STRUCTURED PREDICTION OF DENSE MAPS BETWEEN GEOMETRIC DOMAINS

Emanuele Rodolà[*]

Sapienza University of Rome

ABSTRACT

We introduce a new framework for learning dense correspondence between deformable geometric domains such as polygonal meshes and point clouds. Existing learning based approaches model correspondence as a labelling problem, where each point of a query domain receives a label identifying a point on some reference domain; the correspondence is then constructed a posteriori by composing the label predictions of two input geometries. We propose a paradigm shift and design a structured prediction model in the space of functional maps, linear operators that provide a compact representation of the correspondence. We model the learning process via a deep residual network which takes dense descriptor fields as input, and outputs a soft map between the two given objects. The resulting correspondence is shown to be accurate on several challenging shape correspondence benchmarks.

***Index Terms*—** Spectral geometry, deep learning, functional maps, structured prediction

1. INTRODUCTION

3D acquisition technology has made great progress in the last decade, and is being rapidly incorporated into commercial products ranging from Microsoft Kinect [1] for gaming, to LIDARs used in autonomous cars. An essential building block for application design in many of these domains is to recover 3D shape correspondences in a fast and reliable way. While handling real-world scanning artifacts is a challenge by itself, additional complications arise from non-rigid motions of the objects of interest (typically humans or animals). Most non-rigid shape correspondence methods employ local descriptors that are designed to achieve robustness to noise and deformations; however, relying on such "handcrafted" descriptors can often lead to inaccurate solutions in practical settings. Partial remedy to this was brought by the recent line of works on learning shape correspondence [2, 3, 4, 5, 6, 7, 8]. A key drawback of these methods lies in their emphasis on learning a descriptor that would help in identifying corresponding points, or on learning a labelling with respect to some reference domain. On the one hand, by focusing on the descriptor, the learning process remains agnostic to the way the final correspondence is computed, and costly post-processing steps are often necessary in order to obtain accurate solutions from the learned descriptors. On the other hand, methods based on a label space are restricted to a fixed number of points and rely on the adoption of an intermediate reference model.

[*]Joint work with O. Litany, T. Remez, A. Bronstein, and M. Bronstein.

Fig. 1. Correspondence results obtained by our network model on two pairs of real scans. Corresponding points are assigned the same color. The average error for the left and right pairs is 5.21cm and 2.34cm respectively. Accurate correspondence is obtained despite mesh "gluing" in areas of contact.

Contribution. Our main contributions can be summarized as follows:

- We introduce a new *structured prediction* model for shape correspondence [9]. Our framework allows end-to-end training: it takes base descriptors as input, and returns matches.
- We show that our approach consistently outperforms existing descriptor and correspondence learning methods on several recent benchmarks.

2. RELATED WORK

Probably the first example of learning correspondence for deformable 3D shapes is the "shallow" random forest approach of Rodolà et al. [3]. More recently, Wei et al. [10] employed a classical (extrinsic) CNN architecture trained on huge training sets for learning invariance to pose changes and cloth-

ing. Convolutional neural networks on non-Euclidean domains (surfaces) were first considered by Masci et al. [4] with the introduction of the geodesic CNN model, a deep learning architecture where the classical convolution operation is replaced by an intrinsic (albeit, non-shift invariant) counterpart. The framework was shown to produce promising results in descriptor learning and shape matching applications, and was recently improved by Boscaini et al. [7] and generalized further by Monti et al. [8]. These methods are instances of a broader recent trend of *geometric deep learning* attempting to generalize successful deep learning paradigms to data with non-Euclidean underlying structure such as manifolds or graphs [11].

3. BACKGROUND

Manifolds. We model shapes as two-dimensional Riemannian manifolds $\mathcal{X}$ (possibly with boundary $\partial\mathcal{X}$) equipped with the standard measure $\mathrm{d}\mu$ induced by the volume form. Throughout the paper we will consider the space of functions $L^2(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \mid \langle f, f\rangle_{\mathcal{X}} < \infty\}$, with the standard manifold inner product $\langle f, g\rangle_{\mathcal{X}} = \int_{\mathcal{X}} f \cdot g \,\mathrm{d}\mu$.

The positive semi-definite Laplace-Beltrami operator $\Delta_{\mathcal{X}}$ generalizes the notion of Laplacian from Euclidean spaces to surfaces. It admits an eigen-decomposition $\Delta_{\mathcal{X}}\phi_i = \lambda_i\phi_i$ (with proper boundary conditions if $\partial\mathcal{X} \neq \emptyset$), where the eigenvalues form a discrete spectrum $0 = \lambda_1 \leq \lambda_2 \leq \ldots$ and the eigenfunctions $\phi_1, \phi_2, \ldots$ form an orthonormal basis for $L^2(\mathcal{X})$, allowing us to expand any function $f \in L^2(\mathcal{X})$ as a Fourier series

$$f(x) = \sum_{i\geq 1} \langle \phi_i, f\rangle_{\mathcal{X}} \phi_i(x)\,. \tag{1}$$

Functional correspondence. In order to compactly encode correspondences between shapes, we make use of the functional map representation introduced by Ovsjanikov et al. [12]. The key idea is to identify correspondences by a linear operator $T : L^2(\mathcal{X}) \to L^2(\mathcal{Y})$, mapping functions on $\mathcal{X}$ to functions on $\mathcal{Y}$. This can be seen as a generalization of classical point-to-point matching, which is a special case where delta functions are mapped to delta functions.

The linear operator T admits a matrix representation $\mathbf{C} = (c_{ij})$ with coefficients $c_{ji} = \langle \psi_j, T\phi_i\rangle_{\mathcal{Y}}$, where $\{\phi_i\}_{i\geq 1}$ and $\{\psi_j\}_{j\geq 1}$ are orthogonal bases on $L^2(\mathcal{X})$ and $L^2(\mathcal{Y})$ respectively, leading to the expansion:

$$Tf = \sum_{ij\geq 1} \langle \phi_i, f\rangle_{\mathcal{X}} c_{ji} \psi_j\,. \tag{2}$$

A good choice for the bases $\{\phi_i\}$, $\{\psi_j\}$ is given by the Laplacian eigenfunctions on the two shapes [12, 13], since (by analogy with Fourier analysis) it allows to truncate the series (2) after the first k coefficients – yielding a band-limited approximation of the original map. The resulting matrix $\mathbf{C}$ is a $k \times k$ compact representation of a correspondence between the two shapes, where typically $k \ll n$ (here n is the number of points on each shape).

Functional correspondence problems seek a solution for $\mathbf{C}$, given a set of corresponding functions $f_i \in L^2(\mathcal{X})$ and $g_i \in L^2(\mathcal{Y})$, $i = 1, \ldots, q$, on the two shapes. In the Fourier basis, these functions are encoded into matrices $\hat{\mathbf{F}} = (\langle \phi_i, f_j\rangle_{\mathcal{X}})$ and $\hat{\mathbf{G}} = (\langle \psi_i, g_j\rangle_{\mathcal{Y}})$, leading to the least-squares problem:

$$\min_{\mathbf{C}} \|\mathbf{C}\hat{\mathbf{F}} - \hat{\mathbf{G}}\|_F^2\,. \tag{3}$$

In practice, dense q-dimensional descriptor fields (e.g., HKS [14]) on $\mathcal{X}$ and $\mathcal{Y}$ are used as the corresponding functions.

Label space. Previous approaches at learning shape correspondence phrased the matching problem as a *labelling* problem [3, 4, 6, 7, 8]. These approaches attempt to label each vertex of a given query shape $\mathcal{X}$ with the index of a corresponding point on some reference shape $\mathcal{Z}$ (usually taken from the training set), giving rise to a dense point-wise map $T_{\mathcal{X}} : \mathcal{X} \to \mathcal{Z}$. The correspondence between two queries $\mathcal{X}$ and $\mathcal{Y}$ can then be obtained via the composition $T_{\mathcal{Y}}^{-1} \circ T_{\mathcal{X}}$ [3].

Given a training set $S = \{(x, \pi^*(x))\} \subset \mathcal{X} \times \mathcal{Y}$ of matches under the ground-truth map $\pi^* : \mathcal{X} \to \mathcal{Y}$, label-based approaches compute a descriptor $F_\Theta(x)$ whose optimal parameters minimize the *multinomial regression* loss:

$$\ell_{\mathrm{mr}}(\Theta) = -\sum_{(x,\pi^*(x))\in S} \langle \delta_{\pi^*(x)}, \log F_\Theta(x)\rangle_{\mathcal{Y}}\,, \tag{4}$$

where $\delta_{\pi^*(x)}$ is a delta function on $\mathcal{Y}$ at point $\pi^*(x)$.

Such an approach essentially treats the correspondence problem as one of classification, where the aim is to approximate as closely as possible (in a statistical sense) the correct label for each point. The actual construction of the full correspondence is done *a posteriori* by a composition step with an intermediate reference domain, or by solving the least-squares problem (3) with the learned descriptors as data.

Discretization. In the discrete setting, shapes are represented as manifold triangular meshes with n vertices (in general, different for each shape). The Laplace-Beltrami operator Δ is discretized as a symmetric $n \times n$ matrix $\mathbf{L} = \mathbf{A}^{-1}\mathbf{W}$ using a classical linear FEM scheme [15], where the *stiffness matrix* $\mathbf{W}$ contains the cotangent weights, and the *mass matrix* $\mathbf{A}$ is a diagonal matrix of vertex area elements. The manifold inner product $\langle f, g\rangle$ is discretized as the area-weighted dot product $\mathbf{f}^\top\mathbf{A}\mathbf{g}$, where the vectors $\mathbf{f}, \mathbf{g} \in \mathbb{R}^n$ contain the function values of f and g at each vertex. Note that under such discretization we have $\mathbf{\Phi}^\top\mathbf{A}\mathbf{\Phi} = \mathbf{I}$, where $\mathbf{\Phi}$ contains the Laplacian eigenfunctions as its columns.

4. DEEP FUNCTIONAL MAPS

In this paper we propose an alternative model to the labelling approach described above. We aim at learning point-wise de-

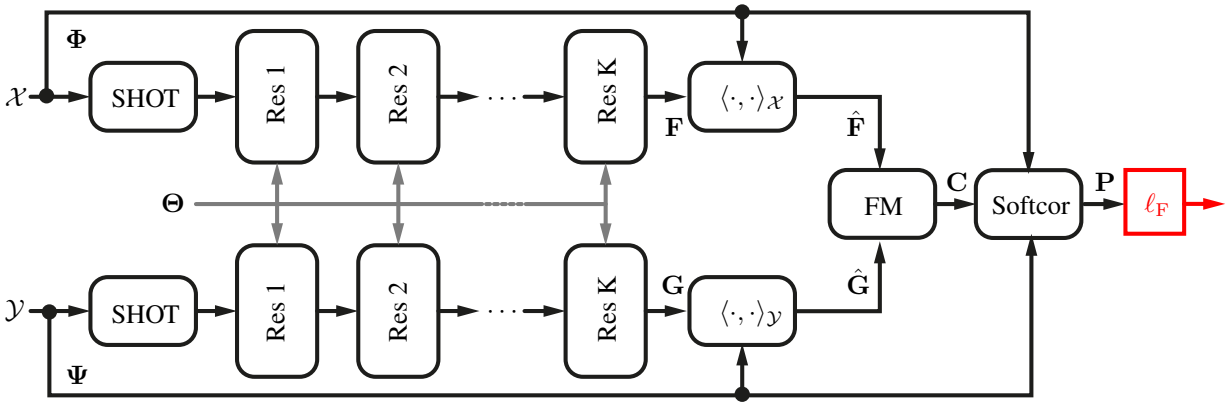

Fig. 2. **FMNet architecture.** Input point-wise descriptors (SHOT [16] in this paper) from a pair of shapes are passed through an identical sequence of operations (with shared weights), resulting in refined descriptors $\mathbf{F}, \mathbf{G}$. These, in turn, are projected onto the Laplacian eigenbases $\mathbf{\Phi}, \mathbf{\Psi}$ to produce the spectral representations $\hat{\mathbf{F}}, \hat{\mathbf{G}}$. The functional map (FM) and soft correspondence (Softcor) layers, implementing Equations (3) and (6) respectively, are not parametric and are used to set up the geometrically structured loss ℓ_{F} (5).

scriptors which, when used in a functional map pipeline such as (3), will induce an accurate correspondence. To this end, we construct a neural network which takes as input existing, manually designed descriptors and improves upon those while satisfying a *geometrically* meaningful criterion. Specifically, we consider the *soft error loss*

$$\ell_{\mathrm{F}} = \sum_{(x,y)\in(\mathcal{X},\mathcal{Y})} P(x,y) d_{\mathcal{Y}}(y, \pi^*(x)) = \|\mathbf{P} \circ \mathbf{D}_{\mathcal{Y}}\|_{\mathrm{F}}\,, \quad (5)$$

where $\mathbf{D}_{\mathcal{Y}}$ is the $n \times n$ matrix of geodesic distances on $\mathcal{Y}$, $\circ$ is the element-wise product, and

$$\mathbf{P} = |\mathbf{\Psi}\mathbf{C}\mathbf{\Phi}^\top \mathbf{A}|^\wedge \quad (6)$$

is a *soft correspondence* matrix, which can be interpreted as the probability of point $x \in \mathcal{X}$ mapping to point $y \in \mathcal{Y}$ (see inset); here, $\mathbf{\Phi}, \mathbf{\Psi}$ are matrices containing the first k eigenfunctions $\{\phi_i\}$, $\{\psi_j\}$ as their columns, $|\cdot|$ acts element-wise, and $\mathbf{X}^\wedge$ is a column-wise normalization of $\mathbf{X}$. In the formula above, the $k \times k$ matrix $\mathbf{C}$ represents a functional map obtained as the least-squares solution to (3) under *learned* descriptors $\mathbf{F}, \mathbf{G}$.

Matrix $\mathbf{P}$ represents a rank-k approximation of the spatial correspondence between the two shapes, thus allowing us to interpret the soft error (5) as a probability-weighted geodesic distance from the ground-truth. This measure, introduced in [17] as an evaluation criterion for soft maps, endows our solutions with guarantees of mapping nearby points on $\mathcal{X}$ to nearby points on $\mathcal{Y}$. On the contrary, the classification cost (4), adopted by existing label-based correspondence learning approaches, considers *equally* correspondences that deviate from the ground-truth, no matter how far. Further, notice that Equation (6) is asymmetric, implying that each pair of training shapes can be used twice for training (i.e., in both directions). Also note that, differently from previous approaches operating in the label space, in our setting the number of effective training examples (i.e. pairs of shapes) increases *quadratically* with the number of shapes in the collection. This is a significant advantage in situations with scarce training data.

We implement descriptor learning using a Siamese residual network architecture [18]. To this network, we concatenate additional *non*-parametric layers implementing the least-squares solve (3) followed by computation of the soft correspondence according to (6). In particular, the solution to (3) is obtained in closed form as $\mathbf{C} = \hat{\mathbf{G}}\hat{\mathbf{F}}^\dagger$, where † denotes the pseudo-inverse operation. The complete architecture (named "FMNet") is illustrated in Fig. 2.

5. UPSCALING TO DENSE MAPS

For increased efficiency, we down-sample the input shapes to 15K vertices by edge contraction [19]. Given two down-sampled shapes $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$, the network predicts a $k \times k$ matrix $\tilde{\mathbf{C}}$ encoding the correspondence between the two. Since this matrix is expressed w.r.t. basis functions $\{\tilde{\phi}_i\}_i, \{\tilde{\psi}_j\}_j$ of the *low-resolution* shapes, it can not be directly used to recover a point-wise map between the full-resolution counterparts $\mathcal{X}$ and $\mathcal{Y}$. Therefore, we perform an upscaling step as follows.

Let $\pi_{\mathcal{X}} : \tilde{\mathcal{X}} \to \mathcal{X}$ be the injection mapping each point in $\tilde{\mathcal{X}}$ to the corresponding point in the full shape $\mathcal{X}$ (this map can be easily recovered by a simple nearest-neighbor search in $\mathbb{R}^3$), and similarly for shape $\mathcal{Y}$. Further, denote by $\tilde{T} : \tilde{\mathcal{X}} \to \tilde{\mathcal{Y}}$ the point-to-point map recovered from $\tilde{\mathbf{C}}$ using the baseline recovery approach of [12]. A map $T : \mathcal{X} \supset \mathrm{Im}(\pi_{\mathcal{X}}) \to \mathcal{Y}$ is obtained via the composition

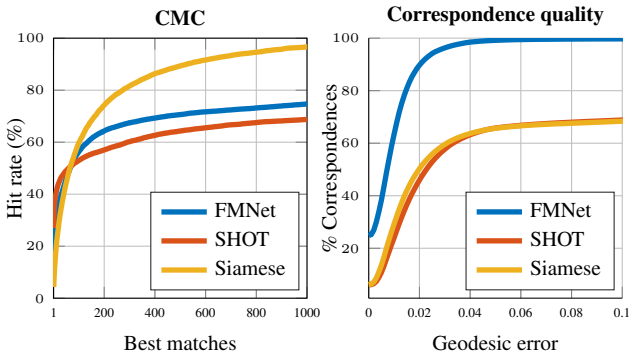

Fig. 3. Comparison between our structured prediction model (FMNet), metric learning (Siamese), and baseline SHOT in terms of CMC (left) and geodesic error (right). While the Siamese model produces better descriptors in terms of proximity (left), these do not necessarily induce a good functional correspondence (right).

$T = \pi_{\mathcal{Y}} \circ \tilde{T} \circ \pi_{\mathcal{X}}^{-1}$. However, while $\tilde{T}$ is dense in $\tilde{\mathcal{X}}$, the map T is *sparse* in $\mathcal{X}$. In order to map *each* point in $\mathcal{X}$ to a point in $\mathcal{Y}$, we construct pairs of delta functions $\delta_{x_i} : \mathcal{X} \to \{0,1\}$ and $\delta_{T(x_i)} : \mathcal{Y} \to \{0,1\}$ supported at corresponding points $(x_i, T(x_i))$ for $i = 1, \dots, |\tilde{\mathcal{X}}|$; note that we have as many corresponding pairs as the number of vertices in the low-resolution shape $\tilde{\mathcal{X}}$. We use these corresponding functions to define the minimization problem:

$$\mathbf{C}^* = \arg\min_{\mathbf{C}} \|\mathbf{C}\hat{\mathbf{F}} - \hat{\mathbf{G}}\|_{2,1}\,, \tag{7}$$

where $\hat{\mathbf{F}} = (\langle \phi_i, \delta_{x_j} \rangle_{\mathcal{X}})$ and $\hat{\mathbf{G}} = (\langle \psi_i, \delta_{T(x_j)} \rangle_{\mathcal{Y}})$ contain the Fourier coefficients (in the full-resolution basis) of the corresponding delta functions, and the $\ell_{2,1}$-norm allows to discard potential mismatches in the data. A dense point-to-point map between $\mathcal{X}$ and $\mathcal{Y}$ is finally recovered from the optimal functional map $\mathbf{C}^*$ by the nearest-neighbor approach of [12].

6. COMPARISON TO METRIC LEARNING

As a proof of concept, we study the behavior of our framework when the functional map layer is removed, and the soft error criterion (6) is replaced with the *siamese* loss [20]:

$$\begin{aligned}\ell_{\mathrm{s}}(\Theta) = &\sum_{x,x^+ \in S} \gamma \|F_\Theta(x) - F_\Theta(x^+)\|_2^2 \\ &+ \sum_{x,x^- \in D} (1-\gamma)(\mu - \|F_\Theta(x) - F_\Theta(x^-)\|_2)_+^2\,,\end{aligned} \tag{8}$$

where $\gamma \in (0,1)$ is a trade-off parameter, $\mu > 0$ is the margin, and $(x)_+ = \max(0, x)$. Here, the sets $S, D \subset \mathcal{X} \times \mathcal{Y}$ constitute the training data consisting of knowingly similar and dissimilar pairs of points respectively. By considering this loss function, we transform our *structured prediction* model into a *metric learning* model. The learned descriptors $F_\Theta(x)$ can be subsequently plugged into (3) to compute a correspondence; this metric learning approach was recently used in a functional map pipeline in [21]. For this test we use FAUST templates [22] as our data and SHOT [16] as an input feature.

From the CMC curves of Fig. 3 (left) we can clearly see that the model (8) succeeds at producing descriptors that attract each other at corresponding points, while mismatches are repulsed. However, as put in evidence by Fig. 3 (right), these descriptors do not perform well when they are used for seeking a dense correspondence via (3). Contrarily, our structured prediction model yields descriptors that are optimized for such a correspondence task, leading to a noticeable gain in accuracy.

Fig. 4. Results of FMNet on the SHREC'16 Partial Correspondence benchmark [23]. Each partial shape is matched to the full shape on the left; the color texture is transferred via the predicted correspondence.

7. MISSING PARTS AND TOPOLOGICAL NOISE

Our framework does not rely on any specific shape model, as it learns from the shape categories represented in the training data. In particular, it does not necessarily require the objects to be complete shapes: different forms of partiality can be tackled if adequately represented in the training set.

We demonstrate this by running our method on the recent SHREC'16 Partial Correspondence challenge [23]. The benchmark consists of hundreds of shapes of multiple categories with missing parts of various forms and sizes; a training set is also provided. We selected the 'dog' class from the 'holes' sub-challenge, being this among the hardest categories in the benchmark. The dataset is officially split into just 10 training shapes, and 26 test shapes. Qualitative examples of the obtained solutions are reported in Fig. 4.

Given the good results on shapes exhibiting missing parts, it may come as a surprise that topological alterations tend to be detrimental – changes of topology can be seen as a form of partiality, and we would expect similar accuracy in such cases. The underlying reason is that Laplacian eigenfunctions are inherently sensitive to topological changes. We propose to mitigate this issue by modeling the expected perturbations using a construction similar to partial functional maps [24], specifically, by incorporating partiality priors into the construction of matrix $\mathbf{C}$ within our structured prediction model.

8. REFERENCES

[1] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[2] R. Litman and A. M. Bronstein, "Learning spectral descriptors for deformable shape correspondence," *Trans. PAMI*, vol. 36, no. 1, pp. 170–180, 2014.

[3] E. Rodolà, S. Rota Bulò, T. Windheuser, M. Vestner, and D. Cremers, "Dense non-rigid shape correspondence using random forests," in *Proc. CVPR*, 2014.

[4] J. Masci, D. Boscaini, M. M. Bronstein, and P. Vandergheynst, "Geodesic convolutional neural networks on Riemannian manifolds," in *Proc. 3dRR*, 2015.

[5] 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*, vol. 34, no. 5, pp. 13–23, 2015.

[6] D. Boscaini, J. Masci, E. Rodolà, M. M. Bronstein, and D. Cremers, "Anisotropic diffusion descriptors," *Computer Graphics Forum*, vol. 35, no. 2, pp. 431–441, 2016.

[7] D. Boscaini, J. Masci, E. Rodolà, and M. M. Bronstein, "Learning shape correspondence with anisotropic convolutional neural networks," in *Proc. NIPS*, 2016.

[8] 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 *Proc. CVPR*, 2017.

[9] 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*, 2017.

[10] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li, "Dense human body correspondences using convolutional networks," in *Proc. CVPR*, 2016.

[11] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond Euclidean data," *arXiv:1611.08097*, 2016.

[12] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, and L. Guibas, "Functional maps: a flexible representation of maps between shapes," *Trans. Graphics*, vol. 31, no. 4, pp. 30:1–30:11, July 2012.

[13] Y. Aflalo, H. Brezis, and R. Kimmel, "On the optimality of shape and data representation in the spectral domain," *SIAM J. Imaging Sciences*, vol. 8, no. 2, pp. 1141–1160, 2015.

[14] J. Sun, M. Ovsjanikov, and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," in *Proc. SGP*, 2009.

[15] M. Meyer, M. Desbrun, P. Schröder, and A. H. Barr, "Discrete differential-geometry operators for triangulated 2-manifolds," *Visualization&Mathematics*, pp. 35–57, 2003.

[16] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of histograms for local surface description," in *Proc. ECCV*, 2010.

[17] A. Kovnatsky, M. M. Bronstein, X. Bresson, and P. Vandergheynst, "Functional correspondence by matrix completion," in *Proc. CVPR*, 2015.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[19] M. Garland and P. S. Heckbert, "Surface simplification using quadric error metrics," in *Proc. SIGGRAPH*, 1997, pp. 209–216.

[20] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. CVPR*, 2006.

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

[22] F. Bogo, J. Romero, M. Loper, and M. J. Black, "FAUST: Dataset and evaluation for 3D mesh registration," in *Proc. CVPR*, June 2014.

[23] L. Cosmo, E. Rodolà, M. M. Bronstein, A. Torsello, D. Cremers, and Y. Sahillioglu, "SHREC'16: Partial matching of deformable shapes," in *Proc. 3DOR*, 2016.

[24] E. Rodolà, L. Cosmo, M. M. Bronstein, A. Torsello, and D. Cremers, "Partial functional correspondence," *Computer Graphics Forum*, vol. 36, no. 1, pp. 222–236, 2017.