Temporally-Coherent Surface Reconstruction via Metric-Consistent Atlases

Jan Bednarik[1] Vladimir G. Kim[2] Siddhartha Chaudhuri[2] Shaifali Parashar[1]
Mathieu Salzmann[1] Pascal Fua[1] Noam Aigerman[2]

[1]EPFL, Lausanne, Switzerland [2]Adobe Research

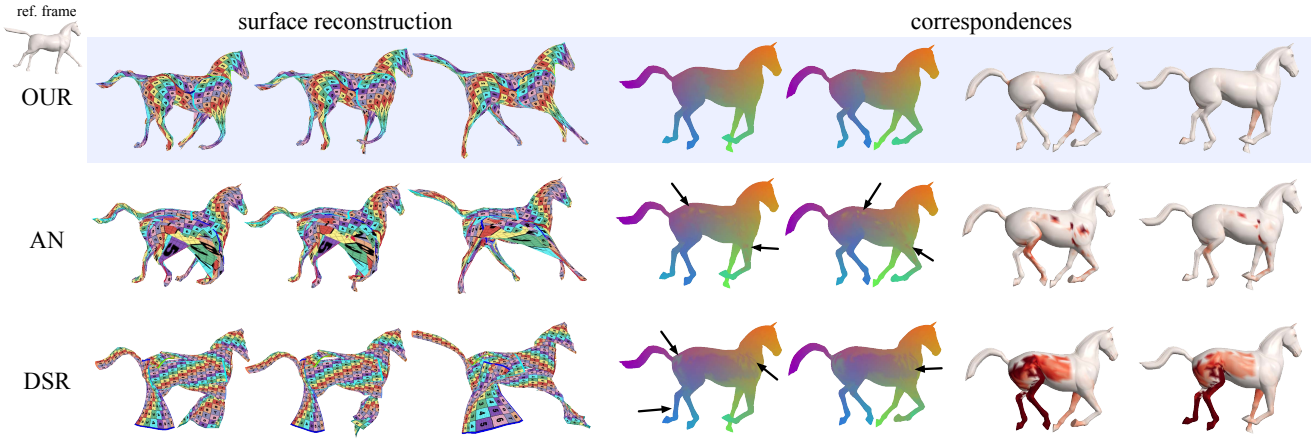

Figure 1. **Temporally-coherent reconstruction and correspondences predicted by our approach (OUR), compared with other atlas-based ones – AtlasNet [19] (AN), and Differential Surface Representation [7] (DSR).** Left: The reconstructed surfaces, textured with a consistent texture to visualize the correspondences between the surfaces. Middle and Right: deviations visualized using a colormap (middle) and a heatmap (right). The competing methods exhibit artifacts and wrong correspondences, while OUR yields reconstructions close to the GT. The black arrows point to inconsistencies, which are absent from our results.

Abstract

We propose a method for the unsupervised reconstruction of a temporally-coherent sequence of surfaces from a sequence of time-evolving point clouds, yielding dense, semantically meaningful correspondences between all keyframes. We represent the reconstructed surface as an atlas, using a neural network. Using canonical correspondences defined via the atlas, we encourage the reconstruction to be as isometric as possible across frames, leading to semantically-meaningful reconstruction. Through experiments and comparisons, we empirically show that our method achieves results that exceed that state of the art in the accuracy of unsupervised correspondences and accuracy of surface reconstruction.

1. Introduction

Applications such as UV-mapping, shape analysis, and partial scan-completion all rely on the availability of a surface representation that is *coherent* across different instances. Namely, the different surfaces should be in correspondence, such that each point on one surface maps to a point with the same semantic meaning on another. In the literature, the most common way to achieve coherence consists of explicitly computing and establishing correspondences between non-coherent input representations, such as 3D meshes [49, 3, 45, 22, 16, 42] or 3D point clouds [24, 21]. This, however, assumes that the input data contains points that can be matched in a semantically-meaningful manner, and in fact only circumvents the true task of retrieving a coherent surface representation.

In this paper, we tackle this problem more directly by learning to reconstruct temporally-coherent surfaces from a sequence of 3D point clouds representing a shape deforming over time. To this end, we rely on the AtlasNet patch-based representation [19] to model the surface underlying the 3D points. However, whereas in the original AtlasNet, any patch can correspond to any part of the surface, we enforce consistency of the patch locations through the whole sequence effectively creating a time-consistent atlas.

To learn atlases that are semantically and temporally consistent, meaning that each 2D point on each 2D atlas patch models the same semantic surface point over time, we leverage differential geometry to require the correspon-

This work was partially carried out while the first author was an intern at Adobe Research and was supported in part by the Swiss National Science Foundation.

dences model a close-to-isometric deformation, for which the metric tensor computed at any surface point remains constant as the shape changes. We translate this into a metric-consistency loss function, which, when minimized, implicitly establishes meaningful point correspondences.

Our approach does not require any ground-truth correspondences, which are usually difficult to obtain. Hence, it is unsupervised and can operate on any shape category *without* a known shape template. Yet, as shown in Fig. 1, it provides reliable correspondences even in cases in which the shapes are complex and the deformations are severe, unlike state-of-the-art methods which tend to break down.

2. Related Work

3D temporal coherence involves both surface reconstruction and correspondence estimation, which are in interplay with one another. Both of these are well-studied, essential tasks in geometry processing, which we review next.

Correspondence estimation commonly assumes that the objects are close to isometric and thus often optimizes for local distance preservation [11, 33, 46]. This can be achieved via local shape descriptors [36, 4, 48, 29], which are in turn used to obtain surface correspondences. Alternatively, obtaining correspondences can be cast as a template fitting problem [31, 56]. This, however is reliant on knowing beforehand what is the class of the shape, and on having a template for this class. Simpler methods [3, 49] have been designed for temporal registration assuming piecewise-rigidity of the shapes. However, these methods generate only region-wise correspondences. In case meshes are given, they can be parameterized into the same 2D common base-domain where correspondences can be optimized [27, 1, 51]. This approach relies on 3D surface (triangulations) given as input, and hence cannot be applied to point-clouds and does not reconstruct surfaces. Taking a cue from this approach, we also use a 2D domain to define the correspondences, but keep the correspondence fixed in 2D, and instead optimize the 3D surface while performing surface reconstruction.

Recently, correspondence estimation has been addressed as a learning problem. Many works use representations such as [36] to retrieve local descriptors and incorporate them in the learning process [22, 16, 45]. Other supervised methods have been proposed, using ground-truth correspondences as training data [44, 32, 10, 35].

Motivated by the fact that obtaining correspondence supervision is expensive, [12, 6] introduced an unsupervised learning framework using triangulated meshes. To avoid meshing, [24, 20] proposed unsupervised learning techniques to extract correspondences directly from point clouds. However, [24] only yields a set of semantically close points without a mechanism to find a unique correspondence, and [20] uses a 3D template.

In contrast to existing methods, our approach yields temporally-coherent surface reconstructions from point clouds and generates meaningful point-wise correspondences. To this end, it learns a unique atlas representation similar to [20] but enforces local metric consistency, which aims to preserve isometry at corresponding points on the output surfaces. Our method is unsupervised and does not require a shape template. Thus, the closest approach to our method is [21], which learns correspondences by enforcing cyclic consistency across multiple shape-triplets. Our extensive comparisons with [21, 19, 7] show that our method consistently outperforms these state-of-the-art techniques.

Surface reconstruction from point clouds has been thoroughly studied in geometry processing. Many non-learning techniques use mathematical tools to reconstruct the surface, e.g., solving the Poisson PDE [26], or using Moving Least Squares [30] to fit points to the surface; see [8] for a survey. Deep learning techniques were first successfully applied to point-cloud reconstruction [39, 41, 17, 24], and afterwards to surfaces, starting with the seminal AtlasNet [20], FoldingNet [54] and their followups [14, 7]. Surfaces can also be reconstructed from learned elementary structures [15]. In [52], an MLP was shown to be effective in reconstruction when optimized to fit a point cloud.

Other representations such as meshes [25, 37] are simple to handle, however require a predesignated triangulation, which is not versatile enough to accommodate for arbitrary shapes with different articulations. Likewise, implicit fields such as SDF's [38, 34] can represent a surface accurately, however the implicit definition does not lend itself to defining correspondences.

In any case, none of these methods target temporally-coherent surface reconstruction.

Metric preservation and shape interpolation are closely related to our approach. Metric preservation is widely used when a low-distortion map between shapes is required, especially in the context of shape interpolation that has long been studied in computer graphics [28, 2, 53]. In recent years several data-driven methods have been proposed for this task [18, 13], but they assume to be given point correspondences and do not infer them. Closer to our work, [43] discussed how to smoothly interpolate between two point clouds, without given correspondences. However, this work focuses solely on interpolating the point clouds without generating meaningful correspondences nor producing a continuous surface.

3. Methodology

3.1. Problem statement and overview

We assume to be given as input a temporal sequence of 3D point clouds $P_1, ..., P_K$. Our output is a corresponding sequence of reconstructed surfaces $\mathcal{S}_1, ..., \mathcal{S}_K$, one for each point cloud, along with a canonical bijective mapping $\Psi_{i,j}$ between the surfaces S_i, S_j, defining temporally-consistent point-to-point correspondences, for any point on one of the reconstructed surfaces.

We use an atlas-based representation with multiple patches similarly to [19], with an atlas ϕ_j representing each surface $\mathcal{S}_j$. This immediately defines a canonical bijective map $\Psi_{i,j}$ between any two surfaces $\mathcal{S}_i, \mathcal{S}_j$ via the shared 2D domain (see Figure 2). We wish to optimize the atlases so that their surfaces satisfy two properties:

1. **Fitting.** Each surface $\mathcal{S}_k$ should model the corresponding point cloud P_k as closely as possible.

2. **Temporal coherence.** Each predefined canonical bijective map $\Psi_{i,j}$ maps semantic parts of the surface correctly between frames (nose is mapped to nose).

Our core observation is that we can achieve this goal in an unsupervised manner, by making the $\Psi_{i,j}$ as isometric as possible, thus encouraging the transition from one frame in the sequence to the next to preserve local shape features, thereby making the reconstructions consistent. Next, we elaborate on the above.

3.2. Atlas-based surface representation

Atlases and canonical surface correspondences. In its most basic form, an *atlas* can be defined as a map ϕ, embedding a 2D domain Ω to a surface in 3D $\phi : \Omega \rightarrow \mathbb{R}^3$, such that the image of ϕ is $\mathcal{S}$ (we use $\Omega = [0, 1]^2$ in all experiments).

Using atlases enables us to define a canonical point correspondence between any two 3D surfaces, $\mathcal{S}_1, \mathcal{S}_2$, described by two atlases, ϕ_1, ϕ_2, see Figure 2. Specifically, we can trivially define a bijective (1-to-1 and onto) correspondence between the two 3D surfaces by defining the point $\phi_1(p) \in \mathcal{S}_1$ to correspond to $\phi_2(p) \in \mathcal{S}_2$, and vice versa, for any point $p \in \Omega$. This correspondence enables us to optimize the atlases to ensure that corresponding points are mapped to the same semantic 3D surface point on $\mathcal{S}_1, \mathcal{S}_2$.

Isometry through metric consistency. We enforce isometry between different atlases. To achieve that we use the *Riemannian metric tensor*. For any point $p = (u, v) \in \Omega$, the metric tensor is expressed in terms of the Jacobian, the matrix $J_\phi \in \mathbb{R}^{3\times 2}$ of partial derivatives of the map ϕ at p, $J_\phi = \left[\frac{\partial \phi}{\partial u}, \frac{\partial \phi}{\partial v}\right]$. Specifically, the metric tensor is defined as

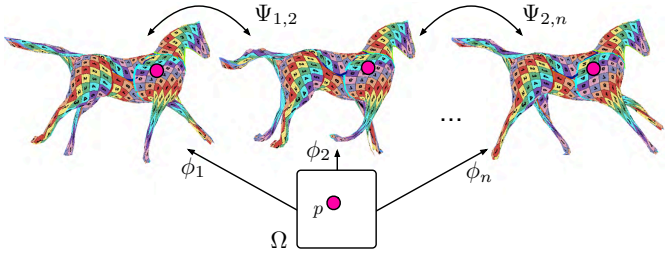

Figure 2. Correspondences defined between three surfaces by the mapping of one point $p \in \Omega$ through three different atlases.

$g = J_\phi(p)^\top J_\phi(p)$. Intuitively, g defines a local inner product between any two vectors $q, r \in \mathbb{R}^2$ as $q^T \cdot g(p) \cdot r$, enabling one to measure local lengths and angles at any point $\phi(p)$ on the surface $\mathcal{S}$.

The rest of the paper can then be completely understood just from the high-level definition of the metric tensor as a descriptor of local geometric quantities.

Given two surfaces as above, using the canonical correspondence just defined, we can compare the metrics of the surfaces, g_{ϕ_1}, g_{ϕ_2}, at corresponding points $\phi_1(p), \phi_2(p)$, and measure the difference between the two, $\|g_{\phi_1}(p) - g_{\phi_2}(p)\|_F$, where $\|\cdot\|_F$ stands for the Frobenius norm. We can now define a metric-consistency energy between the two surfaces as

$$E_{\text{cons}}(\phi_1, \phi_2) = \int_{p\in\Omega} \|g_{\phi_1}(p) - g_{\phi_2}(p)\|_F^2, \qquad (1)$$

which measures the deviation from isometry of the map between the two surfaces the two atlases represent.

3.3. Temporally-coherent surface reconstruction

Atlases via a neural network. To define atlases in a deep learning setting, we follow the standard AtlasNet [19] formulation: The network receives a point $p \in \Omega$, along with a latent code $z \in \mathbb{R}^C$ (where C is the dimension of the latent space) and outputs a 3D point, essentially defining an atlas conditioned on z. Note that, most importantly, all differential quantities introduced in the previous section can be easily inferred for the network's atlases, since the network is a (piecewise) differentiable mapping.

Lastly, we note that instead of relying on a single map ϕ, any number of maps $\phi_1, \phi_2, ..., \phi_j$ can be chosen before optimization, enabling mapping several 2D domains into several 3D patches, whose union forms the complete shape. This poses no change to any of the notions discussed herein, and hence we simply consider the domain Ω and ϕ as aggregating all the patches, their domains and maps, except when explicitly referring to these patches. In all experiments, we used 10 patches.

Given a dataset with K point clouds $P_1, ..., P_K$, we encode each point cloud P_k into a latent code z_k through a PointNet [40] encoder as used by [19]. We denote by ϕ_k

the resulting atlas defined via the code z_k, representing the reconstructed surface.

Loss Functions. To enforce isometry across the sequence, we use a loss function measuring metric consistency between pairs of atlases,

$$\mathcal{L}_{\text{metric}} = \alpha_{\text{mc}} \sum_{(i,j)\in\mathcal{I}} E_{\text{cons}}\left(\phi_i, \phi_j\right), \tag{2}$$

where $\mathcal{I}$ are chosen pairs of surfaces out of all possible pairs, and $\alpha_{\text{mc}} \in \mathbb{R}$ is a hyper-parameter of our approach.

Next, to train the network to reconstruct the given dataset, we follow standard practice in shape reconstruction [20, 15, 7, 14] and use the Chamfer distance (CD) to define the *reconstruction loss*

$$\mathcal{L}_{\text{CD}} = \frac{1}{K}\sum_{k=1}^{K}\left[\int_{p\in\Omega}\min_{q\in P_k}||\phi_k(p)-q||^2 + \sum_{q\in P_k}\min_{p\in\Omega}||\phi_k(p)-q||^2\right]. \tag{3}$$

We then take our final loss to be

$$\mathcal{L} = \mathcal{L}_{\text{CD}} + \mathcal{L}_{\text{metric}}\ . \tag{4}$$

Sampling surface pairs. The metric consistency loss 2 operates on pairs of surfaces defined by $\mathcal{I}$, $(\mathcal{S}_i, \mathcal{S}_j), (i,j) \in \mathcal{I}$. Our assumption is that the shape gradually deforms over time, and hence surfaces in subsequent frames should change close-to-isometrically with respect to one another. Hence we define a "time window" δ, which is a hyper-parameter of our method, and sample pairs of surfaces only if they fall within that window, $(S_i, S_j) : |i-j| \leq \delta$.

3.4. Implementation details

Our method uses the AtlasNet [19] architecture with the same adjustments of [7] for computing the metric (ReLU replaced with Softplus in the decoder; batch normalization layers removed). We use $P = 10$ patches in all experiments.

We use the Adam optimizer with a learning rate $l = 0.001$ and a batch size of 4 for 200000 iterations. We employ a learning rate scheduler which divides the current l by a factor of 10 at 80% and 90% of the training iterations. Following [19, 7], 2500 points are sampled from the UV domain Ω. We set the weight of the loss term $\mathcal{L}_{\text{metric}}$ of Eq. 2 to $\alpha_{\text{mc}} = 0.1$, and choose the value of δ using one sequence as validation subset and then measuring the correspondence metrics m_{sL2}, m_r and m_{AUC}.

At evaluation time, we follow [7] and remove any patch with area smaller than $1/1000$ than the average area of a patch. We sample a given number of available points in each patch as evenly as possible using a simulated annealing based algorithm. Please refer to the supplementary material for all other details.

4. Evaluation

We test our method by reconstructing surfaces from various raw point-cloud sequences of human and animal motions, showing our method naturally adapts to different kinds of data, without any known correspondences between the frames or a reference template shape, and without requiring prior training on any specific category. Please refer to the supplementary material for a video showing the reconstructed sequences of all figures here and other to get a full sense of the accuracy of our method.

Visualization of the correspondences between surfaces. Before continuing, let us explain the technique used to visualize the correspondences between the surfaces. In all figures, to illustrate the temporal consistency of our reconstructions, we use the same texture in the UV space in all frames of the sequence. Hence, corresponding regions are textured with the same checkerboard cells, revealing the accuracy of the correspondences.

Figure 3 shows our temporally-coherent reconstructions for six sequences. Note how our method manages to reconstruct high-curvature regions, such as the elephant's husks and the cat's tail and paws, with both accurate geometry and high correspondence accuracy, tracking the paws as they move. The human models exhibit much more articulated deformations, nonetheless our method tracks the limbs and maintains consistent, meaningful correspondences throughout the sequence. Please refer to the supplementary video to view the animations of the entire sequences.

4.1. Inferring point cloud correspondences

A direct application of our method is inferring point-to-point correspondences on the input point clouds. Namely, for two point clouds we map points from P_i to P_j via euclidean projections between the point clouds and the reconstructed surfaces, using the map $f_{i\to j} = \pi_{P_j} \circ \phi_j \circ \phi_i^{-1} \circ \pi_{\phi_i}$, where $\pi_{\mathcal{X}}$ projects a 3D point to its nearest neighbor on the surface $\mathcal{X}$ and ϕ^{-1} is the inverse mapping which is known implicitly. Specifically, we densely sample N points in the 2D domain Ω and get their 3D counterparts via the learned ϕ and since this is a bijection, we know ϕ^{-1} for these N points.

Aside from being a useful application, it also enables us to evaluate the accuracy of our method w.r.t the ground truth correspondences of the dataset's point clouds. In Figure 4 we visualize the correspondences predicted on the input point clouds using a matching colormap. We show a visualization of the error on the models (Note that the triangulated meshes are only used for visualization), with red

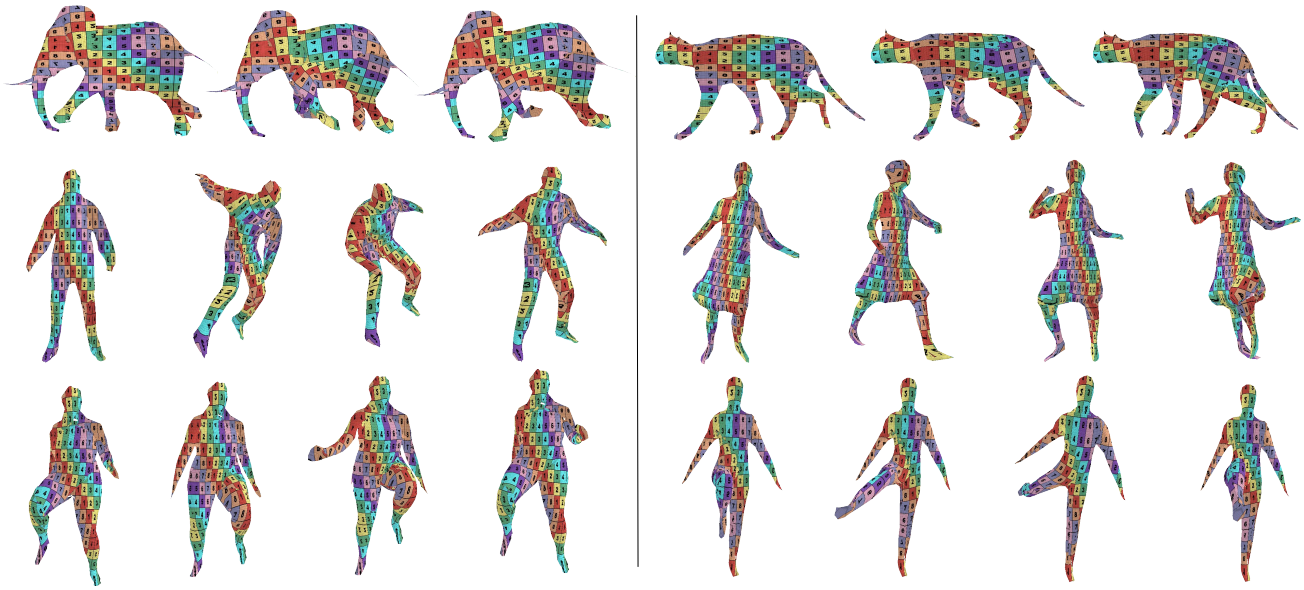

Figure 3. **Our temporally-coherent surface reconstructions, for 6 sequences.** (top to bottom) `elephant` and `cat` from ANIM, `jumping` and `swing` from AMA, `running_on_spot` and `knees` from DFAUST. Note how the reconstructed surfaces have consistent correspondences, as well as accurate geometry.

indicating the magnitude of the error indicated via the colorbars – note most of the error is quite below the maximal values chosen. As is evident from the image, the correspondences we compute are highly accurate and exhibit small to no error. Some drifting can occur in relatively flat regions, such as the woman's thigh, and around very extruded regions like the elephant's feet which are harder to model exactly.

We report quantitative evaluation of the correspondence and reconstruction in Table 1. To evaluate the quality of correspondences, we randomly draw $M = 500$ shape pairs (P_i, P_j) with known ground truth correspondences (p_k, q_k) where $p_k \in P_i$ and $q_k \in P_j$. Each shape has $N = 3125$ points. We report the average error over M pairs, with respect to the metrics described below.

Squared correspondence distance (m_{sL2}). This metric evaluates the error in the predicted inter-surface map $f_{i\to j}$ as $m_{sL2} = \frac{1}{N}\sum_{k=1}^{N} \|f(p_k) - q_k\|^2$.

Normalized correspondence rank (m_r). m_r expresses the rank of a predicted point with respect to all the other points on the target object. Formally $m_r = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N} \mathbb{1}_{\|q_l - q_k\|^2 < \|f(p_k) - q_k\|^2}$.

Area under the percentage of correct keypoints (PCK) curve (m_{AUC}). Following the literature on keypoint classification and correspondences [23, 55], we compute a mean PCK curve in a given range $[d_{\text{min}}, d_{\text{max}}]$ and report the area under that curve (AUC). We set $d_{\text{min}} = 0, d_{\text{max}} = 0.02$ in all our experiments.

Chamfer Distance (CD). This metric is equal to the loss term $\mathcal{L}_{\text{CD}}$ of Eq. 3. Note that this is the only metric that does not evaluate the quality of correspondences but rather of the reconstruction.

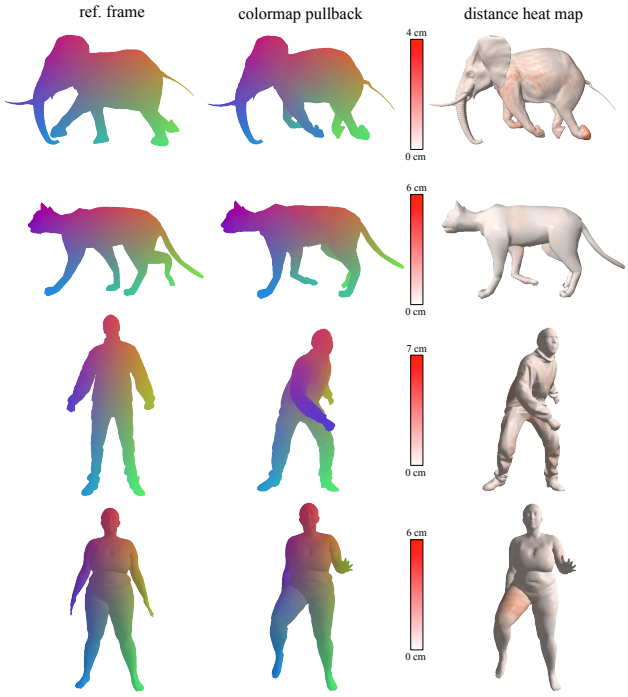

Figure 4. **Our correspondences retrieved on `elephant` and `cat` from ANIM, `jumping` from AMA, `running_on_spot` from DFAUST.** We visualize the correspondences via matching colors and show the error colorcoded as a heat map on the right.

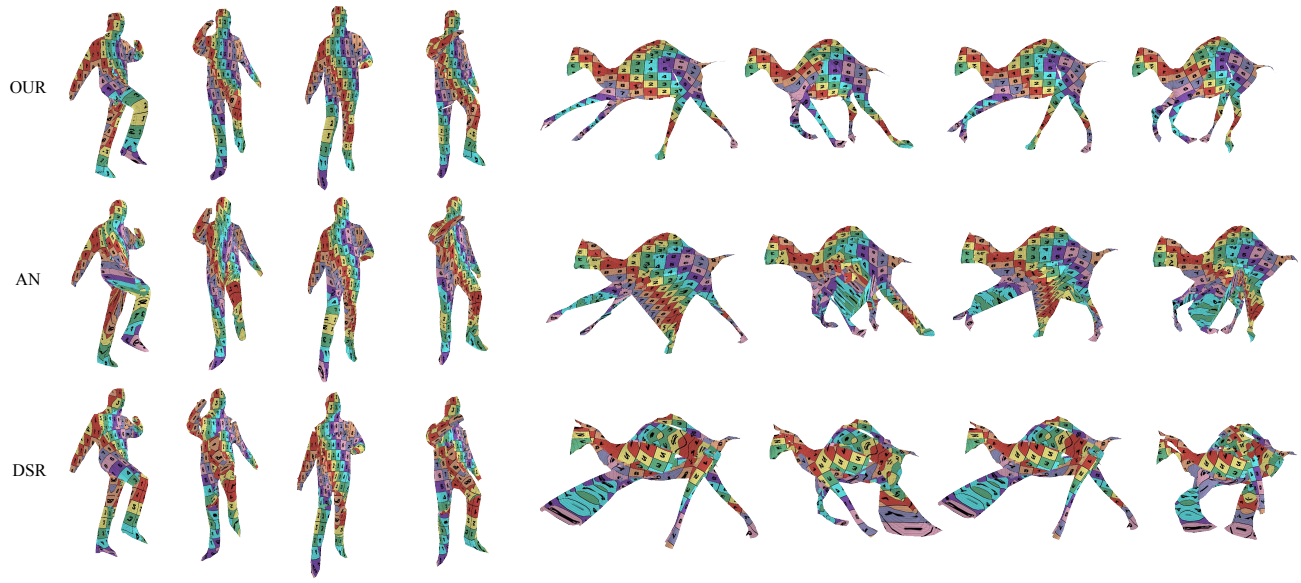

Figure 5. **Comparison of reconstructions of sequences `march_1` (AMA) and `camel` (ANIM) as produced by AN, DSR and OUR.** The other methods struggle to reconstruct the camel, and produce bad correspondences for the human (swapped legs on leftmost frame).

4.2. Datasets

We evaluate our method on 3 datasets of point cloud sequences. **Animals in motion [47, 5] (ANIM)** consists of 4 synthetic mesh sequences of 4-legged animals in stride. We uniformly scale each sequence s.t. the first point cloud fits in a unit cube. **Dynamic FAUST [9] (DFAUST)** is a real-world dataset which contains 14 sequences of 10 unclothed human subjects performing various actions. **Articulated mesh animation [50] (AMA)** is a real-world dataset containing 10 sequences depicting 3 different human subjects performing various actions, however here they are wearing loose-fitting clothes making the surface more intricate and time-varying, hence more challenging for correspondence methods. We pre-process the sequences to align them, by choosing the vertical axis rotation which minimizes the chamfer distance w.r.t. the previous frame.

For each dataset, we use one sequence of the entire data set for validation (`cat` for ANIM, `jumping jacks` for DFAUST, `crane` for AMA). We use this validation sequence to choose the hyper-parameter δ by training our model using $\delta \in [1, 6]$ and choosing the optimal one w.r.t the metrics m_{sL2}, m_r and m_{AUC}. We then report the metrics on the rest of the sequences.

To generate point clouds from these meshes, we perform uniform random sampling to draw 2500 points. We train and evaluate all methods on every sequence (e.g., walking cat or jumping human) individually. In the case of DFAUST, we simultaneously train on all subjects performing the sequence, but still draw point-cloud pairs of the same subject.

4.3. Results and Comparisons

We compare our approach (OUR) to both traditional and deep learning based methods.

Non-rigid ICP is a popular classic technique for shape registration. We use the recent implementation of [23], which we denote as nrICP. We experimented with several ways to use it to match shape pairs and chose the optimal one. Please refer to the supplementary material for details.

Atlas-based methods. As OUR builds on an atlas-based representation, we compare it to the original AtlasNet [19] (AN). We also compare to a more recent method [7] (DSR), which aims to reduce patch distortion, but is unaware of the temporal distortion. As the base architecture of both methods is nearly identical to that of OUR, for fair comparison we train both methods in the same way as summarized in Section 3.4.

Cycle consistent point cloud deformation. The recent method of [21] (CC) learns to align one point cloud to another in order to find correspondences. As the training of CC relies on sampling triplets, we experimented to find the optimal sampling technique for CC from the given sequence. Please refer to the supplementary for details.

All the deep learning based methods (AN, DSR, CC, OUR) are trained on the given sequence and then evaluated on it to retrieve the correspondences.

Figure 5 shows a qualitative comparison between OUR and other atlas-based methods on reconstructing surfaces from point clouds, AN and DSR. As expected, in both sequences our method is more temporally-coherent and the

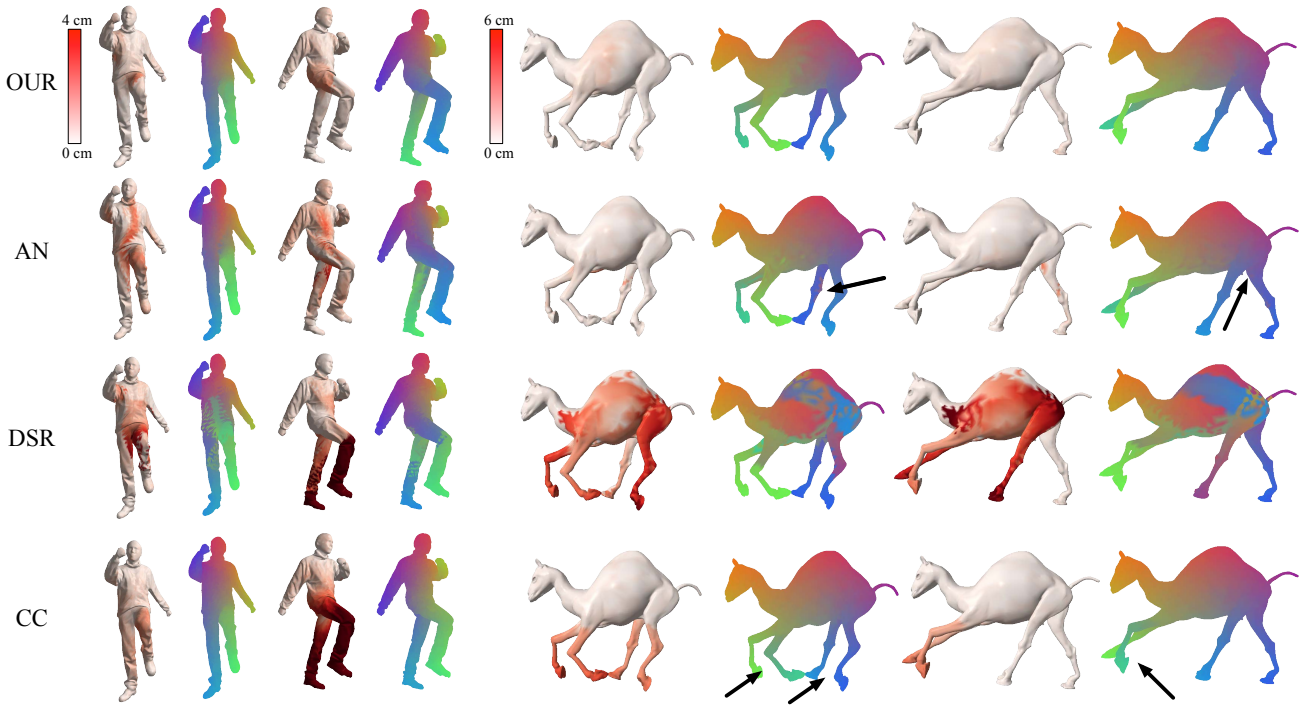

Figure 6. **Comparison of inferred point-cloud correspondences within 2 sequences.** The black arrows indicate point to errors and artifacts in the correspondences, such as swapped legs of the camel as predicted by DSR and CC, and minor problems, such as small but sever local mismatches, such as the knee area of the camel as predicted by AN.

correspondences are more accurate. Note how in the left-most frame of the human sequence, the bending at the knee causes AN to introduce a significant amount of unnecessary distortion, while DSR maps the left leg to the right leg and vice versa. The camel sequence reveals an even more interesting observation: the temporal coherence also acts as a regularizer and makes the reconstruction itself more tight and accurate to the point cloud's geometry, as our method's reconstruction is more true to the input than the competing methods. Please refer to the supplementary video to view the animations of the entire sequences.

In Figure 6, we show a representative qualitative comparison of the correspondences computed by OUR on the input point clouds, with ones inferred by the other techniques. We visualize the correspondences via matching colors, along with the measured correspondence error as a heat map. The correspondences inferred from DSR and CC swap the legs of the camel and the human. AN achieves comparable results to OUR on the camel, but exhibits a non-smooth jump in correspondences across the human's torso.

We report quantitative comparisons w.r.t. all metrics in Table 1, which demonstrates that our method achieves the best correspondence; we also achieve the best reconstruction quality (in terms of CD) over all other methods except for AtlasNet on AMA. Since CC and nrICP do not reconstruct surfaces, we do not report CD for them.

Table 1. **Comparison of OUR to SotA methods on correspondence accuracy and reconstruction quality**. Our method is the most accurate and also yields reconstruction quality competitive with AN.

dataset	model	$m_{sL2} \downarrow$	$m_r \downarrow$	$m_{\text{AUC}} \uparrow$	CD $\downarrow$
ANIM	nrICP	70.32±84.86	5.46±9.52	74.23±13.68	-
	AN	18.40±24.82	0.78±2.85	96.28±1.56	**0.09±0.00**
	DSR	46.43±67.42	3.44±6.71	83.96±9.58	0.19±0.01
	CC	33.84±54.13	2.21±4.75	87.96±7.76	-
	OUR	**11.93±11.00**	**0.30±0.57**	**98.10±0.61**	**0.09±0.00**
AMA	nrICP	150.94±134.31	6.63±10.26	45.40±22.27	-
	AN	86.80±91.28	2.90±6.18	70.07±15.31	**0.30±0.01**
	DSR	123.56±109.92	5.00±7.39	59.69±15.94	62.08±52.50
	CC	74.58±97.98	2.47±6.37	77.07±15.00	-
	OUR	**57.12±65.33**	**1.55±3.90**	**82.29±11.16**	0.32±0.02
DFAUST	nrICP	79.78±118.46	4.09±10.17	74.79±15.90	-
	AN	31.74±43.46	0.90±2.95	91.88±5.84	**0.34±0.06**
	DSR	68.79±61.04	3.76±5.19	78.00±6.25	11.21±2.89
	CC	29.57±65.26	1.12±5.26	94.35±9.82	-
	OUR	**19.81±22.19**	**0.38±1.17**	**96.17±2.31**	**0.34±0.06**

Stress test. In Figure 7, we test the limits of our method by reconstructing an extreme deformation sequence, of a rubber horse deflating. Even under the many foldovers of the model, our method reconstructs the legs as a separate part of the surface, while other baselines clamp different regions together, as can be seen from the bottom view. Please refer to the supplementary video for the entire sequence.

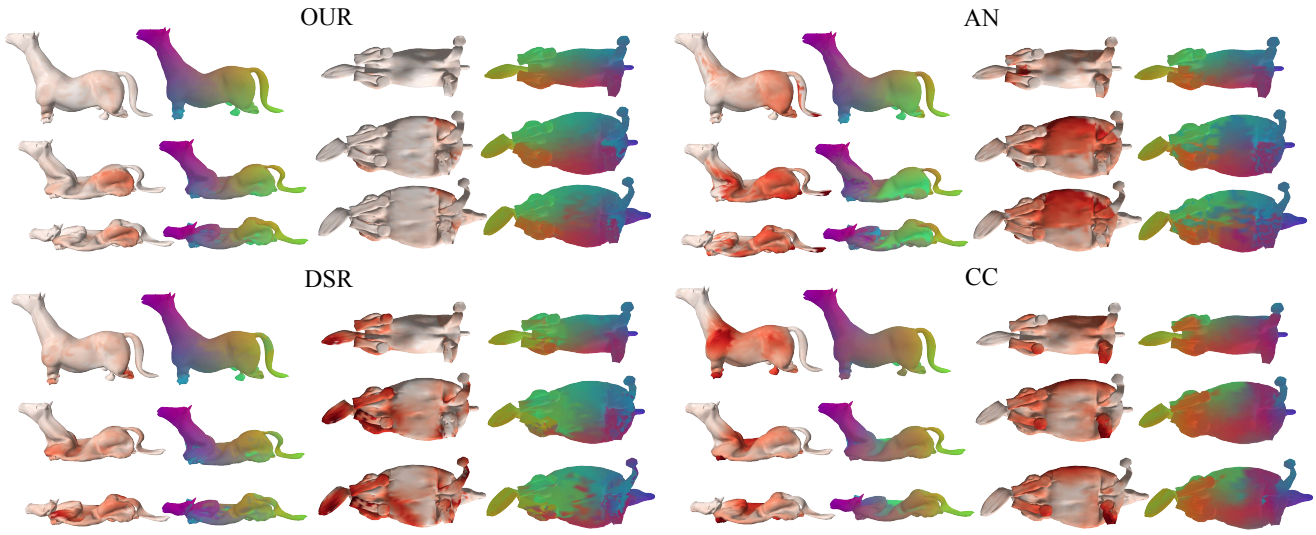

Figure 7. **Stress test on the sequence `horse_collapse` from ANIM, of a horse deflating.** Despite the many self-foldovers, our method still finds accurate correspondences. The other methods fail. For each method we show side and bottom views of 3 frames from the sequence, with correspondence visualized via matching colors, and a heatmap showing correspondence error.

Table 2. **Comparison of different pair sampling strategies.** Switching from our default (*neighbours*) to random leads to deterioration in correspondence accuracy and a slight improvement in CD.

strategy	$m_{sL2} \downarrow$	$m_r \downarrow$	$m_{\text{AUC}} \uparrow$	**CD** $\downarrow$
random	96.82±161.54	3.97±10.27	77.08±16.41	**0.30±0.01**
neighbors	**66.63±103.11**	**2.11±6.75**	**80.24±11.42**	0.31±0.02

Effect of the Sampling Strategy for Training Pairs. Instead of using time-adjacent point-cloud pairs, we can use random pairs instead (*random*). The results of this change are shown in Table 2. We evaluated both strategies on the `crane`, with $\delta = 1$. The results show that *neighbors* clearly yields higher correspondence accuracy. Interestingly, the deterioration in correspondence accuracy lets *random* produce slightly better reconstruction in terms of CD.

Effect of the Metric Consistency Term $\mathcal{L}_{\text{metric}}$. We evaluate the effect of the hyper-parameter α_{mc}, which balances metric consistency and chamfer distance. Results in terms of the correspondence metric m_{sL2} are shown in Table 3, using the validation sequences of all three datasets. Setting α_{mc} too low turns off $\mathcal{L}_{\text{mc}}$ while setting it too high overpowers $\mathcal{L}_{CD}$, which imposes strict isometry and makes the position of the patches ambiguous. Hence, different values may be less or more optimal, depending on the severity of the underlying deformation. $\alpha_{\text{mc}} \in [0.1, 1]$ yields the best results and the variations within that range are small. In all other experiments, we used $\alpha_{\text{mc}} = 0.1$, and we note that a better, automated method to choose α_{mc} may improve our performance further.

Table 3. **The impact of the metric consistency term $\mathcal{L}_{\text{metric}}$ on the resulting accuracy of correspondences.**

α_{mc}	$1e^{-4}$	$1e^{-3}$	$1e^{-2}$	0.1	1	$1e^{1}$	$1e^{2}$	$1e^{3}$
ANIM cat	13.2	11.3	**9.8**	**9.8**	12.5	14.0	15.1	62.5
AMA crane	127.2	232.5	111.8	66.6	**61.0**	102.6	179.3	174.3
DFAUST jacks	35.6	28.0	**23.1**	28.0	30.7	88.9	106.4	194.2

5. Conclusion

We have introduced an atlas-based method that yields temporally-coherent surface reconstructions in an unsupervised manner, by enforcing a point on the canonical shape representation to map to metrically-consistent 3D points on the reconstructed surfaces.

While our method yields better surface correspondences than state-of-the-art surface reconstruction techniques, it shares one shortcoming with these atlas-based methods. The reconstructed patches may overlap, causing imperfections in the reconstructions. Another limitation is that we use heuristics for the hyper-parameters balancing metric-consistency and reconstruction; employing an annealing-like technique which gradually permits more non-isometric deformations may be the next logical step.

We see many future applications to our approach. By replacing Chamfer distance with, e.g., some visual loss, we can apply our method to 2D sequences of images, which we believe could instigate progress in video-based 3D reconstruction. In the context of 3D geometry, our metric-consistency loss targets nearly-isometric deformations, however our framework could easily extend to other distortion measures, such as the conformal one. Studying this for non-isometric reconstruction and matching will be the focus of our future work.

References

[1] Noam Aigerman, Roi Poranne, and Yaron Lipman. Lifted bijections for low distortion surface mappings. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014. 2

[2] Marc Alexa, Daniel Cohen-Or, and David Levin. As-rigid-as-possible shape interpolation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 157–164, 2000. 2

[3] Romain Arcila, Cédric Cagniart, Franck Hétroy, Edmond Boyer, and Florent Dupont. Segmentation of temporal mesh sequences into rigidly moving components. *Graphical Models*, 2013. 1, 2

[4] M. Aubry, U. Schlickewei, and D. Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *ICCVW*, 2011. 2

[5] Gregoire Aujay, Franck Hetroy, Francis Lazarus, and Christine Depraz. Harmonic Skeleton for Realistic Character Animation. In *Eurographics/SIGGRAPH Symposium on Computer Animation*, 2007. 6

[6] Alex Baden, Keenan Crane, and Misha Kazhdan. Möbius Registration. *Computer Graphics Forum*, 2018. 2

[7] Jan Bednarik, Shaifali Parashar, Erhan Gundogdu, Mathieu Salzmann, and Pascal Fua. Shape reconstruction by learning differentiable surface representations. In *CVPR*, 2020. 1, 2, 4, 6

[8] Matthew Berger, Andrea Tagliasacchi, Lee Seversky, Pierre Alliez, Joshua Levine, Andrei Sharf, and Claudio Silva. State of the art in surface reconstruction from point clouds. In *Eurographics 2014-State of the Art Reports*, 2014. 2

[9] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *CVPR*, 2017. 6

[10] D. Boscaini, J. Masci, E. Rodola, and Bronstein M. M. Learning shape correspondence with anisotropic convolutional neural networks. In *NIPS*, 2016. 2

[11] Alexander M. Bronstein, Michael M. Bronstein, and Ron Kimmel. Generalized multidimensional scaling: A framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5):1168–1172, 2006. 2

[12] Qifeng Chen and Vladlen Koltun. Robust nonrigid registration by convex optimization. In *ICCV*, 2015. 2

[13] Luca Cosmo, Antonio Norelli, Oshri Halimi, Ron Kimmel, and Emanuele Rodolà. LIMP: Learning Latent Shape Representations with Metric Preservation Priors. *ECCV*, 2020. 2

[14] Zhantao Deng, Jan Bednarik, Mathieu Salzmann, and Pascal Fua. Better patch stitching for parametric surface reconstruction. In *3DV*, 2020. 2, 4

[15] T. Deprelle, T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry. Learning Elementary Structures for 3D Shape Generation and Matching. In *NeurIPS*, 2019. 2, 4

[16] Nicolas Donati, Abhishek Sharma, and Maks Ovsjanikov. Deep geometric functional maps: Robust feature learning for shape correspondence. In *CVPR*, June 2020. 1, 2

[17] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3D object reconstruction from a single image. In *CVPR*, 2017. 2

[18] Lin Gao, Shu-Yu Chen, Yu-Kun Lai, and Shihong Xia. Data-driven shape interpolation and morphing editing. In *Computer Graphics Forum*, 2017. 2

[19] T. Groueix, M. Fisher, V. Kim, B. Russell, and M. Aubry. Atlasnet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *CVPR*, 2018. 1, 2, 3, 4, 6

[20] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. 3d-coded : 3d correspondences by deep deformation. *ECCV*, 2018. 2, 4

[21] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. Unsupervised cycle-consistent deformation for shape matching. *Computer Graphics Forum*, 2019. 1, 2, 6

[22] Oshri Halimi, Or Litany, Emanuele Rodola, Alex M Bronstein, and Ron Kimmel. Unsupervised learning of dense shape correspondence. In *CVPR*, 2019. 1, 2

[23] Haibin Huang, Evangelos Kalogerakis, Siddhartha Chaudhuri, Duygu Ceylan, Vladimir G. Kim, and Ersin Yumer. Learning local shape descriptors from part correspondences with multiview convolutional networks. *ACM Trans. Graph.*, 2017. 5, 6

[24] Eldar Insafutdinov and Alexey Dosovitskiy. Unsupervised learning of shape and pose with differentiable point clouds. In *Advances in Neural Information Processing Systems*, 2018. 1, 2

[25] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2

[26] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 2

[27] Vladislav Kraevoy and Alla Sheffer. Cross-parameterization and compatible remeshing of 3d models. *ACM Transactions on Graphics (TOG)*, 23(3):861–869, 2004. 2

[28] Aaron W. F. Lee, David Dobkin, Wim Sweldens, and Peter Schröder. Multiresolution mesh morphing. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 343–350, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 2

[29] Chunyuan Li and A Ben Hamza. A multiresolution descriptor for deformable 3d shape retrieval. *The Visual Computer*, 29(6-8):513–524, 2013. 2

[30] Yaron Lipman, Daniel Cohen-Or, and David Levin. Data-dependent mls for faithful surface approximation. In *Proceedings of the fifth Eurographics symposium on Geometry processing*, pages 59–67, 2007. 2

[31] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M.J. Black. Smpl: A skinned multi-person linear model. In *SIGGRAPH Asia*, 2015. 2

[32] Jonathan Masci, Davide Boscaini, Michael M. Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 37–45, 2015. 2

[33] F. Mémoli and S. Sapiro. *A theoretical and computational framework for isometry invariant recognition of point cloud data*. Springer, 2005. 2

[34] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2

[35] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *CVPR*, 2017. 2

[36] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, and L. Guibas. Functional maps: A flexible representation of maps between shapes. *ACM Trans. Graph.*, 2012. 2

[37] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9964–9973, 2019. 2

[38] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[39] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016. 2

[40] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *CVPR*, 2017. 3

[41] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, 2017. 2

[42] Marie-Julie Rakotosaona and Maks Ovsjanikov. Intrinsic point cloud interpolation via dual latent space navigation. In *ECCV*, 2020. 1

[43] Marie-Julie Rakotosaona and Maks Ovsjanikov. Intrinsic point cloud interpolation via dual latent space navigation. *arXiv preprint arXiv:2004.01661*, 2020. 2

[44] E. Rodola, S. Rota Bulo, T. Windheuser, M. Vestner, and D. Cremers. Dense non-rigid shape correspondence using random forests. *CVPR*, 2014. 2

[45] Jean-Michel Roufosse, Abhishek Sharma, and Maks Ovsjanikov. Unsupervised deep learning for structured shape matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[46] M. Salzmann and P. Fua. Reconstructing Sharply Folding Surfaces: A Convex Formulation. In *CVPR*, June 2009. 2

[47] Robert W. Sumner and Jovan Popović. Deformation transfer for triangle meshes. *ACM Trans. Graph.*, 2004. 6

[48] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A Concise and Provably Informative Multi-Scale Signature Based on Heat Diffusion. *Computer Graphics Forum*, 2009. 2

[49] Kiran Varanasi and Edmond Boyer. Temporally Coherent Segmentation of 3D Reconstructions. In *3DPVT 2010 - 5th International Symposium on 3D Data Processing, Visualization and Transmission*, 2010. 1, 2

[50] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.*, 2008. 6

[51] Ofir Weber and Denis Zorin. Locally injective parametrization with arbitrary fixed boundaries. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014. 2

[52] Francis Williams, Teseo Schneider, Claudio Silva, Denis Zorin, Joan Bruna, and Daniele Panozzo. Deep geometric prior for surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10130–10139, 2019. 2

[53] Tim Winkler, Jens Drieseberg, Marc Alexa, and Kai Hormann. Multi-scale geometry interpolation. In *Computer graphics forum*, 2010. 2

[54] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *CVPR*, 2018. 2

[55] Yang You, Yujing Lou, Chengkun Li, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Cewu Lu, and Weiming Wang. Keypointnet: A large-scale 3d keypoint dataset aggregated from numerous human annotations. In *CVPR*, 2020. 5

[56] S. Zuffi and M. J. Black. The stitched puppet: A graphical model of 3d human shape and pose. In *CVPR*, 2015. 2