Leveraging Intrinsic Properties for Non-Rigid Garment Alignment

Siyou Lin Boyao Zhou Zerong Zheng Hongwen Zhang Yebin Liu
Tsinghua University
Beijing, China
linsy21@mails.tsinghua.edu.cn bzhou22@mail.tsinghua.edu.cn zrzheng1995@foxmail.com
zhanghongwen@tsinghua.edu.cn liuyebin@mail.tsinghua.edu.cn

Abstract

We address the problem of aligning real-world 3D data of garments, which benefits many applications such as texture learning, physical parameter estimation, generative modeling of garments, etc. Existing extrinsic methods typically perform non-rigid iterative closest point and struggle to align details due to incorrect closest matches and rigidity constraints. While intrinsic methods based on functional maps can produce high-quality correspondences, they work under isometric assumptions and become unreliable for garment deformations which are highly non-isometric. To achieve wrinkle-level as well as texture-level alignment, we present a novel coarse-to-fine two-stage method that leverages intrinsic manifold properties with two neural deformation fields, in the 3D space and the intrinsic space, respectively. The coarse stage performs a 3D fitting, where we leverage intrinsic manifold properties to define a manifold deformation field. The coarse fitting then induces a functional map that produces an alignment of intrinsic embeddings. We further refine the intrinsic alignment with a second neural deformation field for higher accuracy. We evaluate our method with our captured garment dataset, GarmCap. The method achieves accurate wrinkle-level and texture-level alignment and works for difficult garment types such as long coats. Our project page is https://jsnln.github.io/iccv2023_intrinsic/index.html.

1. Introduction

The research into building realistic animatable human avatars has drawn increasing attention in the past few years. Recent developments have demonstrated that compared with full-body avatars [8, 24, 26, 29, 31, 43, 63, 64], methods utilizing accurately aligned garment scans to explicitly model garments produce much more realistic geometric deformations and/or rendering results [15, 18, 55, 56], and enable or improve a number of downstream tasks such as retargeting [38], texture learning [55, 56], pose-driving garment animation [18] and virtual try-on [9, 32].

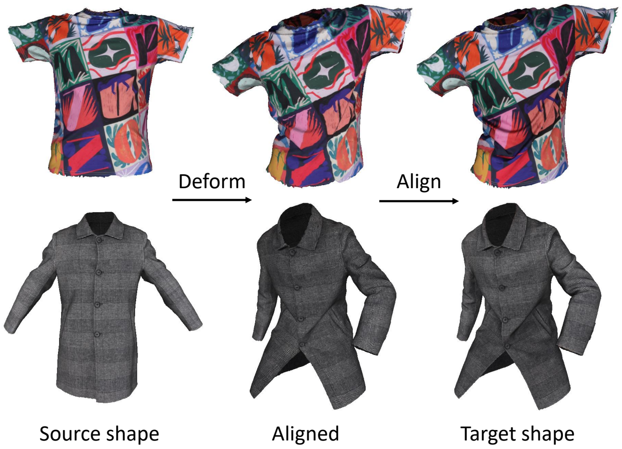

Figure 1. Our method can align deformed versions of a garment to the extent that both texture patterns and geometric details are accurately matched. Various types of garments can be handled, including difficult ones such as coats.

In this paper, we focus on the problem of aligning garments. Existing methods for aligning garments are mostly extrinsic [6, 30, 38, 52, 55, 56, 61]. These methods directly operate in the extrinsic space, *i.e.*, the 3D space, typically by fitting a template shape to target shapes in a non-rigid iterative closest point (ICP) manner. Extrinsic methods easily suffer from incorrect point matches, making it difficult to align details and even the overall shape. This problem becomes more severe for garments, due to their complex deformations as well as articulated motions.

On the other hand, intrinsic methods [2, 12, 19, 27, 33, 37, 42] leverage intrinsic manifold properties, most commonly the eigenfunctions of the Laplace-Beltrami operator (LBO) [35], which are independent of the extrinsic shape and therefore also free from incorrect matches in the extrinsic space. These high-dimensional intrinsic embeddings are also smoother than their extrinsic counterparts, and are easier to align. Current mainstream methods for this purpose are mostly based on the Functional Maps framework [37], where one estimates a linear functional map between two intrinsic embeddings. Unfortunately, directly applying the Functional Maps framework [37] for intrinsic garment alignment still poses challenges. Since a functional map typically needs to be computed using shape descriptor constraints [3, 44, 51], they are more suitable for as-rigid-as-possible deformations. Garments, on the other hand, can stretch, shear and bend, rendering shape descriptors unreliable for functional map estimation. Moreover, the linear assumption of a functional map becomes insufficient for garments due to their non-isometric deformations.

In this work, we resolve these challenges by proposing two neural deformation fields to align the intrinsic embeddings of garments in a coarse-to-fine manner. Our method works in two stages. In the first stage, we use a neural deformation field to obtain a coarse 3D alignment of the source and the target, which induces a functional map to align their intrinsic embeddings. This bypasses the need for descriptors to estimate functional maps as in previous work [2, 12, 27, 37]. Furthermore, we use an intrinsic neural field [21] to implement this deformation. Leveraging intrinsic features makes this deformation field robust to defects such as self-contact or self-intersections of the base template, allowing us to handle difficult garment types, *e.g*., long coats. In the second stage, the intrinsic alignment is further refined with a second neural deformation field. This differs from traditional (linear) functional maps by introducing non-linearity to the Functional Maps framework. We remark that the necessity of introducing non-linearity for garments is rooted in their highly non-rigid and non-isometric deformations. An analysis of the necessity of non-linearity will be provided in the supplementary materials. We demonstrate that our two-stage pipeline can align garment to the extent that both geometry and texture are accurately matched (Fig. 1). To summarize, our contributions include:

- We propose a two-stage pipeline for aligning highly non-rigid 3D garment data, leveraging intrinsic manifold properties and neural deformation fields to achieve high-accuracy results.

- In the first stage, we propose an intrinsic neural deformation field that leverages intrinsic manifold properties. Such intrinsic fields do not suffer from self-contact or self-intersections of the base template. To the best of our knowledge, we are the first to use intrinsic neural fields for modeling deformations.

- In the second stage, we propose another neural deformation field that introduces non-linearity to the Functional Maps framework. This allows us to achieve texture-level high-accuracy alignment. We also provide a theoretical analysis why this is necessary in the supplementary material.

- We collect a dataset of high-quality 3D garment data, including difficult garment types such as long coats.

2. Related Work

The problem of garment alignment falls into the larger category of non-rigid shape alignment/registration. We focus on some highly-related methods in this section. For a complete review we refer the readers to the survey of Deng *et al*. [11].

2.1. General Non-rigid Shape Alignment

Extrinsic methods for shape alignment directly operate in the extrinsic space, *i.e*., the 3D space, typically by deforming a template to a target shape in a non-rigid ICP manner. Common choices for modeling deformations include direct vertex optimization (vertices as free variables) with different regularizations (differential coordinates, local affinity, as-rigid-as-possible, *etc*.) [1, 20, 25, 57], embedded deformations [7, 16, 22, 50, 53] and neural deformation fields [23, 58].

Closely related to our work are neural deformations [23, 58], typically implemented as coordinate-based networks. They enjoy both flexibility and regularity due to being non-linear, continuous and having low-frequency bias [4, 39]. However, coordinate-based neural deformations are continuous w.r.t. the ambient 3D space. If two parts are close in the 3D space, a neural deformation field can hardly pull them apart even if they are geodesically distant. On observing this limitation of coordinate-based networks for other tasks, Koestler *et al*. [21] proposed intrinsic neural fields which are suited for representing fields on manifolds, and exhibited superior results in texture learning.

Intrinsic methods utilize intrinsic embeddings, namely the embeddings derived from the eigenfunctions of the Laplace-Beltrami operator (LBO), to obtain shape correspondences and alignment [2, 12, 19, 27, 33, 37, 42]. Intrinsic embeddings are independent of the extrinsic appearances of the underlying manifolds, and thus do not suffer from incorrect closest point matches faced by extrinsic methods. However, the LBO eigenfunctions of two near-isometric shapes may not be in one-to-one correspondences due to sign ambiguity, eigenvalue switching and multiplicity of eigenvalues [19, 37, 45].

The Functional Maps (FM) framework [37] solves for a linear transformation between the LBO eigenbases of two shapes. Functional map estimation typically requires shape descriptors [3, 44, 51], whose quality is essential in the performance of functional maps. For this reason, a large body of existing work focuses on improving descriptors [2, 10, 12, 17, 27, 41]. However, for highly non-rigid garments, shape descriptors become unreliable. Aside from descriptors, the function map itself can also be refined for better accuracy. Ovsjanikov *et al*. [37] straightforwardly refine a functional map with linear ICP in the intrinsic space. Ren *et al*. [40] improve this procedure by enforcing bijectivity and continuity. Spectral upsampling techniques can obtain functional correspondences in a coarse-to-fine manner [13, 14, 34]. Even though intrinsic shape correspondences have been extensively studied, existing work focuses on near-isometric or as-rigid-as-possible deformations. To the best of our knowledge, there has been no intrinsic method that is devoted to high-accuracy garment alignment, which involves highly non-rigid wrinkle deformations.

2.2. Human and Garment Shape Alignment

There has been extensive research into human performance capture that reconstructs and tracks non-rigid clothed human surface geometries [16, 36, 46, 47, 49, 59]. These methods generally deal with full-body geometries and use tracking only to aid the reconstruction process. Another line of work [5, 30, 31, 29] leverages the parametric body model SMPL [28] to represent clothing as an offset layer of naked body. While this introduces alignment in the sense that all garment deformations share a common body template, such an alignment does not track the actual movement of a point. Moreover, these methods do not segment body and garments, resulting in limited realism.

Despite the progress in reconstructing and animating full-body humans, recent work [18, 48, 55, 56] shows using aligned garment data allows human avatars to be presented with higher realism. ClothCap [38] made an early attempt in this direction by aligning garments from 4D scans using a template segmented from the SMPL [28] model. SimulCap [60] jointly captures the naked body and the deformed garment with multi-layered meshes to achieve plausible results for both body-cloth interaction and non-rigid tracking.

Unlike purely articulated shapes, garments contain detailed wrinkles where one faces an unavoidable trade-off between deformation flexibility and regularity. This difficulty is almost faced by all extrinsic methods. On the other hand, we are unaware of any intrinsic method that is devoted to garment alignment. Our work makes an attempt in this direction by leveraging intrinsic manifold properties for non-rigid garment alignment.

3. Method

We seek to align deformed versions of a garment in different poses. We assume that garment data are presented as 3D triangular meshes with known SMPL [28] pose parameters $\theta \in \mathbb{R}^{72}$. Fig. 2 shows an overview of our pipeline. Let us denote the source shape by M and the target shape by N. In the coarse stage, a smooth template is obtained from the source M, and is then used to align with the target N in the 3D space (Sec. 3.2). This alignment induces a functional map [37], which approximately aligns the intrinsic embeddings. In the refinement stage, we further refine the intrinsic alignment to obtain dense correspondences with a non-linear neural deformation field (Sec. 3.3). Finally, the vertex coordinates of N are transferred to M through the shape correspondences obtained from the intrinsic alignment (Sec. 3.4). Details of network architecture and parameter selection are left to the supplementary material.

3.1. Preliminary: Functional Maps

The Functional Maps framework [37] is a shape correspondence representation that has been extended significantly due to its compactness and flexibility [2, 10, 12, 14, 13, 17, 27, 34, 40, 41]. We briefly review it to introduce terminology and fix notations. Interested readers can refer to [37] for more details. Given two manifolds M and N, together with a point-to-point map $P : M \to N$, for any function $g \in \mathcal{F}(N)$ (the function space on N), we can define a corresponding function on M as $f := g \circ P \in \mathcal{F}(M)$. Thus, any point-to-point map $P : M \to N$ between manifolds *induces* a functional map $P_F : \mathcal{F}(N) \to \mathcal{F}(M),\ g \mapsto g \circ P$ between function spaces. Since any induced functional map must be linear as shown in [37], if two bases Φ^M and Φ^N are chosen for $\mathcal{F}(M)$ and $\mathcal{F}(N)$, respectively, then the functional map (and thus also the point-to-point map), can be represented as a (possibly infinite) matrix. In practice, Φ^M is often chosen to be the eigenfunctions of the Laplace-Beltrami operator (LBO) [35].

In the discrete case, let M be represented by a vertex matrix $V^M \in \mathbb{R}^{n_V \times 3}$ and a face matrix $F^M \in \mathbb{N}^{n_F \times 3}$. A function $f \in \mathcal{F}(M)$ is then a scalar array $f \in \mathbb{R}^{n_V}$. The LBO is then a matrix $L \in \mathbb{R}^{n_V \times n_V}$. We use cotangent weighting [35] to discretize the LBO as $\mathcal{L} = A^{-1}W$, where A_{ii} is the Voronoi area near vertex i and

$$W_{ij} = \begin{cases} -\frac{\cot \alpha_{ij} + \cot \beta_{ij}}{2} & (i,j) \text{ is an edge} \\ \sum_{(i,k) \text{is an edge}} \frac{\cot \alpha_{ik} + \cot \beta_{jk}}{2} & i = j \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Here, α_{ij} and β_{ij} are the angles opposite to the edge (i,j). It is well known that eigenvalues of $\mathcal{L}$ are non-negative, with a natural ordering given by: $\lambda_0 = 0 < \lambda_1 \leq \lambda_2 \leq$ ..., and that the eigenvectors $\{\phi_j\}_{j \geq 0}$ form an orthonormal (w.r.t. the A-inner product) basis of $\mathcal{F}(M)$. Let

$$\Phi^M = [\phi_1, \phi_2, \cdots, \phi_K] \in \mathbb{R}^{n_V \times K} \tag{2}$$

denote the first K non-constant eigenfunctions on M stacked together (omitting K for simplicity). Then the columns of Φ^M spans a subspace of $\operatorname{span}(\Phi^M)$ of $\mathcal{F}(M)$.

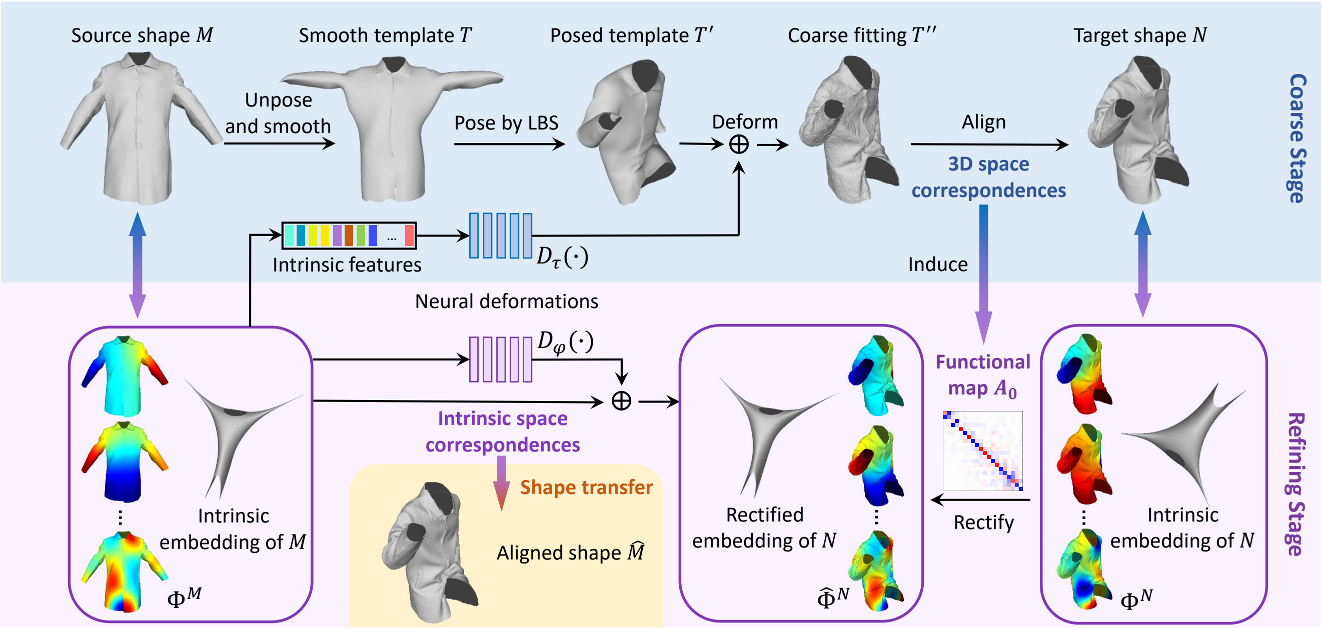

Figure 2. An overview of our pipeline. Given the source M and the target N, we start from the 3D space where a neural deformation field with intrinsic input features deforms the template to align with the target (coarse stage, Section 3.2). This coarse alignment induces a functional map that can bring the intrinsic embeddings of M and N into approximate alignment, which is further refined with a second neural deformation field (refining stage, Section 3.3). Finally, we transfer the 3D vertex coordinates from N to M through the intrinsic space correspondences (shape transfer, Section 3.4).

Let these notations be defined for N similarly, then any functional map $\text{span}(\Phi^N) \rightarrow \text{span}(\Phi^M)$ can be represented as a matrix $C \in \mathbb{R}^{K\times K}$.

3.2. Coarse Alignment in the 3D Space

In this section, we obtain a coarse alignment of the source and the target in the 3D space. The closest point correspondences from this alignment will be used to obtain a function map for aligning intrinsic embeddings. We start from obtaining a rigged smooth template and then use this template as a substitute of M to align with the chosen target mesh (coarse stage in Fig. 2).

Rigged smooth template acquisition Following ClothCap [38], to avoid any potential negative effect of pose-dependent wrinkles in the source mesh, prior to 3D fitting we obtain a smoothed template rigged to the SMPL [28] skeleton from an approximately A-pose scan. Let $\text{LBS}(v, w, \theta)$ denote the linear blend skinning (LBS) transformation as in [28], *i.e.*, it moves a point $v \in \mathbb{R}^3$ in the T-pose space to the pose space (defined by the SMPL [28] pose parameter θ), with $w \in \mathbb{R}^{24}$ being its skinning weights. Let $M = (V^M, F^M)$ denote the source mesh and θ^M its corresponding SMPL pose parameters. The desired smooth template is another mesh $T = (V^T, F^M)$ having the same topology as M, with skinning weights W, s.t. when it is posed to θ, it aligns with M. More specifically, we let V^T minimize the following energy:

$$\begin{aligned} E_1(V^T) \quad = \quad & w_1\|\text{LBS}(V^T, W(V^T), \theta^M) - V^M\|_F^2 \\ & + w_2\|E^T - E^M\|_F^2 + w_3\|L^M V^T\|_F^2, \quad (3) \end{aligned}$$

where w_i are weights to balance different energy terms. Here, we use the diffused skinning field $W(\cdot)$ in [26] to ensure smoothness. E^T and E^M are arrays of edge lengths, which can be computed from (V^T, F^M) and (V^M, F^M). Note that the edges of T and M are in one-to-one correspondence since they have identical face lists. L^M is the uniform Laplacian of M. The first term forces the posed vertices of T to be close to M. The second term enforces edge regularization for M and T to be isometric. The last term enforces smoothness.

Coarse fitting with neural deformation field Having obtained the rigged smooth template T, which has identical topology to that of the source garment mesh M, we use it as a smooth substitute for M to align with any other target garment mesh N. Let θ^N denote the body pose corresponding to N. We first pose T with θ^N to obtain $T' = (V', F^M)$, where

$$V' = LBS(V^T, W(V^T), \theta^N) \quad (4)$$

We then fit T' to N by adding a neural deformation field to it. Note that this deformation field only needs to be defined on the manifold M, not the entire 3D space. We thus let this deformation to be modeled as an intrinsic neural field [21] D_τ (see Fig. 2). Leveraging intrinisic features, *i.e.*, the LBO eigenfunctions, allows us to define a manifold deformation field independent of how it is embedded in the 3D space. More specifically, an intrinsic deformation field does not suffer from self-contact or self-intersection of the posed template T' and allows more flexible deformations. The vertices V' of the posed template T' are further deformed to

$$V'' = V' + D_\tau(\Phi^M), \quad D_\tau : \mathbb{R}^K \to \mathbb{R}^3. \tag{5}$$

Note that n_V is considered as a batch dimension for Φ^M (see Eq. (2)) when writing $D_\tau(\Phi^M)$.[1] Here, τ denotes the optimizable parameters of D_τ. Let V''_b denote the subset of V'' containing only the boundary vertices (same for V^M_b). We minimize the following energy:

$$\begin{aligned} E_2(\tau) &= w_4\text{CD}(V'', V^N) + w_5\text{CD}(V''_\text{b}, V^N_\text{b}) \\ &\quad + w_6\| \max(E^{T''} - E^M, 0)\|_F^2, \end{aligned} \tag{6}$$

where $\text{CD}(\cdot,\cdot)$ denotes the Chamfer distance between point clouds and $E^{T''}$ denotes the edge lengths of the posed and deformed template $T'' = (V'', F^M)$. Note that we clip $E^{T''} - E^M$ by 0 to allow squeezing, since the target shape N may have invisible parts due to occlusion during capture. After τ has been optimized, the deformed template T'' is coarsely fitted to N. This concludes the coarse fitting stage.

3.3. Refinement in the Intrinsic Space

In this section, we seek to accurately align the intrinsic embeddings of M and N. Note that the initially computed intrinsic embeddings can differ greatly due to sign ambiguity and eigenvalue switching [19, 37, 45]. Intuitively, these can lead to the intrinsic embedding of N in Fig. 2 appearing "upside-down". The coarse 3D correspondences from the last stage are thus used to induce a linear functional map A_0 (bottom-right corner of Fig. 2), bringing two intrinsic embeddings into an approximate alignment. However, limited by its linearity, A_0 cannot achieve enough accuracy when the ground truth deformation is non-rigid and non-isometric. We thus use a second neural network (introducing non-linearity) to further refine the intrinsic alignment.

Functional map from coarse 3D correspondences We obtain the functional map A_0 as follows. First, we search for each vertex i in V'' its nearest neighbor $j(i)$ in V^N, giving rise to a point-to-point map $i \mapsto j(i)$. Then we reindex Φ^N to obtain $\hat{\Phi}^M \in \mathbb{R}^{n_V \times K}$ as follows:

$$\hat{\Phi}^M_{ik} := \Phi^N_{j(i),k} \tag{7}$$

As defined in [37], the functional map $A_0 \in \mathbb{R}^{K\times K}$ induced by $i \mapsto j(i)$ satisfies

$$\hat{\Phi}^M A_0 = \Phi^M. \tag{8}$$

In practice this is solved in a least square sense. We then apply A_0 to rectify the intrinsic embedding of N as:

$$\hat{\Phi}^N = \Phi^N A_0. \tag{9}$$

Refining the intrinsic alignment The initial functional map A_0 in (9) can bring the intrinsic embeddings of M and N to an approximate alignment, which we further refine with a second neural deformation network (the refining stage in Fig. 2). Unlike previous work [37] that estimates a linear transformation between Φ^M and $\hat{\Phi}^N$, our neural deformation field in the intrinsic space introduces non-linearity. We remark that it is necessary to introduce non-linearity if high accuracy is desired. A detailed analysis is given in the supplementary material.

Prior to refining the alignment, we follow [42] to scale down ϕ_j to $\phi_j/\sqrt{\lambda_j}$, where λ_j is its corresponding eigenvalue. This is because high-frequency components (corresponding to large eigenvalues) may lead to incorrect closest point matches when computing the Chamfer distances in Eq. (11). We use $\Phi^{M\downarrow}$ denote the eigenfunctions that have been scaled down (similarly defined for Φ^N, $\hat{\Phi}^N$, *etc.*). Then we deform $\Phi^{M\downarrow}$ using a neural network D_φ:

$$\hat{\Phi}^{M\downarrow} = \Phi^{M\downarrow} + D_\varphi(\Phi^{M\downarrow}). \tag{10}$$

Here, φ denotes the parameters of D_φ, which are optimized to minimize:

$$\begin{aligned} E_3(\varphi) &= w_7\text{CD}(\hat{\Phi}^{M\downarrow}, \hat{\Phi}^{N\downarrow}) + w_8\text{CD}(\hat{\Phi}^{M\downarrow}_\text{b}, \hat{\Phi}^{N\downarrow}_\text{b}) \\ &\quad + w_9\|D_\varphi(\Phi^{M\downarrow})\|_F^2. \end{aligned} \tag{11}$$

The first two terms are data terms for the whole shape and for boundaries. Here, CD denotes the obvious extension of Chamfer distance to high-dimensional points, *i.e.*, the averaged squared L_2-distance from each point in one point cloud to its closest point in the other point cloud, computed in both directions. The third term regularizes the magnitude of deformation. Having optimized φ for (11), we obtain accurately aligned intrinsic embeddings $\hat{\Phi}^{M\downarrow}$ and $\hat{\Phi}^{N\downarrow}$, from which dense correspondences can be extracted via nearest-neighbor search in the intrinsic space.

3.4. Shape Transfer

For garment alignment, a triangle mesh having the same topology as M but aligned to N is generally more desired.

[1] Throughout the paper we use the convention that if $f(\cdot) : \mathbb{R}^l \to \mathbb{R}^q$ is a mapping defined on $\mathbb{R}^l$ and $V \in \mathbb{R}^{n\times l}$ is a matrix, then we write $f(V)$ to mean that f applies to each row of V independently, with the results again stacked together as a matrix, *i.e.*, $f(V) \in \mathbb{R}^{n\times q}$.

Method	Chamfer Dist. $(\times 10^{-3})\downarrow$	Cos. Sim. $\uparrow$	LPIPS-AlexNet $(\times 10^{-2})\downarrow$	LPIPS-VGG $(\times 10^{-2})\downarrow$	SSIM $\uparrow$
NDP [23]	14.305	0.885	11.414	11.690	0.865
ClothCap [38]	3.144	0.965	7.874	9.468	0.888
Deep Shells [14]	4.094	0.955	8.471	9.796	0.887
Ours (NoTex)	**2.613**	**0.979**	<u>5.053</u>	<u>7.409</u>	<u>0.901</u>
Ours	<u>2.627</u>	**0.979**	**4.318**	**6.616**	**0.914**

Table 1. Quantitative comparison results. "NoTex" means not using texture information as done in Sec. 3.5. Our method (with or without using texture information) performs notably better than baselines. The best results are in boldface and the second best results are underlined.

A naive shape transfer method is to first find the shape correspondences using their aligned intrinsic embeddings $\hat{\Phi}^{M\downarrow}$ and $\hat{\Phi}^{N\downarrow}$ from the last step, and then transfer the 3D vertex coordinates of N through these correspondences. However, we find this leads to noisy alignment. To reduce noise in the shape transfer process, we utilize 3D coordinates together with Laplacian coordinates.

Leveraging the aligned intrinsic embeddings, we search for each point i in $\hat{\Phi}^{M\downarrow}$ its k nearest neighbors $k\mathrm{NN}(i)$ in $\hat{\Phi}^{N\downarrow}$, with $j(i)$ being the nearest one. Then we define

$$\hat{V}_i^N = \frac{c_{ik} w_{ik}}{\sum_{k \in k\mathrm{NN}(i)} c_{ik} w_{ik}} V_k^N, \tag{12}$$

where w_{ik} is inversely proportional to the L_2 distance between $\hat{\Phi}_i^{M\downarrow}$ and $\hat{\Phi}_k^{N\downarrow}$, and c_{ik} is set to a larger value if i or k corresponds to a boundary vertex. We let the desired vertex coordinates $\hat{V}^M$ for the alignment from M to N minimize:

$$\begin{aligned} E_4(\hat{V}^M) &= \sum_i \|\hat{V}_i^M - \hat{V}_i^N\|^2 \\ &+ \sum_i \|(L^M \hat{V}^M)_i - (L^N V^N)_{j(i)}\|^2, \end{aligned} \tag{13}$$

where L^M and L^N are the uniform Laplacian matrices on M and N. In short, we transfer both 3D coordinates and Laplacian coordinates, and solve for the vertex positions that best satisfy both. Finally, the aligned mesh is obtained as $\hat{M} = (\hat{V}^M, F^M)$, which has the same topology as M, but aligns with N in geometry.

3.5. Texture Alignment

Since the deformations in both stages are neural networks, which enjoy continuity and low-frequency bias [4, 39], the deformations are automatically well-regularized. Texture alignment with neural deformations can easily be achieved by directly concatenating per-vertex RGB colors $C^M \in \mathbb{R}^{n_V \times 3}$ to Φ^M:

$$\widetilde{\Phi}^M = [\beta_1 \Phi^M, \beta_2 C^M] \in \mathbb{R}^{n_V \times (K+3)}. \tag{14}$$

where β_1 and β_2 are parameters to balance geometry and color information. We only concatenate colors when computing the first term in (6) and in (11). Colors are not used as input to D_τ and D_φ, nor are they used in the boundary terms in (6) and (11).

4. Experiments

4.1. Dataset and Baselines

We introduce a new dataset, *GarmCap*, containing high-quality textured 3D garment scans in various poses. This dataset includes four different garments (Fig. 3): G01 (a T-shirt with rich color patterns, 173 scans), G02 (a long coat with black-white strip patterns, 103 scans), G03 (a thick coat with a fading graywhite texture, 101 scans) and G04 (an orange coat, 119 scans). Data in *GarmCap* are garments in static poses, collected in a cage with 128 cameras.

We evaluate our method with the *GarmCap* dataset, and compare with ClothCap [38], Deep Shells [14] and Neural Deformation Pyramid (NDP) [23]. Deep Shells [14] is the most related to our work since it also leverage both 3D-space and intrinsic-space alignments, but in a product space formulation. Also bearing similarity to ours, NDP [23] models neural deformations in a hierarchical manner to achieve coarse-to-fine alignment. Please refer to the supplementary material for more details of the experimental setup.

4.2. Comparisons

We compare our method with the baselines introduced above. Since the *GarmCap* dataset has no ground truth correspondences, we use Chamfer distance and cosine normal similarity to measure geometric similarity. For evaluating correspondence, we directly apply the texture of the source mesh to the aligned mesh and render 32 views per scan, and measure the similarity between rendered images and ground truth renderings. We use perceptual metrics (LPIPS) [62] and structural similarity index measure (SSIM) [54].

In Table 1, our method outperforms others in both geometry and texture alignments. Since the existing baseline implementations do not support using colors, we also test our method without using texture information as in Sec. 3.5 (labeled "NoTex" in Table 1). While doing so leads to a slight performance drop in texture alignment (see the LPIPS and SSIM metrics), there is no performance drop in geometric accuracy. Moreover, even without texture informa-

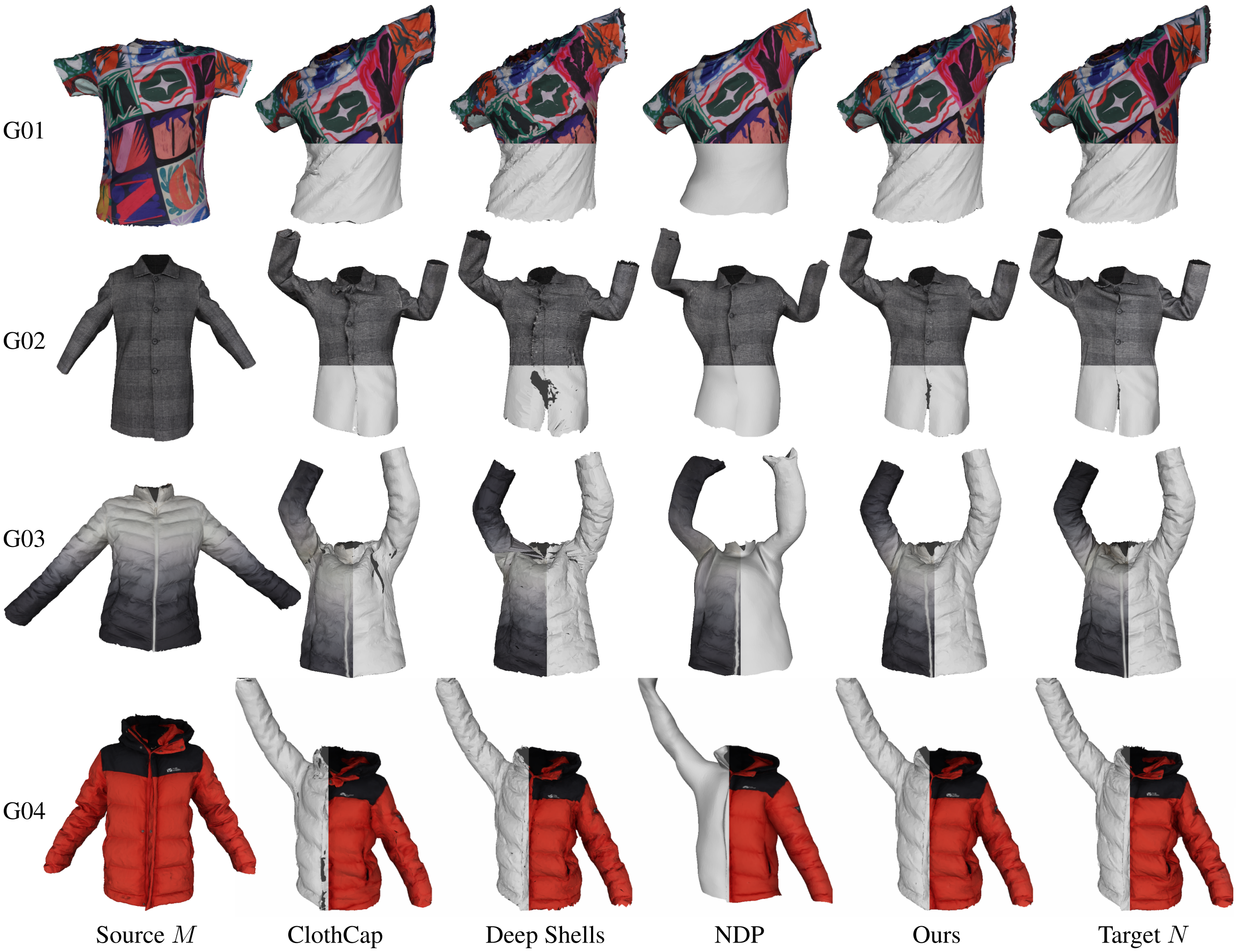

Figure 3. Qualitative comparisons. We exhibit both texture and geometry results. ClothCap [38] is too flexible and produces unnatural distortions and inverted faces. Deep Shells [14] sometimes infers incorrect correspondences, possibly due to its reliance on shape descriptors. NDP [23] is too rigid to produce wrinkles. Our method achieves both geometric and texture alignment.

tion, our method still surpasses all baseline methods. Fig. 3 shows qualitative results of the alignments. While ClothCap [38] and Deep Shells [14] are able to produce geometric details in the alignment, they may fail due to incorrect point matches and lead to visible defects, *e.g.*, indentations on wrinkles (G01), wrongly aligned garment parts (G02), unnatural distortions (G03 and G04), and a number of inverted or self-intersecting faces. Deep Shells also occasionally suffer from symmetry issues (the left arm and the right arm are switched in G03). This is possibly due to its reliance on SHOT descriptors [44], which does not disambiguate the left-right symmetry of garments. The alignment results of NDP [23] does not produce much distortion. However, being too rigid also forbids wrinkles and other details. Our method not only recovers accurate geometric details, but also achieves texture-level alignment. This can be observed from S02: our texture of pocket is accurately overlaid over its corresponding geometry, while for other methods texture shift and distortion distortion are evident.

4.3. Ablation Studies

We conduct ablation studies to evaluate how certain modules affect the overall alignment quality. Note that the performance of each module can vary with garment types, *e.g.*, the coarse-stage intrinsic deformation field handles difficult garments (S02) while texture alignment is more useful for rich-texture garments (S01). We thus report the metrics for each garment separately. We present some metrics for S02 in Table 2 since it is the most difficult case in our dataset that can clearly reflect the effect of different modules. We leave the complete results to the supplementary material due to space limits.

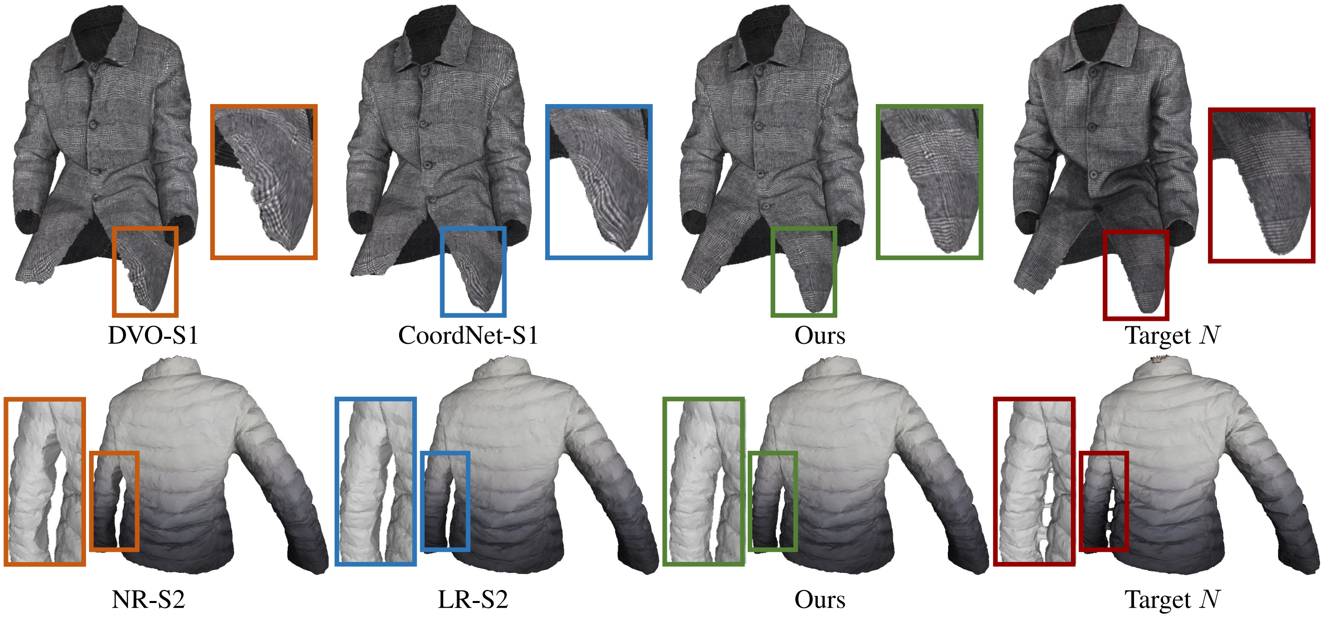

Figure 4. Qualitative ablation studies. First row: using different deformation models for stage 1. Second row: using different refinement method for stage 2.

Different deformation models in stage one We replace the intrinsic deformation field in the coarse stage (stage 1) by direct vertex optimization (DVO-S1), *i.e.*, optimize vertex coordinates as free variables. Note that DVO is essentially the same as ClothCap [38], but with our energy function (6). We also experiment with coordinate-based neural networks in stage 1 (CoordNet-S1). The first row in Fig. 4 shows that both directly optimizing vertices and using coordinate-based networks can introduce unwanted texture distortions. Compared with using a coordinate-based network in the coarse stage, our intrinsic network is better at handling more complicated garment types, namely, long coats (Table 2). Another benefit of using intrinsic deformation networks is the robustness to poor initializations. As shown in Fig. 5, an inaccurate SMPL estimation can lead to a poorly initialized template T', which notably differs from the target shape and even has self-intersections. Since intrinsic features enjoy geodesic continuity on T', the intrinsic deformation network allows us to pull away the self-intersecting parts and still obtain good results. Due to being robust to inaccurate SMPL estimations and template initializations, our method can also be applied loose garments where SMPL-driven the LBS initialization is not fully compatible (Fig. 6).

Different refinement methods in stage two We replace our non-linear neural refinement procedure in the refining stage (stage 2) by linear ICP refinement [37] (denoted as LR-S2, for "linear refinement in stage 2"), and no refinement at all (NR-S2, for "no refinement in stage 2"). The second row in Fig. 4 shows that our non-linear refinement can lead to a more uniform mesh triangulation in the final alignment than other methods. Quantitative results in Table 2 show that our non-linear refining strategy can improve both geometric and rendering results.

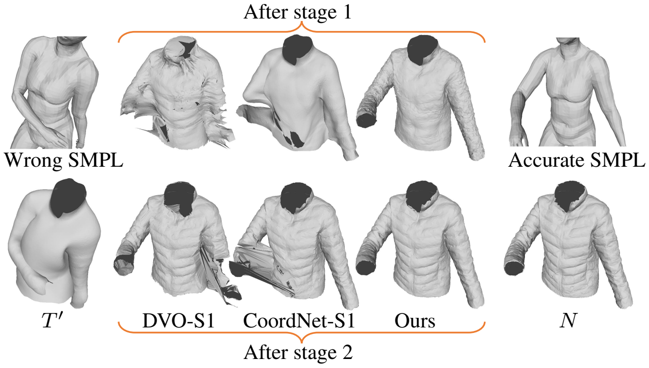

Figure 5. Robustness to inaccurate SMPL estimations. Leftmost column: the template T' initialized by LBS with an inaccurate SMPL. Middle columns: alignment results after stage 1 and stage 2 for different deformation modules used in stage 1. Rightmost column: The target shape N and its corresponding accurate SMPL estimation.

Texture information We also test our pipeline without using texture information in Section 3.5, labeled NoTex in Table. 2. As shown in Table 2, while this does not affect ge-

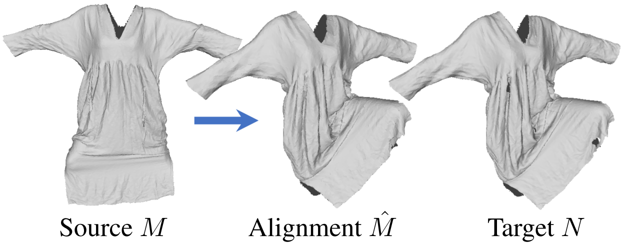

Figure 6. The alignment result for a long dress.

Method	Chamfer Dist. ($\times 10^{-3}$) ↓	LPIPS-VGG ($\times 10^{-2}$) ↓	SSIM ↑
Ours	**2.780**	**8.866**	**0.824**
DVO-S1	2.899	10.868	0.813
CoordNet-S1	2.793	9.188	<u>0.821</u>
LR-S2	2.834	9.146	<u>0.821</u>
NR-S2	2.897	9.344	0.819
NoTex	<u>2.783</u>	<u>9.097</u>	<u>0.821</u>

Table 2. Quantitative results of ablation studies. We report the metrics for S02 (knee-long coat). The best results are in boldface and the second best results are underlined.

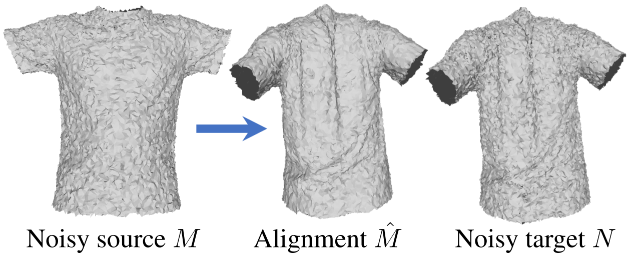

Figure 7. The alignment result for data with synthetic noise.

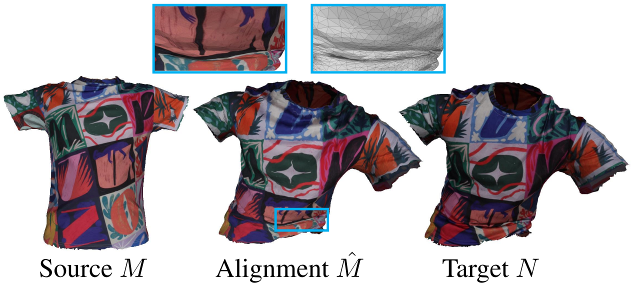

Figure 8. The alignment result for data with deep wrinkles.

ometric accuracy, texture alignment becomes less accurate.

5. Conclusion and Future Work

In this work, we introduce a two-stage pipeline that aligns garments by utilizing intrinsic manifold properties and neural deformation fields. Our intrinsic deformation network for 3D fitting leverages manifold continuity instead of extrinsic 3D continuity, and can thus handle difficult garment types such as long coats, as well as being robust to poor initializations of the base template. Furthermore, our method extends the Functional Maps framework [37] by introducing non-linearity with neural deformation fields, achieving texture-level high accuracy garment alignment where highly non-rigid and non-isometric deformations are present.

Limitations and Failure Cases As discussed in [13], extrinsic methods are often more suitable for shapes with inconsistent topology. While we have also relied on extrinsic (3D) fitting in the coarse stage, our final alignment comes from intrinsic embeddings, which may not be suitable if the target shape is topologically different from the source shape. Moreover, since it is not intuitive how physics-based constraints can be added to the intrinsic space, the output may not always be physically correct. Another limitation is that in both stages we model the alignment between complete shapes that are nicely segmented without noise and self-occlusion. Noisy scans or partial scans, which are very common for garment captures, cannot be perfectly handled yet. Currently, our method does not explicitly handle noise but can still produce a reasonable alignment for noisy data (Fig. 7). If the target shape has deep wrinkles (which is essentially a type of partiality due to self-occlusion), the vertices near the wrinkle would appear smoothly squeezed together as shown in Fig. 8.

Future Work Aligned garment data enable a number of applications, *e.g.*, texture capture or physical parameter estimation. Our method can also be used as a tool for building correspondences for garment datasets to be used for learning tasks, *e.g.*, learning deformations and textures of garments to drape them on human avatars, or learning shape descriptors specific to garments for correspondence inference. Efforts should also be devoted to improve various aspect of our method. In our current setting it is assumed that garment data are nicely segmented scans, without noise and occlusion during capture. Future work should extend this pipeline to support imperfections such as noise, partiality and segmentation failures. Generalizing the LBO eigenfunction computation to point clouds can also make the pipeline more compatible with depth sensor-based captures.

Acknowledgments

This paper is supported by National Key R&D Program of China (2022YFF0902200), the NSFC project No.62125107 and No.61827805.

References

[1] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[2] Souhaib Attaiki, Gautam Pai, and Maks Ovsjanikov. DPFM: Deep partial functional maps. In *2021 International Conference on 3D Vision (3DV)*. IEEE, Dec. 2021.

[3] Mathieu Aubry, Ulrich Schlickewei, and Daniel Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. pages 1626–1633, 11 2011.

[4] Ronen Basri, David Jacobs, Yoni Kasten, and Shira Kritchman. *The Convergence Rate of Neural Networks for Learned Functions of Different Frequencies*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[5] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, December 2020.

[6] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[7] Aljaž Božič, Pablo Palafox, Michael Zollhöfer, Angela Dai, Justus Thies, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. *CVPR*, 2021.

[8] Xu Chen, Yufeng Zheng, Michael J. Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 11594–11604, October 2021.

[9] Toby Chong, I-Chao Shen, Nobuyuki Umetani, and Takeo Igarashi. Per garment capture and synthesis for real-time virtual try-on. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, UIST '21, page 457–469, New York, NY, USA, 2021. Association for Computing Machinery.

[10] Étienne Corman, Maks Ovsjanikov, and Antonin Chambolle. Supervised descriptor learning for non-rigid shape matching. In *Computer Vision - ECCV 2014 Workshops*, pages 283–298, Cham, 2015. Springer International Publishing.

[11] Bailin Deng, Yuxin Yao, Roberto M. Dyke, and Juyong Zhang. A survey of non-rigid 3d registration. *Computer Graphics Forum*, 41(2):559–589, 2022.

[12] Nicolas Donati, Abhishek Sharma, and Maks Ovsjanikov. Deep geometric functional maps: Robust feature learning for shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] Marvin Eisenberger, Zorah Lahner, and Daniel Cremers. Smooth shells: Multi-scale shape registration with functional maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[14] Marvin Eisenberger, Aysim Toker, Laura Leal-Taixé, and Daniel Cremers. Deep shells: Unsupervised shape correspondence with optimal transport. *Advances in Neural Information Processing Systems*, 34, 2020.

[15] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[16] Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, and Qionghai Dai. Robust non-rigid motion tracking and surface reconstruction using l0 regularization. *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[17] Oshri Halimi, Or Litany, Emanuele Rodola, Alex M. Bronstein, and Ron Kimmel. Unsupervised learning of dense shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[18] Oshri Halimi, Tuur Stuyck, Donglai Xiang, Timur Bagautdinov, He Wen, Ron Kimmel, Takaaki Shiratori, Chenglei Wu, Yaser Sheikh, and Fabian Prada. Pattern-based cloth registration and sparse-view animation. *ACM Transactions on Graphics (TOG)*, 41(6), nov 2022.

[19] Hajar Hamidian, Zichun Zhong, Farshad Fotouhi, and Jing Hua. Surface registration with eigenvalues and eigenvectors. *IEEE Transactions on Visualization and Computer Graphics*, 26(11):3327–3339, Nov 2020.

[20] Peng Huang, C. Budd, and A. Hilton. Global temporal registration of multiple non-rigid surface sequences. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, page 3473–3480, USA, 2011. IEEE Computer Society.

[21] Lukas Koestler, Daniel Grittner, Michael Moeller, Daniel Cremers, and Zorah Lähner. Intrinsic neural fields: Learning functions on manifolds. In *European Conference on Computer Vision (ECCV)*, pages 622–639, Cham, 2022. Springer Nature Switzerland.

[22] Hao Li, Bart Adams, Leonidas J. Guibas, and Mark Pauly. Robust single-view geometry and motion reconstruction. *ACM Transactions on Graphics (TOG)*, 28(5):1–10, dec 2009.

[23] Yang Li and Tatsuya Harada. Non-rigid point cloud registration with neural deformation pyramid. In *Advances in Neural Information Processing Systems*, 2022.

[24] Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. Posevocab: Learning joint-structured pose embeddings for human avatar modeling. In *ACM SIGGRAPH Conference Proceedings*, 2023.

[25] Miao Liao, Qing Zhang, Huamin Wang, Ruigang Yang, and Minglun Gong. Modeling deformable objects from a single depth camera. In *2009 IEEE 12th International Conference on Computer Vision*, pages 167–174, 2009.

[26] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *European Conference on Computer Vision (ECCV)*, page 210–228, Berlin, Heidelberg, 2022. Springer-Verlag.

[27] Or Litany, Tal Remez, Emanuele Rodola, Alex Bronstein, and Michael Bronstein. Deep functional maps: Structured prediction for dense shape correspondence. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015.

[29] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16082–16093, 2021.

[30] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6469–6478, 2020.

[31] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021.

[32] Sahib Majithia, Sandeep N. Parameswaran, Sadbhavana Babar, Vikram Garg, Astitva Srivastava, and Avinash Sharma. Robust 3d garment digitization from monocular 2d images for 3d virtual try-on systems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3428–3438, January 2022.

[33] Diana Mateus, Radu Horaud, David Knossow, Fabio Cuzzolin, and Edmond Boyer. Articulated shape matching using laplacian eigenfunctions and unsupervised point registration. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[34] Simone Melzi, Jing Ren, Emanuele Rodolà, Abhishek Sharma, Peter Wonka, and Maks Ovsjanikov. Zoomout: Spectral upsampling for efficient shape correspondence. *ACM Transactions on Graphics (TOG)*, 38(6), nov 2019.

[35] Mark Meyer, Mathieu Desbrun, Peter Schröder, and Alan H. Barr. Discrete differential-geometry operators for triangulated 2-manifolds. In *Visualization and Mathematics III*, pages 35–57, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

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

[37] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: A flexible representation of maps between shapes. *ACM Transactions on Graphics (TOG)*, 31(4), jul 2012.

[38] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4), jul 2017.

[39] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 09–15 Jun 2019.

[40] Jing Ren, Adrien Poulenard, Peter Wonka, and Maks Ovsjanikov. Continuous and orientation-preserving correspondences via functional maps. *ACM Transactions on Graphics (TOG)*, 37(6), dec 2018.

[41] Jean-Michel Roufosse, Abhishek Sharma, and Maks Ovsjanikov. Unsupervised deep learning for structured shape matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[42] Raif M. Rustamov. Laplace-beltrami eigenfunctions for deformation invariant shape representation. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*, SGP '07, page 225–233, Goslar, DEU, 2007. Eurographics Association.

[43] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2886–2897, 2021.

[44] Samuele Salti, Federico Tombari, and Luigi Di Stefano. Shot: Unique signatures of histograms for surface and texture description. *Computer Vision and Image Understanding*, 125:251–264, 2014.

[45] Yonggang Shi, Rongjie Lai, Danny J J Wang, Daniel Pelletier, David Mohr, Nancy Sicotte, and Arthur W Toga. Metric optimization for surface analysis in the laplace-beltrami embedding space. *IEEE transactions on medical imaging*, 33(7):1447—1463, July 2014.

[46] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1395, 2017.

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

[48] Zhaoqi Su, Liangxiao Hu, Siyou Lin, Hongwen Zhang, Zhang Shengping, Justus Thies, and Yebin Liu. Caphy: Capturing physical properties for animatable human avatars. *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[49] 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 *European Conference on Computer Vision (ECCV)*, pages 246–264. Springer, 2020.

[50] Robert W. Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics (TOG)*, 26(3):80–es, jul 2007.

[51] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based

on heat diffusion. *Computer Graphics Forum*, 28(5):1383–1392, 2009.

[52] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *European Conference on Computer Vision (ECCV)*, page 1–18, Berlin, Heidelberg, 2020. Springer-Verlag.

[53] Edgar Tretschk, Ayush Tewari, Michael Zollhöfer, Vladislav Golyanik, and Christian Theobalt. DEMEA: Deep Mesh Autoencoders for Non-Rigidly Deforming Objects. *European Conference on Computer Vision (ECCV)*, 2020.

[54] Zhou Wang, Alan Bovik, Hamid Sheikh, Student Member, and Eero Simoncelli. Image quality assessment: From error measurement to structural similarity. *IEEE Trans. Imgage Process.*, 13, 11 2003.

[55] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, Yaser Sheikh, Jessica Hodgins, and Chenglei Wu. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Transactions on Graphics (TOG)*, 41(6), nov 2022.

[56] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics (TOG)*, 40(6), dec 2021.

[57] Shuntaro Yamazaki, Satoshi Kagami, and Masaaki Mochimaru. Non-rigid shape registration using similarity-invariant differential coordinates. In *Proceedings of the 2013 International Conference on 3D Vision*, 3DV '13, page 191–198, USA, 2013. IEEE Computer Society.

[58] Guandao Yang, Serge Belongie, Bharath Hariharan, and Vladlen Koltun. Geometry processing with neural fields. In *Advances in Neural Information Processing Systems*, volume 34, pages 22483–22497. Curran Associates, Inc., 2021.

[59] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7287–7296, 2018.

[60] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap: Single-view human performance capture with cloth simulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5504–5514, 2019.

[61] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

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

[64] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4), 2023.