</td>
</tr>
<tr>
<td>Ours (w/o PGM)</td>
<td>0.766</td>
<td>0.00037</td>
<td>0.874</td>
<td>0.037</td>
</tr>
<tr>
<td>Ours</td>
<td>0.785</td>
<td>0.00032</td>
<td>0.883</td>
<td>0.034</td>
</tr>
<tr>
<td>Ours (w/ GT PS)</td>
<td>0.809</td>
<td>0.00021</td>
<td>0.894</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Table 2. Ablation studies. We evaluate the effects of a learned pose encoder initialization (PE), using our part grid mask (PGM) in test-time fitting, and we additionally study the impact of ground truth part segmentation (GT PS) of input depth maps. Our part-driven shape and pose spaces provide significant robustness against lack of pose initializations, and leveraging part mask guidance for optimization additionally improves performance.

### 4.2. Ablations

**Robustness to Pose Code Initialization.** We observe that our part-aware disentanglement of shape and pose provides significant robustness in pose tracking, and in particular, maintains robustness in the absence of pose encoder initialization (which may not be available in scenarios such as generalizing to different sensor inputs). We demonstrate this in Table 2, comparing with a variant of our approach without pose encoder initialization (w/o PE) and instead using random initialization for pose codes. This results in very poor initial pose estimates, with an effectively random initial set of deformation fields. While this poor initialization results in slightly impaired performance, our structurally-guided SPAMs nonetheless can recover a significant portion of the reconstruction and tracking performance of using the pose encoder initialization. In contrast, we observe that the full-shape shape and pose encoding of NPMs* fails to recover from poor pose initializations. We refer to the supplementary material for further qualitative visualizations.

**Part-based Grid Masking.** Table 2 additionally evaluates the effect of leveraging part information as a part grid mask (PGM) during test-time optimization; this local part focus enables more accurate reconstruction and tracking.

**Limitations.** Our structured parametric modeling of deformable objects enables robust model fitting to challenging monocular depth sequences, but maintains several limitations. For instance, our local pose spaces do not characterize potential high-level motion priors given by motion very far from the local region (e.g., one hand moving back more often occurs with the other hand moving forward than not), which could provide additional global context. Additionally, very fine-scale sharp details can become oversmoothed across the global optimization, which could potentially be characterized with perceptually-oriented measures.
5. Conclusion

In this work we have introduced SPAMs, a deformable object representation where non-rigid object motion is structurally decomposed into part-based disentangled representations of shape and pose. Our structured characterization of object movement can be leveraged at test time to fit to input depth sequences of unseen shapes, leveraging part-based correspondences to establish robust optimization to fit to the input sequence. Our experiments show significantly improved robustness in reconstruction and tracking, particularly in scenarios of challenging, complex motions in observed depth sequences. We believe that this representation will be useful in a variety of spatio-temporal tasks.

Acknowledgments: This work was done during an internship at Meta. We thank Yuanlu Xu for informative discussions, as well as TUM’s support from the Bavarian State Ministry of Science and the Arts, coordinated by the Bavarian Research Institute for Digital Transformation (bidt).
References


[27] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J Black. The power of points for modeling humans in clothing. In ICCV, 2021. 3


