Robust Body Shape Correspondence with Anthropometric Landmarks

Yibo JIAO*[1], Chang SHU[2], Dinesh K. PAI[1]
[1] University of British Columbia, Vancouver BC, Canada;
[2] National Research Council Canada, Canada

https://doi.org/10.15221/22.17

Abstract

We propose a method to improve the robustness of state-of-art learning-based methods for finding point-to-point correspondences of 3D human models with anthropometric landmarks. Specifically, current deep learning-based methods generally focus on intrinsic, local, properties of body shapes, which lack extrinsic global information. Thus, these methods are challenged by matching ambiguities, for instance, due to the bilateral symmetry of human body shapes. We demonstrate our method with an unsupervised learning-based method, DeepShells [5]. Our work introduces a landmark supervision method based on the Shells by adding linear soft constraints to minimize this problem that we term the "intrinsic feature ambiguity problem." To that end, we derive a simple but efficient pipeline that better distinguishes self-similarities yet has similar overall matching quality.

Keywords: shape matching, deep learning, anthropometry

1. Introduction

Anthropometric landmarks are the most important locations that define human shape correspondences. However, current work on shape correspondence seldom uses landmark information, neither for prediction nor for validation. Popular human scan datasets like FAUST [3] and SCAPE [1] lack ground truth data for landmarks. For interclass matching, and matching shapes after large deformation, landmarks are the only reliable validation. Therefore, anthropometric landmarks are critical for automatically computing point-to-point correspondences for body shapes.

Early deep learning-based methods for dense shape matching like FSPM [8] and 3D-CODED [6] requires ground truth about deformation and noise of 3D shapes. Fortunately, Smooth Shells [4] and Deep Shells [5] (collectively "the Shells") proposed unsupervised algorithms that do not need ground truth labels and still handles various types of noise and deformations.

Both methods use a hierarchical matching algorithm that iteratively aligns approximated shapes in a coarse-to-fine manner, thus the Shells need relatively good initial alignments for optimization. However, finding a good enough initial alignment is challenging because self-similarities like symmetries and self-touching cases for 3D scans are difficult to detect. Smooth shells [4] used surrogate-based MCMC sampling for initialization, which generates a large number of proposed alignments and selects the one with the lowest cost. Deep Shells [5] used the same hierarchical matching algorithm but introduced a fully differentiable cost function and used learning to find refined local features to initialize the matching pipeline.

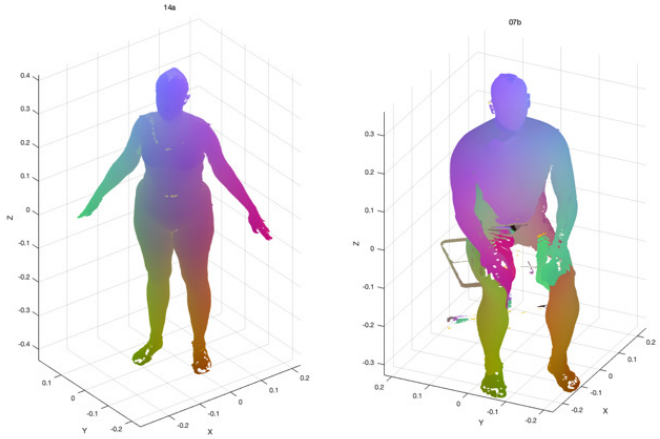

Fig. 1. Deep Shells with a bad initialization is challenged by self-similarities. This trained model aligns the upper body of the two shapes on the opposite side

Although taking advantage of deep learning made Deep Shells more computationally efficient, the learned refined local features were not guaranteed to disambiguate self-similarities because Deep Shells only randomly samples a subset of points on the input shapes for training and it used SHOT [10] descriptors which are very local and do not include extrinsic global information. A failure case for a trained Deep Shell model is shown in Fig. 1.
In this work, we used human scan and landmark information from the CAESAR [7] dataset. CAESAR identifies 73 anthropometric landmarks on each human model. The ground truth positions of these landmarks are utilized in our method as user-defined constraints and embedded into the training pipeline. Constraints of landmark losses penalize alignments that match landmarks on the other side of the body. An optimal alignment trained with these constraints tends to better distinguish self-symmetries. To that end, we converted the original unsupervised learning method to a weakly-supervised learning approach with only 73 landmark positions. We demonstrated that adding landmark constraints is not only beneficial for distinguishing intrinsic symmetries that DeepShells suffered from but also maintains the quality of overall matching. Our method is considered as a semi-automatic approach since it only requires a small number of ground truth labels and still takes advantage of unsupervised learning for efficiency. Addressing the intrinsic feature ambiguity problems also improves the accuracy of landmark prediction. We further showed that in addition to more accurate dense correspondences, our method also has higher accuracy in matching landmarks.

2. Background and Related Work

2.1. Smooth Shells and MCMC

Self-similarities were previously solved [4] by running surrogate-based Markov Chain Monte Carlo sampling initialization. This method samples various initial alignments τ and pick the alignment with minimal energy. The energy in this initialization step is:

$$E_{init}(P, C, \tau) = \|PX_K^* - Y_K\|_F^2$$

Where X_K^*, Y_K are the approximated source and target shapes sampled by K points. P is correspondences between shapes X, Y and (C, τ) are alignments to be initialized. C is the functional map [9] and τ is the displacement parameter, i.e., point-wise translations. Since the correspondence P computed by nearest neighbor searching is in extrinsic coordinates, this energy only requires an optimal τ. Ideally, if we sample a sufficient number of alignments and pick a reasonable threshold of energy to select optimal alignment, $(\tau_{best}, X_{best}^*)$ are guaranteed to initialize the matching and find P at the coarsest level. However, exploring initial poses $\tau_{prop} \in R^{K\times 3}$ are very computationally costly, thus Deep Shells used learning to replace this step to improve efficiency.

2.2. Deep Shells and Spectral Convolution

Deep Shells replaces the costly initialization in Smooth Shells by learning refined features. Instead of searching τ, Deep Shells takes SHOT descriptors as input and uses spectral convolution to get spectral information. Then spectral filters are learned to extract refined local features in order to find initial correspondences. The energy in this initialization layer is [5]:

$$E_{init}(\pi) = \int_{X\times Y} \|G^X(x) - G^Y(y)\|_2^2 d\pi(x, y) - \lambda H(\pi) \qquad (1)$$

Where X, Y are descriptors inputs and G are learned features computed by learned spectral filters, π is initial soft correspondences, which is a probability matrix indicating point-wise matching probabilities. With the same hierarchical matching algorithm in Smooth Shell followed by this initialization layer, spectral filters are learned to result optimal π_{init}. However, learned spectral features are unable to faithfully distinguish self-similarities because training spectral filters relies on input descriptors, and SHOT descriptors are unstable, very local, and highly depend on the mesh structure. In addition, Deep Shells only samples a small part of the points to train because of the computational cost. Deep Shells has worse performances when the training shapes are in high resolution, and running full-resolution without sampling is still costly.

2.3. Anthropometry

There is a long history of research in anthropometry, measuring human bodies using different technologies including 3D body scanning shapes. The highly influential CAESAR project [7] created a dataset of different demographic groups, with both traditional measurements and 3D scans measurements. In addition, 73 pre-specified anthropometric landmarks were measured in standard

postures by experts [7]. Such landmarks were proven to be critical to deformation and correspondence of human body shapes [2]. Subsequently there have been many efforts to create anthropometric data bases in both academia (e.g., SCAPE [1] and the FAUST [3]) and industry (e.g., http://www.sizenorthamerica.com/). However, localizing landmarks on human body is very time-consuming, fortunately, there were other work focused on automatic localization.

Automatic locating algorithms for anthropometric landmarks are proposed in the prior work [2], in which the problem is formulated as a probabilistic inference problem. In this prior work, a Markov network is trained to localize all 73 anthropometric landmarks on human scans. This could potentially provide us the ground truth labels for landmark supervision, and embedding this network as a pre-processing step in our data driven pipeline could make our method fully unsupervised. However, there are no axiomatic methods for locating such landmarks on the human body, and large errors were found in prior works. The traditional way of locating such landmarks relies on placing markers on human bodies prior to scanning. Such markers are measured and placed by experts, which are time-consuming

2.4. Motivation and Goal

The trade-off between training resolution and quality of refinements of learned local features in Deep Shells motivates our method. Instead of training with all points, we try to sample points that are more representative for local geometrical features and easier to obtain ground truth labels. Such points are called anthropometric landmarks. We take advantage of an unsupervised learning method that does not require point-to-point truth labels and we supervise 73 located such landmarks to help distinguish self-similarities.

The main goal of this work is to reduce cases of symmetrical problems of Deep Shells by learning better local feature using landmark supervision and still maintain the overall quality of matching and computation time especially for denser meshes. In this work, we use the ground truth of landmarks in CAESAR [7] dataset rather than auto-located positions to get better performance.

3. Method

3.1. Locating Anthropometric Landmarks

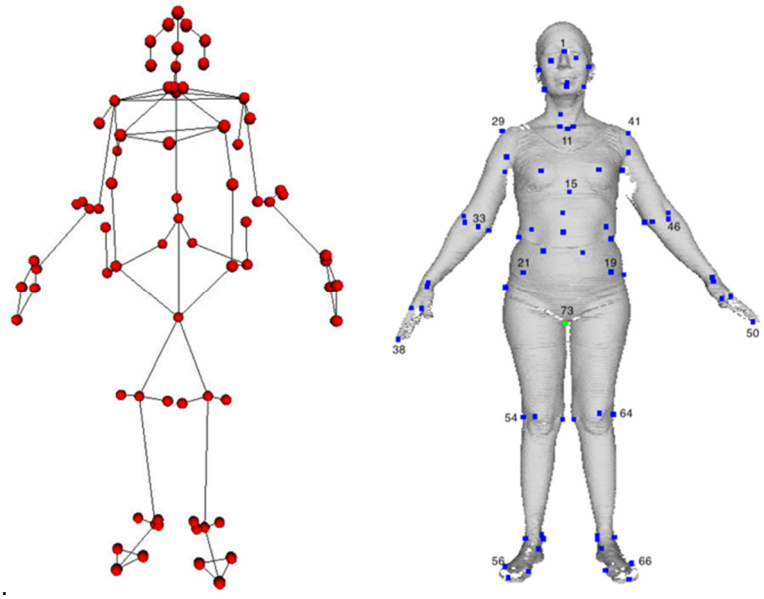

Fig. 2. Structure of landmark graph [2] and indexed landmarks on human scans.

An example of the structure and located landmarks on scans is shown in Fig. 2. [2] provided a full-list of indexed anthropometric landmarks with formal names and we adapt the same index from this prior work. In this work, we define human shape with a mesh, with vertex array X, and landmarks on the shape as points $l_i^{(X)}, i = 1, \dots, 73$. For simplicity, we assume that a landmark is represented by a vertex in X, though it is easy to relax this assumption to allow any position on the shape. Given a correspondence P that maps from the source shape X to the target shape Y, and a landmark i, define the landmark projected from X to Y as $P(l_i^{(X)})$. This is the matched position of landmark X on shape Y.

The landmark loss is defined by the error of localization between the matched position and the ground truth landmark position on target shape Y:

$$L_i(X,Y) = \left\| P\left(l_i^{(X)}\right) - l_i^{(Y)} \right\|_2^2$$

However, with differential settings in Deep Shells, correspondence P is replaced by a soft correspondence matrix Π. We define the soft projection as:

$$\Pi\left(l_i^{(X)}\right) = \pi^T Y, \pi \in \Pi(X,Y) = row_{l_i^{(X)}}(\Pi),$$

where we take probability vector π that indicates the soft correspondence that projects $l_i^{(X)}$ onto Y. Taking inner product will give us a barycentric extrinsic coordinate with probabilistic weights. Thus, differentiable landmark loss function for landmark i is:

$$L_i(X,Y) = \left\| \Pi\left(l_i^{(X)}\right) - l_i^{(Y)} \right\|_2^2$$

A visual illustration for this landmark projection and landmark loss is shown in Fig. 3.

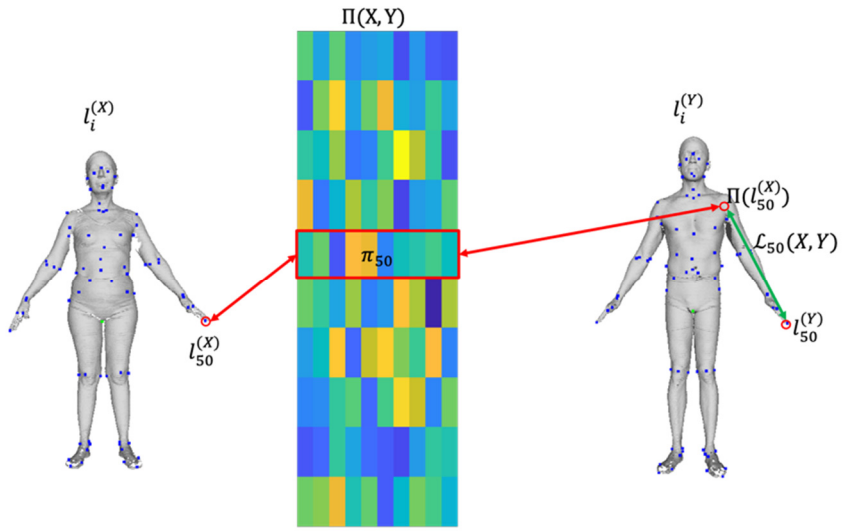

Fig. 3. Example of landmark projection and landmark distance loss function for landmark 50: Lt. Dactylion

3.2. Assumption

Recall that our goal is to minimize cases with the "intrinsic feature ambiguity problem", and symmetry is one of the leading causes. We assume that large landmark losses are only caused by two possible situations: either overall matching is erroneous or the matching has a symmetrical problem.

As shown in Fig. 3, if the large landmark loss value is caused by matching landmarks to an area that do not have similar local features, this is the case of erroneous matching because the learned spectral filters cannot distinguish local features. This case should not be common because Deep Shells has proven to have good quality of matching for benchmark datasets. However, if the Shell is challenged by the symmetrical problem and matches landmarks on the opposite side of the human shape, that indicates the learned parameters are able to distinguish local geometry features but not to distinguish ambiguity caused by self-similarities because of the symmetrical properties of human bodies.

3.3. Energy Formulation with Landmark Constraints

To apply landmark constraints, we simply add landmark losses to the original Deep Shell energy in equation (1) for optimization. These constraints need to be applied on both the initialization layer and hierarchical matching steps. For initialization, we formulate the energy as:

$$min\ \boldsymbol{E_{init}}(\boldsymbol{\pi}) \quad \boldsymbol{s.t.} \sum_{\boldsymbol{i=1}}^{\boldsymbol{73}} \boldsymbol{w_i L_i}(\boldsymbol{X^*}, \boldsymbol{Y}) \leq \boldsymbol{\epsilon}\ ,$$

which is rewritten as:

$$min\ \boldsymbol{E_{init}}(\boldsymbol{\pi}) + \boldsymbol{\mu} \textstyle\sum_{\boldsymbol{i=1}}^{\boldsymbol{73}} \boldsymbol{w_i L_i}(\boldsymbol{X^*}, \boldsymbol{Y})\ ,$$

where we assign different weights $\boldsymbol{w_i}$ for each landmark because landmarks that are further from the center axis are more affected by symmetrical problem, and the hyperparameter μ controls the hardness of the constraint. In this way, we force the initialization layer to learn spectral filters that produce initial correspondences with minimal landmark loss in order to solve symmetrical alignment in the initialization step.

For hierarchical matching steps, similar addition of landmark loss is applied:

$$\boldsymbol{E(\pi, C, \tau)} \;=\; \boldsymbol{E_{init}(\pi)} \;+ \boldsymbol{\mu} \sum_{\boldsymbol{i=1}}^{\boldsymbol{73}} \boldsymbol{w_i L_i(X^*(C, \tau), Y)}$$

3.4. Pipeline Overview

Fig. 4. gives an overview of the method. For a given pair of training shapes, we firstly locate anthropometric landmarks as part of the pre-processing. In the initialization layer, instead of randomly sampling points, we sample nearest points around each landmark. Landmark loss constraints are then applied in both initialization layers and hierarchical matching steps to resolve symmetrical ambiguities.

For testing a pair of shapes, we input SHOT features and use the learned spectral filters in the training stage to compute the soft correspondence, then run the same matching algorithm in the training pipeline.

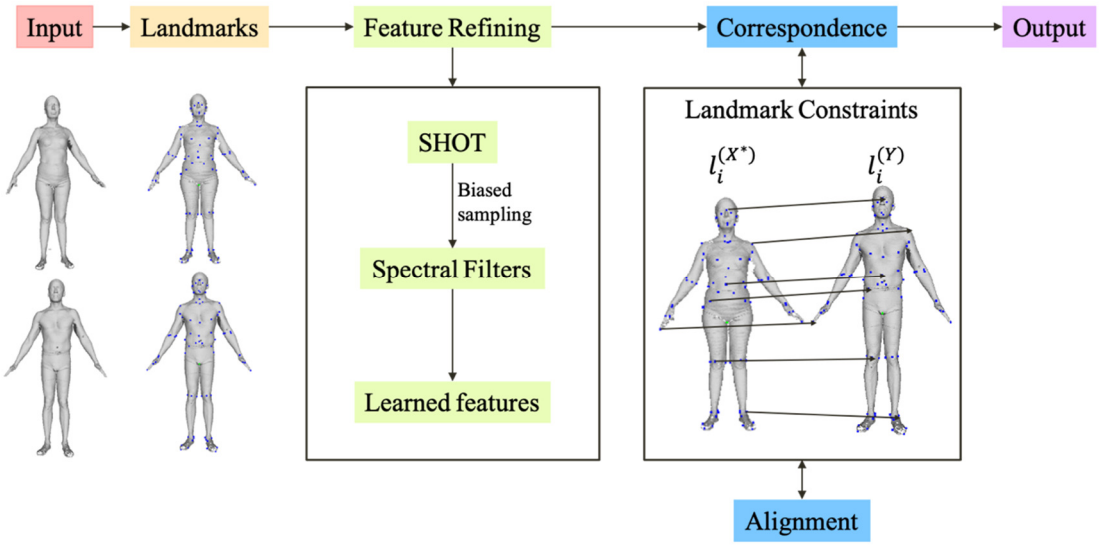

Fig. 4 Pipeline overview of Deep Shells with landmark supervision

4. Results

4.1. Implementation Details

We implemented the network based on Deep Shells. In addition to synthetic human mesh dataset like the SCAPE [1] and the FAUST [3], we also train our model with the CAESAR [7] dataset, which contains real-world scans with much higher resolution(more than 200K vertices per shape). We normalized the meshes by centering the shape on the origin and re-sizing the meshes to have same overall surface area.

4.2. Computational Efficiency Comparison

One of the main goals of our work is to make training more efficient with higher resolution shapes. We trained our network and comparable networks on 3 datasets using a GPU cluster with 4 GPUs. We recorded the time of running one pair of matching and initialization. Our model is trained with sampling 20 nearest points around each landmark; thus 1460 points are selected. See Table. 1. For runtime comparisons. Our method has similar running time compared to Deep Shells because landmark operations are efficient and the optimization is a least square problem.

Table 1. Runtime comparison, DSFR stands for Deep Shells in full resolution and DSRS is Deep Shells with 1000 random sampling, time in seconds.

	DSFR	DSRS	Ours
FAUST	680	221	230
SCAPE	613	189	201
CAESAR	1343	372	367

4.3. Symmetrical Problem Results

Our method solves symmetrical problems during matching. We trained our model using 25 shapes (625 pairs), tested on 30 shapes (900 pairs) and computed the maximum error of localization of landmarks for test results, relatively large errors indicate symmetrical problems and we record the number of matching pairs that had such result. See Table 2. for symmetrical cases comparison and Fig. 5. is an example case where Deep Shells failed and our method worked for the same pair of shapes. We note that our method had no failures for scans in similar poses, and the remaining failure cases were all for drastic pose changes (standing to sitting). We chose the scan that has extreme degrees of noise and partiality.

Table 2. Number of cases that models failed to distinguish symmetries

	DSFR	Ours
FAUST	98	23
SCAPE	103	17
CAESAR	178	25

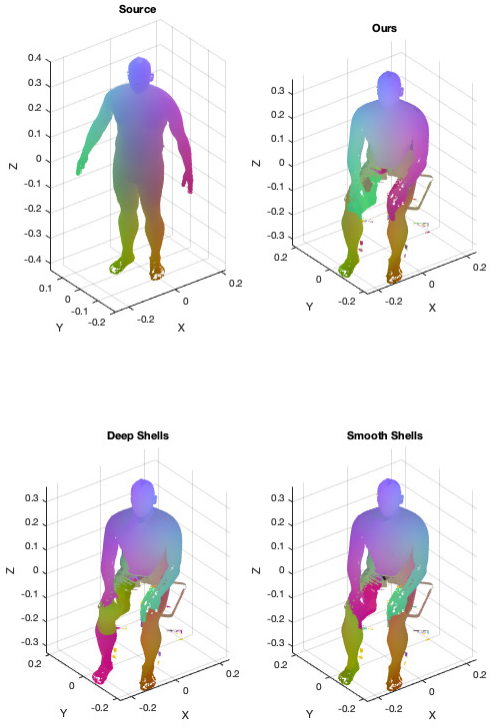

Fig. 5. An example of noisy and partial scan from matching, both Deep Shells and Smooth Shells were failed to distinguish symmetries while ours worked.

4.4. Overall Matching Quality

Our method outperforms at solving symmetrical problems, and at the same time, it is also able to obtain similar overall high-quality correspondences that are comparable with the Shells [5]. We evaluate matching quality by recording average error for landmarks for all tested data pairs in CAESAR dataset [7], because FAUST and SCAPE dataset does not provide ground truth labels for point-to-point correspondences. See Fig. 6. for details.

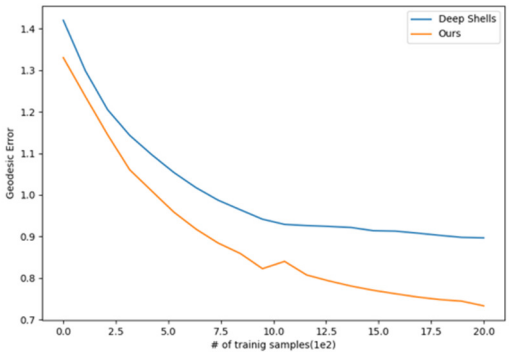

Fig. 6. A quantitative comparison of our method and Deep Shells, we train different number of samples and show average geodesic errors tested with same 900 testing scans

4.5. Limitation

Although our method is more efficient and solves symmetrical cases, like all other shape matching methods, self-touching is still a challenging problem. In addition, our method is highly dependent on the precision of landmark localizations, thus we need a much more reliable algorithm to automatically locate landmarks in order to train with synthetic shape datasets. Furthermore, obtaining landmarks from non-human scans are difficult. In this case, pure unsupervised methods are better options for interclass shape matching problems. The matching may improve with a more judicious choice of $\boldsymbol{\mu}$, or treatment of explicit constraints, which we leave for future work.

5. Conclusion

We propose a stable and simple method based on Deep Shells that solves symmetrical problems by optimizing problems with landmark constraints. Our method takes advantage of spectral convolution for feature extraction and learns more refined features that can distinguish self-similarities and use landmark supervision to constrain optimizations. We show that constraining landmark losses not only helps to disambiguate but also maintain similar quality of matching and with the same efficiency.

References

[1] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. In *ACM SIGGRAPH 2005*, pages 408-416, 2005.

[2] Zouhour Ben Azouz, Chang Shu and Anja Mantel, Automatic locating of anthropometric landmarks on 3d human models. In *Third International Symposium on 3D Data Processing, Visualization and Transmission(3DPVT'06)*, pages 750-757, 2006

[3] Federica Bogo, Javier Romero, Matthew Loper, and Michael J Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3794-3801, 2014

[4] Marvin Eisenberger, Zorah Lahner, and Daniel Cremers. Smooth Shells: Multi-scale shape registration with functional maps. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12265-12274, 2020.

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

[6] Thibault Groueix, Matthew Fisher, Valdmir G. Kim, Bryan C. Russel, and Mathieu Aubry. 3d-coded: 3d correspondences by deep deformation. 2018.

[7] Robinette Kathleen, Daanen H, and Paquet Eric. The Caesar project: a 3-d surface anthropometry survey. *3-D Digital Imaging and Modelling*, pages 380-386, 1999.

[8] Or Litany, Emanuele Rodola, Alex Bronstein, and Michael Brotein. Fully spectral partial shape matching. 36(2):1681-1707, 2021.

[9] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher and Leonidas Guibas. Functional maps: a flexible representation of maps between shape. *ACM Transactions on Graphics(TOG)*, 31(4):1-11,2012.

[10] Federico Tombari, Samuele Salti, and Luigi Di Stefano. 3unique signatures of histogram for local surface description. 16(9):356-369, 2010