Supplementary Material: Deformation-Guided Unsupervised Non-Rigid Shape Matching

Aymen Merrouche[1]
aymen.merrouche@inria.fr
João Regateiro[2]
joaoregateiro@gmail.com
Stefanie Wuhrer[1]
stefanie.wuhrer@inria.fr
Edmond Boyer[1]
edmond.boyer@inria.fr

[1] Inria centre
University Grenoble Alpes
Grenoble, France
[2] Interdigital
Cesson-Sévigné, France

This supplementary material provides implementation details of our method in Section 1, more details on the datasets used in Section 2, ablation studies in Section 3, and additional quantitative and qualitative results in Section 4.

1 Implementation Details

1.1 Patch Extraction

To compute multi-resolution surface patches, we consider a greedy approach based on the furthest point sampling strategy inspired by [8]. Starting from a randomly selected vertex x_1, we compute the geodesic distance map U_{x_1} to all other vertices on the mesh. Then, given that we have a set of vertices $S_n = \{x_1, .., x_n\}$, and their distance map U_n, we select the new vertex x_{n+1} to be the furthest vertex from S_n. We compute $U_{x_{n+1}}$, the distance map from x_{n+1} and update $U_{n+1} = \min(U_n, U_{x_{n+1}})$ and add x_{n+1} to S_n to get S_{n+1}. We stop the algorithm when a target number of points is reached. To get multiple patch resolutions we stop the algorithm at increasing numbers of target points.

We select $L+1$ patch resolutions where patches at level 0 are restricted to vertices and level L represents the coarsest patch level. For each hierarchical level, the selected samples are used as patch centers $\mathcal{C}_l = (c_i^l \in \mathbb{R}^3)_{1 \leq i \leq n_l}$ and their corresponding Voronoi cells on the mesh as patches $(P_i^l)_{1 \leq i \leq n_l}$ for $l = 0, \dots, L$.

In all of our experiments, we extract 4 patch resolutions : all the mesh vertices, 800, 200 and 50 patches. Figure 1 shows an example of these surface patches.

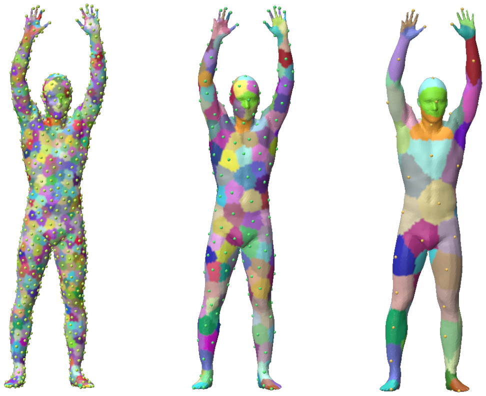


Figure 1: Example of surface patches. From left to right : 800, 200 and 50 patches along with their centers on a mesh with $\approx 13k$ vertices.

1.2 Architecture Detail

Pooling and Unpooling The extracted surface patches are not strictly hierarchical in the sense that vertices of a patch at level l can belong to multiple patches at coarser levels $l+1$. We use these surface patches for pooling and unpooling operations in our architectures. To pool features from patches $(P_i^l)_{1\leq i\leq n_l}$ of level l to coarser patches $(P_i^{l+1})_{1\leq i\leq n_{l+1}}$ of level $l+1$ we proceed in two step:

1. We first unpool to the vertex level (level 0) such that each vertex is associated with the features of the patch it belongs to in level l.

2. We then employ max-pooling to go to patch level $l+1$.

To unpool features from patches $(P_i^{l+1})_{1\leq i\leq n_{l+1}}$ of level $l+1$ to finer patches $(P_i^l)_{1\leq i\leq n_l}$ of level l we proceed similarly. We first unpool to the vertex level (level 0) such that each vertex is associated with the features of the patch it belongs to in level $l+1$. We then employ max-pooling to go to patch level l.

Feature Extractor In both the association and the deformation networks, we use identical feature extractors based on hierarchical graph convolutional network FeaStConv operators [9]. In all cases, we fixed the number of attention heads of this operator to 9. Figure 2 illustrates the feature extractor architecture.

Association Network Figure 3 details the architecture of the association network. We fix the temperature parameter of the softmax operator to 10^{-2}.

Deformation Network Figure 4 details the architecture of the deformation network. It uses a deformation decoder module that outputs per-patch rotation and translation parameters detailed in Figure 5.

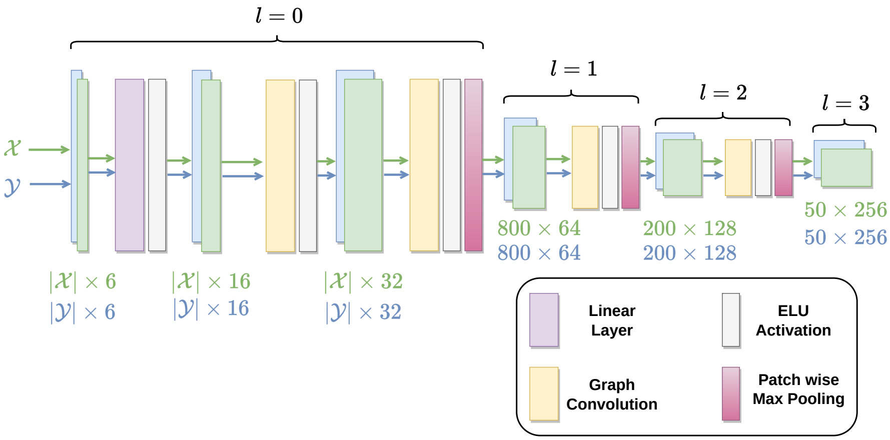


Figure 2: Feature extractor architecture. Inputs $\mathcal{X}$ and $\mathcal{Y}$ are decomposed into a hierarchy of 800, 200 and 50 patches and fine-to-coarse features are extracted. Input features at the vertex level are 3D coordinates and normals.

1.3 Training and Inference

Loss Weights The final loss per level l is the weighted combination of five self-supervised criteria, as given in Equation (5) of the main paper. Table 1 shows the weights used to train the network. The geodesic criterion was not used at the vertex level (weight fixed to 0) as it involves matrix multiplications that are computationally prohibitive. The matching loss and the rigidity loss are somehow in opposition: The rigidity loss promotes isometric deformations whereas the matching loss promotes deformations that reflect the association matrix. Their weights are fixed so as to satisfy the association matrices while preserving the spatial continuity at the patch borders.

Weight \ Level	$l=0$	$l=1$	$l=2$	$l=3$
λ_g^l	0.0	1.0	1.0	1.0
λ_c^l	10.0	10.0	10.0	10.0
λ_r^l	10.0	10.0	10.0	10.0
λ_m^l	2.0	2.0	2.0	2.0
λ_{ri}^l	200.0	100.0	50.0	10.0

Table 1: Loss weights for every level.

Training Our network is trained using Adam [4] with gradient clipping. The learning rate is fixed to 1×10^{-3} for the first epoch, 5×10^{-4} between the 2^{nd} and 10^{th} epochs and 2.5×10^{-4} after the 10^{th} epoch. Our model takes 372 epochs to train on the Extended FAUST training set. We select the model that achieves the smallest loss on a validation set.

Inference At test time we allow the network to specialise for each new shape pair to improve the matching. This is achieved by resuming the training of the selected model on a training set restricted to the two input shapes, with the same fixed architecture, optimization

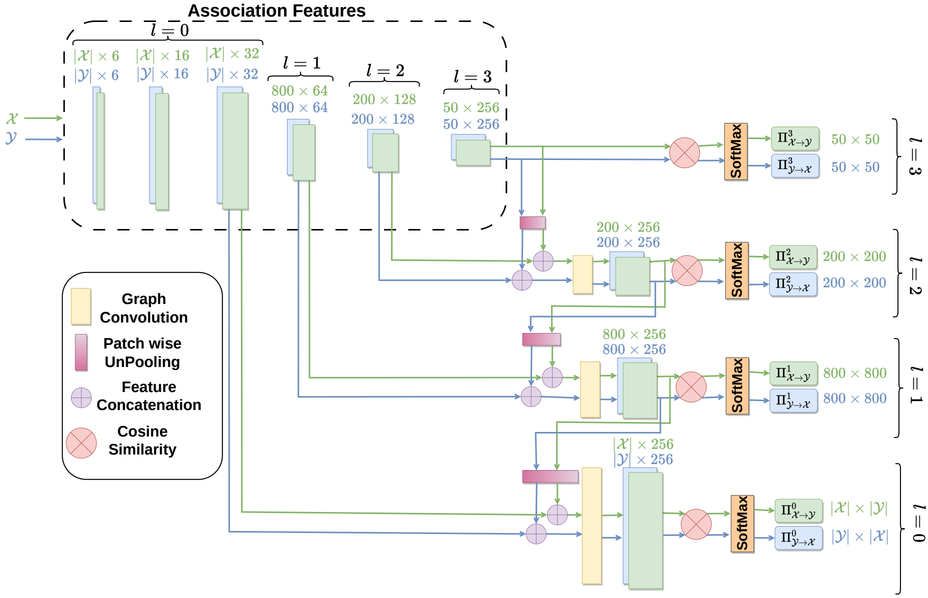

Association Features
$l=0$
$|\mathcal{X}| \times 6$
$|\mathcal{Y}| \times 6$
$|\mathcal{X}| \times 16$
$|\mathcal{Y}| \times 16$
$|\mathcal{X}| \times 32$
$|\mathcal{Y}| \times 32$
$l=1$
800×64
800×64
$l=2$
200×128
200×128
$l=3$
50×256
50×256
$\mathcal{X}$
$\mathcal{Y}$
SoftMax
$\Pi^3_{\mathcal{X}\to\mathcal{Y}}$
$\Pi^3_{\mathcal{Y}\to\mathcal{X}}$
50×50
50×50
$l=3$
200×256
200×256
$\Pi^2_{\mathcal{X}\to\mathcal{Y}}$
$\Pi^2_{\mathcal{Y}\to\mathcal{X}}$
200×200
200×200
$l=2$
800×256
800×256
$\Pi^1_{\mathcal{X}\to\mathcal{Y}}$
$\Pi^1_{\mathcal{Y}\to\mathcal{X}}$
800×800
800×800
$l=1$
$|\mathcal{X}| \times 256$
$|\mathcal{Y}| \times 256$
$\Pi^0_{\mathcal{X}\to\mathcal{Y}}$
$\Pi^0_{\mathcal{Y}\to\mathcal{X}}$
$|\mathcal{X}| \times |\mathcal{Y}|$
$|\mathcal{Y}| \times |\mathcal{X}|$
$l=0$
Graph Convolution
Patch wise UnPooling
Feature Concatenation
Cosine Similarity

Figure 3: Architecture of the association network. Given meshes $\mathcal{X}$ and $\mathcal{Y}$ it outputs coarse-to-fine association maps at every level of the patch hierarchy, $\Pi^l_{\mathcal{X}\to\mathcal{Y}}$ and $\Pi^l_{\mathcal{Y}\to\mathcal{X}}$ for every level l.

technique and hyper-parameters. We refer to this specialisation as fine tuning in the main paper. In all our experiments we fix the number of epochs for fine tuning to 50.

2 Data

We detail below how ground truth correspondences were obtained to evaluate our approach.

Pre-processed Data For our experiments on pre-processed data, we used the extended FAUST dataset [1] which is composed of meshes with the same template connectivity. In order to remove the connectivity consistency, which can strongly bias the matching, we remeshed all shapes individually and created the 3 test sets mentioned in the main paper (sec 4.1). To get the ground truth correspondences for vertices on the remeshed shapes, we revert to the connectivity consistent meshes by searching, for each vertex on a remeshed shape, the triangle on the original mesh, with minimal distance along the vertex normal direction. As a result of the connectivity changes, such distance can be large and we discard in the evaluation vertices for which this distance is higher than $2/10$ of the remeshed shape's mean edge length. This corresponds to 4.87% of the vertices for uniform remeshings to $5k$ vertices, 1.2% for uniform remeshing to $15k$ vertices and 1.01% for curvature adapted remeshings to $5k$ vertices. A similar strategy is used for meshes altered with topological noise. In this case, 7.69% of the vertices are discarded in the evaluation.

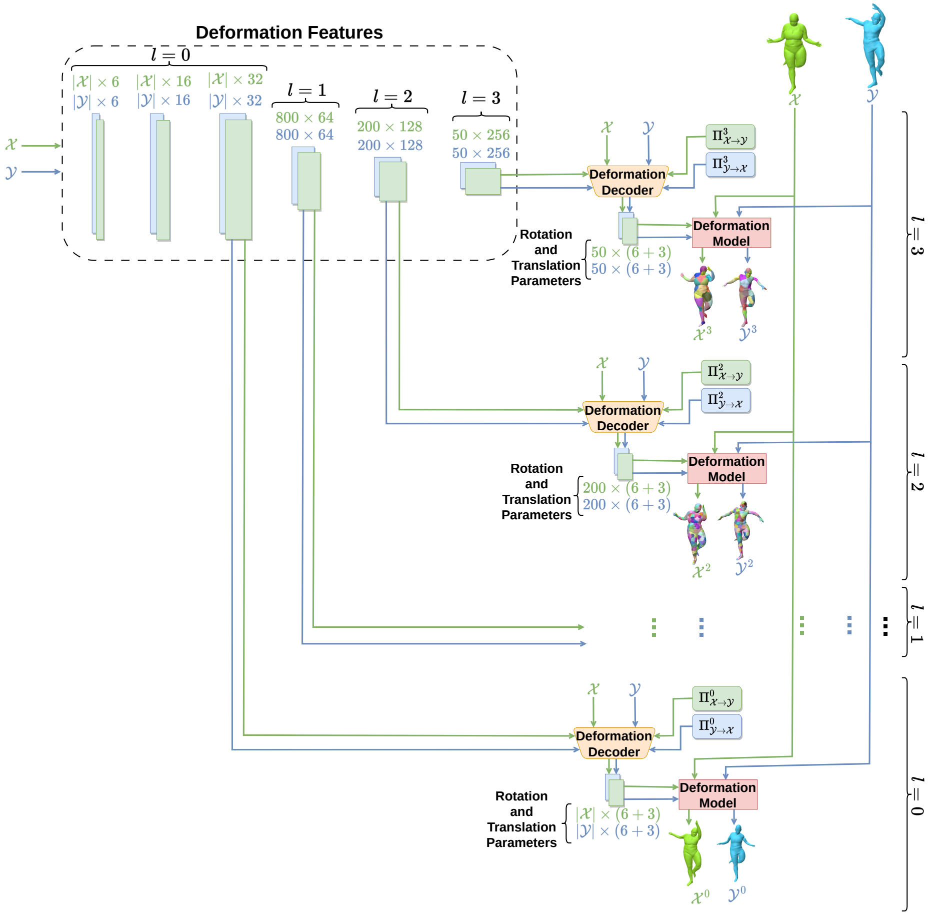


Figure 4: Architecture of the deformation network. Given $\mathcal{X}$, $\mathcal{Y}$, $\Pi^l_{\mathcal{X}\to\mathcal{Y}}$, and $\Pi^l_{\mathcal{Y}\to\mathcal{X}}$ it estimates the 3D deformations $\mathcal{X}^l$ and $\mathcal{Y}^l$ of $\mathcal{X}$ and $\mathcal{Y}$ for every level l.

Raw 3D Acquisition We also experimented our matching approach with raw 3D scans, *e.g.* [7]. In this case, ground truth matchings were obtained by fitting the SMPL model [6] to the scans and considering closest vertices on the template in the normal direction. Again here, vertices with distances to the template higher than 2 times the scan's mean edge length are discarded in the evaluation. This corresponds in practice to 8.04% for the naked raw acquisition dataset and 17.42% for the clothed one.

3 Ablation Studies

We present ablation studies that evaluate the respective benefits of the main components of our approach, i.e. the hierarchical modeling and the deformation model constraint. The impact of fine tuning is also given. To assess the hierarchical modeling, we use only two

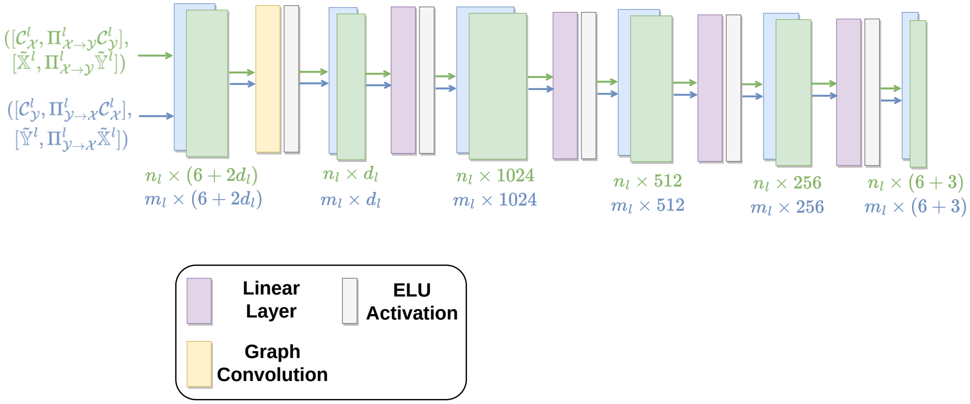

$([\mathcal{C}^l_{\mathcal{X}}, \Pi^l_{\mathcal{X}\to\mathcal{Y}}\mathcal{C}^l_{\mathcal{Y}}], [\tilde{\mathbb{X}}^l, \Pi^l_{\mathcal{X}\to\mathcal{Y}}\tilde{\mathbb{Y}}^l])$
$([\mathcal{C}^l_{\mathcal{Y}}, \Pi^l_{\mathcal{Y}\to\mathcal{X}}\mathcal{C}^l_{\mathcal{X}}], [\tilde{\mathbb{Y}}^l, \Pi^l_{\mathcal{Y}\to\mathcal{X}}\tilde{\mathbb{X}}^l])$
$n_l \times (6+2d_l)$
$m_l \times (6+2d_l)$
$n_l \times d_l$
$m_l \times d_l$
$n_l \times 1024$
$m_l \times 1024$
$n_l \times 512$
$m_l \times 512$
$n_l \times 256$
$m_l \times 256$
$n_l \times (6+3)$
$m_l \times (6+3)$
Linear Layer
ELU Activation
Graph Convolution

Figure 5: Architecture of the deformation decoder. It is composed of a graph convolution followed by an MLP.

levels of hierarchy: the mesh vertices and 800 patches. To assess the deformation model constraint, the network is restricted to the association network, losses involving the deformations, i.e. matching and rigidity loss, are discarded. We trained the models on extended FAUST and tested on both extended FAUST and the raw 3D acquisition data with everyday clothing. Tab. 2 shows the results and the number of epochs required to train each model. The complete model achieves the best results on both pre-processed and raw acquisition data. Note that the dimensionality reduction using the hierarchical modeling of associations is essential to the unsupervised learning and allows for much faster training. Note also that the deformation model improves the quality of the matching and that the fine tuning improves the results across all models.

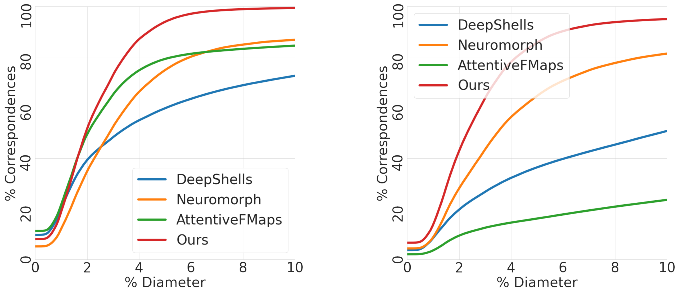
Hierarchical Feature Space	Deformation Model	Fine Tuning	Number of Epochs to Train	Pre-processed data		Raw 3D acquisitions	
				Cycle Geodesic Error	Mean Geodesic Error	Cycle Geodesic Error	Mean Geodesic Error
✓	✓	✓	372	**0.0259**	**4.09**	**0.0463**	**6.02**
✓	✓	✗	372	0.0352	4.99	0.0980	9.13
✗	✓	✓	803	0.0653	14.01	0.0556	9.41
✗	✓	✗	803	0.0855	14.49	0.0890	11.26
✓	✗	✓	311	0.0260	4.24	0.0519	6.44
✓	✗	✗	311	0.0336	4.88	0.1247	10.77

Table 2: Ablation tests for the hierarchy in feature space, the deformation model and the fine tuning.

To prove the added benefit of the Self-Reconstruction Criterion that ensures that each patch, on the shape itself, is identified to avoid many-to-one matches, we present an ablation where we both train and test the models on extended FAUST. Tab. 3 shows the results.

Self Reconstruction Criterion	Fine Tuning	Cycle Geodesic Error	Mean Geodesic Error
✓	✗	**0.0352**	**4.99**
✗	✗	0.0390	6.90

Table 3: Ablation test of the Self-Reconstruction Criterion.



(a) Naked raw acquisition data. (b) Clothed raw acquisition data.

Figure 6: Comparison to state-of-the-art on raw acquisition data. Percentage of correct correspondences within a certain geodesic error tolerance radius.

4 Additional Results

4.1 Additional Quantitative Evaluation

Fig. 6 shows cumulative error plots giving the percentage of correct correspondences within a certain tolerance radius of geodesic error for the raw acquisitions in the naked regime in (a) and with everyday clothing in (b). On naked raw acquisitions, our method achieves 99% of exact matches when we tolerate errors smaller then 8.1% of the geodesic diameter, where the second best i.e. Neuromorph achieves 85.1%. On clothed raw acquisitions, which is the most challenging test dataset, our method is more accurate both when considering details (i.e. small errors) and global alignments (i.e. large errors).

4.2 Additional Qualitative Comparisons

Fig. 7 shows qualitative comparisons on pre-processed meshes in the first row, pre-processed meshes with topological noise in the second row (the left heel is glued to the right calf and the left arm is glued to the head on the target shape), naked raw 3D scans in the third row and clothed raw 3D scans in the fourth row. The target is color coded and the colors are transferred using the correspondences as estimated by the different methods. Ours is both locally and globally accurate in all four cases and is robust to the presence of hairs, clothes and severe topological noises (see the last row). DeepShells makes global alignment errors (*e.g*. the arms are flipped in the first row, the belly area also in the other rows). Neuromorph suffers from local distortions and makes local errors (*e.g*. see the right hand in all rows). AttentiveFMaps is globally and locally accurate on the pre-processed meshes in the first row but fails in the presence of topological noise which is ubiquitous in raw scans (*e.g*. see the full body in the second and fourth rows and the left arm area in the third row).

4.3 Qualitative Deformation Results

Fig. 8 shows an example of deformed shapes output by our network at every hierarchical level. The top row shows the deformation in the $\mathcal{X} \rightarrow \mathcal{Y}$ direction and the bottom row shows the deformation in the $\mathcal{Y} \rightarrow \mathcal{X}$ direction. The deformed shapes at the finest level, i.e. the vertex level are close to the target deformation shapes ($\mathcal{X}^0 \approx \mathcal{Y}$ and $\mathcal{Y}^0 \approx \mathcal{X}$). In coarser levels, the deformation approximates the pose (global alignment) while in finer levels, where the rigidity constraint is weaker, it approximates the body shape (local alignment). This deformation is the induced alignment in 3D that guides the matching output of our method as shown in the top row of Fig. 7.

References

[1] Jean Basset, Adnane Boukhayma, Stefanie Wuhrer, Franck Multon, and Edmond Boyer. Neural human deformation transfer. In *2021 International Conference on 3D Vision (3DV)*, pages 545–554. IEEE, 2021.

[2] M. Eisenberger, D. Novotny, G. Kerchenbaum, P. Labatut, N. Neverova, D. Cremers, and A. Vedaldi. Neuromorph: Unsupervised shape interpolation and correspondence in one go. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[3] Marvin Eisenberger, Aysim Toker, Laura Leal-Taixé, and Daniel Cremers. Deep shells: Unsupervised shape correspondence with optimal transport. *Advances in Neural information processing systems*, 33:10491–10502, 2020.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[5] Lei Li, Nicolas Donati, and Maks Ovsjanikov. Learning multi-resolution functional maps with spectral attention for robust shape matching. In *Advances in Neural Information Processing Systems*, 2022.

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

[7] Mathieu Marsot, Stefanie Wuhrer, Jean-Sébastien Franco, and Stephane Durocher. A structured latent space for human body motion generation. In *Conference on 3D Vision*, 2022.

[8] Gabriel Peyré and Laurent D Cohen. Geodesic remeshing using front propagation. *International Journal of Computer Vision*, 69:145–156, 2006.

[9] Nitika Verma, Edmond Boyer, and Jakob Verbeek. Feastnet: Feature-steered graph convolutions for 3d shape analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2598–2606, 2018.

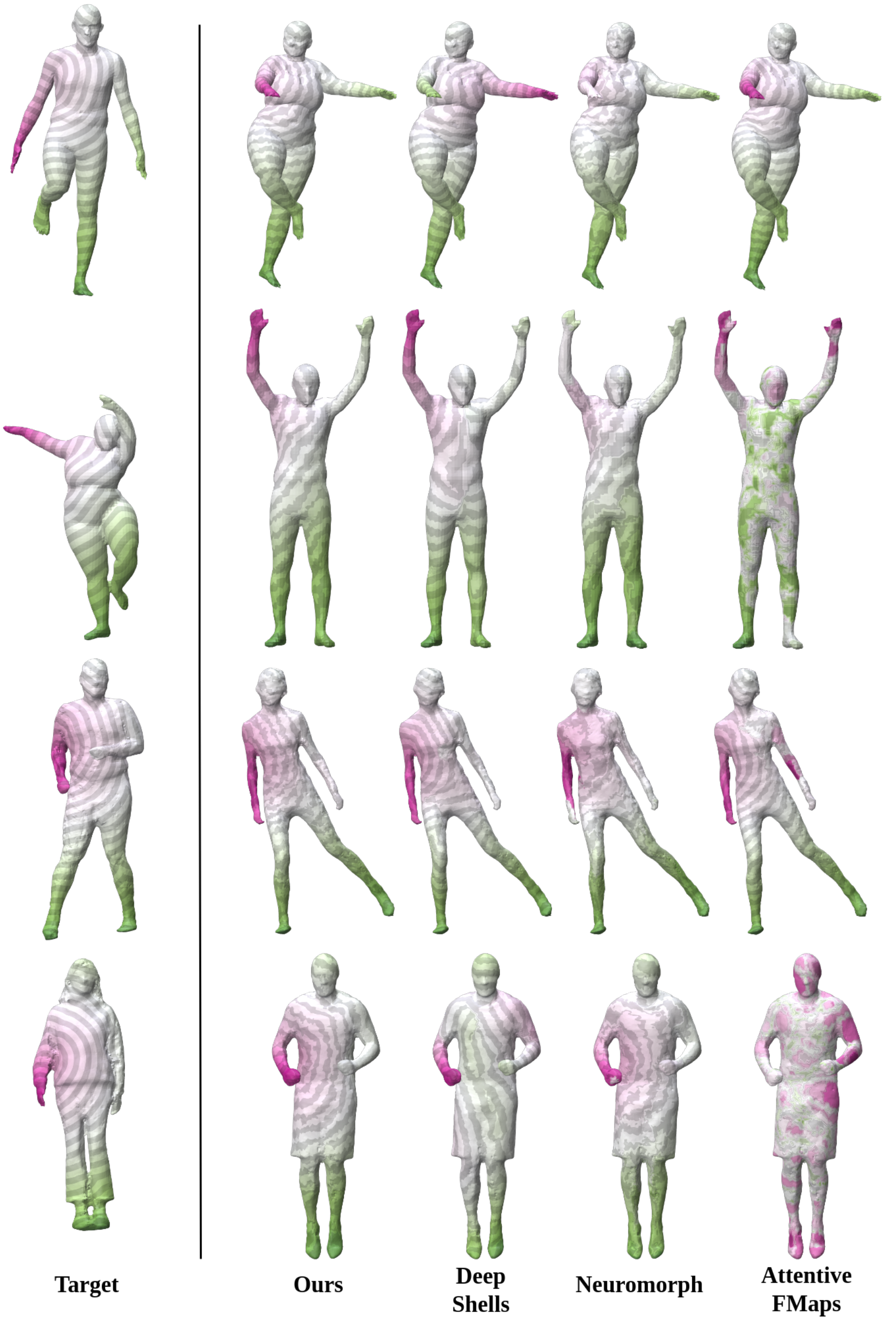


Figure 7: Comparisons with Deep Shells [3], Neuromorph [2] and AttentiveFMaps [5] on pre-processed human meshes (first and second rows) and on raw human 3D scans (third and forth rows). Each point on the target mesh (left) is assigned a color, which is transferred to the source mesh (right) using the correspondences estimated by the different methods.

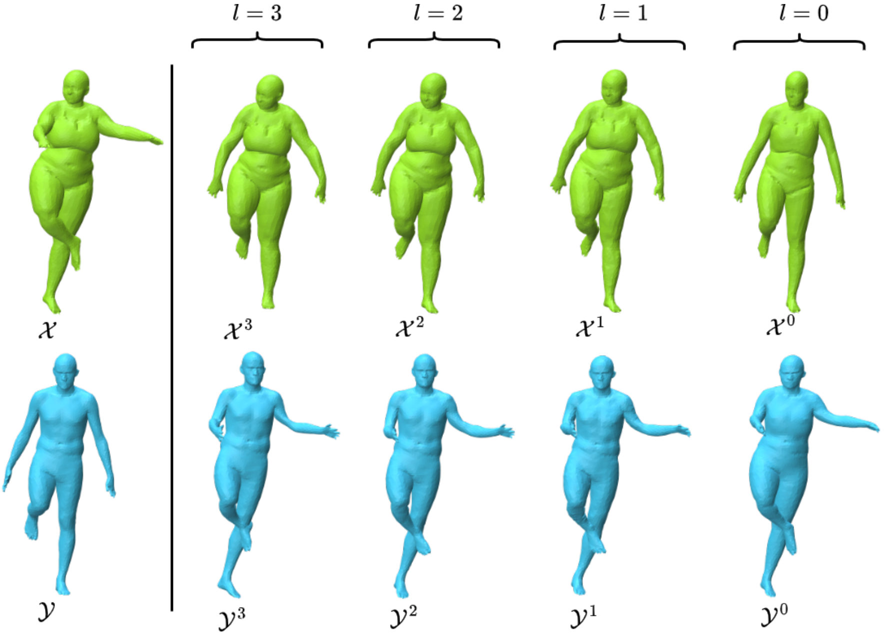


Figure 8: Deformation examples output by our network for every level in the hierarchy. The top row shows the deformation results in the $\mathcal{X} \rightarrow \mathcal{Y}$ direction and the bottom row shows the deformation results in the $\mathcal{Y} \rightarrow \mathcal{X}$ direction.