GFrames: Gradient-Based Local Reference Frame for 3D Shape Matching

Simone Melzi
University of Verona
simone.melzi@univr.it

Riccardo Spezialetti
University of Bologna
riccardo.spezialetti@unibo.it

Federico Tombari
TU Munich
tombari@in.tum.de

Michael M. Bronstein
Imperial College London / USI
m.bronstein@imperial.ac.uk

Luigi Di Stefano
University of Bologna
luigi.distefano@unibo.it

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

Abstract

We introduce GFrames*, a novel local reference frame (LRF) construction for 3D meshes and point clouds. GFrames are based on the computation of the intrinsic gradient of a scalar field defined on top of the input shape. The resulting tangent vector field defines a repeatable tangent direction of the local frame at each point; importantly, it directly inherits the properties and invariance classes of the underlying scalar function, making it remarkably robust under strong sampling artifacts, vertex noise, as well as non-rigid deformations. Existing local descriptors can directly benefit from our repeatable frames, as we showcase in a selection of 3D vision and shape analysis applications where we demonstrate state-of-the-art performance in a variety of challenging settings.*

1. Introduction

Computing correspondence between 3D shapes (in particular, meshes and point clouds) is a key task in computer graphics and vision, recently becoming increasingly relevant due to the availability of off-the-shelf depth sensors such as Microsoft Kinect or Intel RealSense and large-scale 3D datasets. 3D shape matching underlies applications such as robotic grasping and manipulation, scene understanding for augmented reality, obstacle avoidance and path planning for driver-less cars, to mention a few. Effective pipelines addressing this task rely upon the definition of a compact representation of the local geometry of the objects involved. Such point *descriptors* are expected to be compact, local, and distinctive. While point descriptors have been traditionally hand-crafted, recent proposals attempted to learn them by leveraging recent advances in deep learning [7, 20, 3, 24].

At the heart of most descriptors typically lies the construction of a *local reference frame* (LRF), a local system of Cartesian coordinates at each point, with respect to which the local geometric structure around that point is encoded. The effectiveness of the descriptor directly depends on the reliability of its underlying LRF; in turn, the quality of the latter is determined by its *invariance* to transformations that can be observed in the data. Most LRFs exploit some geometric properties of the local neighbors, such as the covariance matrix of the 3D coordinates of the neighborhood.

In this paper, we propose *GFrames*, a novel LRF that is demonstrably robust to severe sampling artifacts, vertex noise, and object deformation. Key to our method is the definition of the tangent component as the intrinsic gradient

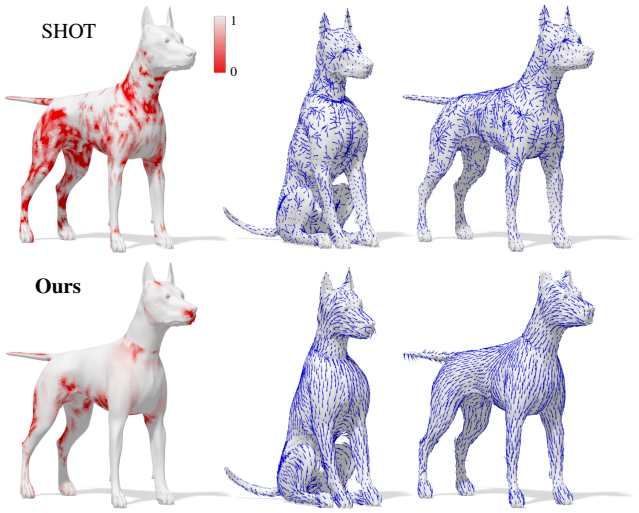


Figure 1. Comparison of LRF repeatability measured as mean cosine error on two non-rigid poses of the dog shape. We compare with the de-facto standard SHOT [34]. *Left*: The heat map encodes local frame alignment growing from red (gross misalignment) to white (perfectly aligned LRFs). *Right*: The computed LRFs; we only show the x axes for visualization purposes.

of a scalar function defined on the 3D object. The choice of the function directly determines the invariance classes of the resulting LRF; by doing so, we crucially shift the key difficulties of directly dealing with the object geometry to the simpler manipulation of a vector space of real-valued functions. The intrinsic construction further makes our LRF a natural choice in the more challenging non-rigid setting (see Figure 1). GFrames can be used as-is to improve existing descriptors and provide a robust choice in applications requiring a repeatable frame.

Our key **contributions** can be summarized as follows:

- We introduce a novel, theoretically principled LRF for 3D shapes that is remarkably robust to *sampling*, and that can be made provably invariant to *non-rigid* near-isometric transformations;
- We provide simple algorithms for its robust computation on triangle *meshes* as well as *point clouds*, and demonstrate its effectiveness on datasets addressing deformable matching of meshes as well as rigid point cloud registration;
- We showcase our construction in several classical 3D vision applications in challenging settings, on both synthetic and real-world data, where we achieve a gap in performance compared with the state-of-the-art.

We make a special effort of comparing on benchmarks used both in computer graphics and vision, which include tasks such as deformable matching of complete meshes (common in the former community) and registration of partial point clouds (from the latter community), demonstrating in both cases the effectiveness of our approach. With this, we aim to bridge the gap between the two worlds and propose a method that can be used broadly across the board.

2. Related work on LRFs

A robust and repeatable LRF is a key component for most 'handcrafted' 3D local descriptors, such as fast point feature histograms [30], exponential mapping [25], SHOT [34], ROPS [13], USC [33], and point signatures [8], to name just a few. Furthermore, robust local frames have a crucial role in recent geometric deep learning approaches [7], constructing non-Euclidean analogies of CNNs on meshes through local patch operators [20, 3, 24].

Local descriptors making use of local frames tend to be very sensitive to misalignment between LRFs at corresponding feature points, causing the performance of surface matching pipelines to be notably dependent on the repeatability of the adopted LRF (see [26, 27] for an extensive study). On the other hand, local descriptors that *do not* exploit an LRF are either not distinctive enough [1], costly to compute [28], or suffer from poor performance in the presence of noise and missing parts [32].

For a given 3D shape $\mathcal{M}$, an LRF $\mathcal{L}(p)$ at point $p \in \mathcal{M}$ is an orthogonal set of unit vectors (moving frame):

$$\mathcal{L}(p) = \{\hat{\mathbf{x}}(p), \hat{\mathbf{y}}(p), \hat{\mathbf{z}}(p)\} \tag{1}$$

satisfying the right-hand rule $\hat{\mathbf{y}} = \hat{\mathbf{z}} \times \hat{\mathbf{x}}$. Following [35], we distinguish between LRFs depending on whether they are based on *covariance analysis* or *geometric attributes*.

The former family includes methods that define the axes in $\mathcal{L}(p)$ as eigenvectors of the 3D covariance matrix between points lying within a spherical support of radius $r > 0$ centered at p, denoted by $B_r(p) = \{s \in \mathcal{M} : \|p - s\|_2 < r\}$. Inherent to such methods is the *sign ambiguity* of eigenvectors, making it hard to define repeatable directions; thus, efforts have largely concentrated on the reliable disambiguation of the axes sign. In [25], no disambiguation takes place, and the axis $\hat{\mathbf{x}}$ is simply defined as the principal eigenvector projected onto the tangent plane defined by the normal $\hat{\mathbf{n}}(p)$ (assumed to be given as input). In [22], all the three axes are given directly by the eigenvectors; however, here $\pm\hat{\mathbf{z}}$ is disambiguated by evaluating the two inner products $\langle \hat{\mathbf{n}}, \pm\hat{\mathbf{z}} \rangle$ and keeping the sign yielding a positive number. Axis $\hat{\mathbf{x}}$ nevertheless remains ambiguous. The LRF proposed with the SHOT descriptor [34] employs a different covariance matrix, where the contributions of the points in $B_r(p)$ are weighted by their distance to p. Sign ambiguity is addressed by choosing the sign that makes the eigenvector consistent with the majority of the measurements [5]; in practice, this results in the $\hat{\mathbf{x}}$ axis pointing in the direction of greater sample density. Similarly, in the ROPS descriptor [13] the axis $\hat{\mathbf{x}}$ is made to point in the direction of greater mesh density.

Methods based on geometric attributes determine $\hat{\mathbf{x}}$ by identifying a reference point $q \in B_r(p)$ within the support region, and then projecting the difference vector $q - p$ onto the tangent plane at p. Within this family, methods mainly differ by the geometric criterion used to select q. As an early example, point signatures [8] first fit a plane to the boundary path $\gamma = \mathcal{M} \cap B_r(p)$; the reference is then selected as the point $q \in \gamma$ with the largest positive distance to the fitted plane. In [26], a tangent plane is fitted to the entire $B_r(p)$; its normal vector $\hat{\mathbf{z}} = \pm\hat{\mathbf{n}}$ is disambiguated by taking the sign yielding a positive inner product with the average normal of the points in $B_r(p)$. The reference is then taken as the point $q \in B_{r'>r}(p)$ having the largest angular deviation with respect to $\hat{\mathbf{n}}$. The method of [27] follows a similar approach, whereas q is selected as the point exhibiting the largest signed distance (rather than angle) to the tangent plane. To our knowledge, the latter approach is the current state-of-the-art for computing repeatable LRFs, and hence is chosen as our baseline in the experimental section.

Finally, several deep learning-based 3D descriptors have been proposed in recent years, with most of them relying upon fixed LRFs in order to achieve rotation invariance.

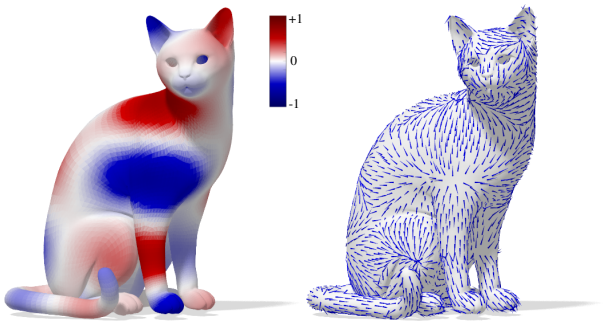


Figure 2. A scalar field on shape $\mathcal{M}$, and its intrinsic gradient ∇f.

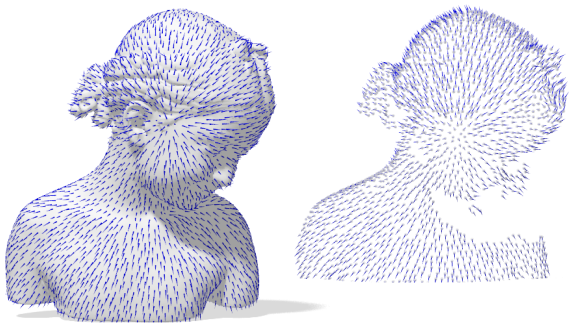


Figure 3. Gradient estimation on a triangle mesh (left) and on a point cloud of a partial scan (right). Our approach only needs a notion of a tangent space to be applied to any given representation.

Such cases include CGF-32 [15], PFFNet [11], and metric learned SHOT [9] which all deploy the LRF of [34]. In ACNN [4, 3] and MoNet [24] architectures, the local patches are oriented using the principal curvature direction. PointNet [29] uses a spatial transformer network [14] to predict a rigid transformation to canonically align the input data, while PCPNet [12] applies the transformer locally.

Based upon these considerations, we argue that the definition of a robust and repeatable LRF is still considered an open and challenging problem that underpins many existing approaches and is ubiquitous in many applications.

3. Background

In this paper, we consider 3D shapes represented as meshes or point clouds. To this end, we start with a continuous mathematical model and then discuss the discretization.

Manifolds. We assume that our shapes arise from the sampling of 2-dimensional Riemannian manifolds (surfaces) $\mathcal{M}$, possibly with boundary $\partial\mathcal{M}$, embedded into $\mathbb{R}^3$. Locally around each point $x \in \mathcal{M}$, the manifold is homeomorphic to the *tangent plane* $T_p\mathcal{M}$; the disjoint union of all such planes forms the *tangent bundle* $T\mathcal{M}$. We further equip the manifold with a *Riemannian metric*, defined as an inner product $\langle \cdot, \cdot \rangle_{T_p\mathcal{M}} : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ on the tangent plane depending smoothly on p. Functions of the form $f : \mathcal{M} \to \mathbb{R}$ and $F : \mathcal{M} \to T\mathcal{M}$ are referred to as *scalar-* and *(tangent) vector fields*, respectively. Properties expressed solely in terms of the metric are called *intrinsic*. In particular, *isometric* deformations of the manifold (such as a change in pose) preserve all intrinsic structures.

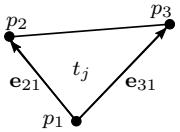
Intrinsic gradient. In classical calculus, derivatives describe how a function f changes with an infinitesimal change of its argument x. Due to the lack of vector space structure on the manifold (meaning that we cannot add two points, *i.e.*, an expression like "$p + dp$" is meaningless), one needs to define the *differential* of f as an operator $\mathrm{d}f : T\mathcal{M} \to \mathbb{R}$ acting on tangent vector fields. At each point p, the differential is a linear functional $\mathrm{d}f(p) = \langle \nabla f(p), \cdot \rangle_{T_p\mathcal{M}}$ acting on tangent vectors $F(p) \in T_p\mathcal{M}$, which models a small displacement around p. The change of the function value as the result of this displacement is given by applying the differential to the tangent vector, $\mathrm{d}f(p)F(p) = \langle \nabla_{\mathcal{M}} f(p), F(p) \rangle_{T_p\mathcal{M}}$. This can be thought of as an extension of the notion of directional derivative, where the linear operator $\nabla_{\mathcal{M}} f : L^2(\mathcal{M}) \to L^2(T\mathcal{M})$ is called the *intrinsic gradient*, and is similar to the classical gradient defining the direction of the steepest change of the function at a point, with the only difference that the direction is now a tangent vector; see Figure 2 for an example.

Discretization. Let us now assume the manifold is sampled at n points $p_1, \dots, p_n$, being the most basic representation of the shape called a *point cloud*. Equipping it further with a simplicial structure with edges $\mathcal{E}$ and triangular faces $\mathcal{F}$ yields a triangular mesh, which we assume to be a (discrete) manifold. Scalar functions $f : \mathcal{M} \to \mathbb{R}$ are represented as vectors $\mathbf{f} = (f(p_1), \dots, f(p_n))^\top$ encoding the value of f at each point. Following standard practice, functions are assumed to behave linearly between neighboring points (within each triangle in the case of meshes).

On meshes, the discrete intrinsic gradient ∇f yields tangent vector fields defined on the mesh *triangles*; on each triangle t_j, it is computed as a 3D vector

$$\nabla f(t_j) = \begin{pmatrix} \mathbf{e}_{21} & \mathbf{e}_{31} \end{pmatrix} \begin{pmatrix} E & F \\ F & G \end{pmatrix}^{-1} \begin{pmatrix} f(p_2) - f(p_1) \\ f(p_3) - f(p_1) \end{pmatrix} \quad (2)$$

where $E = \|\mathbf{e}_{21}\|^2$, $F = \langle \mathbf{e}_{21}, \mathbf{e}_{31} \rangle$, and $G = \|\mathbf{e}_{31}\|^2$ (see inset below for the notation).



On point clouds, the intrinsic gradient is discretized as follows. For each point p, we first estimate its tangent space by locally fitting a plane to points within radius r around p. These points are projected onto the plane, where they are locally meshed into a triangle patch $\mathcal{P}$ using Delaunay triangulation. We then take the weighted average $\nabla f(p) = \frac{1}{\sum A(t_j)} \sum_{t_j \in \mathcal{P}} A(t_j) \nabla f(t_j)$, where $A(t_j)$ denotes the area of triangle t_j.

We remark that, with this procedure, only a *local* reconstruction is carried out at each point p, and then thrown

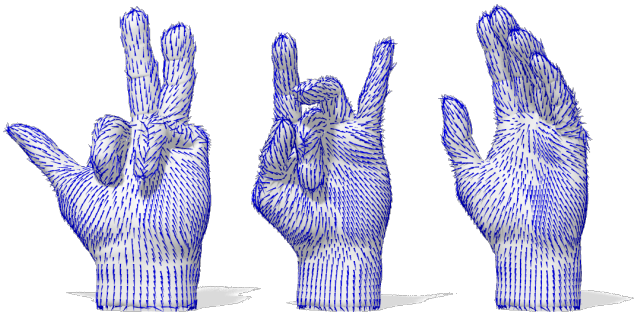


Figure 4. The $\hat{\mathbf{x}}$ axis of our LRF on different hand poses. In this example, repeatability is almost ideal due to the repeatability of the chosen scalar function $f(p_i) = \frac{1}{n}\sum_{j=1}^{n} d(p_i, p_j)$ equal to the average geodesic distance [37] from each point to all the others.

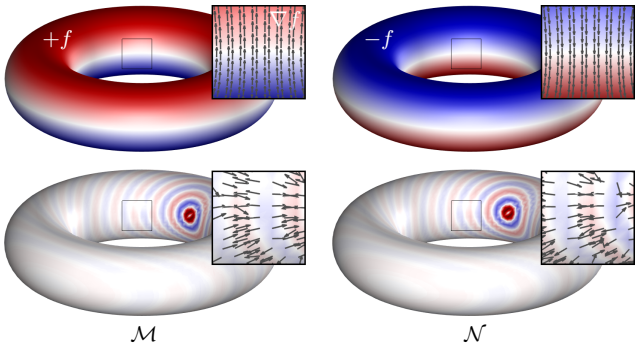


Figure 5. Sign flips of f (top row) lead to reversed axes in our LRF. In the bottom row, two high-frequency functions which are not exactly repeatable on $\mathcal{M}$ and $\mathcal{N}$ lead to local axis flips.

away once $\nabla f(p)$ is estimated. This brings additional robustness and efficiency in the presence of clutter or large point clouds; see Figure 3 for an example. Finally, normals on point clouds are estimated via total least squares [23]; for triangle meshes, the normal $\hat{\mathbf{n}}(p)$ at a vertex p is computed as the area-weighted average of the normals of the triangles sharing the vertex p.

4. Proposed local reference frame

Our technique is based upon the construction of tangent vector fields as *gradients* of scalar functions $f : \mathcal{M} \to \mathbb{R}$. We compute the *average gradient* of f around p as:

$$\mathbf{x}(p) := \frac{1}{\sum_{t_j \in \mathcal{N}_r(p)} A(t_j)} \sum_{t_j \in \mathcal{N}_r(p)} A(t_j) \nabla f(t_j) \qquad (3)$$

where $\mathcal{N}_r(p)$ is the set of triangles within distance r from p.

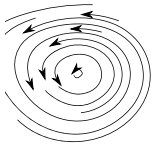
While it brings resilience to noise, the averaging process does not guarantee orthogonality to the normal vector $\hat{\mathbf{n}}(p)$. We thus project $\mathbf{x}(p)$ onto the plane identified by $\hat{\mathbf{n}}(p)$ and rescale the projection to unit norm, leading to the moving frame $\mathcal{L}_f(p) = \{\hat{\mathbf{x}}(p), \hat{\mathbf{y}}(p), \hat{\mathbf{z}}(p)\}$:

$$\hat{\mathbf{x}}(p) := (\mathbf{x}(p) - (\mathbf{x}(p)^\top \hat{\mathbf{n}}(p))\hat{\mathbf{n}}(p))_{\|\,\|} \qquad (4)$$

$$\hat{\mathbf{y}}(p) := \hat{\mathbf{z}}(p) \times \hat{\mathbf{x}}(p) \qquad (5)$$

$$\hat{\mathbf{z}}(p) := \hat{\mathbf{n}}(p) \qquad (6)$$

where $(\cdot)_{\|\,\|}$ denotes vector normalization and the notation $\mathcal{L}_f$ emphasizes that the definition of the LRF depends on the choice of the scalar function f.



Note that the gradient ∇f is guaranteed *curl-free* (*i.e.*, it never behaves like a vortex, see inset). This is desirable so as to reduce LRF inconsistency.

Choice of the function. Eq. (3) requires f to be differentiable; this is always true in our case, due to the assumption of piecewise-linearity. A separate question concerns the presence of singular points (where $\nabla f(p) = \mathbf{0}$). In the particular case of closed (genus zero) surfaces, these are unavoidable due to the Poincaré-Hopf ("Hairy Ball") Theorem stating that the only surface with nowhere vanishing tangent vector field is torus-like (genus 1); as we will show in our experiments, however, such points are rare and do not affect the overall quality of the LRF.

The choice of the function plays a role in determining the invariance class induced by the LRF, and is task-dependent: for instance, in order to achieve invariance to 3D rotations, it is sufficient that the function does not depend on the position of the object in space (an example is mean curvature). We will provide several possible choices in Section 6.

Descriptor steering. Constructing local descriptors $d : \mathcal{M} \to \mathbb{R}^k$ on top of a smooth frame field $\mathcal{L}_f$ can be seen as "steering" the descriptor field d along a given flow. For shape matching and registration applications, this fact can be exploited by designing flows using prior knowledge (in the form of sparse input correspondence). Specifically, given a single point-wise match $(x^*, y^*) \in \mathcal{M} \times \mathcal{N}$, the simple Euclidean distance from x^* (resp. y^*) to all other points in $\mathcal{M}$ (resp. $\mathcal{N}$) have compatible gradients (Fig. 3), making our LRFs an ideal choice in correspondence pipelines.

5. Properties

We list some of the key properties that make our proposed LRFs suitable for challenging settings. Additional properties depend on the choice of the underlying function, and will be explored in the experimental section.

Invariance properties depend on the choice of the scalar function f; for example, mean curvature endows the LRF with rotation invariance, and Gaussian curvature with invariance to isometric deformations. If available, it is also possible to use color texture as f. The chosen function must be repeatable only up to a global scale, since $\nabla \alpha f = \alpha \nabla f$ for any $\alpha \in \mathbb{R}$ and the normalization (4) make all options automatically scale-invariant; see Figure 4 for an example.

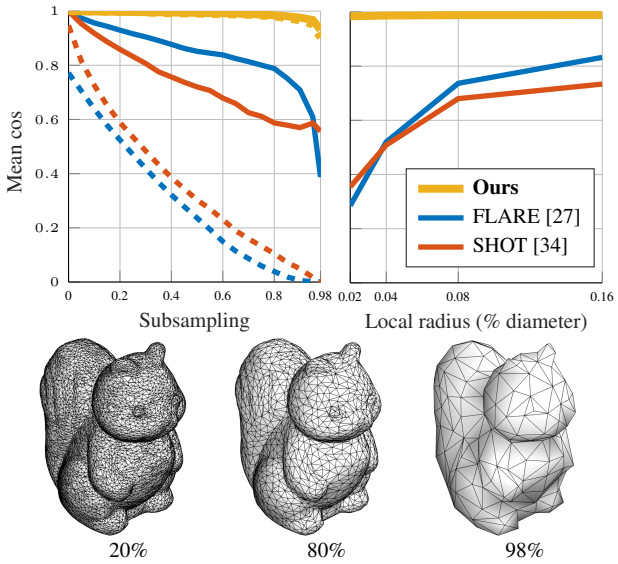


Figure 6. *Top left*: LRF repeatability under increasing subsampling, from 0% (no subsampling) to 98% (severe). We report results obtained with local radius $r = 0.02$ (dashed) and $r = 0.16$ (solid); all shapes have unit diameter. *Top right*: Comparisons at increasing radius, averaged over all subsampling levels. *Bottom*: Example of subsampled shapes used in these tests.

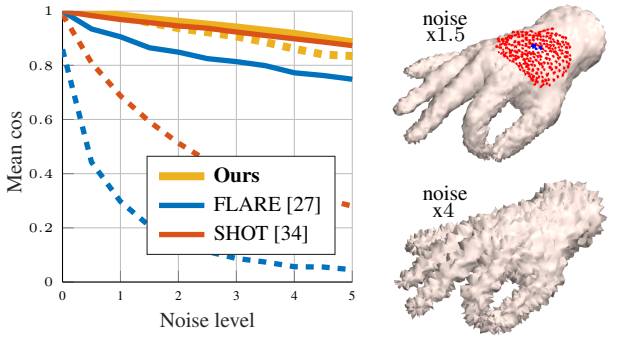


Figure 7. LRF repeatability at increasing surface noise (expressed as a multiplier of mesh resolution), obtained with radius $r = 0.02$ (dashed) and $r = 0.16$ (solid). Our results are better than FLARE and comparable with SHOT while using a much smaller radius; for comparison, on the top hand we plot the neighborhood at $r = 0.02$ (in blue) and $r = 0.16$ (in red). Due to the use of much smaller radius, our LRFs are much more robust to clutter and partiality.

Sign ambiguity is arguably the key issue of existing LRFs (see Section 2), with a direct impact on their reliability. Our frames do not suffer from sign ambiguity unless the sign of f is flipped (Figure 5, top), or if f contains high frequency oscillations (Figure 5, bottom).

Robustness to sampling is another central weakness of many state-of-the-art LRFs [34, 27]. To the best of our knowledge, none of the existing methods can deal with strong differences in sampling (arising, for example, when matching a CAD model to a 2.5D scan) or severe subsampling. We run a full quantitative comparison with such methods in Figure 6, using the repeatability measure defined in Sec. 6.1; for these and the following tests, we average over a representative dataset of six different shape classes (cat, centaur, dog, hand, human, squirrel) of varying resolution (ranging from 6K to 28K vertices).

Robustness to noise is achieved by averaging the gradient over a local neighborhood (3). Crucially, our LRFs also leverage the smoothness of the *function* $f : \mathcal{M} \rightarrow \mathbb{R}$ in addition to the smoothness of the 3D object $\mathcal{M}$ itself. This way, we shift the problem of dealing with a noisy geometric domain to a far easier task of choosing a smooth enough function on top of it. In Figure 7 we show a full quantitative evaluation at increasing amounts of surface noise.

Symmetry disambiguation is another property unique to our method. Choosing an asymmetric f (*e.g.*, distance to a point) directly endows descriptors constructed on top of $\mathcal{L}_f$ with symmetry-awareness. The lack of such property is considered a big drawback in shape analysis applications, leading to ambiguous solutions in most top-performing shape matching pipelines.

6. Applications

We evaluate GFrames in different applications and settings, including rigid registration and deformable matching.

6.1. LRF repeatability and rigid shape matching

Data. We use real scans of 4 objects (*Armadillo*, *Bunny*, *Buddha*, *Dragon*), from the Stanford 3D Scanning Repository [10], acquired with a Cyberware 3030 MS scanner. Some views of these objects are depicted in Figure 8. Ground-truth transformations are available for all the scans.

Evaluation. We evaluate the proposed method through the repeatability of the LRF and the efficiency of the descriptor built on the LRF. LRF repeatability at corresponding points on different shapes is assessed via the *MeanCos* and *ThCos* metrics [27]. The former represents the alignment error of both the $\hat{\mathbf{x}}$ and $\hat{\mathbf{z}}$ axes, while the latter indicates whether the two reference systems are aligned. More specifically, *ThCos* is the percentage of LRFs with a *MeanCos* value above a certain threshold (we used the value of 0.97).

For each of M scans for a given test model, we extract a set of uniformly distributed keypoints, and take all possible $N = \frac{M(M-1)}{2}$ view pairs (V_i, V_j). Due to partial overlap, a keypoint belonging to V_i may have no correspondence in V_j. Hence, given the ground-truth transformations $\mathbf{G}_i, \mathbf{G}_j$ bringing V_i, V_j into a canonical frame, we compute the set:

$$\mathcal{O}_{i,j} = \{k_i : \|\mathbf{G}_i k_i - \mathcal{N}(\mathbf{G}_i k_i, \mathbf{G}_j V_j)\| \leq \epsilon_{ovr}\}, \quad (7)$$

containing the keypoints in V_i that have a corresponding

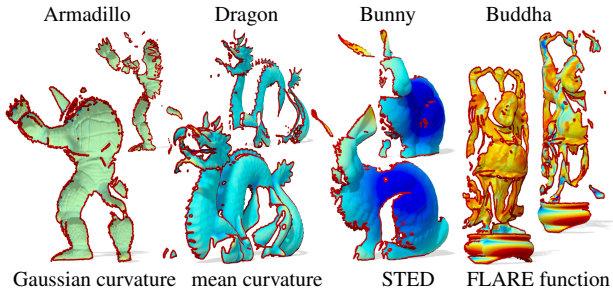


Figure 8. Example views from the Stanford repository. On each object we plot one of the four scalar functions used for the rigid matching experiments. Note how, despite baseline curvatures appear almost constant, they still exhibit enough gradient to outperform the SHOT LRF in most of our tests (compare with Figure 9).

point in V_j. Here, $\mathcal{N}(\mathbf{G}_i k_i, \mathbf{G}_j V_j)$ denotes the nearest neighbor of $\mathbf{G}_i k_i$ in the transformed view $\mathbf{G}_j V_j$, and ϵ_{ovr} is set to 2.5ρ[1]. If the number of points in $\mathcal{O}_{i,j}$ is less than 20% of the keypoints in V_i, the pair (V_i, V_j) is discarded due to insufficient overlap; otherwise, keypoint correspondences are established via nearest neighbor search in $\mathbb{R}^3$. Then, given a pair of corresponding keypoints $(k_i, k_j) \in V_i \times V_j$, we compute their LRFs and evaluate their repeatability according to the *MeanCos* and *ThCos* metrics.

Choice of the scalar function. The freedom of choosing f is a big advantage, allowing us to better adapt to the task at hand. As baseline choices, we use the aforementioned mean (***Ours mean***) and Gaussian (***Ours Gauss***) curvature. In addition, we use the following two functions:

STED (sum of total Euclidean distances), simply defined as the sum $f(x_i) = \sum_{j=1}^{n} \|x_i - x_j\|_2$.

FLARE: originally proposed in [27], it is computed at each point p as the average of the signed distances to the tangent plane defined by the normal $\hat{\mathbf{n}}(p)$, computed only on a subset of points lying at the periphery of the support region.

An example of each scalar function is shown in Figure 8.

In Figure 9, we compare our LRF construction to the ones used in the state-of-the-art SHOT [34] and FLARE [27]. Our scalar functions result effective in achieving a repeatable LRF. ***Ours FLARE*** is consistently better than SHOT and FLARE LRFs on both metrics, while ***Ours STED*** outperforms them in terms of MeanCos. Note how ***Ours FLARE*** always outperforms the original FLARE method, highlighting the usefulness of relying on gradient-based LRFs for better repeatability. ***Ours STED*** tends to be less repeatable in terms of ThCos; the *STED* function is more sensitive to scan overlap, as we noticed a significant improvement with the increase of the overlap.

A qualitative comparison on two views of a room from the RGB-D SLAM Dataset [31] is further shown in Fig. 11,

[1]Point cloud resolution ρ is the average Euclidean distance from each point to its nearest neighbor.

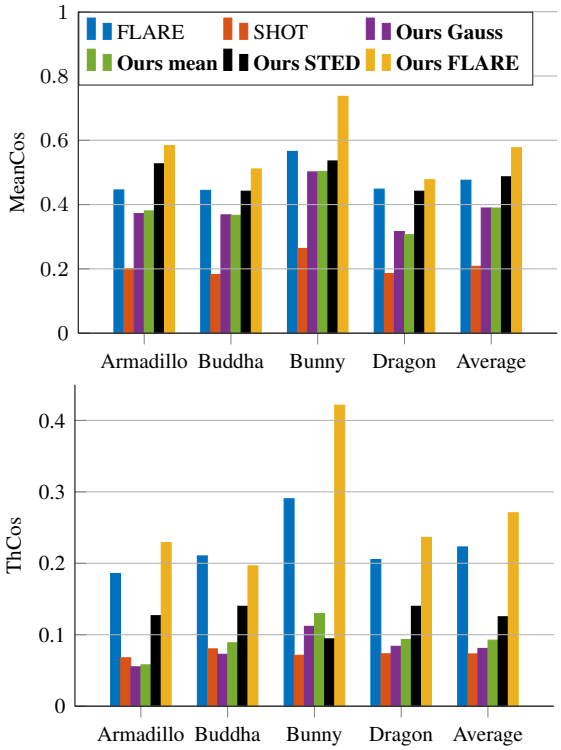


Figure 9. LRF repeatability on the Stanford dataset (the higher the better). Here, SHOT denotes the LRF of the SHOT descriptor.

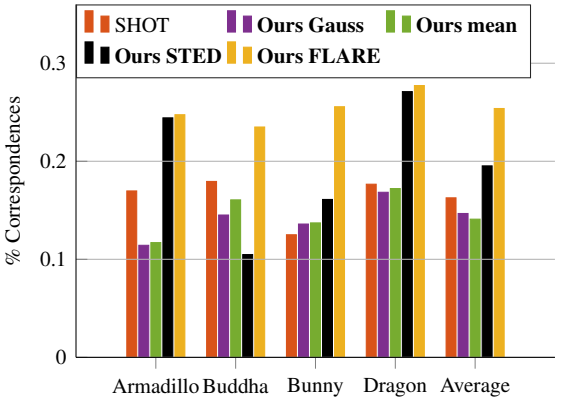


Figure 10. Descriptor matching results using the SHOT *descriptor* computed on different LRFs (among which the SHOT LRF itself). The y-axis denotes the percentage of matches whose Euclidean distance from the ground truth is less than 7mm.

confirming the large improvement produced by our approach also on this type of real-world data. We refer to the supplementary material for additional examples and details.

Finally, Fig. 10 reports a comparison in terms of descriptor matching, where the SHOT descriptor is used on top of each LRF. These results confirm the trend of the previous

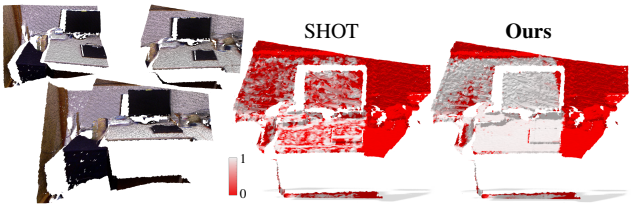


Figure 11. LRF repeatability on two views of a room (depicted on the left; their alignment is on the bottom). MeanCos error is encoded as a heat map, growing from white to red. Most of the error of our LRFs comes from incomplete overlap of the two views.

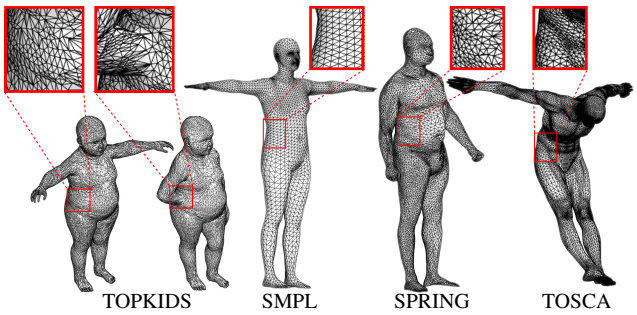


Figure 12. Representative data used in the deformable matching tests. TOPKIDS exhibit topological gluing at self-contacts (arm touching the body). Shapes from SMPL, SPRING, and TOSCA are used in *cross-dataset* matching experiments; the zoom-ins highlight the difference in mesh density and connectivity.

tests; ***Ours STED*** and ***Ours FLARE*** exhibit the best accuracy, with the former having larger error on Buddha, which has smaller overlap compared to the other objects.

To stress our robustness in multiple contexts, we additionally evaluate our method on the *Angel* point clouds from the recent laser scan dataset [15] (more results are in the supplemetary material due to lack of space). On average we obtain (MeanCos, ThCos, %Correspondence); SHOT: 0.19, 0.06, 0.13; ***Ours STED***: 0.69, 0.16, 0.25.

6.2. Deformable shape matching

Data. We adopt real-world as well as synthetic datasets: TOSCA [6] (7 classes of synthetic animal and human meshes undergoing non-rigid deformations), FAUST [2] (100 scanned meshes of 10 human subjects in different poses), TOPKIDS [17] (15 synthetic human meshes in different poses, with severe topological artifacts in areas of self-contact). Examples of these shapes are shown in Figure 12. All datasets come with ground-truth correspondence; for cross-dataset experiments, the ground-truth is estimated using the shape registration method FARM [19].

Evaluation. As a baseline, we use the original SHOT descriptor and compare it to SHOT descriptors constructed on top of our LRFs. Pointwise correspondence is established by nearest neighbors in descriptor space, and evaluated according to the Princeton protocol [16], computing the percentage of matches that fall into geodesic radius r (represented as a fraction of the shape geodesic diameter) from the ground truth correspondence.

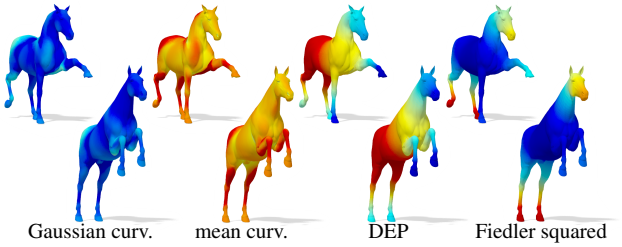


Figure 13. The four scalar functions used in the deformable setting. Their gradient has few singular points, which do not strongly affect the quality of the resulting LRF.

Choice of the scalar function. We adopt the same baseline as the rigid setting (mean and Gaussian curvature) plus two functions specific to this task (see examples in Figure 13):

Fiedler vector is the first non-constant eigenfunction of the Laplace-Beltrami operator of the surface. Except for a sign ambiguity (simply solved by taking the square of the function), it is fully intrinsic and thus invariant to isometries.

Discrete Time Evolution Process (DEP) is a recent intrinsic point descriptor that is stable to non-isometric distortions, missing parts, and topological noise [21]. It is similar to the *average geodesic distance* [37], but was shown to be more robust for non-isometric deformations. We compute DEP using biharmonic distances as done in [19].

We consider four different settings of deformable shape matching (see Figure 14):

Isometric deformations. We test on 8 pairs of deformable animals (TOSCA, dog and horse categories) and 20 scans of a human subject in different poses (FAUST intra). We report an improvement of $\sim 10\%$ over the de-facto LRF.

Non-isometric deformations. We test on 20 pairs of different poses and different *subjects* (FAUST inter), demonstrating similar performance to the previous setting.

Topological noise. We evaluate on 15 poses of a synthetic human undergoing severe topological variations, *e.g.*, gluing hands to the body (TOPKIDS). Performance here is worse than on previous datasets; for instance, the Fiedler vector is directly affected by topological gluing happening over long distances. Despite this, in this challenging setting, the advantage of our model (***Ours DEP*** and ***Ours Fied***) over the baseline is even more pronounced.

Different connectivity and resolution. We compose a hybrid dataset of human shape pairs from SMPL [18], TOSCA and SPRING [36]; see the last three columns of Figure 12 for examples. This experiment is particularly challenging due to the differences in mesh connectivity and density among the various models. Such a setting is a notoriously

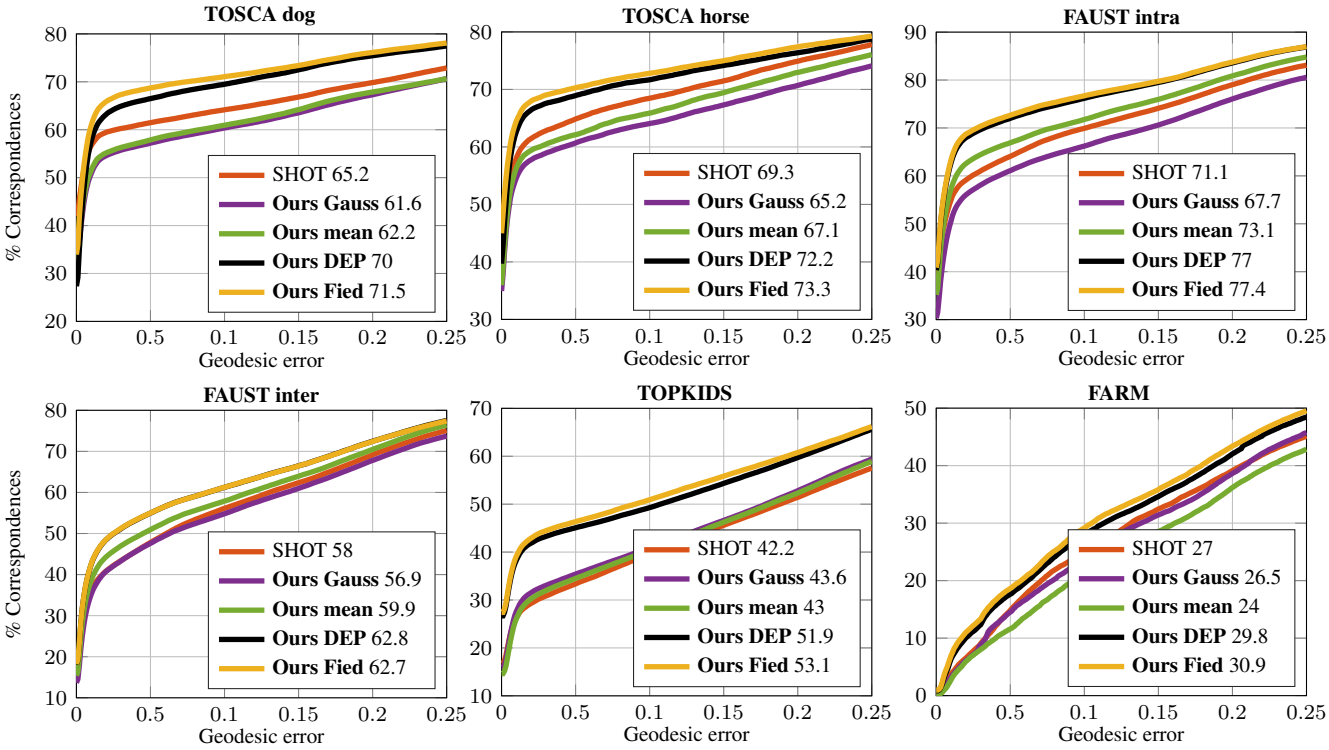


Figure 14. Error rates for deformable matching on different datasets. The y-axis represents the percentage of matches for which the geodesic distance from the ground truth is less than the value reported on the x-axis. The numbers in the legend denote the AUC.

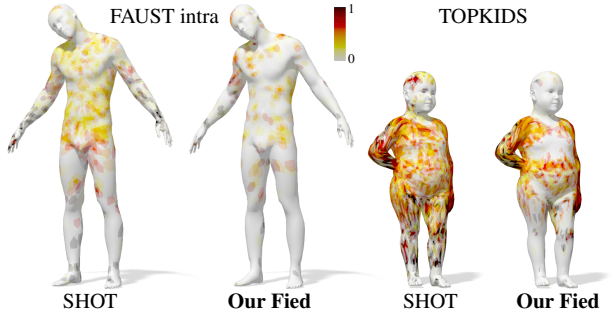


Figure 15. Qualitative comparisons on a standard (left) and challengin (right) case. Pointwise matching error is encoded as a heatmap, growing from white to dark red.

tough nut for existing LRFs and local descriptors, and is not frequently considered in existing benchmarks. Nevertheless, we still outperform the baseline.

In Fig. 15 we show the matching error in one standard (FAUST) and one challenging (TOPKIDS) case. Finally, on TOSCA we also evaluate the LRF repeatability (MeanCos) over all 64 pairs of the dog and horse classes. SHOT achieves an average score of 0.76, while ***Ours Fied*** reaches up to 0.90 (close to ideal). A qualitative evaluation of this result is shown for a dog pair in Fig. 1 using ***Ours DEP***.

7. Conclusions

We introduced GFrames, a new local reference frame for 3D shape matching applicable to meshes and point clouds. Our construction is based on the computation of the tangent component of the LRF as the intrinsic gradient of a scalar function on the surface; different designs are possible depending on the task, as we showcased on a selection of relevant problems in 3D computer vision and shape analysis. The flexibility of our approach lies in the freedom of choosing a scalar function on top of which a stable LRF, and in turn repeatable descriptors, can be constructed. On the other hand, the main limitation lies in the requirement for the chosen function to have limited high frequency content, which may lead to unstable gradients; this excludes, for instance, the adoption of highly detailed texture or oscillatory functions obtained, *e.g.*, by wave propagation. As a promising avenue of future work, we envision the adoption of GFrames in deep learning pipelines, where the scalar function itself may be learned in an end-to-end fashion.

Acknowledgments

We thank Arianna Rampini for useful discussions. ER is supported by the ERC Grant No. 802554 (SPECGEO). MB is partially supported by ERC Grant No. 724228 (LEMAN), Google Research Faculty awards, the Royal Society Wolfson Research Merit award and Rudolf Diesel industrial fellowship at TU Munich.

References

[1] Andrea Albarelli, Emanuele Rodolà, and Andrea Torsello. Loosely distinctive features for robust surface alignment. In *European Conference on Computer Vision*, pages 519–532. Springer, 2010.

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

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

[4] Davide Boscaini, Jonathan Masci, Emanuele Rodolà, Michael M Bronstein, and Daniel Cremers. Anisotropic diffusion descriptors. *Computer Graphics Forum*, 35(2):431–441, 2016.

[5] Rasmus Bro, Evrim Acar, and Tamara G Kolda. Resolving the sign ambiguity in the singular value decomposition. *J. Chemometrics*, 22(2):135–140, 2008.

[6] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. *Numerical Geometry of Non-Rigid Shapes*. Springer Science & Business Media, 2008.

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

[8] Chin Seng Chua and Ray Jarvis. Point signatures: A new representation for 3d object recognition. *IJCV*, 25(1):63–85, 1997.

[9] Luca Cosmo, Emanuele Rodolà, Jonathan Masci, Andrea Torsello, and Michael M Bronstein. Matching deformable objects in clutter. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2016.

[10] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312. ACM, 1996.

[11] Haowen Deng, Tolga Birdal, and Slobodan Ilic. PPF-FoldNet: Unsupervised learning of rotation invariant 3d local descriptors. *arXiv preprint arXiv:1808.10322*, 2018.

[12] Paul Guerrero, Yanir Kleiman, Maks Ovsjanikov, and Niloy J Mitra. PCPNet learning local shape properties from raw point clouds. *Computer Graphics Forum*, 37(2):75–85, 2018.

[13] Yulan Guo, Ferdous A Sohel, Mohammed Bennamoun, Jianwei Wan, and Min Lu. Rops: A local feature descriptor for 3d rigid objects based on rotational projection statistics. In *Proc. ICCSPA*, 2013.

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Proc. NIPS*, 2015.

[15] Marc Khoury, Qian-Yi Zhou, and Vladlen Koltun. Learning compact geometric features. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 153–161. IEEE, 2017.

[16] Vladimir G Kim, Yaron Lipman, and Thomas Funkhouser. Blended Intrinsic Maps. *TOG*, 30(4):79, 2011.

[17] Z. Lähner, E. Rodolà, M.M. Bronstein, D. Cremers, O. Burghard, L. Cosmo, A. Dieckmann, R. Klein, and Y. Sahillioğlu. Shrec'16: Matching of deformable shapes with topological noise. *Eurographics Workshop on 3D Object Retrieval, EG 3DOR*, pages 55–60, 2016.

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *TOG*, 34(6):248:1–248:16, 2015.

[19] Riccardo Marin, Simone Melzi, Emanuele Rodolà, and Umberto Castellani. Farm: Functional automatic registration method for 3d human bodies. *arXiv:1807.10517*, 2018.

[20] J. Masci, D. Boscaini, M. Bronstein, and P. Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proc. 3dRR*, 2015.

[21] Simone Melzi, Maks Ovsjanikov, Giorgio Roffo, Marco Cristani, and Umberto Castellani. Discrete time evolution process descriptor for shape analysis and matching. *TOG*, 37(1):4:1–4:18, Jan. 2018.

[22] Ajmal Mian, Mohammed Bennamoun, and Robyn Owens. On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes. *IJCV*, 89(2-3):348–361, 2010.

[23] Niloy J Mitra and An Nguyen. Estimating surface normals in noisy point cloud data. In *Proc. Symp. Computational Geometry*, 2003.

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

[25] John Novatnack and Ko Nishino. Scale-dependent/invariant local 3d shape descriptors for fully automatic registration of multiple sets of range images. In *Proc. ECCV*, 2008.

[26] Alioscia Petrelli and Luigi Di Stefano. On the repeatability of the local reference frame for partial shape matching. In *Proc. ICCV*, 2011.

[27] Alioscia Petrelli and Luigi Di Stefano. A repeatable and efficient canonical reference for surface matching. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 403–410. IEEE, 2012.

[28] Helmut Pottmann, Johannes Wallner, Qi-Xing Huang, and Yong-Liang Yang. Integral invariants for robust geometry processing. *Computer Aided Geometric Design*, 26(1):37–60, 2009.

[29] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. *Proc. CVPR*, 1(2):4, 2017.

[30] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *Proc. ICRA*, 2009.

[31] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[32] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. *Computer Graphics Forum*, 28(5):1383–1392, 2009.

[33] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique shape context for 3d data description. In *Proc. 3DOR*, 2010.

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

[35] Jiaqi Yang, Yang Xiao, and Zhiguo Cao. Toward the repeatability and robustness of the local reference frame for 3d shape matching: An evaluation. *IEEE Trans. Image Processing*, 27(8):3766–3781, 2018.

[36] Y. Yang, Y. Yu, Y. Zhou, S. Du, J. Davis, and R. Yang. Semantic parametric reshaping of human body models. In *Proc. 3DV*, 2014.

[37] Eugene Zhang, Konstantin Mischaikow, and Greg Turk. Feature-based surface parameterization and texture mapping. *ACM Trans. Graph.*, 24(1):1–27, Jan. 2005.