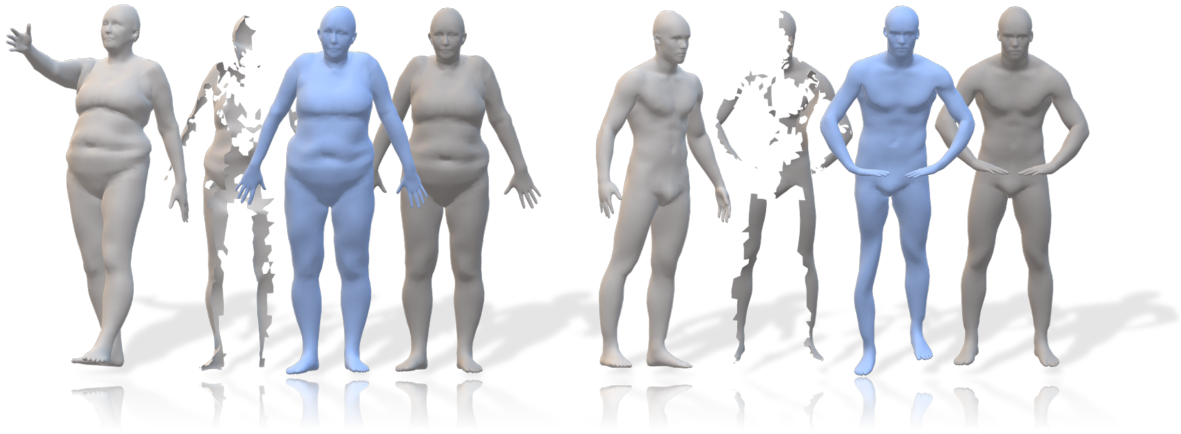
The Whole Is Greater Than the Sum of Its Nonrigid Parts



Left to right: input reference shape, input part, output completion, and the ground truth full model.

Oshri Halimi
Technion, Israel
oshri.halimi@gmail.com

Ido Imanuel
Technion, Israel
ido.imanuel@gmail.com

Or Litany
Stanford University
or.litany@gmail.com

Giovanni Trappolini
Sapienza University of Rome
igiovanni.trappolini@uniroma1.it

Emanuele Rodolà
Sapienza University of Rome
rodola@di.uniroma1.it

Leonidas Guibas
Stanford University
guibas@cs.stanford.edu

Ron Kimmel
Technion, Israel
ron@cs.technion.ac.il

Abstract

According to Aristotle, a philosopher in Ancient Greek, "the whole is greater than the sum of its parts". This observation was adopted to explain human perception by the Gestalt psychology school of thought in the twentieth century. Here, we claim that observing part of an object which was previously acquired as a whole, one could deal with both partial matching and shape completion in a holistic manner. More specifically, given the geometry of a full, articulated object in a given pose, as well as a partial scan of the same object in a different pose, we address the problem of matching the part to the whole while simultaneously reconstructing the new pose from its partial observation. Our approach is data-driven, and takes the form of a Siamese autoencoder without the requirement of a consistent vertex labeling at inference time; as such, it can be used on unorganized point clouds as well as on triangle meshes. We demonstrate the practical effectiveness of our model in the applications of single-view deformable shape completion and dense shape correspondence, both on synthetic and real-world geometric data, where we outperform prior work on these tasks by a large margin.

1. Introduction

Aristotle, a philosopher in Ancient Greek announced that *"the whole is greater than the sum of its parts"*. This axiomatic observation was narrowed down to human perception of planar shapes by Gestalt psychology school of thought in the twentieth century. A central principle of Gestalt theory is the principle of reification, claiming that humans perception contains more spatial information than can be extracted merely from the sensory stimulus, and giv-

ing rise to the view that the mind generates the additional information based on verbatim aquired patterns.

From a practical perspective, advances in robotics and augmented and virtual reality technologies often rely upon the ability to render a scene from novel views, manipulate its content, and add physical constraints. This often requires the completion of geometric structures from partial data. To that end, we would like to have a systematic and universal way to complete a partial shape into its full counterpart, a problem frequently referred to as *shape completion* in the geometry processing community.

Shape completion is an ill-posed problem by definition, as one must address the question of what can be considered a legitimate completion. An attempt to answer this question can be addressed in a statistical manner. One can consider many instances of partial shapes and their corresponding completions, defining a statistical relation between a part and the geometric structure from which it was cropped, and by which the whole structure can be recovered in a statistically optimal manner. The limitation of the statistical approach reveals in cases of extreme partiality, where the correlation between the partial shape and the whole structure is too weak to enable accurate reconstruction of details in the missing regions.

In this paper, we introduce a *deterministic* shape completion framework. Given a partial observation, the method returns a reliable reconstruction of a full, realistic object. At a first glance, at least in a rigid setting, this task may seem impossible: How can we guarantee that such a reconstruction reliably describes the part that was hidden from the viewing direction? However, one soon realizes that the problem can be formulated differently in the *non-rigid* case. Since two non-rigidly related shapes could share the same intrinsic geometry [22, 14, 10, 31, 29], each shape holds information about the other. We therefore pose the alternative question: Given a full object Q and a partial view P *in a different pose*, can we reconstruct a *full* version of P by borrowing geometric information from Q? In this paper we address precisely this question. Our goal can thus be seen as a combination of two complementary tasks: (1) partial shape matching, and (2) non-rigid shape completion. While the two tasks are often addressed separately, we claim that considering their coupling provides powerful means to deal with both.

Contribution. In this paper we propose a novel formulation unifying deformable shape completion and partial shape matching. Specifically, given a full and a partial shape related by a non-rigid transformation, our objective is to deform the full shape to best fit the partial data. To compute this deformation, we train an encoder-decoder network. Once a completion for the part is predicted, the part-to-full correspondence can be trivially recovered using nearest-neighbor search. Our main contributions can be summarized as follows:

1. We introduce a deep Siamese architecture to tackle non-rigid alignment between a shape and its partial scan;
2. To the best of our knowledge, the proposed method is the first that addresses shape completion and partial correspondence in one framework under *extreme* partiality;
3. The proposed method is *efficient*, taking less than a second to provide both outputs without requiring any time-consuming post-processing steps.

2. Related work

Our problem setting is closely related to multiple research directions in the shape analysis and geometric deep learning communities. In an early attempt to use one pose in order to geometrically reconstruct another, Devir *et al*. [20] considered mapping a clean shape in a given pose onto the same noisy shape in a different pose. Elad and Kimmel were the first to treat shapes as metric spaces [21, 22]. In fact, it was the first approach of matching shapes by their spectra, specifically, second order moments of embedding the intrinsic metric into a Euclidean one via classical scaling. It involved comparing the spectra of the click's laplacian. Bronstein *et al*. [9, 10, 13, 8, 11, 14] dealt with partial matching of articulated objects in various scenarios, including pruning of the intrinsic structure while accounting for cuts.

Shape completion. Recovering a complete shape from partial measurements is a longstanding research problem that comes in many flavors. In the context of deformable shapes, early efforts focused on completion based on geometric priors [37] or reoccurring patterns [13, 39, 61, 41]. These methods are not suited for severe partiality. For such cases model-based techniques are quite popular, *i.e*., category-specific parametric morphable models that can be fitted to the partial data [4, 23, 45, 1]. Model-based shape completion was demonstrated for keypoint input [2], and was recently proven to be quite useful for recovering 3D body shapes from 2D images [66, 65, 28, 73]. Parametric morphable models [4], coupled with axiomatic image formation models were used to train a network to reconstruct face geometry from images [57, 56, 62]. Still, much less attention has been given to the task of fitting a model to a partial 3D point cloud. Recently, Jiang *et al*. [35] tackled this problem using a skeleton-aware architecture. However, their approach works well when full coverage of the underlying shape is given.

Deep learning of surfaces. Following the huge success of convolutional neural networks in images, in recent

years, the geometry processing community has been rapidly adopting and designing computational modules suited for such data. The main challenge is that unlike images, geometric structures like surfaces come in many types of representations, and each requires a unique handling. Early efforts focused on a simple extension from a single image to multi-view representations [63, 69]. Another natural extension are 3D CNN on volumetric grids [70]. A host of techniques for mesh processing were developed as part of a research branch termed *geometric deep learning* [15]. These include graph-based methods [67, 68, 32], intrinsic patch extraction [48, 7, 49], and spectral techniques [42, 30]. Point cloud networks [53, 54] have recently gained much attention. Offering a light-weight computation restricted to sparse points with a sound geometric explanation [36], these networks have shown to provide a good compromise between complexity and accuracy, and are dominating the field of 3D object detection [52, 71], semantic segmentation [24, 72], and even temporal point cloud processing [17, 44]. For generative methods, recent implicit and parametric methods have demonstrated promising results [27, 50].

Inspired by the recent success of [26] in encoding non-rigid shape deformations using a point cloud network, here, we also choose to use a point cloud representation. Importantly, while the approach presented in [26] predicts alignment of two shapes, it is not designed to handle severe partiality, and assumes a fixed template for the source shape. Instead, we show how to align arbitrary input shapes and focus on such a partiality.

Partial shape matching. Dense non-rigid shape correspondence [38, 16, 42, 30, 59, 12, 18] is a key challenge in 3D computer vision and graphics, and has been widely explored in the last few years. A particularly challenging setting arises whenever one of the two shapes has missing geometry. This setting has been tackled with moderate success in a few recent works [58, 43, 55], however it largely remains an open problem whenever the partial shape exhibits severe artifacts or large, irregular missing parts. In this paper we tackle precisely this setting, demonstrating unprecedented performance on a variety of real-world and synthetic datasets.

3. Method

3.1. Overview

We represent shapes as point clouds $S = \{s_i\}_{i=1}^{n_s}$ embedded in $\mathbb{R}^3$. Depending on the setting, each point can carry additional semantic or geometric information encoded as feature vectors in R^d.

Given a full shape $Q = \{q_i\}_{i=1}^{n_q}$ and its partial view in a different pose $P = \{p_i\}_{i=1}^{n_p}$, our goal is to find a nonlinear function $F : \mathbb{R}^3 \to \mathbb{R}^3$ aligning Q to P[1]. If $R = \{r_i\}_{i=1}^{n_r}$ is the (unknown) full shape such that $P \subset R$, ideally we would like to ensure that $F(Q) = R$. Thus, the deformed shape $F(Q)$ acts as a proxy to solve for the correspondence between the part P and the whole Q. By calculating for every vertex in P its nearest neighbor in $R \approx F(Q)$, we trivially obtain the mapping from P to Q as well.

Clearly, the deformation function F depends on the input pair of shapes (P, Q). We model such dependency by considering a parametric function $F_\theta : \mathbb{R}^3 \to \mathbb{R}^3$, where θ is a latent encoding of the input pair (P, Q). We implement this idea via an encoder-decoder neural network, and learn the space of parametric deformations from example pairs of partial and complete shapes, together with full uncropped version the partial shape, serving as the ground truth completion.

Our network is composed of an encoder E and a generator F_θ. The encoder takes as input the pair (P, Q) and embeds it into a latent code θ. To map points from Q to their new location, we feed them to the generator along with the latent code. Our network architecture shares a common factor with 3D-CODED architecture [25], namely the deformation of a one shape based on the latent code of the another shape. However our pipeline is designed with the goal to merge two different sources of information into the reconstructed model, resulting in an accurate performance under extreme partiality. Specifically, in our architecture the deformation is applied to the varying full shape based on the mutual relation between the pair of input shapes, as represented in the latent code. In Appendix A.1 we perform an analysis where we train our network in a fixed-template setting, similar to 3D-CODED and demonste the advantage of our paradigm. In what follows we first describe each module, and then give details on the training procedure and the loss function. We refer to Figure 1 for a schematic illustration of our learning model.

3.2. Encoder

We propose the adoption of a Siamese pair of encoders, each producing a global shape descriptor (respectively θ_{part} and θ_{whole}). The two codes are then concatenated so as to encode the information of the specific pair of shapes, $\theta = [\theta_{part}, \theta_{whole}]$.

Our choice for the internal architecture in the Siamese pair is based on a preliminary ideal requirement: the encoder should *injectively* map each pair of shapes into a latent code. In other words, we require that each shape should be accurately reconstructed from its latent code. This requirement guarantees that for a fixed full shape Q and two different parts P_1 and P_2, there would be a way to deform

[1] In our setting, we assume that the pose can be inferred from the partial shape (*e.g.*, an entirely missing limb would make the prediction ambiguous), hence the deformation function F is well defined.

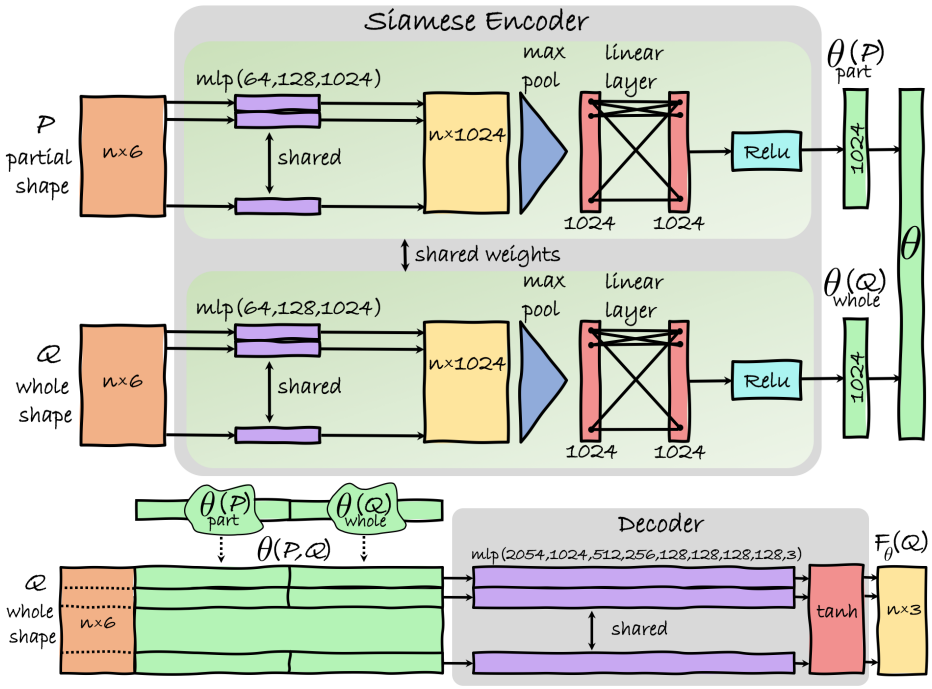


Figure 1. **Network Architecture**. Siamese encoder architecture at the top, and the decoder (generator) architecture at the bottom. A shape is provided to the encoder as a list of 6D points, representing the spatial and unit normal coordinates. The latent codes of the input shapes $\theta_{part}(P)$ and $\theta_{whole}(Q)$ are concatenated to form a latent code θ representing the input pair. Based on this latent code, the decoder deforms the full shape by operating on each of its points with the same function. The result is the deformed full shape F_θ(Q).

the full shape differently by using $\theta(P_1, Q)$ or $\theta(P_2, Q)$, based on the input part. Importantly, the ability to recover each shape from its latent code guarantees that the concatenation of the individual codes fully encodes the input pair, accommodating the use of the same encoder in each Siamese branch, instead of two different encoders.

Encouraged by recent methods [27, 26] that showed detailed reconstruction from PointNet latent code, we use a PointNet architecture for our encoder as well. Specifically, each encoder within the Siamese pair applies a multilayer perceptron to each 3D point of the input shape, with hidden dimensions $(64, 128, 1024)$, followed by a maxpool operation over the input points, leading to a 1024-dimensional vector. Finally, we apply a linear layer of size 1024 and a ReLu activation function. Hence, each shape in the input pair is represented by a latent code $\theta_{whole}, \theta_{part}$ of size 1024 respectively. We concatenate these to a joint representation θ of size 2048.

In practice, we find it helpful to also use the normal vector coordinates as additional input features to each of the Siamese network encoders, making each input point a 6D point. The normals were computed from the connectivity, for mesh input and approximated without the connectivity in case of point cloud input, as described in the experimental section. The normal vector field is especially helpful to disambiguate contact points of the surface and prevent contradicting requirements on the deformation function.

3.3. Generator

Given the code θ, representing the partial and full shapes, the generator has to predict the deformation function F_θ to be applied to the full shape Q. We realize F as a Multi-Layer Perceptron (MLP) to approximate the functional relation between an input point q_i on the full shape Q, and the corresponding output point r_i on the ground truth completed shape. The MLP operates pointwise on each tuple (q_i, θ), where the shape context θ is kept fixed for each input pair. The result is the destination location $F_\theta(q_i) \in \mathbb{R}^3$, for each input point of the full shape Q. This generator architecture allows, in principle, to calculate the output reconstruction in a flexible resolution, by providing the generator a full shape with some desired output resolution. In detail, the generator consists of 9 layers of hidden dimensions $(2054, 1024, 512, 256, 128, 128, 128, 128, 3)$, followed by a hyperbolic tangent activation function.

3.4. Training Procedure

In our experiments we used training examples that were prepared from datasets of full human shapes, see details in the data preparation section 4.1. Each of these datasets contain 3D models of different subjects in various poses. Each of our training examples is composed of a triplet (P, Q, R):

a partial shape P, a full shape in a different pose Q and a ground truth completion R. The shape Q and R were sampled from the same subject in two different poses. In each training example, the partial shape P is extracted from R by rendering its depth map in a random viewpoint angle, with the azimuth changing between 0° and 360° and the elevation angle kept at 0°. These synthetic projections aim to approximate partiality realizations commonly occurring in depth sensors. By their nature, these partial shapes still hold most of the information needed to determine the degrees of freedom with respect to the pose, making the pose reconstruction a well-posed problem. The training examples $(P_n, Q_n, R_n)_{n=1}^N$ were provided in batches to the Siamese Network, where N is the size of the train set. Each input pair is fed to the encoder to receive the latent code $\theta(P_n, Q_n)$ and the reconstruction $F_{\theta(P_n,Q_n)}(Q_n)$ is determined by the generator. This reconstruction is subsequently compared against the ground-truth reconstruction R_n using the loss defined in the next subsection 3.5.

3.5. Loss function

Essentially, the loss definition should reflect the visual plausibility of the reconstructed shape. Measuring such a quality analytically is a difficult problem. In this paper we adopt a naive measurement based on the Euclidean proximity between the ground-truth and the reconstruction. Formally, we define the loss as,

$$\mathcal{L}(P, Q, R) = \sum_{i=0}^{n_q} \left\| F_{\theta(P,Q)}(q_i) - r_i \right\|^2, \qquad (1)$$

where $r_i = \pi^*(q_i) \in R$ is the ground-truth matched point of $q_i \in Q$, and $\pi^* : Q \to R$ is the ground-truth mapping between the full shape Q and the ground-truth reconstruction R. To promote the preservation of shape details we measure the Euclidean distance between the coordinates in $\mathbb{R}^3$ as well as the Euclidean distance between the normal vectors evaluated at each point, interpreting q_i and r_i in the expression above as points in $\mathbb{R}^6$. Specifically, $q_i = (\vec{x}_{qi}, \alpha\vec{n}_{qi}), r_i = (\vec{x}_{ri}, \alpha\vec{n}_{ri}) \in \mathbb{R}^6$ are given by the concatenation of the coordinates vectors and the unit normal vector of each point in Q and R, respectively. The constant $\alpha > 0$ scales the normal vectors with respect to the Euclidean coordinates. The notation $\vec{n}_{qi}$ denotes the normal of the surface Q at point q_i, while $\vec{x}_{qi}$ denotes the coordinates of that point in $\mathbb{R}^3$. Similarly, $\vec{n}_{ri}$ denotes the normal of the surface R at point r_i, while $\vec{x}_{ri}$ denotes the coordinates of that point in $\mathbb{R}^3$. In spite of its simple form, empirically we observed that this loss's convergence produces high quality reconstructions.

3.6. Implementation considerations

Our implementation is available at https://github.com/OshriHalimi/shape_completion. The network was trained with each batch containing 10 triplet examples (P, Q, R), using the PyTorch [51] ADAM optimizer with a learning rate of 0.001 and a momentum of 0.9. We used a scale factor of $\alpha = 0.1$ for the normal vector. The network was trained for 50 epochs, each containing $10,000$ random triplet examples. The input shapes were translated such that their center of mass lies at the origin and the Iterative Closest Point algorithm [3] was further applied on the network output to perfectly align the axes with respect to the partial shape. Finally, to calculate the partial correspondence for each point in the partial shape, we retrieve its nearest-neighbor in the aligned reconstruction.

4. Experiments

The proposed method simultaneously tackles two important tasks in nonrigid shape processing and analysis, shape completion and partial shape matching. We emphasize that the suggested framework gracefully handles severe partiality. Prior efforts either addressed one of these tasks or attempted to address both only at mild partiality conditions. To thoroughly evaluate our performance, in this section we test our method on each of these tasks and compare with prior art. After a description of the datasets utilized, we present results of shape completion from a single view , and non-rigid partial correspondence. Finally, we show performance on real scanned data.

4.1. Datasets

We utilize two datasets of human shapes for training and evaluation, FAUST [5] and AMASS [46]. Both datasets were generated by fitting SMPL parametric body model [45] to raw scans. The second took the approach one step further by fitting those parameters to motion capture data. These datasets are quite different in size and variability. FAUST is a relatively small set of 10 subjects posing at 10 poses each. We follow previous methods, and test our method in partial shape matching and shape completion tasks using 10 projected views of these models. AMASS is currently the largest and most diverse dataset of human motion designed specifically for deep learning applications. It was generated by curating 15 archived datasets of marker-based optical motion capture data and unifying them into a shared statistical model SMPL+H [60]. As such, it can provide a much richer resource for evaluating the generalization ability of shape alignment and matching techniques. To this end, we used AMASS to create a large set of single-view projections. Specifically, we sampled every 100th frame from all provided sequences and rendered single-view projections from 10 equally spaced azimuth angles (elevation was kept fixed) using pyRender [34]. Keeping the original data splits, our dataset consists a total of $110K$, $10K$, and $1K$ full shapes for train, vali-

dation and test, respectively; and 10 times that size for the partial shapes. Note that at train time we randomly mix and match full shapes and their parts which drastically increases the effective set size.

4.2. Methods in comparison

A recent exploration of the problem of deformable shape completion for arbitrary partiality can be found in Litany *et al*. [40]. They suggest to find a completion via optimization in a learned shape space. Note that in that paper, the task was defined as a completion from a partial view without explicit access to a full model. Moreover, their solution requires a preliminary step of running a partial shape matching algorithm, which is slow due to the optimization at inference time. 3D-CODED [26] performs a template based alignment to an input shape in two stages: fast inference and a slow refinement through optimization. It is designed for pairs which are either full or witness to mild partiality, hence, to make the comparison more meaningful we adjust their loss at the refinement phase to the more suited directional Champfer distance. FARM [47] is also an alignment-based solution which has shown impressive results on shape completion and dense correspondences. It builds on the SMPL [45] human body model due to its compact parameterization, yet, we found it to be very slow to converge (up to 30 min for a single shape) and prone to getting trapped in local minima. 3D-EPN [19] is a rigid shape completion method based on a voxelization approach, utilizing a 3D CNN. Results are converted to a mesh via computation of an isosurface. A classic Poisson reconstruction [37] is also provided as a naïve baseline with no access to additional data other than the partial input.

4.3. Evaluation metrics

As previously discussed, it is challenging to define an analytic measurement for completion quality. Therefore we provide 5 different measurements, each reflecting a different perspective of the final completion as summarized in tables 1,2. First we report the mean square error (MSE) of the Euclidean distance between each point in the reconstructed shape and its ground truth mapping. This measure is reported only for template alignment methods for which the correspondence between the template and the ground truth reconstruction is defined. Next, we report the MSE of directional Chamfer distances: from the ground-truth to the completion and vice versa; The former measures the coverage of the ground-truth shape by the completion while the later penalizes outliers of the completion; We report the sum of both as full the Chamfer distance. Finally we measure the absolute error in the volume of the completion divided by the volume of the ground truth as a measure for volume deformation.

4.4. Single view completion

We evaluate the proposed method on the task of deformable shape completion on two datasets: FAUST and AMASS.

FAUST projections We follow the evaluation protocol proposed in [40] and summarize the completion results of our method and prior art in Table 1. As can be seen our network performs a much more accurate completion. Contrary to optimization based methods [40, 25, 47] which are very slow at inference time, our feed-forward network performs inference in less than a second. To better appreciate the quality of our reconstructions, in Figure 3 we visualize several completions attained with various methods. Note the accurate preservation of intricate details which were completely lost in previous methods.

AMASS projections Using our generated set of partial shapes from AMASS described in 4.1, we compare our method with two recent methods based on shape alignment: 3D-CODED [26], and FARM [47]. As described in 4.2, 3D-CODED is based on a fixed template, and is not trained to handle severe partiality. It thus serves as a lower bound for our proposed method. FARM, on the other hand, was build for the same setting as ours. We summarize the results in Table 2. As can be seen, our method outperforms the two baselines by a large margin in all reported metrics. Note that on some of the examples (about 30%) FARM crashed during the optimization. We therefore only report the errors on its successful runs. Visualizations of several completions are shown in Figure 2.

4.5. Non-rigid partial correspondences

Finding dense correspondences between a full shape and its deformed parts is still a very much open research topic. Here we propose a solution in the form of alignment between the full shape and the partial shape, allowing for the recovery of the the correspondence by a simple nearest neighbour search. As before, we evaluate this task on both FAUST and AMASS data.

FAUST projections On the FAUST projections dataset, we compare with both alignment-based methods, FARM and 3D-CODED, as well as 3 methods designed to directly recover correspondences, i.e. without performing shape completion : MoNet [49], and two 3-layered Euclidean CNN baselines, trained on either SHOT [64] descriptors or depth maps. Results are reported in Figure 4. The test set consists of a total of 200 shapes: 2 subjects at 10 different poses and 10 projected views. The direct matching baselines solve a classification problem for each shape vertex to a template shape. Differently, 3D-CODED has its own

Error	Euclidean distance GT and reconstruction [cm]	Volumetric err. mean $\pm$ std [%]	Directional Chamfer distance GT to reconstruction [cm]	Directional Chamfer distance reconstruction to GT [cm]	Full Chamfer distance [cm]
Poisson [37]	23.73	24.8 ± 23.2	7.3	3.64	10.94
3D-EPN [19]	23.5	89.7 ± 33.8	4.52	4.87	9.39
3D-CODED [25]	35.50	21.8 ± 0.3	11.15	38.49	49.64
FARM [47]	35.77	43.08 ± 20.4	9.5	3.9	13.4
Litany *et al*. [40]	7.07	9.24 ± 8.62	2.84	2.9	5.74
Ours	**2.94**	**7.05 $\pm$ 3.45**	**2.42**	**1.95**	**4.37**

Table 1. **FAUST Shape Completion**. Comparison of different methods with respect to errors in vertex position and shape volume.

Error	Euclidean distance GT and reconstruction [cm]	Volumetric err. mean $\pm$ std [%]	Directional Chamfer distance GT to reconstruction [cm]	Directional Chamfer distance reconstruction to GT [cm]	Full Chamfer distance [cm]
3D-CODED [25]	36.14	**14.84 $\pm$8.02**	13.65	35.35	49
FARM [47]	27.75	49.42 ± 29.12	11.17	5.14	16.31
Ours	**6.58**	27.62 ± 15.27	**4.86**	**3.06**	**7.92**

Table 2. **AMASS Shape Completion**. Comparison of different methods with respect to errors in vertex position and shape volume.

template used at train time. Note that since 3D-CODED was not build for severe partiality, we adjust its loss to a one-sided Chamfer distance in the refinement stage. Our method and FARM both require a complete shape, which we chose as the null pose of each of the test examples. Due to slow convergence and unstable behavior of FARM we only kept 20 useful matching results on which we report the performance. As can be seen from Figure 4, our method outperforms prior art by a significant margin. This result is particularly interesting since it demonstrates that even though we solve an alignment problem, which is a strictly harder problem than correspondence, we receive better results than methods that specialize in the latter. At the same time, looking at the poor performance demonstrated by the other alignment based methods, 3D-CODED and FARM, we conclude that simply solving an alignment problem is not enough and the details of our method and training scheme allow for a substantial difference.

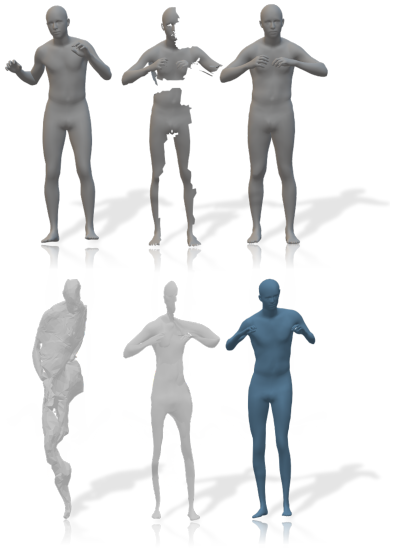


Figure 2. **AMASS Shape Completion**. At the top from left to right: full shape Q, partial shape P, ground truth completion R. At the bottom from left to right: reconstructions of FARM [47], 3D-CODED [25] and ours.

4.6. Real scans

To evaluate our method in real world conditions, we test it on raw measurements taken during the preparation of the Dynamic FAUST [6] dataset. This use case nicely matches our setting. These are partial scans of a subject for which we have a complete reference shape at a different pose. We pre-process the input point cloud by extending it with estimation of its point normals using the method presented in [33]. The point cloud and the reference shape are subsequently inserted into a network pre-trained on FAUST. The template, raw scan, and our reconstruction are shown, from left to right, in Figure 6. We show our result both as the recovered point cloud as well as the recovered mesh using the template triangulation. As apparent from the figure, this is a challenging test case as it introduces several properties not seen at test time: a point cloud without connectivity leads to noisier normals, scanner noise, different point density and extreme partiality (note the missing bottom half of the shapes). Despite all these, the proposed network was able to recover the input quite elegantly, preserving shape details and mimicking the desired pose. In the rightmost column we report a comparison with Litany *et al*. [40]. Note that while [40] was trained on Dynamic FAUST, our network trained on FAUST which is severely constrained in its pose variability. The result highlights that our method favors realism and details in appearance over pose accuracy.

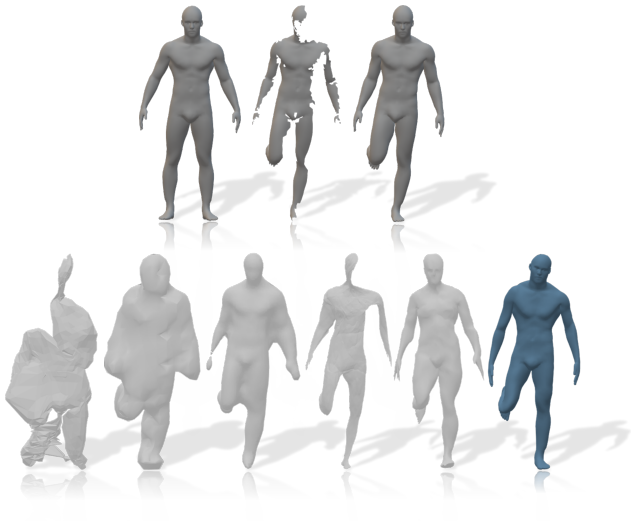


Figure 3. **FAUST Shape Completion**. At the top from left to right: full shape Q, partial shape P, ground truth completion R. At the bottom from left to right: reconstructions from FARM [47], 3D-EPN [19], Poisson [37], 3D-CODED [26], Litany *et al* [40] and ours.

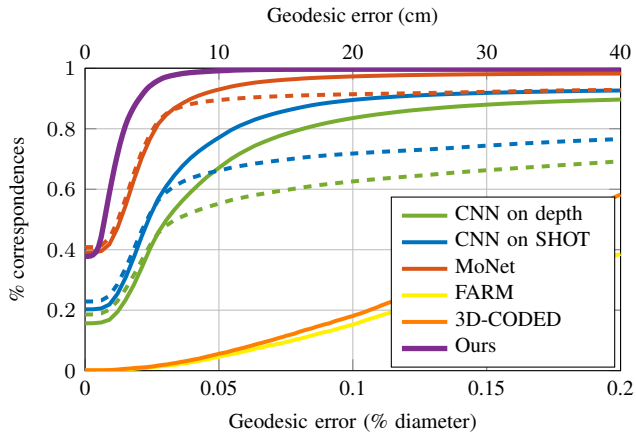


Figure 4. **Generalization error on the FAUST dataset**. Dashed line and solid line in the same color indicate performance before and after refinement, respectively. Note that our method doesn't require refinement, contributing to its computational speed.

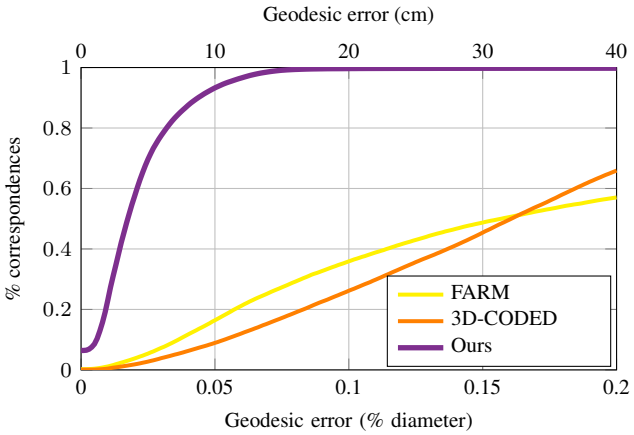


Figure 5. **Generalization error on the Amass dataset**

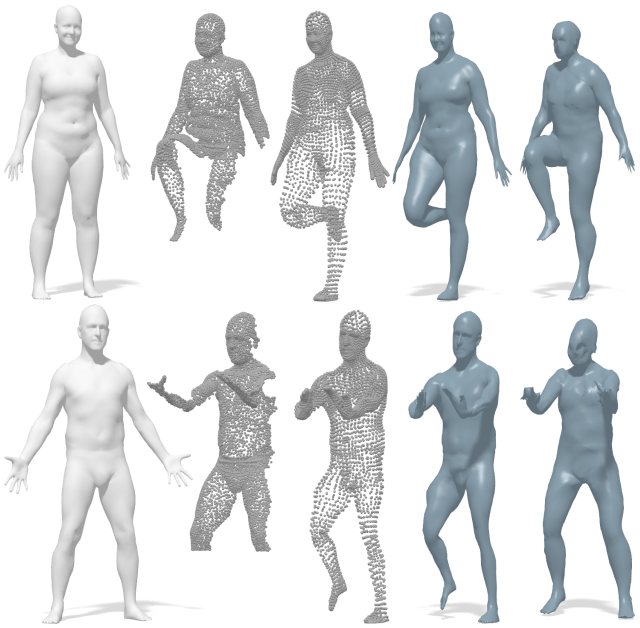


Figure 6. **Completion from real scans from the Dynamic Faust dataset** [6]. From left to right: Input reference shape; input raw scan; our completed shape as a point cloud; and as mesh; completion from Litany *et al*. [40].

5. Concluding remarks

We have demonstrated that the problem of partial matching can be treated in a holistic manner when trying to fit a given part to a whole restricted to the pose inflicted by the part. Our data-driven solution is based on learning the space of distortions linking parts at various poses to whole shapes in other poses. As a result, at test time we are able to match unseen pairs of parts and whole shapes at different poses. In this paper we focused on human shapes. This is mainly due to the availability of rich data sources. That said, our method is not restricted to a particular class of shapes. In the future we plan to explore other types of data. This also implies studying techniques for improved sample efficiency. From Ancient Greek holistic philosophy, through modern psychology explanations of the human brain perception of shapes, we have demonstrated that computational matching procedures could benefit from the same axiomatic assumption stating that indeed *the whole is larger than the sum of its parts*.

Appendix

In this supplementary we provide

1. Analysis of our network; We provide an ablation experiment introspecting the influence of the full shape Q on the network reconstructions. Additionally, we provide robustness analysis of our trained network in Section A.

2. Additional visualizations of the network reconstructions in Section B.

3. Visualizations of the dense correspondence results from the partial shape to the full shape in Section C.

A. Analysis

A.1. Comparison with a fixed template baseline

As described in the main manuscript, in order to predict the completion of a partial shape P, our method requires a full reference shape Q of the same subject in an arbitrary pose. We motivate this setting by a requirement for a completion that is faithful to the subject shape. This is different from previous completion methods which can only approximate or hallucinate missing details.

Here we would like to support this claim experimentally, by comparing with a baseline which uses a fixed template. Specifically, instead of providing a full shape Q of the same subject as the partial shape P, we provide a *fixed* full template T for all inputs. With this modification, the ablation network is trained with the triplets $\{(P_n, T, R_n)\}_{n=1}^{N}$, where N is the size of the training set. At inference time, we use the same template T to make a prediction for a given input part P. We chose the template to be the first subject from the FAUST Projections dataset, in its null pose. Both the original and the fixed-template networks were trained on the FAUST Projections training set, with identical parameters and for the same number of epochs, as described in Section 3.6 in the paper. Table 3 summarizes the prediction errors of both methods, Figure 9 compares the partial correspondence results and Figure 8 shows visual comparison. The results clearly show the benefit of utilizing the shared geometry between the part and a full non-rigid observation of it. In particular, we receive a noticeable improvement in correspondence prediction as well as a lower reconstruction error across all metrics. Perhaps more importantly, Figure 8 demonstrates the main motivation of our framework: a completion that respects the fine details of the underlying shape. To further emphasize this effect, we magnify the face regions of each shape, showing the loss in detail achieved with the alternative training method.

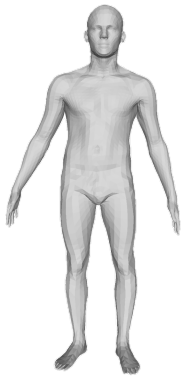


Figure 7. Constant template used in ablation fixed-template experiment

Figure 8 implies how powerful our method is when it comes to the reconstruction of fine details, such as the facial structure and delicate body features. We verify that acquiring access to a full observation in inference time can significantly improve the reliability of the reconstruction for a network trained to utilize such information. In the absence of this full observation at inference time, the ablation network can only utilize the input part and the acquired statistics of the training examples, encoded in the network weights. While this later information can be used for coarse completion, we evidence it is not sufficient for accurate completion.

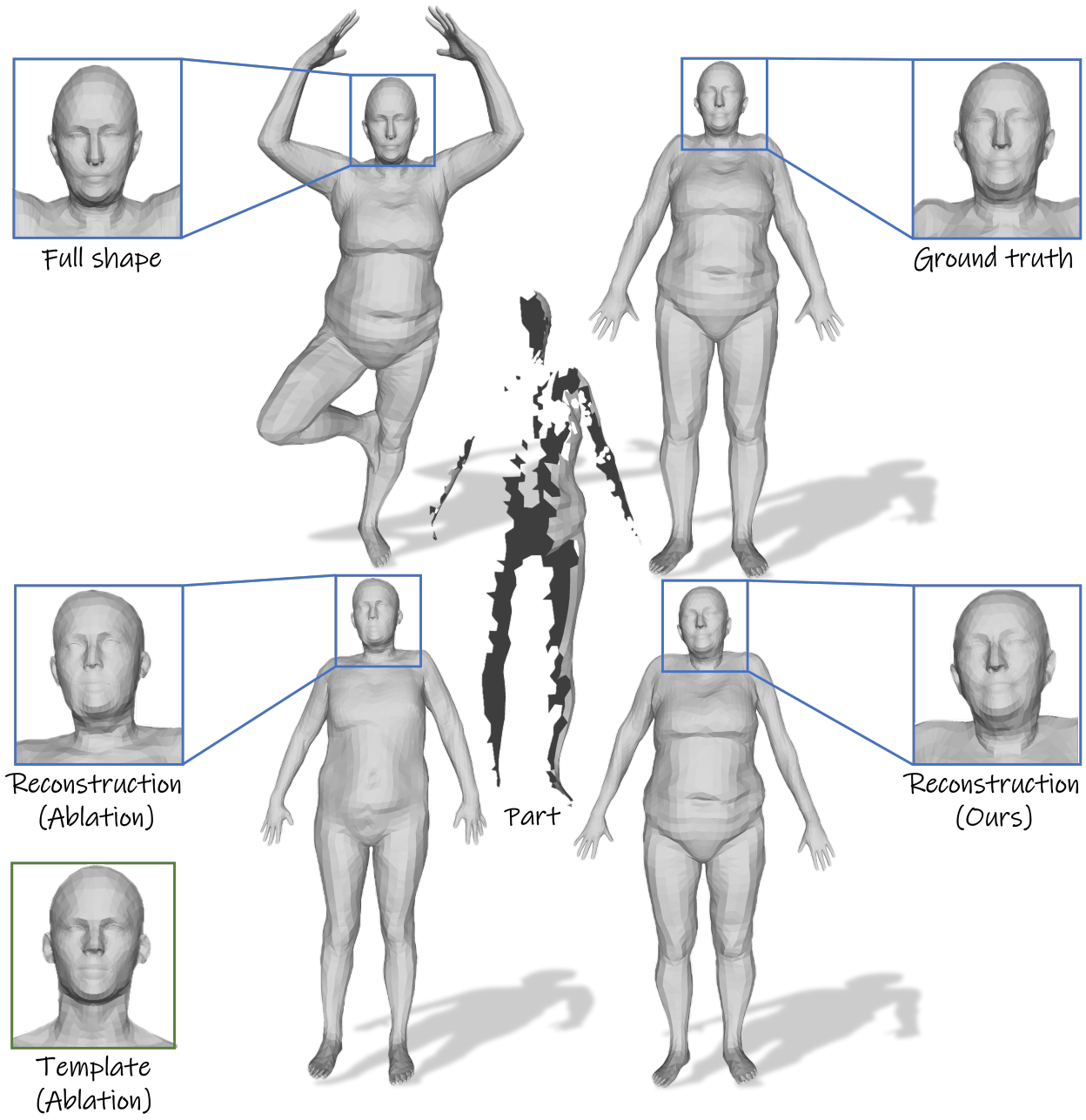


Figure 8. Comparison with fixed-template ablation experiment.

A.2. Robustness Analysis

We turn to analyze the robustness and stability of our proposed method, in hopes of shedding light of its possible applicability in real world conditions. Three specific aspects of the method were inspected empirically, each allowing for a realization of some non-optimal condition commonly found in real scans. The following experiments utilize a network trained over the FAUST train set. The realization is provided over a test-set of 200 single-view projected scans produced from 10 azimutal viewpoints around 2 human subjects exhibiting 10 different poses. The relevant full shapes were taken from the FAUST dataset, and are completely disjoint from our train set, as they contain unique subjects and poses. Each scan P is matched with all possible poses Q of the same subject, achieving a total of 2000 inputs. We utilize a descriptive partial set of the evaluation metrics proposed in section 4.3 of the paper to evaluate each experiment.

Error	Euclidean distance GT and reconstruction [cm]	Volumetric err. mean $\pm$ std [%]	Directional Chamfer distance GT to reconstruction [cm]	Directional Chamfer distance reconstruction to GT [cm]	Full Chamfer distance [cm]
Ablation	3.74	17.63 ± 7.41	3.00	2.32	5.32
Ours	**2.94**	$\mathbf{7.05 \pm 3.45}$	**2.42**	**1.95**	**4.37**

Table 3. **Comparison with Fixed-Template Ablation Experiment.** We evaluate our method against an ablation experiment, repeating exactly the same training except of one significant difference: instead of providing the full shape Q_n as described in the main paper, we provided a *constant* full template T in each of the training examples $\{(P_n, T, R_n)\}_{n=1}^{N}$. The template T is used in inference as well, to predict the completion of a given input part P. We report the prediction errors on FAUST test set, while both networks were trained on FAUST train set. The first and second rows summarize the ablation errors and our method errors, respectively.

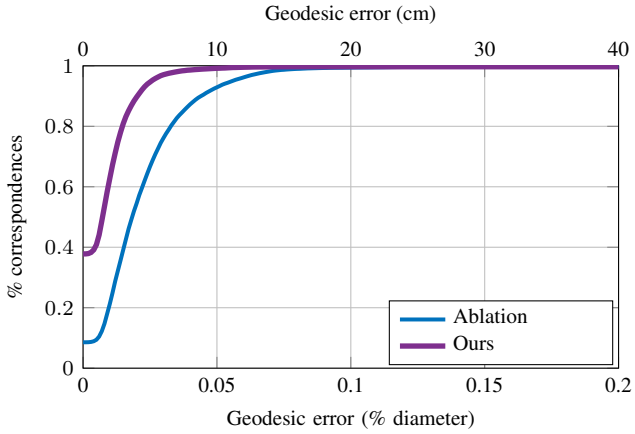


Figure 9. Comparison with fixed-template ablation experiment. Partial correspondence error evaluated on FAUST Projections dataset.

Residual Noise In this experiment, we attempt to emulate various artifacts commonly found in segmented depth scans. We corrupt the vertices of each partial input shape with various degrees of additive white Gaussian noise, with standard deviations in the range [0-4] cm. The corrupted partial shapes are fed to the network, together with the full shapes. Averaged reconstruction statistics are displayed graphically in Figure 10. As apparent from the figure, the method accuracy only slightly declines with the increase of the noise.

Downsampling We address the network's ability to infer on partial shapes with decreasing degrees of resolution. For each partial shape in the mentioned test set, we decimate at random some percentage of the existing vertices, and infer on the resultant set. As can be seen in Figure 11, even under a majority decimation of the vertices, the proposed network is able to recover well the ground truth shape.

Projection Angle Finally, we examine the dependency of our network to the projection angle. We note that due to the different projection angles and poses, it is not unreasonable that some angles hold a higher degree of information

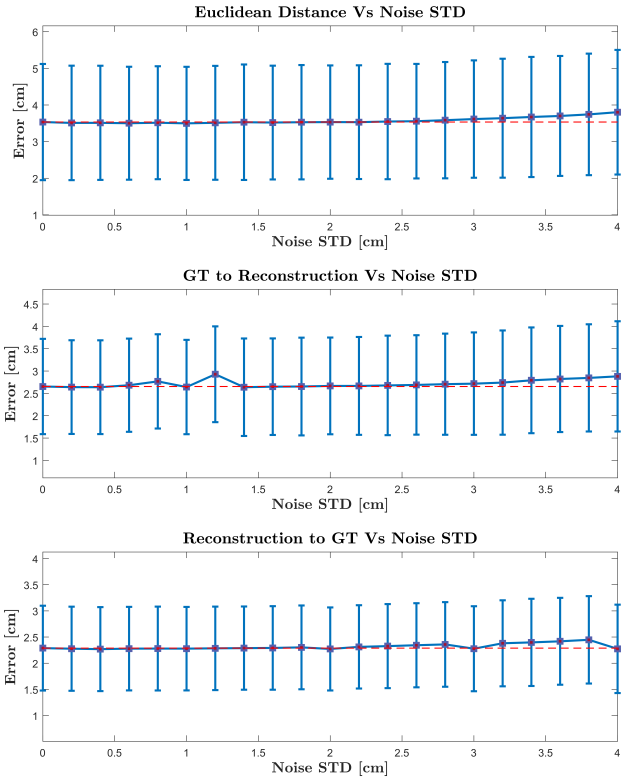


Figure 10. **Robustness to Noise.** Three reconstruction metrics evaluated on completions originating from corrupted partial shapes with varying levels of additive white Gaussian noise. A baseline with the evaluation realized with no noise is marked with a dashed red line.

relevant for reconstruction than others. Ideally, we would like to enable the network a reliable reconstruction at every angle, regardless if the information seen is the back, front or sides of a shape. We partition the 2000 completions received over the test set into their corresponding projection angles, and accumulate the errors over each partition. The result is displayed in Figure 12. The received error distribution is close to uniform, attributing to the method's azimutal invariancy.

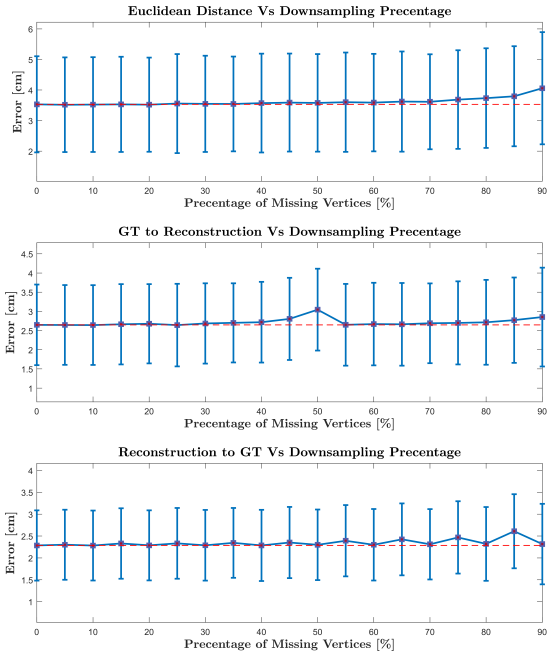


Figure 11. **Robustness to Downsampling.** Three reconstruction metrics evaluated on completions originating from decimated partial shapes with varying levels of vertex erasure. A baseline with the evaluation realized with no decimation is marked with a dashed red line.

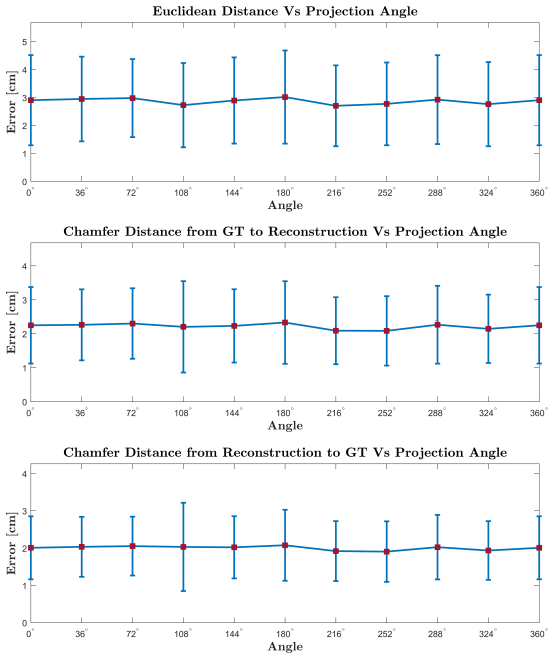


Figure 12. **Robustness to Projection Angle.** Three reconstruction metrics evaluated on different groups of the test-set, partitioned by the projection angle. We note a close to uniform distribution over the different angles, attributing to a azimutal invariancy.

B. Additional Visualizations

Here we provide additional reconstructions that were not included in the main paper in order to save space. Figure 13 and Figure 14 visualize our network predictions for examples from **FAUST Projections** and **AMASS Projections**, respectively.

C. Non-Rigid partial correspondence

Figure 15 visualizes the dense correspondence between the input partial and full shape. As explained in the paper, we achieve this by using the network reconstruction as a proxy; For every point in the partial shape we calculate the nearest neighbor point in the reconstruction allowing us a recovery of a mapping between the partial shape to the reconstructed shape, which is by construction also the mapping between the part and the full input shape. In **Section 4.5** of the paper we evaluated the predicted correspondence numerically for **FAUST Projections** and **AMASS Projections** datasets, providing geodesic error graphs for both, in **Figure 4** and **Figure 5**, respectively. For completion, we show the results also qualitatively here.

References

[1] Brett Allen, Brian Curless, Zoran Popović, and Aaron Hertzmann. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 147–156. Eurographics Association, 2006. 2

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM transactions on graphics (TOG)*, volume 24, pages 408–416. ACM, 2005. 2

[3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992. 5

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proc. Computer Graphics and Interactive Techniques*, pages 187–194, 1999. 2

[5] Federica Bogo, Javier Romero, Matthew Loper, and Michael J Black. FAUST: Dataset and Evaluation for 3d Mesh Registration. In *Proc. CVPR*, 2014. 5

[6] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 7, 8

[7] Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 3189–3197, 2016. 3

[8] A. M. Bronstein, M. M. Bronstein, A.M. Bruckstein, and R. Kimmel. Matching two-dimensional articulated shapes us-

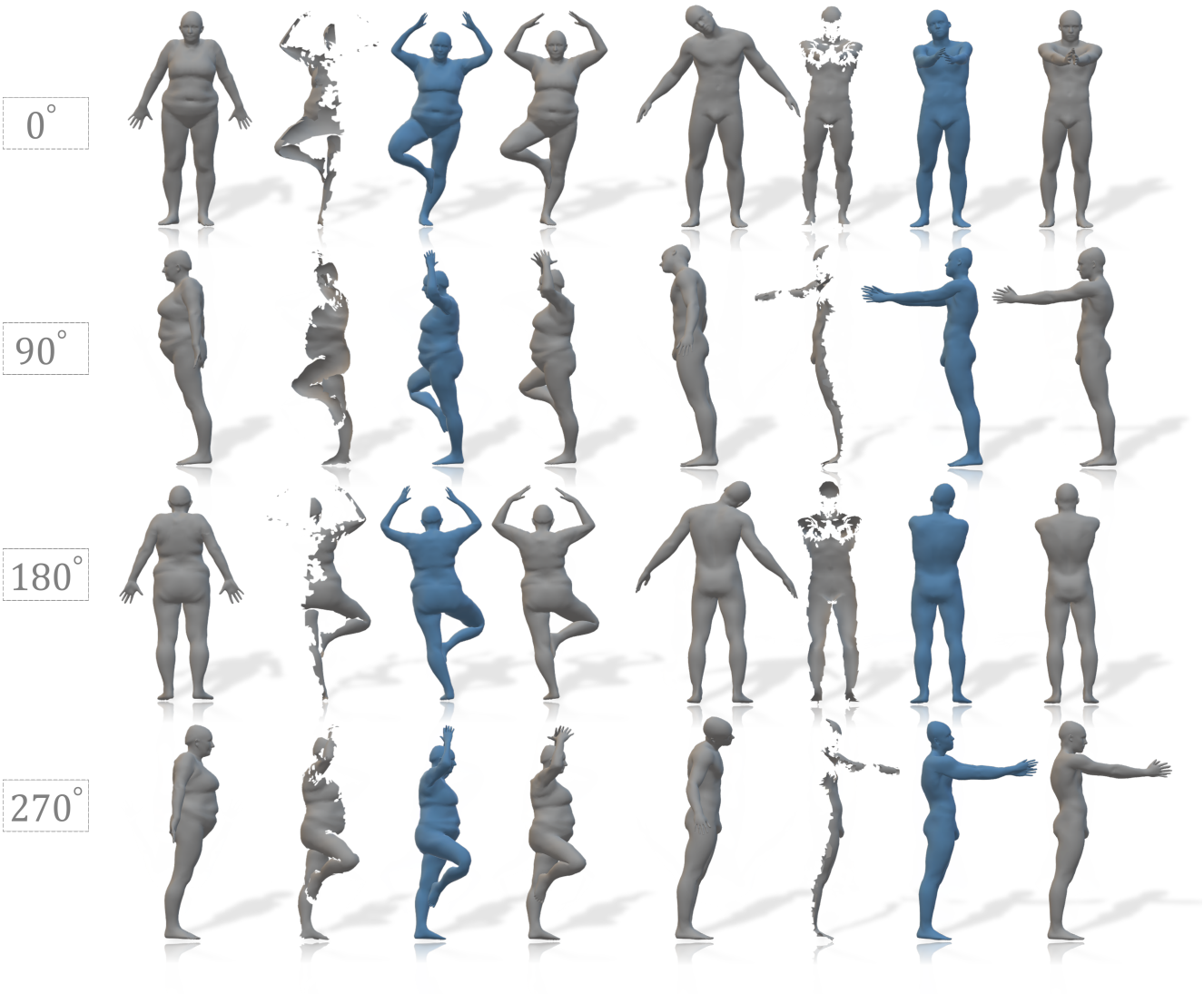


Figure 13. **Predicted completions, FAUST Projections.** Each column shows a completion for a different subject, while each row provides a different perspective on the reconstructed 3D model. From left to right: full input shape Q, input part P, predicted completion $F_{\theta(P,Q)}(Q)$, ground truth completion R.

ing generalized multidimensional scaling. In *Proc. of Articulated Motion and Deformable Objects (AMDO)*, 2006. 2

[9] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Expression-invariant 3d face recognition. In *Proc. Audio & Video-based Biometric Person Authentication (AVBPA), Lecture Notes in Comp. Science 2688, Springer*, 2003. 2

[10] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Three-dimensional face recognition. *International Journal of Computer Vision*, 64(1):5–30, 2005. 2

[11] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Face2face: an isometric model for facial animation. In *Conf. on Articulated Motion and Deformable Objects (AMDO)*, 2006. 2

[12] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *PNAS*, 103(5):1168–1172, 2006. 3

[13] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Robust expression-invariant face recognition from partially missing data. In *Proc. ECCV, Graz, Austria*, May 2006. 2

[14] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Expression-invariant representations of faces. *IEEE Trans. Image Processing*, 16(1):188–197, 2007. 2

[15] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. 3

[16] Qifeng Chen and Vladlen Koltun. Robust nonrigid registration by convex optimization. In *Proc. ICCV*, 2015. 3

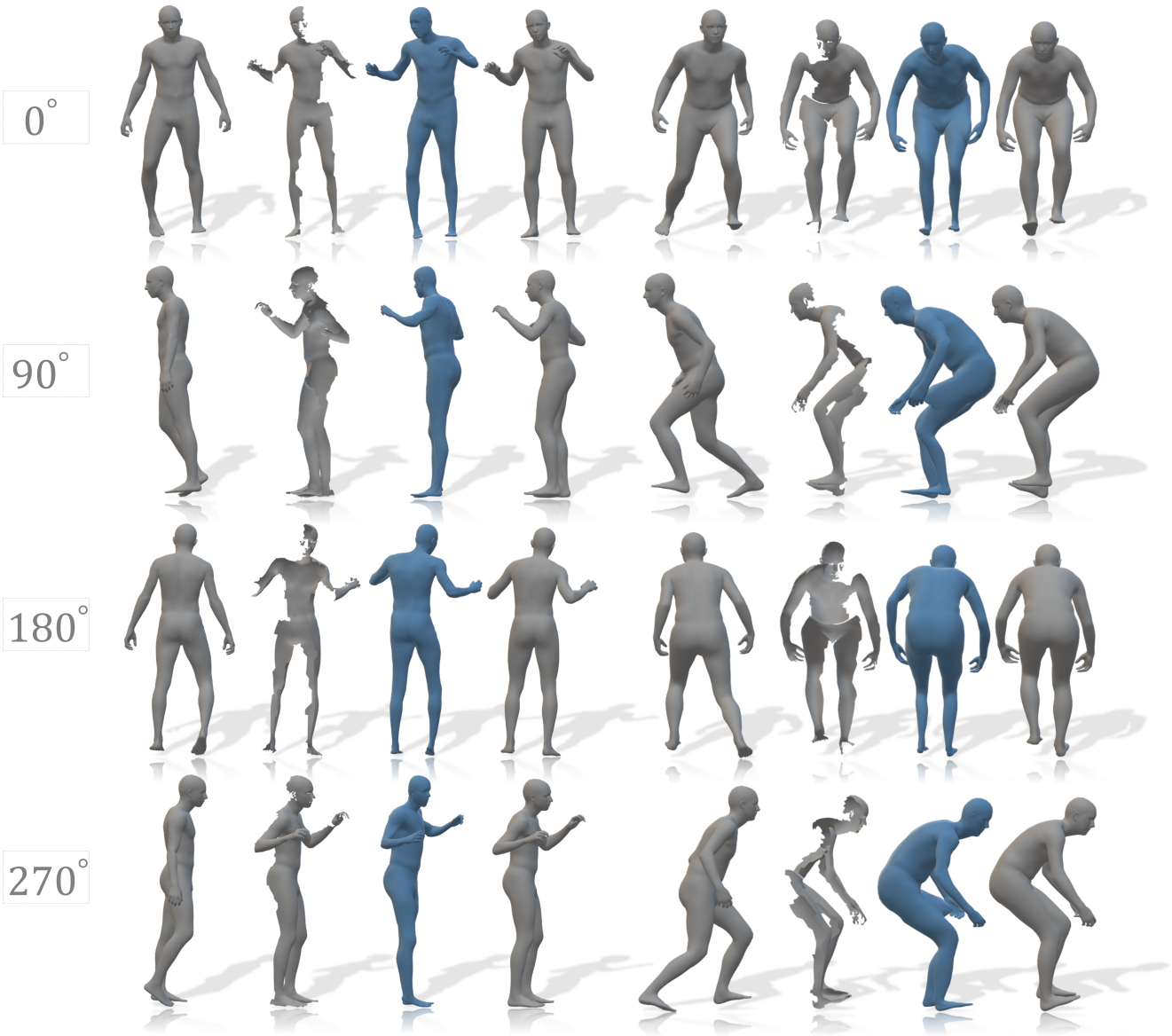


Figure 14. **Predicted completions, AMASS Projections.** Each column shows a completion for a different subject, while each row provides a different perspective on the reconstructed 3D model. From left to right: full input shape Q, input part P, predicted completion $F_{\theta(P,Q)}(Q)$, ground truth completion R.

[17] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. *arXiv preprint arXiv:1904.08755*, 2019. 3

[18] Luca Cosmo, Mikhail Panine, Arianna Rampini, Maks Ovsjanikov, Michael M Bronstein, and Emanuele Rodolà. Isospectralization, or how to hear shape, style, and correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7529–7538, 2019. 3

[19] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3D-encoder-predictor cnns and shape synthesis. *arXiv:1612.00101*, 2016. 6, 7, 8

[20] Y. Devir, G. Rosman, M. M. Bronstein, A. M. Bronstein, and R. Kimmel. On reconstruction of non-rigid shapes with intrinsic regularization. In *Proc. of Workshop on Nonrigid Shape Analysis and Deformable Image Alignment (NORDIA)*, 2009. 2

[21] Asi Elad and Ron Kimmel. Bending invariant representations for surfaces. In *Proc. of CVPR'01, Hawaii*, December 2001. 2

[22] Asi Elad and Ron Kimmel. On bending invariant signatures for surfaces. *IEEE Trans. on Pattern Analysis and Machine*

Figure 15. **Non-Rigid partial correspondence.** Left and right columns show the dense correspondence for FAUST Projections and AMASS Projections, respectively. From left to right: full input shape Q, our network completion $F_{\theta(P,Q)}(Q)$ and partial input shape P. Corresponding points are indicated by the same color.

Intelligence (PAMI), 25(10):1285–1295, 2003. 2

[23] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Lüthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. *arXiv preprint arXiv:1709.08398*, 2017. 2

[24] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9224–9232, 2018. 3

[25] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. 3D-CODED : 3D correspondences by deep deformation. In *ECCV*, 2018. 3, 6, 7

[26] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 3D-coded: 3D correspondences by deep deformation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 230–246, 2018. 3, 4, 6, 8

[27] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. Atlasnet: A papier-Mâchè approach to learning 3D surface generation. *arXiv preprint arXiv:1802.05384*, 2018. 3, 4

[28] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 2

[29] Oshri Halimi and Ron Kimmel. Self functional maps. In *2018 International Conference on 3D Vision (3DV)*, pages 710–718. IEEE, 2018. 2

[30] Oshri Halimi, Or Litany, Emanuele Rodola, Alex M Bronstein, and Ron Kimmel. Unsupervised learning of dense shape correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4370–4379, 2019. 3

[31] Oshri Halimi, Dan Raviv, Yonathan Aflalo, and Ron Kimmel. Computable invariants for curves and surfaces. *Processing, Analyzing and Learning of Images, Shapes, and Forms*, 20:273, 2019. 2

[32] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: A network with

an edge. *ACM Transactions on Graphics (TOG)*, 38(4):90, 2019. 3

[33] Hugues Hoppe, Tony DeRose, Tom Duchamp, John McDonald, and Werner Stuetzle. *Surface reconstruction from unorganized points*, volume 26. ACM, 1992. 7

[34] Jingwei Huang, Yichao Zhou, Thomas Funkhouser, and Leonidas Guibas. Framenet: Learning local canonical frames of 3d surfaces from a single rgb image. *arXiv preprint arXiv:1903.12305*, 2019. 5

[35] Haiyong Jiang, Jianfei Cai, and Jianmin Zheng. Skeleton-aware 3d human shape reconstruction from point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5431–5441, 2019. 2

[36] Mor Joseph-Rivlin, Alon Zvirin, and Ron Kimmel. Momen^{e}t: Flavor the moments in learning to classify shapes. In *Proc. of IEEE Int. Conference on Computer Vision (CVPR) Workshops*, 2019. 3

[37] Michael Kazhdan and Hugues Hoppe. Screened Poisson surface reconstruction. *TOG*, 32(3):29, 2013. 2, 6, 7, 8

[38] Vladimir G. Kim, Yaron Lipman, and Thomas A. Funkhouser. Blended intrinsic maps. *Trans. Graphics*, 30(4), 2011. 3

[39] Simon Korman, Eyal Ofek, and Shai Avidan. Peeking template matching for depth extension. In *Proc. CVPR*, 2015. 2

[40] Or Litany, Alex Bronstein, Michael Bronstein, and Ameesh Makadia. Deformable shape completion with graph convolutional autoencoders. *CVPR*, 2018. 6, 7, 8

[41] Or Litany, Tal Remez, and Alex Bronstein. Cloud dictionary: Sparse coding and modeling for point clouds. *arXiv:1612.04956*, 2016. 2

[42] Or Litany, Tal Remez, Emanuele Rodolà, Alex M Bronstein, and Michael M Bronstein. Deep functional maps: Structured prediction for dense shape correspondence. In *Proc. ICCV*, volume 2, page 8, 2017. 3

[43] Or Litany, Emanuele Rodolà, Alex M Bronstein, and Michael M Bronstein. Fully spectral partial shape matching. *Computer Graphics Forum*, 36(2):247–258, 2017. 3

[44] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. Meteornet: Deep learning on dynamic 3d point cloud sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9246–9255, 2019. 3

[45] 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):248, 2015. 2, 5, 6

[46] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. 5

[47] Riccardo Marin, Simone Melzi, Emanuele Rodolà, and Umberto Castellani. Farm: Functional automatic registration method for 3d human bodies. In *Computer Graphics Forum*. Wiley Online Library, 2018. 6, 7, 8

[48] Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 37–45, 2015. 3

[49] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodolà, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5425–5434. IEEE, 2017. 3, 6

[50] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *arXiv preprint arXiv:1901.05103*, 2019. 3

[51] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[52] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. *arXiv preprint arXiv:1904.09664*, 2019. 3

[53] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proc. CVPR*, 2017. 3

[54] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv:1706.02413*, 2017. 3

[55] Arianna Rampini, Irene Tallini, Maks Ovsjanikov, Alex M Bronstein, and Emanuele Rodolà. Correspondence-free region localization for partial shape similarity via hamiltonian spectrum alignment. *arXiv preprint arXiv:1906.06226*, 2019. 3

[56] Elad Richardson, Matan Sela, Roy Or-El, and Kimmel Ron. Learning detailed face reconstruction from a single image. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Hawaii, Honolulu*, 2017. 2

[57] Elad Richardson, Matan Sela, and Kimmel Ron. 3D face reconstruction by learning from synthetic data. In *4th Int. Conf. on 3D Vision (3DV) Stanford University, CA, USA*, 2016. 2

[58] Emanuele Rodolà, Luca Cosmo, Michael M Bronstein, Andrea Torsello, and Daniel Cremers. Partial functional correspondence. In *Computer Graphics Forum*, volume 36, pages 222–236. Wiley Online Library, 2017. 3

[59] Emanuele Rodolà, Samuel Rota Bulo, Thomas Windheuser, Matthias Vestner, and Daniel Cremers. Dense non-rigid shape correspondence using random forests. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4184, 2014. 3

[60] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017. 5

[61] Kripasindhu Sarkar, Kiran Varanasi, and Didier Stricker. Learning quadrangulated patches for 3D shape parameterization and completion. *arXiv:1709.06868*, 2017. 2

[62] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Int. Conf. Comp. Vision (ICCV), Venice, Italy*, 2017. 2

[63] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proc. CVPR*, 2015. 3

[64] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *International Conference on Computer Vision (ICCV)*, pages 356–369, 2010. 6

[65] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018. 2

[66] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017. 2

[67] Nitika Verma, Edmond Boyer, and Jakob Verbeek. Dynamic filters in graph convolutional networks. *arXiv:1706.05206*, 2017. 3

[68] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):146, 2019. 3

[69] Lingyu Wei, Qixing Huang, Duygu Ceylan, Etienne Vouga, and Hao Li. Dense human body correspondences using convolutional networks. In *Proc. CVPR*, 2016. 3

[70] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D shapenets: A deep representation for volumetric shapes. In *Proc. CVPR*, 2015. 3

[71] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2018. 3

[72] Y Y Ben-Shabat, M Lindenbaum, and Fischer A. 3D point cloud classification and segmentation using 3D modified fisher vector representation for convolutional neural networks. *arXiv preprint arXiv:1711.08241*, 2017. 3

[73] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018. 2