Matching Deformable Objects in Clutter

Luca Cosmo
Università Ca' Foscari Venezia
luca.cosmo@unive.it

Emanuele Rodolà
USI Lugano
emanuele.rodola@usi.ch

Jonathan Masci
USI Lugano
jonathan.masci@usi.ch

Andrea Torsello
Università Ca' Foscari Venezia
torsello@dais.unive.it

Michael M. Bronstein
USI Lugano / Tel Aviv University / Intel
michael.bronstein@usi.ch

Abstract

We consider the problem of deformable object detection and dense correspondence in cluttered 3D scenes. Key ingredient to our method is the choice of representation: we formulate the problem in the spectral domain using the functional maps framework, where we seek for the most regular nearly-isometric parts in the model and the scene that minimize correspondence error. The problem is initialized by solving a sparse relaxation of a quadratic assignment problem on features obtained via data-driven metric learning. The resulting matching pipeline is solved efficiently, and yields accurate results in challenging settings that were previously left unexplored in the literature.

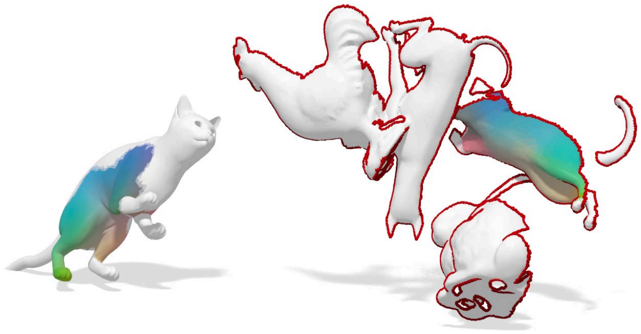


Figure 1. Our method allows to densely match a given 3D model (left) to a cluttered, partial scene (right), where both model and scene are allowed to deform non-rigidly. In this figure, corresponding points have same color whereas white denotes no match.

1. Introduction

Shape matching and object recognition are widely researched areas in 3D computer vision, with applications ranging from reconstruction to surveillance. On the one hand, shape matching concerns the problem of determining a dense correspondence between two given objects. On the other hand, object recognition consists in locating, and at the same time putting into correspondence a template model within a given scene which contains the object of interest. A particularly challenging instance of this problem arises when the object to be sought is allowed to deform in a *non-rigid* fashion – a common scenario, for instance, in robotics applications, where one has to locate a reference model within a dynamic environment acquired in 3D.

Despite the conceptual similarities, however, 3D shape matching and recognition have been tackled separately and under different assumptions. Deformable matching techniques assume the absence of additional objects (clutter); conversely, object-in-clutter methods rely on the scene to contain a *rigidly* transformed instance of the model. These constraints severely limit the usefulness of either family of approaches in practical scenarios.

In this paper, we rule out all the previous assumptions and consider the full problem of deformable object-in-clutter recognition and matching. For given model and scene, both of which are allowed to deform non-rigidly, we jointly determine the object location in the scene and solve for a dense correspondence between the two.

1.1. Related work

Shape matching. Deformable shape matching is an active area of research, with steady progress being made over the years (see, *e.g.*, the recent survey [44]). Our paper builds upon the functional map representation, recently introduced by Ovsjanikov *et al.* [30] as a tool for matching nearly isometric shapes. The key idea is to move from identifying a map between manifolds to identifying a linear operator (the *functional map*) between functional spaces defined over the shapes. By an appropriate choice of bases (Ovsjanikov *et al.* proposed the Laplacian eigenfunctions of the two manifolds), the operator admits a matrix representation that can compactly encode a map relating the two shapes. Desirable

978-1-5090-5407-7/16 $31.00 © 2016 IEEE
DOI 10.1109/3DV.2016.10

properties of the map can then be easily phrased as linear constraints given a known set of corresponding functions on the two shapes [30]. The framework was extended in several follow-up works by introducing a prior on the diagonal structure of the matrix [31, 18], by constructing coupled bases via joint diagonalization of Laplacians [18], and by introducing geometric structure in the correspondence matrix so as to achieve smoothness and localization [19].

All the previous works require the shapes to be given as full 3D models, which are assumed to be similar on the whole. Most related to our paper is the recent work of Rodolà *et al*. [34]. The authors show how the functional maps framework can be adapted to deal with situations in which *one* of the two objects is allowed to have missing parts, *i.e*., the overlap between the two shapes is known and equals the surface area of the partial shape. The method was the best performing in the recent SHREC'16 benchmark [10]. Differently, in this paper we do not assume known overlap between the two 3D objects (model and scene in our case), but rather we seek for approximately isometric subregions in both. This allows us to deal with missing parts as well as spurious geometry (clutter) in both 3D models.

Object detection in clutter. The problem of 3D object detection in cluttered scenes has been tackled for several years by the computer vision community (see [12] for a recent survey). To date, the most successful approaches couple the detection of local rotation-invariant surface features [3] together with some geometric consistency criterion to drive the matching. Candidate matches are first selected according to the similarity of the local features; the set of matches is then pruned by excluding geometrically inconsistent candidates. Popular consistency criteria include four-points congruent sets [2] and minimum pairwise Euclidean distortion [32]. Machine learning techniques have also been proposed to learn optimal local features in the presence of clutter [48], or to learn the consistency function itself [8, 16].

Despite their excellent performance in several computer vision tasks, however, all these methods fail completely when the objects undergo non-rigid deformations.

Point set registration. In the realm of robust point set registration [9], several techniques have been proposed that allow to deal with deformation, occlusion, outliers, and moderate clutter [29, 15, 21, 24]. Differently from the previous approaches, the central focus of these methods lies in the calculation of a parametrized *transformation* relating two given point sets, often represented as probability densities. The point-to-point correspondence is then obtained via softassign as a by-product of the alignment procedure. These methods do not scale well with the size of the point sets (typically limited to a few hundred points). Furthermore, they rely on an initial alignment to be given, and performance degrades drastically with significant clutter.

1.2. Contribution

We propose a method for dense matching of deformable objects in cluttered 3D scenes. To our knowledge, this is the first attempt at solving this problem in a fully deformable setting. Our key contributions are summarized as follows:

- We propose a data-driven feature learning approach to derive low-dimensional feature descriptors in the presence of clutter, occlusion, and deformable objects.
- We show how the presence of clutter and missing parts in the scene affect the spectral representation of the correspondence, and use it as a prior to drive the matching process.
- We introduce for the first time a complete pipeline for matching deformable objects in clutter.

The rest of the paper is organized as follows. In Section 2 we overview the mathematical preliminaries in spectral geometry and functional correspondence. Section 3 describes our method, while Section 4 gives the implementation details. Experimental results and applications are presented in Section 5, and finally, Section 6 concludes the paper.

2. Background

We model shapes as two-dimensional Riemannian manifolds $\mathcal{M}$ (possibly with boundary) equipped with an intrinsic distance function $d_{\mathcal{M}}$ and the standard area element $\mathrm{d}\mu$. The intrinsic gradient $\nabla_{\mathcal{M}}$ and Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ generalize the corresponding notions from Euclidean spaces to manifolds. In analogy to the Euclidean case, the Laplacian $\Delta_{\mathcal{M}}$ provides us with the means to extend Fourier analysis to manifolds; it admits an eigen-decomposition

$$\Delta_{\mathcal{M}}\phi_i(x) = \lambda_i\phi_i(x) \qquad x \in \mathrm{int}(\mathcal{M}) \tag{1}$$
$$\langle \nabla_{\mathcal{M}}\phi_i(x), \hat{n}(x)\rangle = 0 \qquad x \in \partial\mathcal{M}, \tag{2}$$

with Neumann boundary conditions (2), where $\hat{n}$ is the normal vector to the boundary. Here, $0 = \lambda_1 \leq \lambda_2 \leq \dots$ are eigenvalues and $\phi_1, \phi_2, \dots$ are the corresponding eigenfunctions. Note that due to the isometry invariance of the Laplacian, nearly-isometric shapes will have approximately the same eigenfunctions (up to sign) and eigenvalues. Further note that, since in our setting we deal with shapes made of multiple connected components (*i.e*., 3D scenes), the eigenvalue $\lambda_1 = 0$ can have high multiplicity.

Since the eigenfunctions of the Laplacian form an orthonormal basis of $L^2(\mathcal{M}) = \{f : \mathcal{M} \to \mathbb{R} \mid \int_{\mathcal{M}} f^2 d\mu < \infty\}$, *i.e*., the space of square-integrable functions on the manifold $\mathcal{M}$, any function $f \in L^2(\mathcal{M})$ can be represented via the Fourier series expansion

$$f(x) \quad = \quad \sum_{i\geq 1} \langle f, \phi_i \rangle_{\mathcal{M}} \phi_i(x) \,, \tag{3}$$

where we use the standard $L^2(\mathcal{M})$ inner product defined as $\langle f, g\rangle_{\mathcal{M}} = \int_{\mathcal{M}} fg d\mu$.

Functional correspondence. We build our matching pipeline upon the functional maps framework of Ovsjanikov *et al*. [30]. The main idea is to identify correspondences between shapes by a linear operator $T : L^2(\mathcal{M}) \to L^2(\mathcal{N})$, mapping functions on $\mathcal{M}$ to functions on $\mathcal{N}$. This can be seen as a generalization of classical point-to-point matching, as the latter constitutes a special case where one maps delta functions to delta functions.

Since T is a linear operator, it admits a matrix representation $\mathbf{C} = (c_{ij})$ with coefficients computed as follows. Let $\{\phi_i\}_{i\geq 1}$ and $\{\psi_i\}_{i\geq 1}$ be orthogonal bases respectively on $L^2(\mathcal{M})$ and $L^2(\mathcal{N})$, and let $f \in L^2(\mathcal{M})$ be arbitrary. Then

$$\begin{aligned} Tf &= T\sum_{i\geq 1}\langle f,\phi_i\rangle_{\mathcal{M}}\phi_i = \sum_{i\geq 1}\langle f,\phi_i\rangle_{\mathcal{M}} T\phi_i \\ &= \sum_{ij\geq 1}\langle f,\phi_i\rangle_{\mathcal{M}} \underbrace{\langle T\phi_i,\psi_j\rangle_{\mathcal{N}}}_{c_{ji}}\psi_j\,. \end{aligned} \tag{4}$$

A particularly convenient choice for the bases $\{\phi_i\}_{i\geq 1}$, $\{\psi_i\}_{i\geq 1}$ is given by the Laplacian eigenfunctions on the two shapes, as originally proposed in [30]. By analogy with Fourier analysis, this choice allows to truncate the series (4) after the first k coefficients as a low-pass approximation of the original map, giving rise to a $k \times k$ matrix $\mathbf{C}$ encoding the functional correspondence. Further, if the functional map T is built on top of a near-isometry, one obtains $c_{ij} = \langle T\phi_i, \psi_j\rangle_{\mathcal{N}} \approx \pm\delta_{ij}$ since near-isometric shapes have corresponding eigenfunctions (up to sign). This results in matrix $\mathbf{C}$ being diagonally dominant, since $c_{ij} \approx 0$ if $i \neq j$.

Partial functional correspondence. Let us now assume to be given a full shape $\mathcal{M}$ and a *partial* shape $\mathcal{N}$ that is approximately isometric to some (unknown) sub-region $\mathcal{M}' \subset \mathcal{M}$. Recently, Rodolà *et al*. [34] showed that for each "partial" eigenfunction ψ_j of $\mathcal{N}$ there exists a corresponding "full" eigenfunction ϕ_i of $\mathcal{M}$ for some $i \geq j$, such that $c_{ij} = \langle T\phi_i, \psi_j\rangle_{\mathcal{N}} \approx \pm 1$, and zero otherwise. Note that differently from the previous case (full-to-full), where the approximate equality holds for $i = j$, here the inequality $i \geq j$ induces a *slanted*-diagonal structure on matrix $\mathbf{C}$. In particular, the authors showed that the angle of the diagonal can be directly and conveniently estimated from the area ratio of the two surfaces. The precomputed angle is then used as a prior on $\mathbf{C}$ to drive the matching process [34].

3. Our method

Our setting greatly differs from the full-to-full [30, 31] and part-to-full [34] cases, in that we allow both shapes to have missing parts and additional *clutter*.

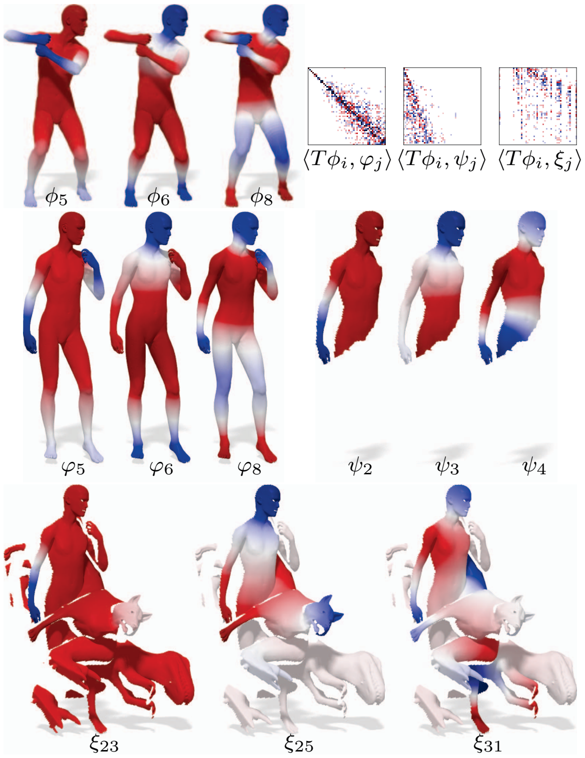


Figure 2. Correspondence among Laplacian eigenfunctions in different scenarios. The model (top left) is matched to a full shape, a partial shape, and a cluttered scene respectively. The resulting matrices $\mathbf{C}$, which encode the correspondence among eigenfunctions, are shown on the top right. In the traditional *full-to-full* setting, the eigenfunctions have a 1-1 index correspondence among model and query (middle left). In the *part-to-full* setting, each eigenfunction of the part has a corresponding index on the model, but not vice-versa (middle right). In the presence of clutter, only a sparse subset of eigenfunctions on the scene have an approximately corresponding eigenfunction on the model (bottom).

Problem statement. Input to our method is a 3D object $\mathcal{M}$ (the *model*) and a 3D surface $\mathcal{S}$ (the *scene*) in which the model *may* appear up to deformation. The scene $\mathcal{S}$ may contain additional clutter, and only a partial view of the model is captured; conversely, the model $\mathcal{M}$ is clutter-free, but only part of it is matchable to the scene. We aim to determine a subset of the scene that is approximately isometric to some sub-region of the model. The output of our method consists of (i) approximately isometric parts $\mathcal{M}' \subseteq \mathcal{M}$, $\mathcal{S}' \subseteq \mathcal{S}$, and (ii) a functional map $T : L^2(\mathcal{M}') \to L^2(\mathcal{S}')$ encoding the correspondence between the parts. We represent each part as a binary indicator function on the respective shape, and call these the *segmentation functions* [46].

Method overview. Our matching pipeline consists of three major stages, as summarized below.

- In the first stage, we introduce a local feature learning approach for deformable object-in-clutter. This proved to be a necessary step, due to the lack of robust point descriptors for this challenging task (Section 3.1);

- The learned descriptors are used to initialize a L_1 variant of a quadratic assignment problem. The output of this step is a sparse collection of few, possibly noisy point-to-point matches (Section 3.2);
- Finally, we show how to extend the functional maps framework to deal with partiality and clutter. The sparse matches from the previous step are used as an initialization, while the learned descriptors are used as a data term (Section 3.3).

3.1. Descriptor learning

Point descriptors are ubiquitous tools in shape analysis and 3D computer vision. They can be broadly classified into two families: rotation-invariant local descriptors, usually robust to clutter and missing parts [3], and isometry-invariant descriptors [39, 4], designed to be robust to non-rigid transformations. Bridging the gap between the two is an open challenge that has so far eluded analysis: To date, no descriptor is capable to deal with clutter, missing parts, and non-rigid deformations simultaneously. In this paper we propose a learning-based approach, motivated by the success of recent methods such as [23, 47, 25, 5, 48].

Metric learning. Let $\{\mathcal{M}_i\}$ be a collection of shapes, with points $x \in \mathcal{M}_i$ represented as d-dimensional vectors corresponding to some initial choice of an input descriptor. Our task is to learn an embedding function $F(x)$ onto some latent descriptor space, where similar points (matches) lie close to each other, while dissimilar points (mismatches) are separated by a safe margin. Similarly to [25], we model the learning process upon the network architecture for metric learning of [6, 13], which has been shown to easily adapt to various computer vision tasks [20, 41, 26].

Assume to be given a set of knowingly similar and dissimilar pairs of points, respectively S and D, and let $F_\Theta(x)$ be modeled as a deep neural network with trainable parameters Θ. We minimize the *siamese* loss function [13]:

$$L_{\mathrm{s}}(\Theta) = \sum_{x,x^+ \in S} \gamma \|F_\Theta(x) - F_\Theta(x^+)\|_2^2 + \sum_{x,x^- \in D} (1-\gamma)(m_s - \|F_\Theta(x) - F_\Theta(x^-)\|_2)_+^2 \quad (5)$$

where γ is a trade-off parameter, m_s is the margin, and $(x)_+ = \max(0, x)$.

For increased robustness we additionally consider a global distance distribution penalty [20, 40], namely:

$$L_{\mathrm{g}}(\Theta) = \sigma_\Theta^+ + \sigma_\Theta^- + (m_g + \mu_\Theta^+ - \mu_\Theta^-)_+ \,. \quad (6)$$

This term enforces the distribution of positive and negative distances to have small variances $\sigma_\Theta^+, \sigma_\Theta^-$ (peaked distributions), and means $\mu_\Theta^+, \mu_\Theta^-$ which are at least m_g apart. The complete loss function is thus given by: $L(\Theta) = L_{\mathrm{s}}(\Theta) + L_{\mathrm{g}}(\Theta)$.

Boundary effects. It is worth mentioning that, differently from previous work [43, 3, 32], in this paper we do *not* avoid vertices close to shape boundary, *i.e.*, all model and scene points are treated equally across the entire pipeline. We did not observe any undesired boundary effect resulting from this choice, and were able to match boundary points accurately (see Fig. 7 for some examples). We attribute this behavior to the regularizing effect of the functional representation (Section 3.3).

3.2. Sparse matching

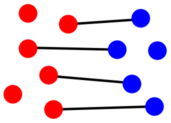
We consider the following L_1-regularized version of the quadratic assignment problem (QAP):

$$\max_{\mathbf{x} \geq 0} \mathbf{x}^\top \mathbf{S} \mathbf{x} \quad \text{s.t. } \mathbf{x}^\top \mathbf{1} = 1 \,, \quad (7)$$

where matrix $\mathbf{S} \in \mathbb{R}_+^{q \times q}$ encodes the compatibility among pairs of candidate matches (see Eq. (8)), while $\mathbf{x} \in [0,1]^q$ is a weight vector for the set of q candidate matches. Note that problem (7) allows the two shapes to have different point density, as the L_1 relaxation does not enforce bijectivity. Candidate matches are constructed by considering pairs $(x, y) \in \mathcal{M} \times \mathcal{S}$ having similar descriptors (see Section 4). As a compatibility function for matrix $\mathbf{S}$, we use:

$$s((x,y),(x',y')) = \exp(-\sigma(d_{\mathcal{M}}(x,x') - d_{\mathcal{S}}(y,y'))^2) \,, \quad (8)$$

where $\sigma > 0$ is a weight and $d_{\mathcal{M}}$ and $d_{\mathcal{S}}$ denote the geodesic distance functions on the respective shapes. Note that $s = 1$ whenever $d_{\mathcal{M}}(x,x') = d_{\mathcal{S}}(y,y')$, whereas $s < 1$ for different distance values. In other words, Eq. (8) encodes the metric distortion of the input pair of candidate matches.



The L_1 constraints in (7) have the effect of producing sparse solutions [33], a desirable outcome given the large presence of unmatchable parts (clutter) in our problem. Further, by following a simple strategy in the construction of $\mathbf{S}$ [37, 32], local solutions to (7) are *guaranteed* to be injective, *i.e.*, no one-to-many matches appear at the optimum (see inset for an example).

Problem (7) can be regarded as an extension of the approaches described in [33, 32] to deformable object-in-clutter. The output of this step is a sparse set of (possibly noisy) one-to-one matches $(x, y) \in \mathcal{M} \times \mathcal{S}$ with minimum metric distortion.

3.3. Functional correspondence

We define our deformable object-in-clutter problem by using the functional maps formalism. The following observation is crucial throughout the paper, and is at the core of our formulation:

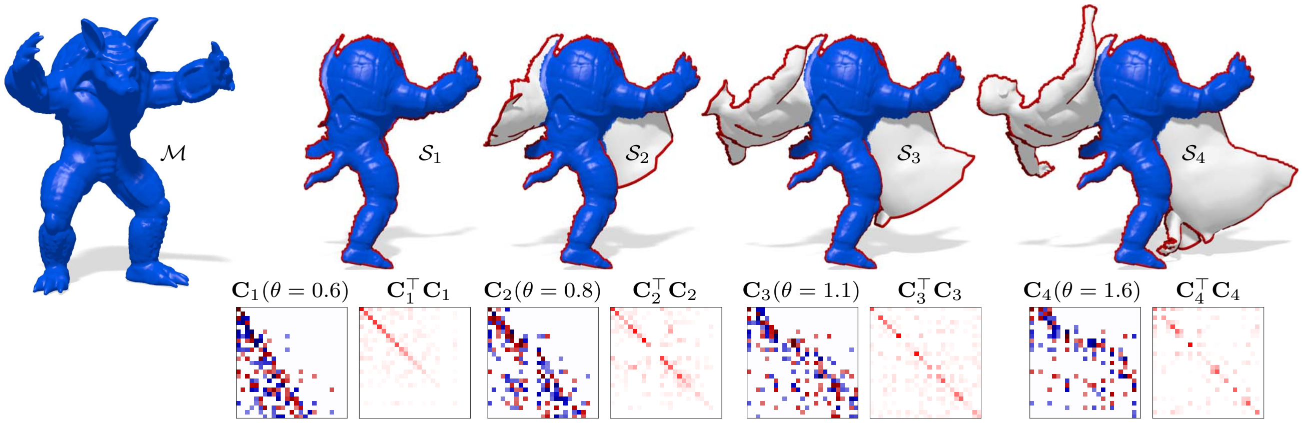

$\mathcal{M}$
$\mathcal{S}_1$
$\mathcal{S}_2$
$\mathcal{S}_3$
$\mathcal{S}_4$
$\mathbf{C}_1(\theta = 0.6)$
$\mathbf{C}_1^\top \mathbf{C}_1$
$\mathbf{C}_2(\theta = 0.8)$
$\mathbf{C}_2^\top \mathbf{C}_2$
$\mathbf{C}_3(\theta = 1.1)$
$\mathbf{C}_3^\top \mathbf{C}_3$
$\mathbf{C}_4(\theta = 1.6)$
$\mathbf{C}_4^\top \mathbf{C}_4$

Figure 3. Functional maps at increasing amounts of clutter. The model $\mathcal{M}$ is matched to scenes $\mathcal{S}_1$-$\mathcal{S}_4$, giving rise to the matrices of spectral coefficients $\mathbf{C}_1$-$\mathbf{C}_4$. Observe how the dominant slope of $\mathbf{C}_i$ (denoted by θ) varies with clutter, moving from the lower- to the upper-triangular part of the matrix. The rank of $\mathbf{C}_i$ decreases as more and more clutter is introduced, a fact that is manifested in empty rows and columns in $\mathbf{C}_i$, and in the sparse diagonal structure on $\mathbf{C}_i^\top \mathbf{C}_i$. The zero-clutter pair $(\mathcal{M}, \mathcal{S}_1)$ is the setting considered in [34].

Motivation. In the presence of clutter, it is still possible to find eigenfunctions ϕ_i on $\mathcal{M}$ for *some* indices i, having corresponding eigenfunctions ψ_j on $\mathcal{S}$ for some indices j. There is a key difference with what we have seen in Section 2: While in the full-to-full case we had correspondence for $i = j$ and in the part-to-full case for $i \geq j$, here the correspondence among indices cannot be reliably predicted. The diagonal slant of $\mathbf{C}$, which identifies the pairs (i, j) for which $c_{ij} = \langle T\phi_i, \psi_j \rangle_{\mathcal{S}} \neq 0$, is now an *unknown* that we need to optimize for. In particular, we expect $c_{ij} \neq 0$ only for a sparse set of indices, *i.e.*, matrix $\mathbf{C}$ will have empty rows and columns. See Figs. 2 and 3 for examples.

It is worth mentioning that the slant of $\mathbf{C}$ directly encodes the amount of overlap between model and scene [34], hence providing an important prior for the correspondence. Litany *et al.* [22] recently showed (for non-rigid puzzles) that the slant can be simply estimated as the area ratio of the two objects, in our case $\frac{\text{area}(\mathcal{M})}{\text{area}(\mathcal{S})}$. However, this property fails to hold when the amount of clutter is significant.

Functional object-in-clutter. Let $\Phi \in \mathbb{R}^{|\mathcal{M}| \times k}$, $\Psi \in \mathbb{R}^{|\mathcal{S}| \times k}$ be two matrices containing as columns the first k Laplacian eigenfunctions of $\mathcal{M}$ and $\mathcal{S}$ respectively, and let matrices $\mathbf{F} \in \mathbb{R}^{|\mathcal{M}| \times d}$, $\mathbf{G} \in \mathbb{R}^{|\mathcal{S}| \times d}$ contain dense d-dimensional descriptor fields on model and scene. Our aim is to solve for the functional map $\mathbf{C}$ between $\mathcal{M}$ and $\mathcal{S}$, the angle $\theta \in \mathbb{R}$ encoding the diagonal slope of $\mathbf{C}$, and the (soft) segmentation functions $u : \mathcal{M} \to [0, 1]$, $v : \mathcal{S} \to [0, 1]$ identifying the corresponding regions on model and scene. We therefore consider the unconstrained problem:

$$\min_{\mathbf{C}, \theta, u, v} \|\mathbf{C}\mathbf{A}(\eta(u)) - \mathbf{B}(\eta(v))\|_{2,1} \tag{9}$$

$$+\|\mathbf{C}\Phi^\top \eta(u) - \Psi^\top \eta(v)\|_2^2, \tag{10}$$

$$+\rho_{\text{corr}}(\mathbf{C}, \theta) + \rho_{\text{part}}(u, v)\,. \tag{11}$$

Data term. Here, $\mathbf{A}(\eta(u))$, $\mathbf{B}(\eta(v)) \in \mathbb{R}^{k \times d}$ contain the spectral coefficients of $\mathbf{F}$ and $\mathbf{G}$ masked by the respective segmentations u and v, *i.e.*, for each column of $\mathbf{A}(\eta(u))$ we write $\mathbf{a}_i = \Phi^\top(\eta(u) \circ f_i)$, and similarly for $\mathbf{B}$. Note that we apply the saturation function $\eta(t) = \frac{1}{2}(\tanh(2t - 1) + 1)$ in order to keep the range of u, v within $[0, 1]$ [34]. As descriptor fields, we use the learned 32-dimensional descriptors of Section 3.1, in addition to indicator functions supported at the few matches obtained as in Section 3.2. The $L_{2,1}$ norm allows to handle possible mismatches arising from the QAP. The L_2 summand in the data term simply asks for the functional map to correctly transfer the segmentation functions.

Regularizer for C. We adopt the following regularization terms for the correspondence:

$$\rho_{\text{corr}}(\mathbf{C}, \theta) = \mu_1 \sum_{i \neq j} (\mathbf{C}^\top \mathbf{C})_{ij}^2 + \mu_2 \sum_i |\mathbf{C}^\top \mathbf{C}|_{ii} \tag{12}$$

$$+ \mu_3 \|\mathbf{C} \circ \mathbf{W}(\theta)\|_{\text{F}}^2\,. \tag{13}$$

The μ_1- and μ_2-terms require $\mathbf{C}^\top \mathbf{C}$ to be as diagonal as possible, with *sparse* diagonal. This induces empty rows and columns in $\mathbf{C}$, hence reinforcing its slanted diagonal structure (see Fig. 3). In addition, the two terms promote area preservation, as for area-preserving functional maps it has been shown that $\mathbf{C}^\top \mathbf{C} = \mathbf{I}$ [30, 34]. Note however that this identity only holds in the full-to-full case: Since our $\mathbf{C}$ has empty rows and columns, we only require the identity to hold approximately at a sparse set of elements. This corresponds to requiring the matched parts to have equal area. Finally, $\mathbf{W}(\theta)$ is a diagonal mask parametrized on the slope θ, requiring a similar structure on $\mathbf{C}$ (here $\circ$ is element-wise product). See Section 4 for the implementation details.

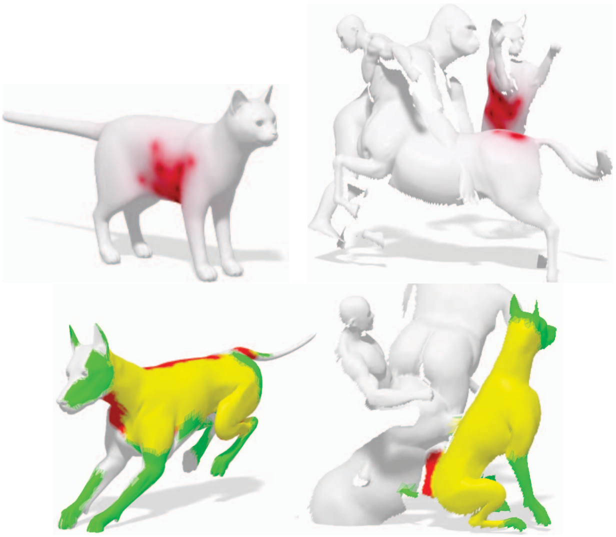


Figure 4. *Top*: Initialization of u and v from sparse matches. *Bottom*: Model and scene parts detected by our method. We show ground truth (green), detected (red), and the intersection (yellow).

Regularizer for u, v. For part regularization we use:

$$\begin{aligned}\rho_{\text{part}}(u,v) = \mu_4 \left(\int_{\mathcal{M}} \|\nabla_{\mathcal{M}} \eta(u)\| dx + \int_{\mathcal{S}} \|\nabla_{\mathcal{S}} \eta(v)\| dx \right) \\ + \mu_5 \left(\int_{\mathcal{M}} \eta(u) dx - \int_{\mathcal{S}} \eta(v) dx \right)^2 \\ - \mu_6 \left(\int_{\mathcal{M}} \eta(u) dx + \int_{\mathcal{S}} \eta(v) dx \right) . \qquad (14)\end{aligned}$$

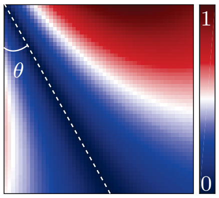
The μ_4-terms encode the boundary length of the segmentation functions on the respective shapes (here $\nabla_{\mathcal{M}}$ denotes the intrinsic gradient operator on $\mathcal{M}$ and similarly for $\mathcal{S}$), following the spirit of the Mumford-Shah functional [28]; penalizing boundary length has the effect of producing contiguous regions, as expected in our setting. The μ_5-term requires the two parts to have same area, while the μ_6-term controls the size of the parts (a large weight will promote large areas and viceversa). Note that this term is necessary in order to avoid the trivial solution $u = 0, v = 0, \mathbf{C} = \mathbf{0}$.

4. Implementation

Descriptor learning. We use 544-dimensional SHOT descriptors [43] as input feature vectors (other choices are possible), and output 32-dimensional dense descriptor fields. We model the function F_Θ as a deep feed-forward network with 3 highway blocks B of 5 fully-connected layers each (also known as res-net) [38, 14]. Each layer implements $y = \sigma(Wx + b)$, where σ is the point-wise ReLU function and W, b are optimization variables, initialized randomly. The chosen network architecture is $B128$ – $B64$ – $B32$, where BN denotes a block of 5 layers with ReLU activation and dimensionality equal to N. The parameters of the loss (5), (6) are set to $m_g = 1$, $m_s = 5$, and $\gamma = 0.5$. Training is performed with batches of 1K samples, uniformly drawn from D and S. The network was modeled in Tensorflow [1], and learning was performed using RMSProp [42].

Sparse matching. To keep problem (7) tractable, we only consider 1000 equally spaced samples (according to the Euclidean metric) on $\mathcal{M}$ and $\mathcal{S}$. For each sampled $y \in \mathcal{S}$, candidate matches are constructed by selecting the 10 closest samples in $\mathcal{M}$ in descriptor space. This results in 10^4 well-spread candidate matches in total. A value of $\sigma = 10^{-2}$ is set in Eq. (8) for computing $\mathbf{S}$. Since in practice only about half of each model is depicted in the scenes, matches that are more than $\frac{1}{2}\text{diam}(\mathcal{M})$ apart are prohibited.

Problem (7) is solved via the infection-immunization algorithm [37]. We follow a multi-start strategy: For each dimension $i = 1, \ldots, q$ we set the entry $\mathbf{x}_i^0 = 1$ and the remaining entries to zero (this concentrates all the mass on a vertex of the simplex), and run the optimization starting at this point. We do so for all dimensions, and keep the solution with largest objective. The final solution is fed to the elastic net solver [36] to increase the number of matches.



Functional correspondence. The Laplace-Beltrami operator is discretized using the classical cotangent scheme [27]; we used $k = 100$ eigenfunctions for the spectral representation of the functional map. The diagonal mask $\mathbf{W}(\theta)$ is constructed using the formula $w_{ij} = (j - i\theta)(1 + \theta^2)e^{-\sigma\sqrt{i^2+j^2}}$, where the exponential term (we use $\sigma = 0.1$) is responsible for the off-diagonal spread at higher frequencies (see inset for $|\mathbf{W}(\theta)|$). Problem (9) is minimized by block-coordinate descent, alternating between $\{\mathbf{C}, \theta\}$ and $\{u, v\}$. Each block is updated via non-linear conjugate gradient. We initialize $\mathbf{C}$ as a matrix of zeros, θ as the area ratio $\frac{\text{area}(\mathcal{M})}{\text{area}(\mathcal{S})}$ [22], and functions u, v as sums of Gaussians supported on the initial matches (see Fig. 4 top). We used the following weights: $\mu_1 = 5$; $\mu_2 = 1$; $\mu_3 = 0.1$; $\mu_4 = 100$; $\mu_5 = 10$; $\mu_6 = 1$.

Conversion to point-to-point map. The found functional map $\mathbf{C}^*$ is converted to a pointwise map using the standard nearest-neighbor approach of [30]. Specifically, for each point $y \in \mathcal{S}$ such that $v^*(y) \approx 1$ we consider its corresponding column in $\Psi^\top$, and look for the closest column in $\mathbf{C}^*\Phi^\top$ in the L_2 sense. If the point $x \in \mathcal{M}$ associated to this column is such that $u^*(x) \approx 1$, this is marked as the matching point; otherwise, y is left unmatched. Note that more sophisticated recovery methods such as [35, 45] cannot be applied due to the presence of clutter.

5. Results

Data. Due to the difficulty of producing a real-world dataset with reliable ground-truth, in our experiments we employ

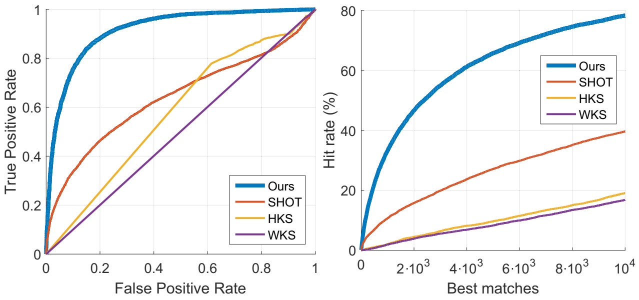


Figure 5. ROC (left) and CMC (right) curves for our 32-dim descriptors, 100-dim HKS/WKS, and 544-dim SHOT. Learning clearly improves the quality of the descriptors while drastically reducing their dimensionality.

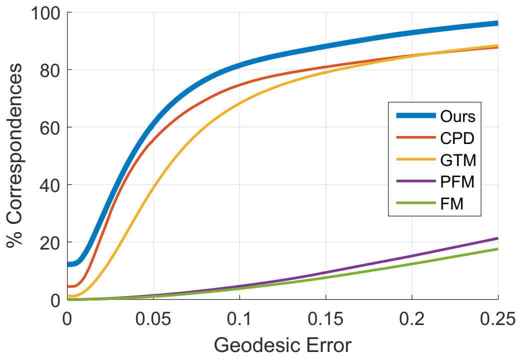


Figure 6. Comparisons with top-performing methods in shape matching (FM), partial matching (PFM), rigid object-in-clutter (GTM), and point set registration (CPD). Note that the geodesic error only measures accuracy within the *correctly* detected parts, *i.e.*, matches that do not hit the correct scene part are not counted.

	Ours	**FM** [30]	**PFM** [34]	**GTM** [32]	**CPD** [29]
$\mathcal{M}$	0.69 / 0.48	0.31 / 1.00	0.30 / 1.00	0.60 / 0.32	0.72 / 0.21
$\mathcal{S}$	0.76 / 0.54	0.30 / 1.00	0.30 / 0.81	0.72 / 0.44	0.75 / 0.25

Table 1. Detection accuracy of each method (Precision / Recall). Note that the high recall of FM and PFM is due to the two methods matching the *entire* scene to the model, giving $|\mathcal{S}' \cap \mathcal{S}'_{\text{gt}}| = |\mathcal{S}'_{\text{gt}}|$. All methods are dense except for CPD ($\sim$3K matches per scene).

a synthetic dataset, introduced in [32] for rigid object-in-clutter. Since the dataset also makes use of deformable objects from TOSCA [7], in order to adapt the data to our non-rigid setting we replace the rigid query models with deformed versions thereof.

The dataset is composed of 35 synthetic scenes captured from arbitrary view points with a virtual camera, and each scene contains 3 to 5 objects. Additional 115 scenes are used for descriptor training. The complete model set is composed of 16 rigid plus 3 non-rigid object classes, and the latter are used as queries in our evaluation (*cat, dog, centaur*). We consider 5 queries (near-isometric deformations) for each of the 3 classes and match them towards the scenes containing that class, resulting in 150 matching problems in total. Note that we use *different* model deformations for the training and test sets.

In order to avoid identical meshings and make the dataset more challenging, each scene and model is independently remeshed to 10K vertices by edge contractions [11].

Evaluation measures. We define two error measures:

Geodesic error. We measure correspondence quality according to the Princeton benchmark protocol [17]. For each point $y \in \mathcal{S}$ that belongs to the query model, assume to be given a match $(x, y) \in \mathcal{M} \times \mathcal{S}$, whereas the ground-truth correspondence is (x_{gt}, y). Then, the inaccuracy of the match is measured by the (normalized) geodesic error:

$$\epsilon(y) \quad = \quad \frac{d_{\mathcal{M}}(x, x_{\text{gt}})}{\text{area}(\mathcal{M})^{1/2}}\,, \tag{15}$$

and has units of normalized geodesic length on $\mathcal{M}$ (ideally, zero). The value $\epsilon(y)$ is averaged over all instances $(\mathcal{M}, \mathcal{S})$. We plot cumulative curves showing the percent of matches which have error smaller than a variable threshold.

Note that if an algorithm produces a match (x, y) where $y \in \mathcal{S}$ does not actually correspond to a model point, then the error $\epsilon(y)$ is undefined as there exists no ground-truth match (x_{gt}, y). For this reason, we can only measure the geodesic error for portions of the scene that actually belong to the model.

Precision-Recall. Detection accuracy is quantified with standard Precision vs Recall curves. Let $\mathcal{S}', \mathcal{S}'_{\text{gt}} \in \mathcal{S}$ be respectively the detected and ground-truth parts in the scene, and let $|\cdot|$ denote area. Precision and Recall are defined as the area ratios $P = \frac{|\mathcal{S}' \cap \mathcal{S}'_{\text{gt}}|}{|\mathcal{S}'|}$ and $R = \frac{|\mathcal{S}' \cap \mathcal{S}'_{\text{gt}}|}{|\mathcal{S}'_{\text{gt}}|}$. In words, Precision measures the percentage of detected area that is correct; Recall measures the percentage of ground-truth area that is captured by the detection (Fig. 4 bottom).

5.1. Local features

The learned 32-dimensional descriptors are evaluated using *receiver operating characteristic* (ROC) and *cumulative matching characteristic* (CMC) curves. We compare against 544-dim SHOT [43], 100-dim HKS [39] and 100-dim WKS [4] using the settings proposed by the authors.

Fig. 5 (left) shows ROC curves (true vs. false positive rate) computed on 10K random point pairs drawn from the test set (the ratio of positive and negative random examples is approximately uniform). In Fig. 5 (right) we show the CMC curves. Each curve evaluates the probability (y-axis) of finding the correct match within the first k best matches (x-axis). Matches are obtained as L_2-nearest neighbors in descriptor space. We see that our learned descriptors are much more accurate, in particular at small k, allowing us to construct good candidate matches in the QAP step.

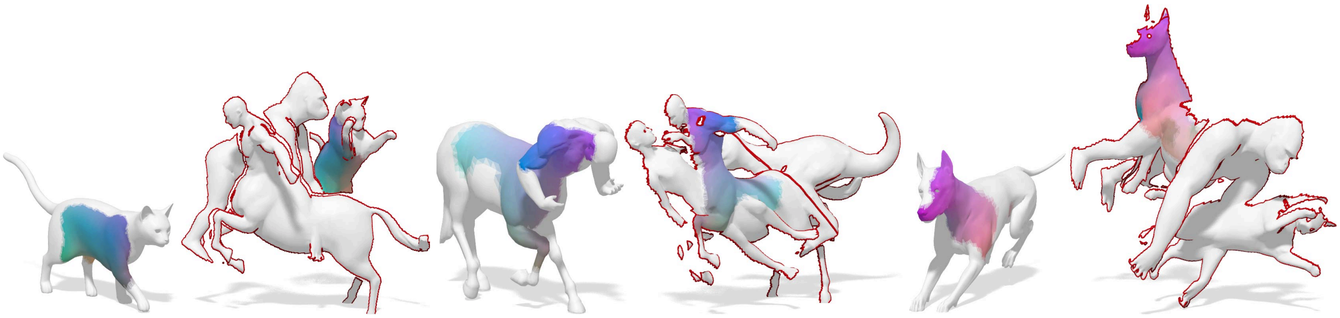


Figure 7. Some solutions obtained by our pipeline for deformable object-in-clutter. Corresponding points have same color; white color denotes no match. Observe how the correspondence is accurate also for points close to scene boundary.

5.2. Comparisons

While there is an abundance of methods for rigid matching with clutter, to our knowledge there are no existing pipelines tackling our non-rigid setting. In order to position us within the current landscape, in Fig. 6 and Table 1 we compare with state-of-the-art methods in shape matching (FM) [30], deformable partial matching (PFM) [34], rigid object-in-clutter (GTM) [32], and point set registration (CPD) [29]. For a fair comparison, FM and PFM are provided with the initial sparse matches produced by our pipeline, and GTM uses our learned descriptors. The rigid transformation computed by GTM is used as initial alignment for CPD. Qualitative examples for our method are shown in Fig. 7; a failure case is shown in Fig. 8.

As can be seen from the plots, FM and PFM are unable to deal with the presence of clutter (by design). The low accuracy of PFM is motivated by its area-preservation requirement [34]. The relatively high accuracy of GTM spurs from the presence of piecewise-rigid deformations in the dataset: Rigid parts (*e.g.*, the dog head) are matched quite accurately, and CPD refines the GTM solutions even further.

In the light of these results we would like to emphasize that our approach is, at its heart, a *spectral* method – given the challenging setting we considered in this paper, we find the accuracy achieved by our method quite remarkable.

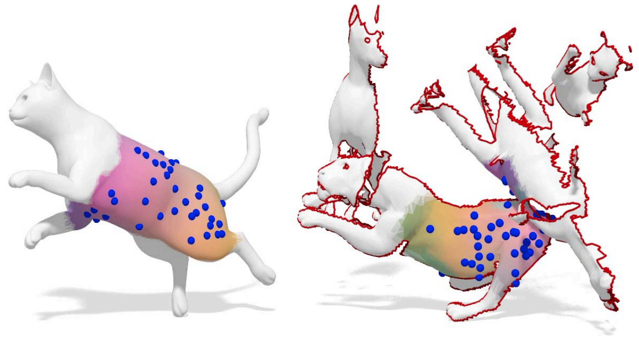


Figure 8. A typical failure case of our method. The cat model is matched to a lion due to similar local appearance and approximate isometry. The blue dots represent the initial QAP matches.

5.3. Runtime

We implemented our method in C++/Matlab[1] and executed on an Intel i7-5820K 3.30GHz cpu with 6 cores.

The sparse matching step takes ∼10s on average, comprehensive of similarity computation and multi-start optimization of (7). Solving for the functional correspondence and for the parts takes ∼1m including conversion to point-to-point map. The average end-to-end runtime of our pipeline is thus less than 2m per matching problem.

6. Discussion and conclusions

We introduced a novel matching pipeline for the deformable object-in-clutter problem, and showed how a combination of feature learning, sparse point matching, and functional correspondence can lead to accurate results. A peculiar aspect of our approach lies in its “hybrid” nature: It makes use of spatial- as well as frequency-domain tools to drive the matching, and we demonstrated how the two aspects complement each other. We surmise that this combination of techniques is a necessary means to tackle realistic settings, where disparate forms of nuisance are present.

Limitations. The main limitation of our method lies in its reliance on local features to initialize the pipeline. While this is a common trait to most existing approaches, our learning-based method requires the availability of labeled data that can be difficult to produce. Defining robust, *local*, isometry-invariant features in a purely unsupervised fashion is a particularly interesting future research direction.

Second, the presence of clutter and missing parts affects the computation of geodesics (Eq. (8)), resulting in distorted distance values and thus wrong initial matches in heavily occluded scenes. Geodesic distances also prohibit us from matching all visible parts of a model if this is fragmented in multiple disconnected components – this is not a problem for *rigid* matching, where Euclidean distance is used.

Acknowledgments. The authors wish to thank O. Litany and V. Ganapathi-Subramanian for useful discussions. ER, JM and MB are supported by the ERC Starting Grant No. 307047 (COMET), Google Faculty Research Award, and nVidia equipment grant.

[1] Code and data are available for download at http://www.dais.unive.it/~cosmo/deformableclutter/

References

[1] M. Abadi, A. Agarwal, P. Barham, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. *Software available from tensorflow.org*, 2015. 6

[2] D. Aiger, N. J. Mitra, and D. Cohen-Or. 4-points congruent sets for robust pairwise surface registration. *TOG*, 27(3):85, 2008. 2

[3] L. Alexandre. 3d descriptors for object and category recognition: a comparative evaluation. In *Proc. IROS Workshops*, pages 1–6, October 2012. 2, 4

[4] M. Aubry, U. Schlickewei, and D. Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *ICCV Workshops*, pages 1626–1633, 2011. 4, 7

[5] D. Boscaini, J. Masci, E. Rodolà, M. M. Bronstein, and D. Cremers. Anisotropic diffusion descriptors. *Computer Graphics Forum*, 35(2), 2016. 4

[6] J. Bromley et al. Signature verification using a "Siamese" time delay neural network. In *Proc. NIPS*. 1994. 4

[7] A. Bronstein, M. Bronstein, and R. Kimmel. *Numerical Geometry of Non-Rigid Shapes*. Springer, 2008. 7

[8] T. S. Caetano, J. J. McAuley, L. Cheng, Q. V. Le, and A. J. Smola. Learning graph matching. *TPAMI*, 31(6):1048–1058, June 2009. 2

[9] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *Proc. CVPR*, volume 2, pages 44–51, 2000. 2

[10] L. Cosmo, E. Rodolà, M. M. Bronstein, et al. SHREC'16: Partial matching of deformable shapes. In *Proc. 3DOR*, 2016. 2

[11] M. Garland and P. S. Heckbert. Surface simplification using quadric error metrics. In *Proc. SIGGRAPH*, pages 209–216, 1997. 7

[12] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan. 3d object recognition in cluttered scenes with local surface features: A survey. *TPAMI*, 36(11):2270–2287, Nov 2014. 2

[13] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Proc. CVPR*, 2006. 4

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015. 6

[15] B. Jian and B. C. Vemuri. Robust point set registration using gaussian mixture models. *TPAMI*, 33(8):1633–1645, Aug 2011. 2

[16] A. Kanezaki, E. Rodolà, D. Cremers, and T. Harada. Learning similarities for rigid and non-rigid object detection. In *Proc. 3DV*, volume 1, pages 720–727, Dec 2014. 2

[17] V. G. Kim, Y. Lipman, and T. A. Funkhouser. Blended intrinsic maps. *TOG*, 30(4):79, 2011. 7

[18] A. Kovnatsky, M. Bronstein, A. Bronstein, K. Glashoff, and R. Kimmel. Coupled quasi-harmonic bases. *Comput. Graph. Forum*, 32(2pt4):439–448, 2013. 2

[19] A. Kovnatsky, M. M. Bronstein, X. Bresson, and P. Vandergheynst. Functional correspondence by matrix completion. In *Proc. CVPR*, 2015. 2

[20] B. Kumar, G. Carneiro, and I. Reid. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. *arXiv preprint arXiv:1512.09272*, 2015. 4

[21] W. Lian, L. Zhang, and D. Zhang. Rotation-invariant nonrigid point set matching in cluttered scenes. *TIP*, 21(5):2786–2797, May 2012. 2

[22] O. Litany, E. Rodolà, A. M. Bronstein, M. M. Bronstein, and D. Cremers. Non-rigid puzzles. *Computer Graphics Forum*, 35(5), 2016. 5, 6

[23] R. Litman and A. M. Bronstein. Learning spectral descriptors for deformable shape correspondence. *TPAMI*, (99):1–1, 2014. 4

[24] J. Ma, J. Zhao, and A. L. Yuille. Non-rigid point set registration by preserving global and local structures. *TIP*, 25(1):53–64, Jan 2016. 2

[25] J. Masci, D. Boscaini, M. M. Bronstein, and P. Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. *arXiv:1501.06297*, 2015. 4

[26] J. Masci, M. M. Bronstein, A. Bronstein, and J. Schmidhuber. Multimodal similarity-preserving hashing. *PAMI*, 36(4):824–830, 2014. 4

[27] M. Meyer, M. Desbrun, P. Schröder, and A. H. Barr. Discrete differential-geometry operators for triangulated 2-manifolds. *Visualization&Mathematics*, pages 35–57, 2003. 6

[28] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure and Applied Math.*, 42(5):577–685, 1989. 6

[29] A. Myronenko and X. Song. Point set registration: Coherent point drift. *TPAMI*, 32(12):2262–2275, Dec 2010. 2, 7, 8

[30] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, and L. Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Trans. Graph.*, 31(4):30:1–30:11, July 2012. 1, 2, 3, 5, 6, 7, 8

[31] J. Pokrass, A. M. Bronstein, M. M. Bronstein, P. Sprechmann, and G. Sapiro. Sparse modeling of intrinsic correspondences. *Computer Graphics Forum*, 32(2pt4):459–468, 2013. 2, 3

[32] E. Rodolà, A. Albarelli, F. Bergamasco, and A. Torsello. A scale independent selection process for 3d object recognition in cluttered scenes. *International Journal of Computer Vision*, 102(1-3):129–145, 2013. 2, 4, 7, 8

[33] E. Rodolà, A. Bronstein, A. Albarelli, F. Bergamasco, and A. Torsello. A game-theoretic approach to deformable shape matching. In *Proc. CVPR*, pages 182–189, June 2012. 4

[34] E. Rodolà, L. Cosmo, M. M. Bronstein, A. Torsello, and D. Cremers. Partial functional correspondence. *Computer Graphics Forum*, 2016. 2, 3, 5, 7, 8

[35] E. Rodolà, M. Moeller, and D. Cremers. Point-wise map recovery and refinement from functional correspondence. In *Proc. VMV*, 2015. 6

[36] E. Rodolà, A. Torsello, T. Harada, Y. Kuniyoshi, and D. Cremers. Elastic net constraints for shape matching. In *Proc. ICCV*, pages 1169–1176, December 2013. 6

[37] S. Rota Bulò and I. M. Bomze. Infection and immunization: A new class of evolutionary game dynamics. *Games and Economic Behavior*, 71(1):193–211, January 2011. 4, 6

[38] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In C. Cortes, N. D. Lawrence, D. D.

Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2377–2385. Curran Associates, Inc., 2015. 6

[39] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Proc. SGP*, 2009. 4, 7

[40] J. Svoboda, J. Masci, and M. M. Bronstein. Palmprint identification via discriminative index learning. In *Proc. ICPR*, December 2016. 4

[41] G. W. Taylor, I. Spiro, C. Bregler, and R. Fergus. Learning invariance through imitation. In *Proc. CVPR*, pages 2729–2736. IEEE, 2011. 4

[42] T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012. 6

[43] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In *Proc. ECCV*, pages 356–369, 2010. 4, 6, 7

[44] O. Van Kaick, H. Zhang, G. Hamarneh, and D. Cohen-Or. A survey on shape correspondence. *Computer Graphics Forum*, 30(6):1681–1707, 2011. 1

[45] M. Vestner, R. Litman, A. Bronstein, E. Rodolà, and D. Cremers. Bayesian inference of bijective non-rigid shape correspondence. *arXiv:1607.03425*, 2016. 6

[46] F. Wang, Q. Huang, M. Ovsjanikov, and L. J. Guibas. Unsupervised multi-class joint image segmentation. In *Proc. CVPR*, pages 3142–3149, June 2014. 3

[47] T. Windheuser, M. Vestner, E. Rodolà, R. Triebel, and D. Cremers. Optimal intrinsic descriptors for non-rigid shape analysis. In *Proc. BMVC*, 2014. 4

[48] A. Zeng, S. Song, M. Nießner, M. Fisher, and J. Xiao. 3dmatch: Learning the matching of local 3d geometry in range scans. *arXiv preprint arXiv:1603.08182*, 2016. 2, 4