Unsupervised Learning of Dense Shape Correspondence

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

Or Litany
Facebook AI Research
orlitany@fb.com

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

Alex Bronstein
Technion, Israel
bron@cs.technion.ac.il

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

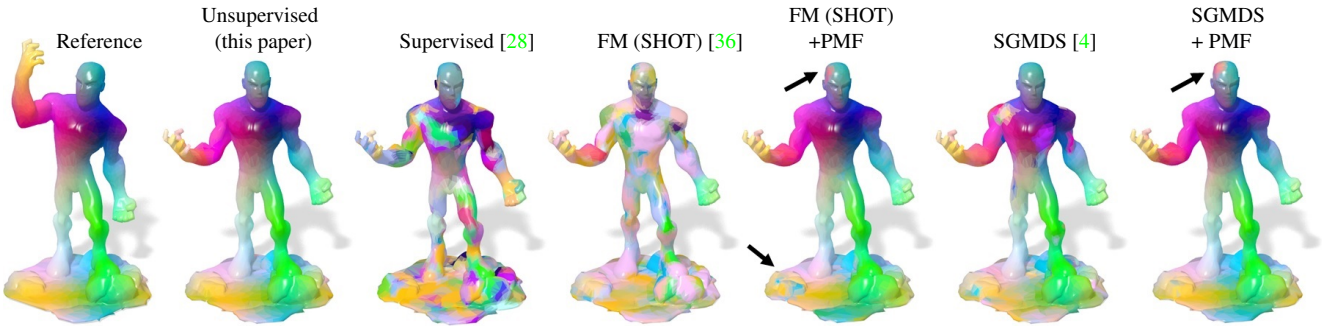

Figure 1: Dense correspondence between articulated objects obtained with the proposed unsupervised loss, optimized on a single (unlabeled) example. Our method is compared with the state-of-the-art *supervised* network pre-trained on human shapes, as well as with two axiomatic methods, employing a post processing algorithm [49] on the axiomatic results. See Section 5.1 for more details. Correspondence is visualized by colors mapped from the leftmost *reference* shape.

Abstract

We introduce the first completely unsupervised correspondence learning approach for deformable 3D shapes. Key to our model is the understanding that natural deformations, such as changes in pose, approximately preserve the metric structure of the surface, yielding a natural criterion to drive the learning process toward distortion-minimizing predictions. On this basis, we overcome the need for annotated data and replace it by a purely geometric criterion. The resulting learning model is class-agnostic, and is able to leverage any type of deformable geometric data for the training phase. In contrast to existing supervised approaches which specialize on the class seen at training time, we demonstrate stronger generalization as well as applicability to a variety of challenging settings. We showcase our method on a wide selection of correspondence benchmarks, where the proposed method outperforms other methods in terms of accuracy, generalization, and efficiency.

1. Introduction

The problem of finding accurate dense correspondence between non-rigid shapes is fundamental in geometry processing. It is a key component in applications such as deformation modeling, cross-shape texture mapping, pose and animation transfer to name just a few. Dense deformable shape correspondence algorithms can be broadly categorized into two families. The first can be referred to as *axiomatic* or *model-based* for which a certain geometric assumption is asserted and pursuit for by some numerical scheme. Modeling assumptions attempt to characterize the action of a class of deformations on some geometric quantities commonly referred to as *descriptors*. Such geometric quantities often encode local geometric information in the vicinity of a point on the shape (point-wise descriptors) such as normal orientation [47], curvature [37], and heat [45] or wave [6] propagation properties. Another type of geometric quantities are the global relations between pairs of points (pair-wise descriptors), which include geodesic [21, 14, 4], diffusion [17, 12] or commute time [50] distances. Given a pair of shapes, a dense map between them is sought to minimize the discrepancy between such descriptors. While the minimization of the point-wise discrepancies can be formulated as a linear assignment problem (LAP) and solved efficiently for reasonable scales, the use of pair-wise descriptors leads to a quadratic assignment problem (QAP) that is

unsolvable for any practical scales. Numerous approximations and heuristics have been developed in the literature to alleviate the computational demand of QAPs.

The second family of correspondence algorithms is *data-driven* and takes advantage of modern efficient machine learning tools. Instead of axiomatically modeling the class of deformations and the geometric properties of the shapes of interest, these methods infer such properties from the data themselves. Among such approaches are learnable generalizations of the heat kernel signature [31], as well as models interpreting correspondence as a labeling problem [41]. Other recent methods generalize CNNs to non-Euclidean structures for learning improved descriptors [35, 10]. A recent method based on extrinsic deformation of a null-shape was introduced in [23]. A common denominator of these approaches is the *supervised* training regime – they all rely on examples of ground truth correspondences between exemplar shapes.

A major drawback of this supervised setting is the fact that in the case of 3D shape correspondence the ground truth data are scarce and expensive to obtain. For example, despite being restricted to a single shape class (human bodies), the MPI FAUST scanning and labeling system [8] required substantial manual labor and considerable financial costs. In practice, labeled models are expected to be just a small fraction of the existing geometric data, bringing into question the scalability of any supervised learning algorithm.

1.1. Contribution

We propose an unsupervised learning scheme for dense 3D deformable shape correspondence based on a purely geometric criterion. The suggested approach bridges between the model-based and the data-driven worlds by learning point-wise descriptors that result in correspondences minimizing pair-wise geodesic distance distortion [21, 14]. The unsupervised loss is intimately related to the *spectral generalized multidimensional scaling* [4] model. The correspondence is solved by functional maps framework [36], totally avoiding the computational burden of the pair-wise methods. The point-wise descriptors are learned on a surrogate task, only approximately characterizing the real data, which deviate from the asserted isometric deformation model. Still, the method shows excellent generalization capabilities exceeding the supervised counterparts without ever seeing examples of ground truth correspondences. To the best of our knowledge, this is the first unsupervised approach applied to the geometric shape correspondence problem.

A major advantage of the proposed framework is when the data themselves are scarce, in extreme conditions we might have only one pair of shapes that we would like to match and we do not have a training dataset that contains similar shapes. While a supervised scheme depends on a relatively large amount of labelled data to deduce a generalizing model, with the unsupervised network we can simply optimize on a single pair of shapes that by itself contains two training samples, one in each direction of the correspondence. Our experiments required only a few iterations that take just a couple of minutes to run. As a result we obtain an accurate matching between the shapes, see Figure 1. For a trained network the inference phase takes less than a second. We believe that this strategy has its own merits as a replacement of the existing computationally expensive methods that are based on pair-wise descriptors. The framework can be interpreted as a fusion between the previously proposed generalized multidimensional scaling (GMDS) [4, 14] and the FMNet network architecture [28]. Here we use the discretization of the pair-wise geodesic distance distortion, as suggested in [4] and justified in [34].

2. Background

2.1. Minimum distortion correspondence

We model shapes as Riemannian 2-manifolds $\mathcal{X}$ equipped with a distance function $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ induced by the standard volume form $\mathrm{d}x$. An *isometry* is a map $\pi : \mathcal{X} \to \mathcal{Y}$ satisfying, for any pair $x_1, x_2 \in \mathcal{X}$:

$$d_{\mathcal{X}}(x_1, x_2) = d_{\mathcal{Y}}(\pi(x_1), \pi(x_2)) \,. \tag{1}$$

Correspondence seeking approaches optimize for a map π satisfying the distance preservation criterion (1). In practical applications, only approximate realizations of an isometry are expected; thus, one is interested in finding a distortion-minimizing map of the form

$$\pi^* = \operatorname*{arg\,min}_{\pi:\mathcal{X}\to\mathcal{Y}} \iint_{x_1,x_2\in\mathcal{X}} \left(d_{\mathcal{X}}(x_1, x_2) - d_{\mathcal{Y}}(\pi(x_1), \pi(x_2))\right)^2 \mathrm{d}a_{x_1} \mathrm{d}a_{x_2} \tag{2}$$

In the discrete setting, we assume manifolds $\mathcal{X}, \mathcal{Y}$ to be represented as triangle meshes sampled at n vertices each. Minimum distortion correspondence takes the form of a quadratic assignment problem (QAP), where the minimum is sought over the space of $n \times n$ permutation matrices. Several studies have tried to reduce the complexity of this QAP at the cost of getting an approximate solution via subsampling [46, 39], hierarchical matching [14, 51, 38, 20] or convex relaxations [3, 16]. However complicated to solve, the minimum distortion criterion (2) is axiomatic and does not require any annotated correspondences, making it a natural candidate for an unsupervised learning loss.

2.2. Descriptor learning

A common way to make the optimization of (2) more efficient is by restricting the feasible set to include only potential matches among points with similar descriptors. By doing so, one shifts the key difficulty from optimizing a highly

non-linear objective to designing deformation-invariant local point descriptors. This has been an active research goal in shape analysis in the last few years, with examples including GPS [42], heat and wave kernel signatures [45, 6], and the more recent geodesic distance descriptors [44]. In 3D vision, several *rotation*-invariant geometric descriptors have been proposed [47, 25]. Despite their lack of invariance to isometric deformations, the adoption of extrinsic descriptors has been advocated in deformable settings [40] due to their locality and resilience to boundary effects.

Handcrafted descriptors suffer from an inherent drawback of requiring manual tuning. Learning techniques have thus been proposed to define descriptors whose invariance classes are learned from the data. Early examples include approaches based on decision forests and metric learning [31, 41, 19]; more recently, several papers have proposed an adaptation of deep learning models to non-Euclidean domains, achieving dramatic improvement. In [33, 10, 35] learnable local filters were introduced based on the notion of patch operator. In [28] a task driven approach was taken instead, where the network learns descriptors which excel at the task at hand in a supervised manner. As we will show in the sequel, our approach replaces the penalty of the latter model with the one optimized for in the SGMDS model [4], completely removing the need for supervision.

2.3. Functional maps

The notion of *functional map* was introduced in [36] as a tool for transferring functions between surfaces without the direct manipulation of a point-to-point correspondence. Let $\mathcal{F}(\mathcal{X}), \mathcal{F}(\mathcal{Y})$ be real-valued functional spaces defined on top of $\mathcal{X}$ and $\mathcal{Y}$ respectively. Then, given a bijection $\pi : \mathcal{X} \to \mathcal{Y}$, the functional map $T : \mathcal{F}(\mathcal{X}) \to \mathcal{F}(\mathcal{Y})$ is a linear mapping acting as

$$T(f) = f \circ \pi^{-1} . \tag{3}$$

The functional map T admits a matrix representation with respect to orthogonal bases $\{\phi_i\}_{i\geq 1}, \{\psi_i\}_{j\geq 1}$ on $\mathcal{X}$ and $\mathcal{Y}$ respectively, with coefficients $\mathbf{C} = (c_{ij})$ calculated by

$$T(f) = \sum_{ij} \langle \phi_i, f \rangle \underbrace{\langle T\phi_i, \psi_j \rangle}_{c_{ji}} \psi_j . \tag{4}$$

While the functional maps formalism makes no further requirements on the chosen bases, a typical choice is the Laplace-Beltrami eigenbasis, where the justification for the optimality of this choice can be found in [2]. Truncating these series to k coefficients, one obtains a band-limited approximation of the functional correspondence T. Specifically, the map

$$P : x \mapsto \sum_{i,j} c_{ji} \phi_i(x) \psi_j , \tag{5}$$

also referred to as a *soft map*, will assign to each point $x \in \mathcal{X}$ a function concentrated around $y = \pi(x)$ with some spread.

To solve for the matrix $\mathbf{C}$, linear constrains are derived from the knowledge of knowingly corresponding functions on the two surfaces. Corresponding functions are functions that preserve their value under the mapping T. Given a pair of corresponding functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ with coefficients $\hat{\mathbf{f}} = \{\langle \phi_i, f \rangle\}_i$ and $\hat{\mathbf{g}} = \{\langle \psi_j, g \rangle\}_j$ in the bases $\{\phi_i\}$ and $\{\psi_j\}$ respectively, the correspondence imposes the following linear constraint on $\mathbf{C}$

$$\hat{\mathbf{g}} = \mathbf{C}\hat{\mathbf{f}}. \tag{6}$$

Each pair of such corresponding functions is translated into a linear constraint.

Suppose there exists an operator receiving a shape $\mathcal{X}$ and producing a set of *descriptor functions* on it. Let us further assume that given another shape $\mathcal{Y}$, the operator will produce a set of corresponding functions related by the latent correspondence between $\mathcal{X}$ and $\mathcal{Y}$. In other words, applying the above operator on the said pair of shapes produces a set of pairs of corresponding functions (f_i, g_i), each pair comprising f_i defined on $\mathcal{X}$ and g_i on $\mathcal{Y}$. We stack the corresponding coefficients $\hat{\mathbf{f}}_i$ and $\hat{\mathbf{g}}_i$ into the columns of the matrices $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$. The functional map matrix $\mathbf{C}$ is then given by the (least squares, or otherwise regularized) solution to the system

$$\hat{\mathbf{G}} = \mathbf{C}\hat{\mathbf{F}}. \tag{7}$$

Thus, the requirement for specific knowledge of the point-to-point correspondence is replaced by the relaxed requirement of knowledge about functional correspondence.

2.4. Deep functional maps

A significant caveat in the above setting is that, unless the shapes $\mathcal{X}$ and $\mathcal{Y}$ are related by a narrow class of deformations, it is very difficult to construct an operator producing a sufficient quantity of stable and repeatable descriptors. However, such an operator can be *learned* from examples. The aim of the deep functional maps network (FMNet) introduced in [28] was to learn descriptors which, when used in the above system of equations, will induce an accurate correspondence. At training time, FMNet operates on input descriptor functions (e.g. SHOT descriptors), and improves upon them by minimizing a geometric loss that is defined on the soft correspondence derived from the functional map matrix. The differentiable functional map layer (FM), solves the equation (7), with the current descriptor functions in each iteration.

The network architecture described in [28] consists of 7 fully-connected residual layers with exponential linear units (ELU) and no dimentionality reduction. The output of the residual network is a dense vector-valued descriptor. Given

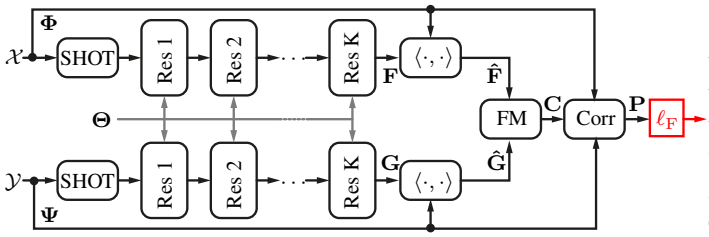

Figure 2: Deep Functional Maps network architecture [28]

two shapes $\mathcal{X}$ and $\mathcal{Y}$, the descriptors are calculated on each shape using the same network, and are projected onto the corresponding truncated LBO bases. The resulting coefficients are given as an input to the functional map (FM) layer that calculates the functional map matrix $\mathbf{C} \in \mathbb{R}^{k\times k}$ according to (7). The following correspondence layer (Corr) produces a soft correspondence matrix $\mathbf{P} \in \mathbb{R}^{n_\mathcal{Y}\times n_\mathcal{X}}$ out of the functional map matrix $\mathbf{C}$,

$$\mathbf{P} = |\mathbf{\Psi C \Phi}^T \mathbf{A}|_{\|\cdot\|}. \tag{8}$$

Where we denoted the number of vertices on the discretized shapes as $n_\mathcal{X}$ and $n_\mathcal{Y}$, and the diagonal matrix $\mathbf{A}$ normalizes the inner products with the discrete area elements of $\mathcal{X}$. The absolute value and the L_2 column normalization, denoted by $_{\|\cdot\|}$, ensure that the values of p_{ji}^2 can be interpreted as the probability of vertex j on shape $\mathcal{Y}$ being in correspondence with vertex i on $\mathcal{X}$. We denote the element-wise square of $\mathbf{P}$ by $\mathbf{Q} = \mathbf{P}\circ\mathbf{P}$, with $\circ$ standing for the Hadamard product.

Treating the i-th column of $\mathbf{Q}$, $\mathbf{q_i}$, as the distribution on the points of $\mathcal{Y}$ corresponding to the point i on $\mathcal{X}$, we can evaluate the expected deviation from the ground truth correspondence $\pi^*(i)$. This is expressed by the second-order moment

$$\mathbb{E}_{j\sim\mathbf{q_i}} d^2_\mathcal{Y}(j,\pi^*(i)) = \sum_{j\in\mathcal{Y}} q_{ji} d^2_\mathcal{Y}(j,\pi^*(i)). \tag{9}$$

where $d_\mathcal{Y}(j,\pi^*(i))$ is the geodesic distance on $\mathcal{Y}$ between the vertex j and the ground truth match $\pi^*(i)$ of the vertex i on $\mathcal{X}$. As usual, this moment comprises a variance and a bias terms; while the former is the result of the band-limited approximation (due to the truncation of the basis), the latter can be controlled. Averaging the above moment over all points on $\mathcal{X}$ leads to the following supervised loss

$$\begin{aligned}\ell_{\text{sup}}(\mathcal{X},\mathcal{Y}) &= \frac{1}{|\mathcal{X}|}\sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{Y}} q_{ji} d^2_\mathcal{Y}(j,\pi^*(i)) \\ &= \frac{1}{|\mathcal{X}|}\left\|\mathbf{P}\circ(\mathbf{D}_\mathcal{Y}\mathbf{\Pi}^*)\right\|^2_{\text{F}},\end{aligned} \tag{10}$$

where $\mathbf{D}_\mathcal{Y}$ denotes the pairwise geodesic distance matrix evaluated for each shape at the pre-processing stage, and $\mathbf{\Pi}^*$ is the ground truth permutation relating between the shapes. The batch loss is the sum of $\ell_{\text{sup}}(\mathcal{X},\mathcal{Y})$ for all the pairs in the minibatch. Training an FMNet follows the standard Siamese setting commonly used for descriptor or metric learning, in which two copies of the network with shared parameters produces the descriptors on $\mathcal{X}$ and $\mathcal{Y}$. From this perspective, the functional map and the soft correspondence layers are parts of the Siamese loss rather than of the network itself.

3. Unsupervised deep functional maps

The FMNet achieves state-of-the-art performance on standard deformable shape correspondence benchmarks as shown in [28]. However, one can argue that the supervised training regime is prohibitive in terms of the amount of the manually annotated data required.

The main contribution of this paper is replacing supervision by pointwise correspondences with standard geometric quantities that do not require annotations, as formulated in [4].

As mentioned before, human pose articulation can be modeled as approximate isometries, that is, the latent correspondence introduces little metric distortion. If two vertices were at some geodesic distance on the source shape, after mapping by the correct correspondence, the distance between corresponding points on the target domain is preserved.

Let $\mathbf{P}$ be the output of the soft correspondence layer of an FMNet; as before, its squared elements $q_{ji} = p^2_{ji}$ are interpreted as probability distributions on $\mathcal{Y}$. In these terms, the ji-th element of the matrix $\mathbf{Q}^T\mathbf{D}_\mathcal{Y}\mathbf{Q}$

$$(\mathbf{Q}^T\mathbf{D}_\mathcal{Y}\mathbf{Q})_{ji} = \sum_{m,n} p^2_{mi} p^2_{nj} d_\mathcal{Y}(m,n) \tag{11}$$

represents the expected distance on $\mathcal{Y}$ between the images of the vertices $i,j\in\mathcal{X}$ under the soft correspondence $\mathbf{P}$.

This allows to define the following *unsupervised* loss

$$\ell_{\text{uns}}(\mathcal{X},\mathcal{Y}) = \frac{1}{|\mathcal{X}|^2}\left\|\mathbf{D}_\mathcal{X} - \mathbf{Q}^T\mathbf{D}_\mathcal{Y}\mathbf{Q}\right\|^2_{\text{F}}. \tag{12}$$

The batch loss is the sum of $\ell_{\text{uns}}(\mathcal{X},\mathcal{Y})$ for all the pairs in the minibatch. This loss measures the L_2 geodesic distance distortion and can be interpreted as the soft correspondence version of the SGMDS loss, see [4]. Note that rather than solving the QAP directly, we propose to train an FMNet using ℓ_{uns}, which promotes the network to generate descriptors for which the resulting soft correspondence minimizes the expected pairwise distance distortion. The unsupervised loss model thereby shares a common theoretical framework with the spectral generalized multidimensional scaling (SGMDS) model. Note, that for the unsupervised loss the whole optimization could have been executed in the spectral domain, as suggested by the SGMDS model

[4]. In that case the *Corr* module in the FMnet architecture presented in Fig. 2 could have been avoided. We plan to explore such architectures in the future.

From the unsupervised perspective all the shapes in the world could constitute a training set. Since the network is ground-truth independent and general to any class of shapes it is expected to improve when new shapes are encountered. The strict train-test separation applied in the supervised regime can be followed in the unsupervised regime depending on the settings; If the training set is representative enough to generalize the test set, the network can learn on the training set and infer on the test set, reducing the processing time per shape. On the other hand, for a non representative training set, the network can improve the learned model by processing the test shapes as well. Learning could in fact be executed even at inference time, as demonstrated in 5.1. For the FAUST scans [7], for example, the authors provide a training set with ground-truth labeling and a disjoint unlabeled test set. While the unsupervised scheme has access to the same data, contrarily to the supervised counterpart, it can leverage the unlabeled test shapes to improve the prediction accuracy.

4. Implementation

We implemented our network in TensorFlow [1], running on a GeForce GTX 1080 Ti GPU. Data preprocessing and correspondence refinement were done in Matlab. We provide a link to the code in the supplementary material.

4.1. Pre-processing

To enable mini-batches of multiple shapes to fit in memory, each shape in the training set was remeshed to between $n \sim 5$K and 7K vertices , by edge contraction [22]. For each remeshed shape $k \sim 70 - 150$ LBO eigenfunctions were calculated as well as a 352-dimensional SHOT descriptor [43], using 10 bins and a SHOT radius roughly chosen to 5% of the maximal pairwise geodesic distance. Geodesic distance matrices $\mathbf{D}$ were estimated using the fast marching method [27]. These quantities constitute the input to the network.

4.2. Network architecture and loss

For a more direct and fair comparison, we adopted the same network architecture as FMNet presented in Fig. 2. The input for each pair of shapes is their $n \times k$ truncated LBO bases $\mathbf{\Phi}$ and $\mathbf{\Psi}$, the $n \times n$ pairwise distance matrices $\mathbf{D}_{\mathcal{X}}$ and $\mathbf{D}_{\mathcal{Y}}$, and the $n \times 352$ SHOT descriptor fields. These are fed to a 7-layer residual network [24] outputting 352-dimensional dense descriptor fields $\mathbf{F}$ and $\mathbf{G}$ on $\mathcal{X}$ and $\mathcal{Y}$ respectively, which can be thought of as non-linearly transformed variants of SHOT. The computed descriptors are then input to the functional map layer, yielding a functional map matrix $\mathbf{C}$ according to (7), followed by a soft correspondence layer producing the stochastic correspondence matrix $\mathbf{P}$ as per Eq. (8). Finally, the unsupervised loss is calculated according to Eq. (12). While in FMNet the loss is calculated on a random sub-sampling of the vertices, we found that this strategy introduces inaccuracies to the descriptor coefficients in the LBO basis; When sub-sampling is used, the network evaluates an *estimate* of the projection coefficients, which quickly becomes inaccurate for descriptors with high-frequency content. To avoid this, in our implementation we perform the projection at full resolution while decreasing the size of the mini-batch to 4–5 pairs of shapes per mini-batch. In all our experiments we used no more than a few thousand (3K–10K) mini-batch iterations.

4.3. Post-processing

Point-wise map recovery. Following the protocol of FMNet, we apply the product manifold filter (PMF) [49], to improve the raw prediction of the network, in the full synthetic shapes settings. We found the geodesic kernel less effective when topological noise exists, e.g. real scans. Additionally, PMF is not well suited to partiality. PMF algorithm takes noisy matches as input, and produces a (guaranteed) bijective and smoother correspondence of higher accuracy as output. The application of PMF boils down to solving a series of linear assignment problems $\arg\max_{\mathbf{\Pi}^t} \langle \mathbf{\Pi}^t, \mathbf{K}_{\mathcal{X}} \mathbf{\Pi}^{t-1} \mathbf{K}_{\mathcal{Y}}^{\top} \rangle_F$, where $\mathbf{\Pi}^t$ ranges over the space of permutations, and $\mathbf{K}_{\mathcal{X}}, \mathbf{K}_{\mathcal{Y}}^{\top}$ are kernel matrices acting as diffusion operators. We refer to [49] for additional details.

Upscaling. Since we operate on remeshed shapes, we finally apply an upscaling step to bring the correspondence back to the original resolution. Again, we follow the procedure described in FMNet [28], namely we solve a functional map estimation problem of the form

$$\mathbf{C}_{\mathrm{up}} = \arg\min_{\mathbf{C}} \|\mathbf{C}\hat{\mathbf{F}}_{\mathrm{up}} - \hat{\mathbf{G}}_{\mathrm{up}}\|_{2,1} \,, \tag{13}$$

where $\hat{\mathbf{F}}_{\mathrm{up}}, \hat{\mathbf{G}}_{\mathrm{up}}$ contain the LBO coefficients (in the full resolution basis) of delta functions supported at corresponding points, extracted from the low resolution map C. The $\ell_{2,1}$ norm (defined as the sum of ℓ_2 norms of the columns) allows to down-weight potential mismatches.

5. Experiments

5.1. Learning to match a single pair

Before delving into training on large datasets, we begin our experimental section with testing one extreme of the shape matching problem: single input pair. Clearly, this is the native environment for classical, non-learning based methods. While learning-based methods have endowed us

with better solutions given large train sets, they are not equipped to handle entirely novel examples. We demonstrate that the unsupervised network can be utilized as an ad-hoc solver for a single pair, producing excellent results, while initialized with random weights. Fig. 1 shows our *unprocessed* network result on a pair of shapes made by an artist. Note that the ground-truth labeling is not provided and therefore a supervised learning-based method can not be fine-tuned on the input pair. Instead we compare with the unprocessed predictions of FMNet pre-trained on human shapes from FAUST. Additionally, we compare with two axiomatic methods [4, 36], using the same number of LBO eigenfunctions (150) and applying the post processing procedure described in 4.3 on the axiomatic results. While our method exhibit superior performance, axiomatic methods runtime exceeded one hour. Conversely, optimizing our network took about 15 minutes. Furthermore, had we been given an additional deformation of the same shape to solve for, any axiomatic method would have to solve the problem from scratch. Differently, as our method had already learned to convert the pair-wise optimization problem to a descriptor matching problem, inference would take about one second! See the supplementary material for *fast inference* experiment.

5.2. Faust synthetic

In this experiment we compare our unsupervised method and its supervised counterpart in the same settings. We show that (a) optimizing for the unsupervised loss results in a correlated decrease of the supervised loss; (b) the unsupervised method achieves the same accuracy as the supervised one. For training our network, we used Faust synthetic human shapes [7] and followed the same dataset split as in [28] where the first 80 shapes of 8 subjects are used for training, and a disjoint set of 20 shapes of 2 other subjects is used for testing. Each training mini-batch contained 4 shape pairs in their full resolution of 6890 vertices. We used the same parameters as in [28], namely, $k = 120$ eigenfunctions and ADAM optimizer with a learning rate of $\alpha = 10^{-3}, \beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$. We used $3K$ training mini-batches. Note that, as in [28], since we train on shape pairs the effective train set size is 6400.

Loss function analysis. Fig. 3 displays the unsupervised loss during the training process (top), alongside with the supervised loss (bottom). Importantly, the unsupervised network had no access to ground truth correspondence. From the graphs, it can be observed that while the optimization target is the unsupervised loss, the supervised loss is decreased as well. This demonstrates nicely that when our underlying assumption of (quasi-) isometric deformations holds, one can replace the expensive supervision altogether with a single axiomatic-driven loss term.

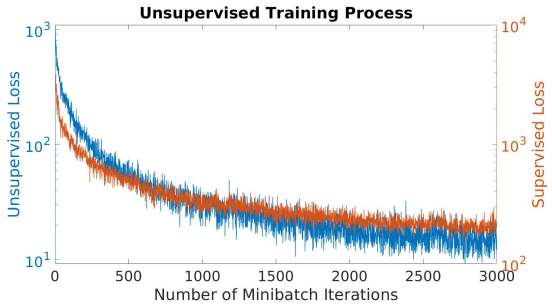

Figure 3: Unsupervised loss (left axis) and supervised loss (right axis) measured during the unsupervised training process, in logarithmic scale.

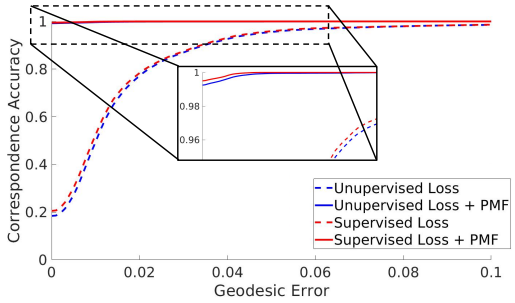

Figure 4: Unsupervised and supervised network results, evaluated on synthetic Faust intra-subject test pairs.

Figure 5: Synthetic Faust texture transfer. Four right models show the predicted matching from the reference model.

Performance comparison. To compare our results with the supervised network, we followed the same training scheme, this time using the supervised loss. We used the 20 test shapes to construct a test dataset of 400 pairs in total; 200 of which are intra-subject pairs, and the other 200 are inter-subject pairs (Note that the matching is directional from source to target, hence this set is not redundant). The intra-subject pairs, are well modeled by isometry while the inter-subject pairs exhibit deviation from isometry. Fig. 4 compares the results for the 200 intra-subject test pairs and Fig. 5 visualizes the calculated correspondences between intra- as well as inter-subject test pairs. Additional visualization is available in the supplementary material.

5.3. Real scans

Traditionally, axiom-based methods were proven useful only in the Computer Graphics regime. One of our goals in introducing learned descriptors is to demonstrate the applicability of our method to real scanned data. To this end, we make use of FAUST real scans benchmark. These are very high-resolution, non-watertight meshes, many of which contain holes and topological noise. We used the whole scanned data in the benchmark. The scans were down-sampled to a resolution of 7K vertices. For each scan the distance matrix was calculated, as well as 352-dimensional SHOT descriptors and $k = 70$ LBO eigenfunctions. Each training mini-batch contained 4 pairs of shapes. We trained our network for $10K$ iterations. The raw network predictions were only upscaled but not refined with PMF, as explained in 4.3. Quantitative results were evaluated through the online evaluation system. With an average and worst case scores of $\mathbf{2.51}$ cm and $\mathbf{24.35}$ cm, respectively, on the intra challenge, our network performs on par with state of the art methods that do not use additional data; namely, FMNet (2.44, 26.16), and Chen et al. [16] (4.86, 26.57). We perform slightly below the recent 3D-CODED method [23] (1.98, 5.18) which uses an additional augmentation of over 200K shapes at training. The same method, when not using additional data achieves worse results by a factor of ≈ 9. For completion, we also trained the network in the same settings, using only the train-set. We noticed a minor change in the performance: [Average error: slight deterioration to **2.82 cm** instead of **2.51 cm**; maximum error: slight improvement to **20.64 cm** instead of **24.35 cm**], keeping the result on-par with FM-Net.

5.4. Generalization

Having an unsupervised loss grants us the ability to train on datasets without given dense correspondences, or even to optimize on individual pairs. Both methods were demonstrated in the previous experiments. In this subsection, however, we would like to pose a different question: what has our network learned by training on a source dataset, and to which extent this knowledge is transferable to a target one. Transferability between training domains is a long-standing research area that has recently re-gained lots of interest, yet it hasen't been explored as much in the shape analysis community. In the scope of this work we focus on transferring from either the synthetic or scanned FAUST shapes to the either (a) human shapes form Dynamic-FAUST, (b) human shapes form SCAPE, and (c) Animal shapes from TOSCA. We show the prediction of the networks that were trained with Faust synthetic or scanned data, evaluated on (a,b,c), without using train samples from these datasets.

Dynamic FAUST is a recent very large collection of human shapes [8], including various sequences of activities. While the shapes are triangulated in the same way as our train set of synthetic FAUST, they significantly differ in pose and appearance. Fig. 8 shows excellent generalization to this set, suggesting that the small set of 80 synthetic FAUST shapes were sufficient to capture the pose and shape variability. For additional visualizations see the supplementary material.

SCAPE [5] also comprises human shapes only. Yet, we've witnessed a quite poor performance using the network trained on synthetic FAUST. By the same reasoning behind the former result, the network might have learned to specialize on synthetic connectivity. To circumvent this, we have tested the network trained on scans, that demonstrate different meshes. Indeed Figure 7 displays good generalization.

TOSCA dataset [15] includes various animal shapes. Impressively, the network trained on *human* scans shows very good performance without ever seeing a single animal shape at train time. In Fig. 9 we compare with several axiomatic models and with a supervised network [41] trained separately on each animal category and show comparable results before pre-processing, and near-perfect results after.

5.5. Partial correspondence

Partial shape correspondence is a notoriously hard problem, and techniques that aim at solving it often require special care [29, 40, 30]. That said, in this experiment we tested the performance of our method under extreme partiality conditions *as is*, namely, without any modification to our network. To this end, we used the challenging "dog with holse" class from [18]. We trained the network on a small set of 10 partial shapes, and evaluated the results 26 test shapes. The network results are shown in Fig. 6. We found that the mismatches occur typically near the boundary of the partial shape. The reason might be the distortion of the SHOT descriptor in these regions.

6. Discussion and conclusions

The main message of the paper is that a properly designed unsupervised surrogate task can replace massive labeling. While we advocate the pure unsupervised approach as a replacement to the supervised one, the two can also be combined in a semi-supervised learning scheme. While we demonstrate that the minimization of geodesic distance distortion achieves good generalization on a variety of benchmarks, local scale variation and topological changes can challenge the classic model and require a proper adaptation. In future studies, we intend to investigate training tasks based on the preservation of more general scale- and conformal-invariant pair-wise geometric quantities, as well as topological properties, e.g. by utilizing pairwise *diffusion* distances, see e.g. [12]. The proposed network exhibits surprisingly high performance on partial correspondence tasks, even though the functional map layer is not

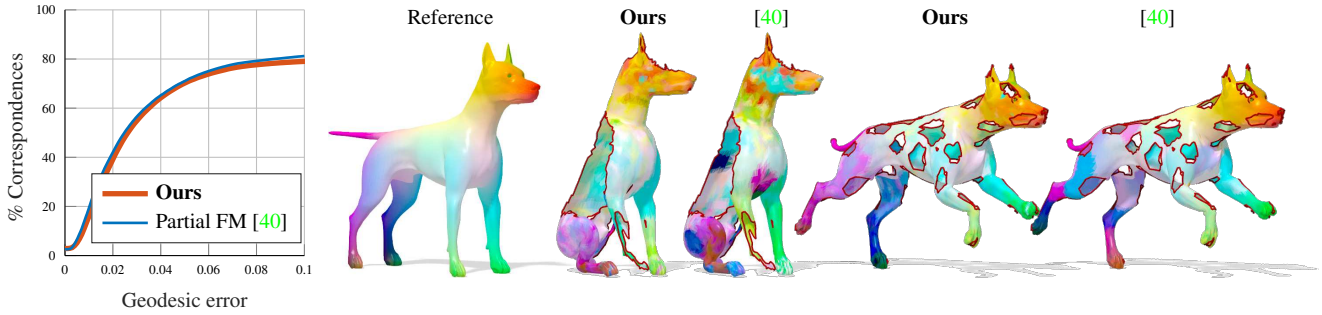

Figure 6: Comparisons on the SHREC'16 benchmark [18] (dog class) for partial matching of deformable shapes. We demonstrate results in line with partial functional maps [40], the current state of the art for this problem. The partial shapes shown on the right are matched to the reference; corresponding points have similar color.

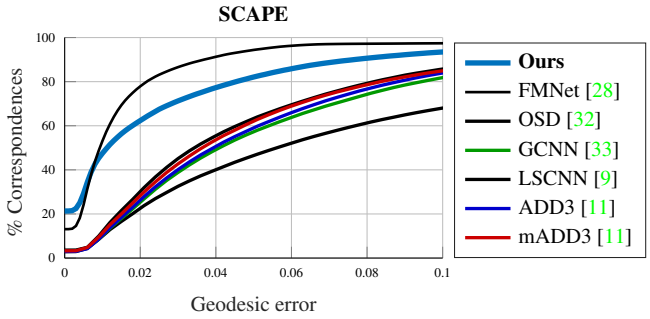

Figure 7: Generalization on SCAPE. For fair comparison, we show our network prediction without PMF refinement.

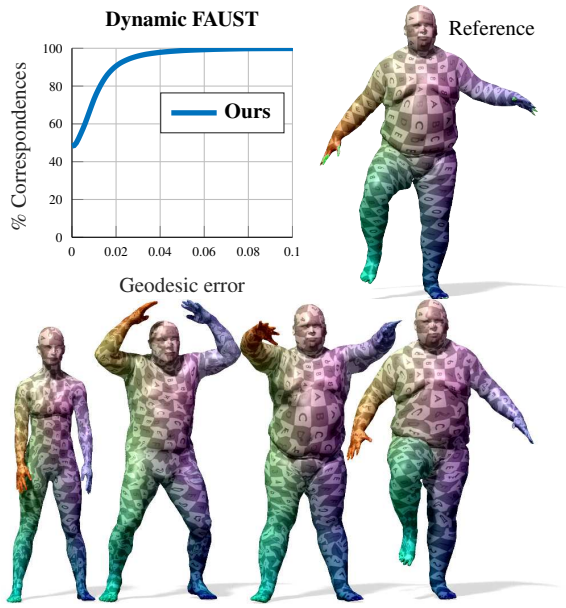

Figure 8: Generalization experiments on Dynamic FAUST. We render the network predictions with PMF refinement.

explicitly designed to treat partial data. Extending it to the partial setting based on the recently introduced partial functional map formalism [40, 29] and its relation to previous explicit efforts [13] will be the subject of further investigation. Finally, we would like to explore additional descriptor

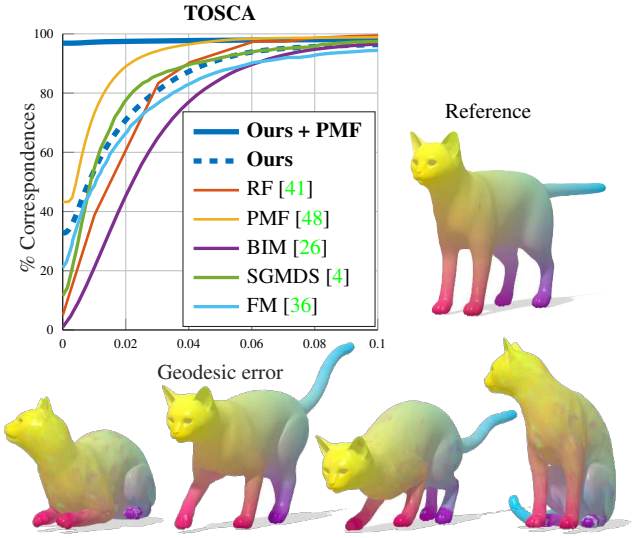

Figure 9: Generalization on TOSCA. We show our network prediction before (dashed curve) and after (solid) PMF refinement. The rendered visualization is *before* refinement.

fields with enhanced properties like increased sensitivity to symmetries, increased robustness to partiality and non rigid deformations. This paper presents a first attempt to create a fully unsupervised learning framework to solve the fundamental problem of non rigid shape correspondence. We believe that the fusion of axiomatic models and deep learning is a promising direction that makes it possible to accommodate the expected future growth of 3D data.

Acknowledgements

We gratefully acknowledge Matteo Sala for providing the models that appear in Fig. 1. Ron Kimmel and Oshri Halimi were supported by Israel Ministry of Science grant no. 3-14719, the Technion Hiroshi Fujiwara Cyber Security Research Center, and the Israel Cyber Bureau. Alex Bronstein was supported by ERC grant no. 335491 (RAPID). Emanuele Rodolà was supported by ERC grant no. 802554 (SPECGEO), and the MIUR under grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of Sapienza University.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 5

[2] Y. Aflalo, H. Brezis, and R. Kimmel. On the optimality of shape and data representation in the spectral domain. *SIAM Journal on Imaging Sciences*, 8(2):1141–1160, 2015. 3

[3] Y. Aflalo, A. Bronstein, and R. Kimmel. On convex relaxation of graph isomorphism. *Proceedings of the National Academy of Sciences*, 112(10):2942–2947, 2015. 2

[4] Y. Aflalo, A. Dubrovina, and R. Kimmel. Spectral generalized multi-dimensional scaling. *IJCV*, 118(3):380–392, 2016. 1, 2, 3, 4, 5, 6, 8

[5] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape Completion and Animation of People. *TOG*, 24(3):408–416, 2005. 7

[6] M. Aubry, U. Schlickewei, and D. Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *Proc. ICCV Workshops*, pages 1626–1633, 2011. 1, 3

[7] F. Bogo, J. Romero, M. Loper, and M. J. Black. FAUST: Dataset and Evaluation for 3d Mesh Registration. In *Proc. CVPR*, 2014. 5, 6

[8] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 7

[9] D. Boscaini, J. Masci, S. Melzi, M. M. Bronstein, U. Castellani, and P. Vandergheynst. Learning class-specific descriptors for deformable shapes using localized spectral convolutional networks. *Computer Graphics Forum*, 34(5):13–23, 2015. 8

[10] D. Boscaini, J. Masci, E. Rodolà, and M. Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 3189–3197, 2016. 2, 3

[11] D. Boscaini, J. Masci, E. Rodolà, M. M. Bronstein, and D. Cremers. Anisotropic diffusion descriptors. In *Computer Graphics Forum*, volume 35, pages 431–441. Wiley Online Library, 2016. 8

[12] A. M. Bronstein, M. Bronstein, M. M. R. Kimmel, and G. Sapiro. A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. *International Journal of Computer Vision*, 89(2-3):266–286, 2010. 1, 7

[13] A. M. Bronstein, M. M. Bronstein, A. M. Bruckstein, and R. Kimmel. Partial similarity of objects, or how to compare a centaur to a horse. *International Journal of Computer Vision*, 84(2):163–183, 2009. 8

[14] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *PNAS*, 103(5):1168–1172, 2006. 1, 2

[15] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008. 7

[16] Q. Chen and V. Koltun. Robust Nonrigid Registration by Convex Optimization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2039–2047, 2015. 2, 7

[17] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS*, 102(21):7426–7431, 2005. 1

[18] L. Cosmo, E. Rodolà, M. M. Bronstein, et al. SHREC'16: Partial matching of deformable shapes. In *Proc. 3DOR*, 2016. 7, 8

[19] L. Cosmo, E. Rodolà, J. Masci, A. Torsello, and M. M. Bronstein. Matching deformable objects in clutter. In *Proc. 3DV*, 2016. 3

[20] A. Dubrovina and R. Kimmel. Matching shapes by eigendecomposition of the Laplace-Beltrami operator. *Proc. 3DPVT*, 2(3), 2010. 2

[21] A. Elad and R. Kimmel. On bending invariant signatures for surfaces. *PAMI*, 25(10):1285–1295, 2003. 1, 2

[22] M. Garland and P. S. Heckbert. Surface simplification using quadric error metrics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 209–216. ACM Press/Addison-Wesley Publishing Co., 1997. 5

[23] T. Groueix, M. Fisher, V. G. Kim, B. Russell, and M. Aubry. 3d-coded : 3d correspondences by deep deformation. In *ECCV*, 2018. 2, 7

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[25] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):433–449, 1999. 3

[26] V. G. Kim, Y. Lipman, and T. A. Funkhouser. Blended intrinsic maps. *Trans. Graphics*, 30(4), 2011. 8

[27] R. Kimmel and J. A. Sethian. Computing geodesic paths on manifolds. *Proceedings of the national academy of Sciences*, 95(15):8431–8435, 1998. 5

[28] O. Litany, T. Remez, E. Rodolà, A. M. Bronstein, and M. M. Bronstein. Deep functional maps: Structured prediction for dense shape correspondence. In *Proc. ICCV*, volume 2, page 8, 2017. 1, 2, 3, 4, 5, 6, 8

[29] O. Litany, E. Rodolà, A. M. Bronstein, and M. M. Bronstein. Fully spectral partial shape matching. *Computer Graphics Forum*, 36(2):247–258, 2017. 7, 8

[30] O. Litany, E. Rodolà, A. M. Bronstein, M. M. Bronstein, and D. Cremers. Non-rigid puzzles. In *Computer Graphics Forum*, volume 35, pages 135–143. Wiley Online Library, 2016. 7

[31] R. Litman and A. M. Bronstein. Learning spectral descriptors for deformable shape correspondence. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):171–180, 2014. 2, 3

[32] R. Litman and A. M. Bronstein. Learning spectral descriptors for deformable shape correspondence. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(1):171–180, Jan. 2014. 8

[33] J. Masci, D. Boscaini, M. Bronstein, and P. Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 37–45, 2015. 3, 8

[34] F. Mèmoli and G. Sapiro. A theoretical and computational framework for isometry invariant recognition of point cloud data. *Foundations of Computational Mathematics*, 5:313–346, 2005. 2

[35] F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. 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. 2, 3

[36] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, and L. Guibas. Functional maps: a flexible representation of maps between shapes. *TOG*, 31(4):1–11, 2012. 1, 2, 3, 6, 8

[37] H. Pottmann, J. Wallner, Q.-X. Huang, and Y.-L. Yang. Integral invariants for robust geometry processing. *Computer Aided Geometric Design*, 26(1):37–60, 2009. 1

[38] D. Raviv, A. Dubrovina, and R. Kimmel. Hierarchical framework for shape correspondence. *Numerical Mathematics: Theory, Methods and Applications*, 6(245-261), 2013. 2

[39] E. Rodolà, A. M. Bronstein, A. Albarelli, F. Bergamasco, and A. Torsello. A game-theoretic approach to deformable shape matching. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 182–189. IEEE, 2012. 2

[40] E. Rodolà, L. Cosmo, M. M. Bronstein, A. Torsello, and D. Cremers. Partial functional correspondence. In *Computer Graphics Forum*, volume 36, pages 222–236. Wiley Online Library, 2017. 3, 7, 8

[41] E. Rodolà, S. Rota Bulo, T. Windheuser, M. Vestner, and D. 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. 2, 3, 7, 8

[42] R. M. Rustamov. Laplace-beltrami eigenfunctions for deformation invariant shape representation. In *Proceedings of the fifth Eurographics symposium on Geometry processing*, pages 225–233. Eurographics Association, 2007. 3

[43] S. Salti, F. Tombari, and L. Di Stefano. Shot: Unique signatures of histograms for surface and texture description. *Computer Vision and Image Understanding*, 125:251–264, 2014. 5

[44] G. Shamai and R. Kimmel. Geodesic distance descriptors. In *CVPR*, pages 3624–3632, 2017. 3

[45] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009. 1, 3

[46] A. Tevs, A. Berner, M. Wand, I. Ihrke, and H.-P. Seidel. Intrinsic shape matching by planned landmark sampling. In *Computer Graphics Forum*, volume 30, pages 543–552. Wiley Online Library, 2011. 2

[47] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In *International Conference on Computer Vision (ICCV)*, pages 356–369, 2010. 1, 3

[48] M. Vestner, Z. Lähner, A. Boyarski, O. Litany, R. Slossberg, T. Remez, E. Rodolà, A. Bronstein, M. Bronstein, R. Kimmel, and D. Cremers. Efficient deformable shape correspondence via kernel matching. In *Proc. 3DV*, 2017. 8

[49] M. Vestner, R. Litman, E. Rodolà, A. M. Bronstein, and D. Cremers. Product manifold filter: Non-rigid shape correspondence via kernel density estimation in the product space. In *CVPR*, pages 6681–6690, 2017. 1, 5

[50] U. Von Luxburg, A. Radl, and M. Hein. Hitting and commute times in large random neighborhood graphs. *The Journal of Machine Learning Research*, 15(1):1751–1798, 2014. 1

[51] C. Wang, M. M. Bronstein, A. M. Bronstein, and N. Paragios. Discrete minimum distortion correspondence problems for non-rigid shape matching. In *Proc. SSVM*, 2011. 2