Self-Supervised Learning for Multimodal Non-Rigid 3D Shape Matching
– SUPPLEMENTARY DOCUMENT –

Dongliang Cao   Florian Bernard
University of Bonn

In this supplementary document we first provide more implementation details of our method. Next, we explain details on the modifications of our approach for partial shape matching. Afterwards, we provide more ablative experiments to demonstrate the advantages of our method. Eventually, we show additional qualitative results of our method.

S1. More implementation details

Our learning framework is implemented in PyTorch and uses the original DiffusionNet implementation[1]. In the context of the functional map framework, we choose the first 80 LBO eigenfunctions as basis functions for complete shape matching. For partial shape matching, we choose the number to be 50 and 30 for the CUTS and HOLES subsets of the SHREC'16, respectively. Similarly, we choose the number to be 30 for the partial view matching. As for deep feature similarity, we use Sinkhorn normalisation with the number of iterations equal to 10 and temperature parameter equal to 0.2. We train our feature extractor with the Adam optimiser with learning rate equal to 10^{-3}. The batch size is chosen to be 8 for SURREAL dataset and 1 for other datasets.

S2. Modifications for partial shape matching

In the context of partial shape matching, the functional map from the complete shape to the partial shape becomes a slanted diagonal matrix. Analogous to DPFM, we regularise the predicted functional maps based on this property. Specifically, for $\mathcal{X}$ being the complete shape and $\mathcal{Y}$ being the partial shape, the unsupervised functional map regularisation can be modified as

$$\mathcal{L}_{\text{bij}} = \|C_{xy}C_{yx} - \mathbf{I}_r\|_F^2, \mathcal{L}_{\text{orth}} = \|C_{xy}C_{xy}^\top - \mathbf{I}_r\|_F^2, \quad \text{(S1)}$$

where $\mathbf{I}_r$ is a diagonal matrix in which the first r elements on the diagonal are equal to 1, and r is related to the surface area ratio between two shapes. To obtain the soft correspondence matrix $\hat{\Pi}_{xy}$, we replace Sinkhorn normalisation by the column-wise softmax operator.

S3. Ablation study

Supervised contrastive learning. One of the key components of our self-supervised loss terms is the unsupervised functional map regularisation. To evaluate its importance, we replace the unsupervised losses in Eq. (5) and Eq. (6) by supervised contrastive loss similar to Eq. (7), i.e.

$$E_{\text{sup}} = -\sum_{(i,j)\in\mathcal{P}} \log \frac{\exp\left(\mathcal{F}_x^i \cdot \mathcal{F}_y^j/\tau\right)}{\sum_{(\cdot,k)\in\mathcal{P}} \exp\left(\mathcal{F}_x^i \cdot \mathcal{F}_y^k/\tau\right)}, \quad \text{(S2)}$$

where $\mathcal{P}$ is the set of matched points between shape $\mathcal{X}$ and shape $\mathcal{Y}$. For this ablation experiment, we consider the same experiment setting as in Sec. 5.2 to avoid over-fitting.

Geo. (×100)	**F (PC)**	**S (PC)**	**S19 (PC)**
with E_{sup}	**1.5** (4.8)	5.2 (6.8)	6.9 (8.1)
Ours	2.0 (**3.5**)	**3.2** (**3.8**)	**4.4** (**6.6**)

Table S1. Quantitative results on the **F**AUST, **S**CAPE and **S**HREC'**19** datasets trained on SURREAL dataset. The **best** results in each column are highlighted.

The quantitative results are summarised in Tab. S1. Notably, our self-supervised approach outperforms the supervised counterpart in most settings. The reason is that the unsupervised functional map regularisation enforces more smooth and consistent correspondences in comparison to the supervised contrastive learning.

Robustness to initial pose. As indicated in the limitation part, our method takes vertex position as input and is thus not rotation-invariant. To be more robust to the choice of initial pose, during training we randomly rotate input shapes as data augmentation, thereby encouraging that the extracted features are less sensitive to the initial pose of the shape. To evaluate the performance, we follow the experiment setting in Sec. 5.1, with the only difference being that here all test shapes are randomly rotated around the vertical axis.

Tab. S2 summarises the quantitative results. We observe that the network performance can be substantially im-

[1] https://github.com/nmwsharp/diffusion-net

Geo. (×100)	**F (PC)**	**S (PC)**	**S19 (PC)**
Ours (w/o aug.)	8.8 (12.0)	14.0 (14.2)	13.9 (14.7)
Ours (w/ aug.)	**4.7 (5.6)**	**5.3 (6.2)**	**6.0 (6.8)**

Table S2. Quantitative results on the **F**AUST, **S**CAPE and **S**HREC'**19** datasets in terms of mean geodesic errors (×100). All test shapes are randomly rotated. The best results in each column are highlighted.

proved by using a random rotation as data augmentation during training. Fig. S1 shows some qualitative results of our method on FAUST dataset with randomly initial poses.

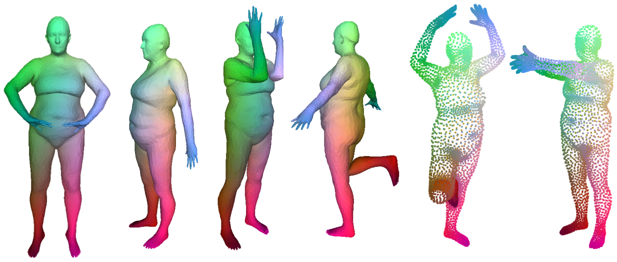

Figure S1. Qualitative results on FAUST dataset with different initial poses for both mesh and point cloud matching.

Robustness to noise. As mentioned in the main paper, previous deep functional map methods predict point maps based on the functional map framework. However, point clouds only admit an inaccurate estimation of LBO eigenfunctions, especially in the presence of noise. Therefore, directly applying such methods to point clouds leads to a large performance drop. In contrast, our method predicts point maps based on the deep feature similarity without relying on the functional map framework.

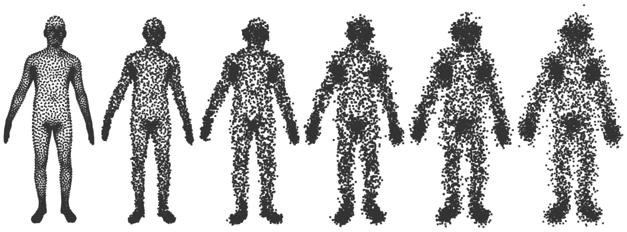

Figure S2. One shape on FAUST dataset with increasing noise magnitude from left to right. The leftmost one is the clean point cloud.

To further evaluate our method's robustness to noise, we add an increasing amount of zero-mean isotropic Gaussian noise to point positions of shapes in the test set of the FAUST dataset. For a fair comparison, we do not train or fine-tune the networks on each noise magnitude. As a proof-of-concept, we choose our method and a simple baseline that is based on our framework but does not use E_{align}, E_{nce} during training for comparison, which is similar to Sec. 6. We note that the simple baseline predicts point maps based on the functional map framework via functional maps conversion. Fig. S2 shows the point cloud with different noise magnitude. Fig. S3 plots the mean geodesic errors on the FAUST dataset w.r.t. the corresponding noise magnitude. We observe that our method achieves better results and is much more robust against noise, especially for large degrees of noise.

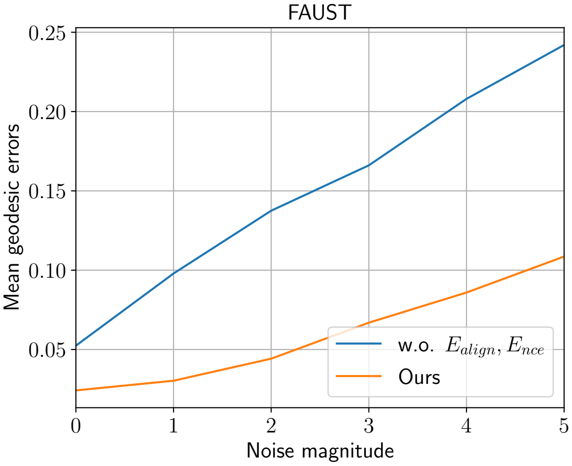

Figure S3. Mean geodesic errors for point cloud matching in different noise magnitudes. Our method achieves more robust point cloud matching based on deep feature similarity.

Robustness to sampling. To evaluate our method's robustness to varying sampling density, we use an anisotropic remeshed version of the FAUST and SCAPE datasets (denoted F_a and S_a). Below we show an example shape pair with varying sampling density and corresponding matching.

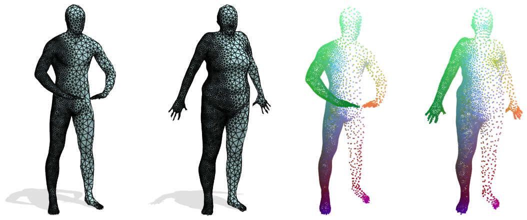

Figure S4. An example shape pair and the corresponding qualitative result of our method on the anisotropic remeshed FAUST dataset.

Tab. S3 shows that both ConsistFMaps and Deep Shells overfit to the sampling density (with SHOT descriptor they overfit both for meshes and point clouds; with vertex position as descriptor they only overfit for point clouds). In contrast, our method is more robust and demonstrates better performance, particularly for point clouds.

Train	FAUST		SCAPE	
Test	F (PC)	F_a (PC)	S (PC)	S_a (PC)
ConsistFMaps (w/ xyz)	2.4 (11.2)	2.9 (12.4)	5.1 (12.3)	5.4 (13.1)
w/ SHOT (original)	**1.5** (16.4)	15.3 (32.1)	**2.0** (18.3)	6.9 (24.8)
Deep Shells (w/ xyz)	1.7 (6.0)	2.7 (7.2)	5.3 (7.8)	5.7 (8.4)
w/ SHOT (original)	1.7 (13.2)	12.0 (18.8)	2.5 (14.1)	10.0 (18.3)
Ours	2.0 (**2.4**)	**2.6** (**3.0**)	3.1 (**4.1**)	**3.3** (**4.4**)

Table S3. Quantitative results on **F**AUST, **S**CAPE and their anisotropic remeshed versions. All methods are trained on the original datasets.

Matching with outliers. Fig. S5 shows an example matching result from real-scanned raw point clouds (by transferring texture). We observe that the extracted DiffusionNet features (see colour-coded shapes on the left and right, which visualise DiffusionNet features projected onto three RGB channels via t-SNE) are degraded due to the outliers. Since we use DiffusionNet, our method carries over this known limitation (of DiffusionNet).

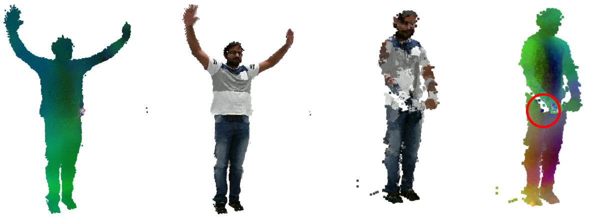

Figure S5. A qualitative result from real-scanned raw point clouds and the corresponding extracted features from DiffusionNet.

Data efficiency. We train our method on the entire SURREAL dataset and summarise the results in Tab. S4. When using significantly more data, our method achieves a (slightly) better cross-dataset generalisation ability. Compared to point cloud matching methods, our method utilises the strong functional map regularisation and explicitly considers multi-modal training, thus requires only a small amount of training data.

|Data|	F (PC)	S (PC)	S19 (PC)
5k	2.0 (3.5)	3.2 (3.8)	4.4 (6.6)
230k	**1.9** (**3.2**)	**3.0** (**3.6**)	**4.0** (**5.8**)

Table S4. Cross-dataset generalisation evaluated on the **F**AUST, **S**CAPE and **S**HREC’**19** datasets and trained on the SURREAL dataset.

S4. More qualitative results

In this section, we provide more qualitative matching results on diverse shape matching datasets, see Figs. S6-S10.

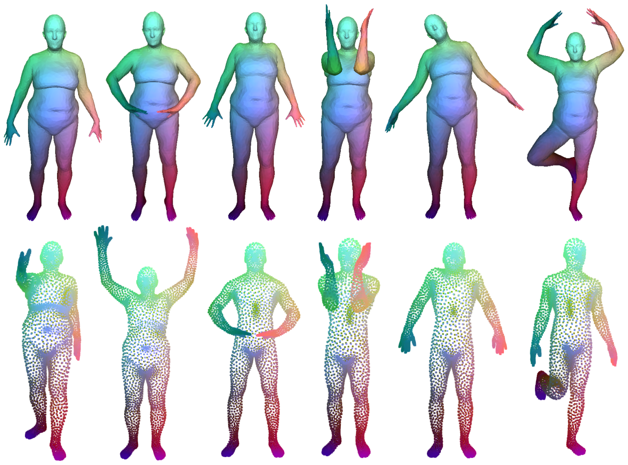

Figure S6. Qualitative results on FAUST dataset of our method applied to both meshes and point clouds. Our method achieves accurate matchings for both modalities.

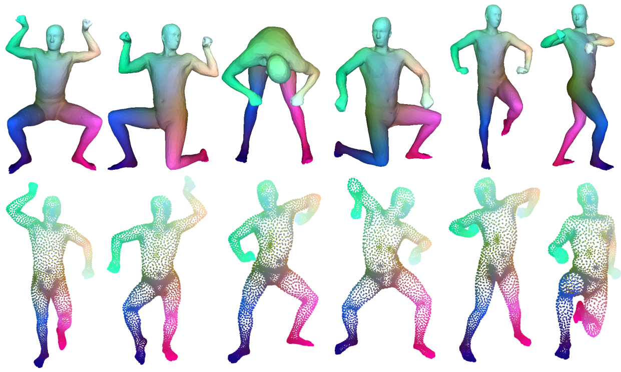

Figure S7. Qualitative results on SCAPE dataset of our method applied to both meshes and point clouds.

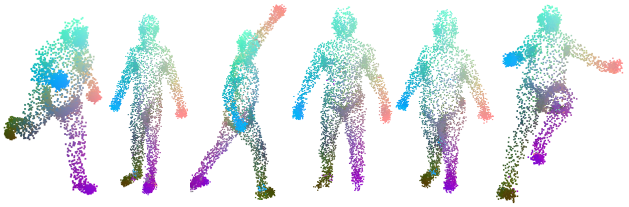

Figure S8. Qualitative results on SHREC’19 dataset of our method applied to noisy point clouds. Our method enables accurate point cloud matching even in the presence of noise.

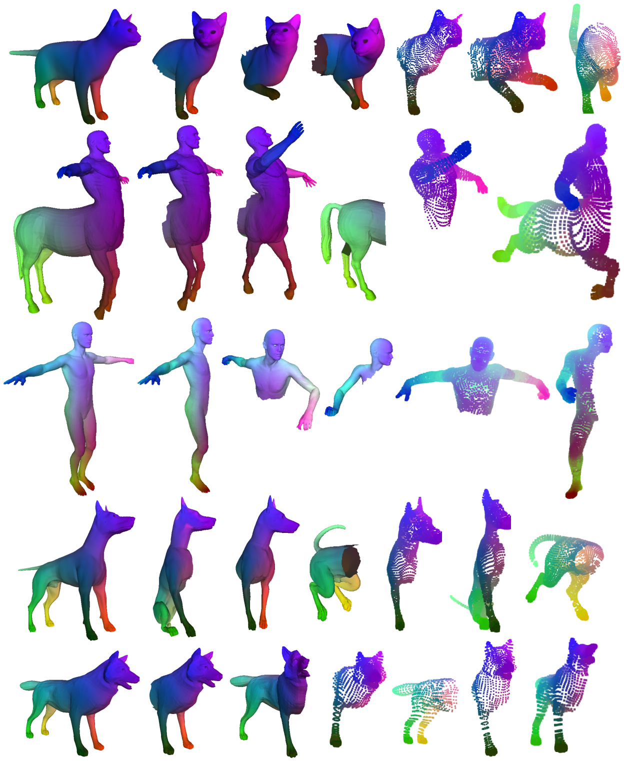

Figure S9. Qualitative partial shape matching results on SHREC'16 dataset of our method applied to both meshes and point clouds. The leftmost one is the complete shape to be matched. Our method enables accurate multimodal partial shape matching.

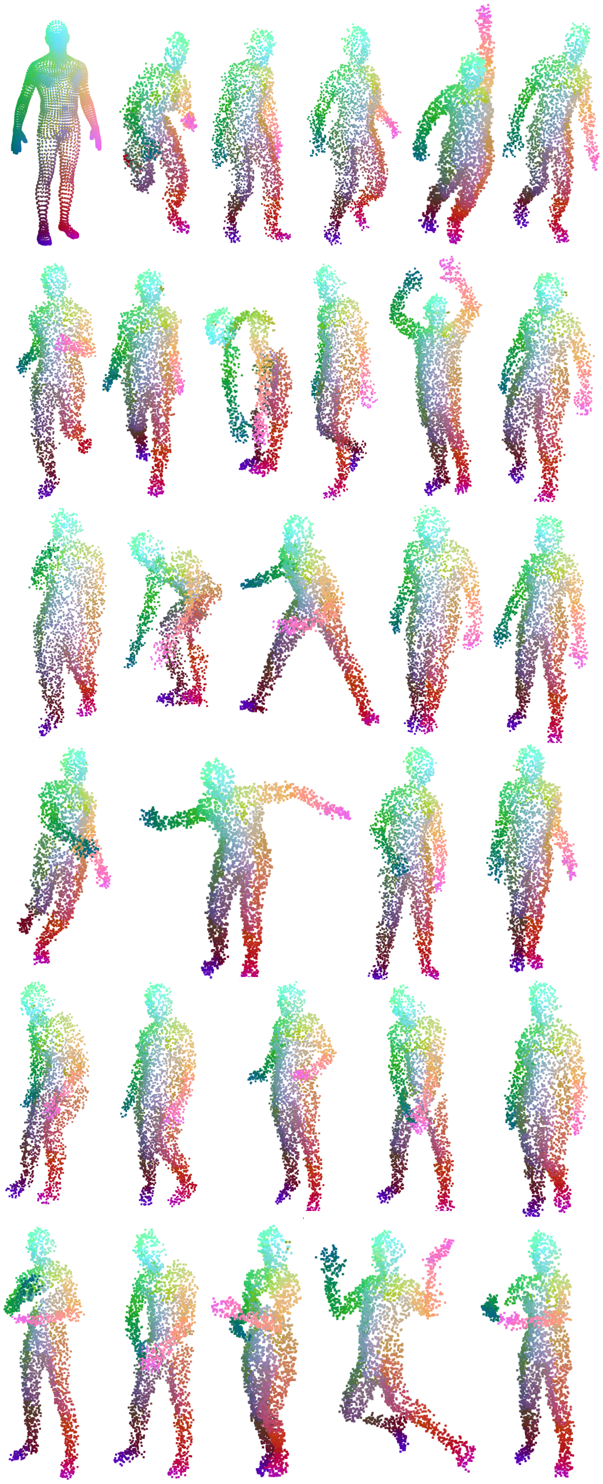

Figure S10. Qualitative partial view matching results on SURREAL-PV dataset of our method applied to noisy partial observed point clouds. The top-left one is the complete shape to be matched. Our method obtains accurate correspondences for partially-observed noisy point clouds with different sampling and disconnected components.