

PaNDaS: Learnable Shape Interpolation Modeling with Localized Control

Supplementary Material

In this supplementary document, we provide details about the datasets we used, the implementation and hardware specifications of our experiments as well as additional results and comparisons with state-of-the-art methods. Finally, we give more examples and details about potential applications.

1. Notations

Symbol	Description	Dim.
\mathcal{S}	Source shape	/
\mathcal{T}	Target deformation of \mathcal{S}	/
$n_{\mathcal{X}}$	Number of vertices of \mathcal{X}	/
$(x_i)_{0 \leq i \leq n_{\mathcal{S}}}$	Vertices of \mathcal{S}	\mathbb{R}^3
$(y_i)_{0 \leq i \leq n_{\mathcal{T}}}$	Vertices of \mathcal{T}	\mathbb{R}^3
$m_{\mathcal{X}}$	Number of faces of \mathcal{X}	/
$(t_j)_j$	Triangles of \mathcal{X}	/
\mathfrak{M}	Region binary mask	$\{0, 1\}^{n_{\mathcal{S}}}$
γ	Shape trajectory (motion)	/
f	Global feature extractor	/
g	Local deformation encoder	/
h	Deformation generator	/
z	Global latent representation	\mathbb{R}^r
$(u_j)_j$	Per-triangle local features	\mathbb{R}^l
$(\omega_j)_j$	Concatenated features	\mathbb{R}^l
$(J_j)_j$	Per-triangle Jacobians	$\mathbb{R}^{3 \times 3}$
\mathbf{V}	Deformation field	/
$(v_i)_{0 \leq i \leq n_{\mathcal{S}}}$	Per-vertex deformation	\mathbb{R}^3

Table 1. Notation summary used throughout the paper.

We employ the notation $(\cdot)_j$ for $(u_j)_{1 \leq j \leq m_{\mathcal{X}}}$.

2. Experimental details

2.1. Datasets

The experiments presented were performed on three different datasets. In Tab. 2, we report the data details in terms of quantity (number of samples in the dataset) and resolution (number of points for a given sample). We also report the train/test split. Our model operates on a relatively low data regime.

Since our framework takes positions and normals as input, all data are rigidly aligned before training and inference.

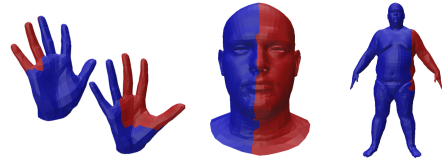
	MANO	DFAUST	COMA
Type	Hand	Body	Face
# vertices	778	6890	3931
# faces	1538	13776	7800
Training samples	828	247	491
Test samples	71	61	92

Table 2. Train / test splits for each dataset.

For DFAUST and COMA, the data is already aligned. For MANO, we perform a Procrustes alignment of the training data (this is done for all compared methods). All other methods have been retrained for fair comparisons.

2.2. Masks used for experiments

In general, the masks used in experiments are computed with a manual 3D box as shown in the introduction figure of the main paper which shows an example of one mask used for partial interpolation. We present in the following figure the masks chosen for the partial interpolation experiment shown in Section 5.3 of the main paper, with red/blue corresponding respectively to values of 1/0.



Section 5.4 in the main paper shows a mix between two facial expressions, body poses (one mask for each side), a partial body pose interpolation (mask on a single side), and a mix of 5 poses (one for each finger) on the hand.

2.3. Adaptation of the competitors for partial interpolation experiments

Here we give more details about how we adapt the most relevant competitors for the specific task of partial shape interpolation:

1. ARAP [6] can be easily adapted by selecting and dragging the vertices subject to the deformations (corresponding to a mask equal to 1). The other vertices are then only constrained by minimizing the As-Rigid-As-Possible energy.
2. Neural Jacobian Fields [1] are modified in order to mask the predicted jacobians (similar to our method but with a different design for the whole framework)
3. VCMC [7] rely on a fixed mesh topology and can only be used for comparison in a registered setting. In this set-

ting, partial interpolations are possible with latent interpolation of the corresponding latent point of the coarse graph (basically corresponding to limbs when used on human body meshes).

2.4. Unregistered metrics formulation

For registered sequences (with vertex-wise correspondence between the predicted sequence and the target sequence), we evaluated our method with the Mean Squared Error (MSE) and the Cosine distance.

For an unregistered sequence of length T , without vertex-wise correspondence between the predicted sequence $\hat{\mathcal{T}}_t$ and the ground-truth sequence \mathcal{T}_t , which is typically the case when we animate raw scans, we rely on averaged unregistered metrics.

For two sets of vertices \mathcal{T}_t and $\hat{\mathcal{T}}_t$, the symmetric Chamfer distance (CD) is defined as follows:

$$d_{CD}(\mathcal{T}_t, \hat{\mathcal{T}}_t) = \frac{1}{n_S} \sum_j \min_i \|y_i - \hat{y}_j\|_2^2 + \frac{1}{n_T} \sum_i \min_j \|y_i - \hat{y}_j\|_2^2, \quad (1)$$

and the Hausdorff distance is defined as

$$d_{HD}(\mathcal{T}_t, \hat{\mathcal{T}}_t) = \max\left\{\max_i \min_j \|y_i - \hat{y}_j\|_2^2, \max_j \min_i \|y_i - \hat{y}_j\|_2^2\right\} \quad (2)$$

Averaging the distances across the sequence yields \mathcal{L}^{CD}

$$\mathcal{L}^{CD} = \frac{1}{T} \sum_t d_{CD}(\mathcal{T}_t, \hat{\mathcal{T}}_t). \quad (3)$$

and \mathcal{L}^{HD}

$$\mathcal{L}^{HD} = \frac{1}{T} \sum_t d_{HD}(\mathcal{T}_t, \hat{\mathcal{T}}_t). \quad (4)$$

2.5. Implementation and hardware details

All PaNDaS models presented in the main paper are trained for 1000 epochs using Adam [4] optimizer with a learning rate of 10^{-4} . In the experiments, for the loss function, the weighting factor λ^n is set as $\lambda^n = 10^{-5}$.

All models training and inferences were performed on a computer running Linux with an Intel Xeon Gold 5218R CPU with 64GB RAM and a NVIDIA Quadro RTX 6000 graphics card with 24Go GPU RAM.

3. Additional experiments

3.1. Predicting static deformations

In this experiment, for each identity in the dataset, we have an identified neutral pose mesh \mathcal{S} . Then, given a target

mesh \mathcal{T} , the model predicts a deformation field v on \mathcal{S} . We observe and measure the quality of the predicted deformed mesh and its distance to the target mesh in Tab. 3. We compared our model with several deep learning methods: Neural Jacobian Fields (NJF) [1] proposes an auto-encoding approach to predict a point-wise deformation. Variant Coefficient MeshCNN (VCMC) [7] is a graph convolution method, LIMP [2] and ARAPReg [3] learn a regularized linear latent shape space of deformations to disentangle identities and poses. The localized deformation representation of PaNDaS surpasses global latent interpolation for large non-linear deformation generations (on MANO and DFAUST) and is competitive for face expression deformations. ARAPReg slightly outperforms PaNDaS on the COMA dataset, an easier setting, as most deformations are quasi-linear.

3.2. Simpler is better

During our investigations when building the framework, we tried several more complex strategies for the feature field construction but found these less efficient, showing that simpler and more direct approaches can be better in this case.

3.2.1. About the masking strategy

In our final framework, the model is not learning from the input mask. We tried variations of our framework with predefined masks used during training. These variations of PaNDaS uses the mask as input and try to learn how to adapt the feature field and the corresponding deformation field from it. In the training loop, we artificially augment the training set with partial shapes given by the input mask. Using either a default segmentation map to create the masks (Dflt.) or with an automatic segmentation algorithm using the HKS descriptor [5], we report the performance for partial deformations of half of the body in Tab. 4. A visualization of the segmentation maps is given in Fig. 1.

In fact, partial interpolation of half of the body is not in the segmentation collection which explains why the other strategies are not performing well: they overfit the masks seen during training. Hence, we conclude that masking is not part of the training procedure for the following reasons: 1) It adds complexity to the training procedure, with a different setup for each shape category: for hands and body, there are clear parts on which restricted interpolations make sense *a priori* (e.g., fingers and limbs), but this is not the case for facial data. 2) Not using masking during training avoids any biases toward predefined training masks, making the method more user-friendly and flexible.

3.2.2. Feature combination strategy

Next, when combining local and global features, we used a simple concatenation of local and global features. We tried more complex strategies, such as cross attention to build ω :

	MANO			DFAUST			COMA		
	MSE ↓	HD ↓	CD (10 ⁻⁵) ↓	MSE ↓	HD ↓	CD (10 ⁻⁵) ↓	MSE ↓	HD ↓	CD (10 ⁻⁵) ↓
LIMP [2]	-	0.049	71.9	-	0.141	0.0047	-	0.0095	1.021
VCMC [7]	0.2240	0.014	5.86	2.92	0.076	0.0017	0.191	0.01400	1.506
ARAPReg [3]	<u>0.1986</u>	0.009	<u>3.29</u>	<u>2.61</u>	<u>0.069</u>	<u>0.0011</u>	0.128	0.0070	0.474
NJF [1]	0.3019	0.019	9.65	5.01	0.126	0.0029	0.238	0.0105	1.435
Ours	0.1422	<u>0.010</u>	2.94	2.23	0.058	0.0009	<u>0.160</u>	<u>0.0086</u>	<u>0.839</u>

Table 3. **Reconstruction errors of poses from a given identity.** We evaluate the quality of the deformations using distances between the predicted deformed mesh and the target mesh.

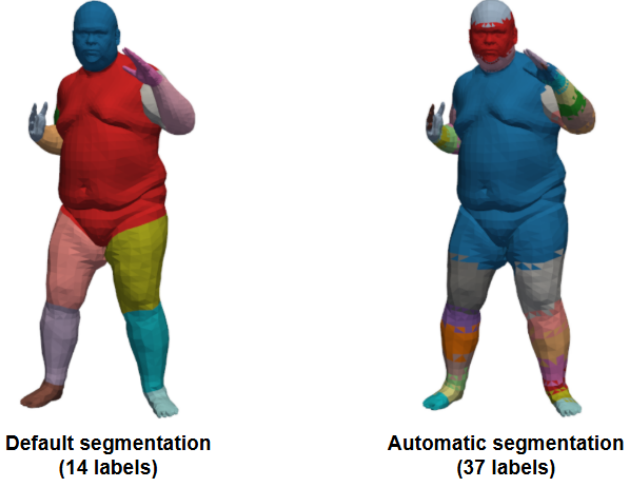


Figure 1. Visualization of the precomputed masks.

	MSE ↓			Cosd ↓		
	Dflt.	Aut.	ours	Dflt.	Aut.	ours
shake	5.0	5.0	2.3	0.12	0.15	0.12
chicken	6.8	5.7	3.9	0.18	0.18	0.13
knees	5.7	4.0	1.4	0.03	0.03	0.01
jumping	8.8	10.3	6.1	0.13	0.16	0.10
Mean	6.6	6.2	3.4	0.12	0.13	0.09

Table 4. **Partial interpolation performance depending on the training strategy.**

$$\omega_j = (\omega_{j,i})_{1 \leq i \leq l+r} = (u_j, CA(u_j^S, u_j^T)) \in \mathbb{R}^{l+r} \quad (5)$$

where CA is the cross-attention between the source mesh features and the target mesh features. Every source feature gets a query and every target feature provides a key and value. This approach proved to be more efficient for the reconstruction, with better final training losses. However, the network generalization is weaker, making the approach suboptimal for latent manipulation and thus the generation of full motions. We display some qualitative exam-

ples in Fig. 2 and report triangle area change along several interpolations from the MANO test set in Tab. 5.

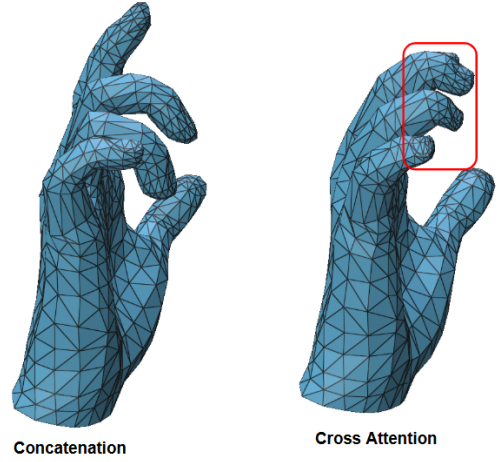


Figure 2. **Mid-interpolation example with different feature combination strategies.** On the left, the model simply concatenates local and global features, on the right, an additional cross attention module projects local features from the target mesh to the source mesh.

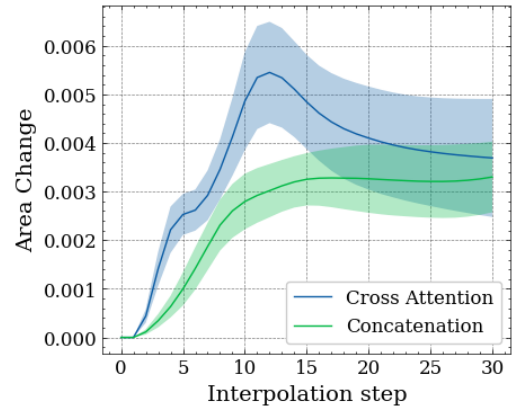


Figure 3. **Average area change for 30 steps latent interpolation.** The lower the curve, the more isometric the deformations are.

Feat. combination	Mean area loss ↓
Concatenation	2.5 ± 0.5
Cross Attention	3.6 ± 0.8

Table 5. Mean area change

3.3. Ablation Study

The following ablation study on the MANO dataset highlights the importance of each of PaNDaS’ components, namely the encoder, decoder, and loss functions. We evaluate the final MSE on the test set after changing the different components. We report all results in Tab. 6. Notably, we observe that DiffusionNet is crucial for the encoding part. Moreover, we observe a large drop in performance without our global aggregation strategy. Similarly, a simple MLP fails to decode properly the shapes. Finally, we tested other loss functions, the simple MSE and using ARAP energy, similarly as in [3]. Notably, adding an ARAP regularization term did not improve either the quality of the deformation or the interpolation quality.

		MSE ↓
<i>Encoder</i>	MLP	0.2456
	DiffusionNet	0.2216
<i>Decoder</i>	MLP	0.2074
	DiffusionNet	0.2074
<i>Loss</i>	\mathcal{L}^{rec}	0.1610
	$\mathcal{L}^{rec} + \text{ARAP}$	0.1790
Ours		0.1422

Table 6. Ablation studies on the MANO dataset. The overall approach of the framework gives satisfying results with simpler components. However, our best results were obtained by combining DiffusionNet for the feature extractor, aggregation by projecting on the eigenvectors, deformations generated with Jacobian fields and loss defined as $\mathcal{L} = \mathcal{L}^{MSE} + \lambda_n \mathcal{L}^n$.

Another important aspect is the novel feature aggregation strategy for the global encoding of the target mesh. The most straightforward approach is a mass-weighted mean of the features over the mesh. However, in Table 7, we report that, for the same size of latent vector, using projections on other eigenvectors of low frequency yield better results. The best configuration is obtained for $s = 4$.

An important note on the experiment: the latent size is fixed (64) and depending on s , we divide the latent vector in equal parts for each eigenvector.

4. Additional qualitative examples

We provide more qualitative results and comparisons of latent interpolations in this section.

		MSE ↓
<i>Aggregation</i>	Mean (s=1)	0.2086
	s=2	0.1860
	s=3	0.1481
	s=4	0.1422
	s=5	0.1902
	s=6	0.2356

Table 7. Ablation studies on the global encoding aggregation.

Full deformations

We first display examples of full deformations generated from latent interpolations in Fig. 4, and a comparison against ARAPReg.

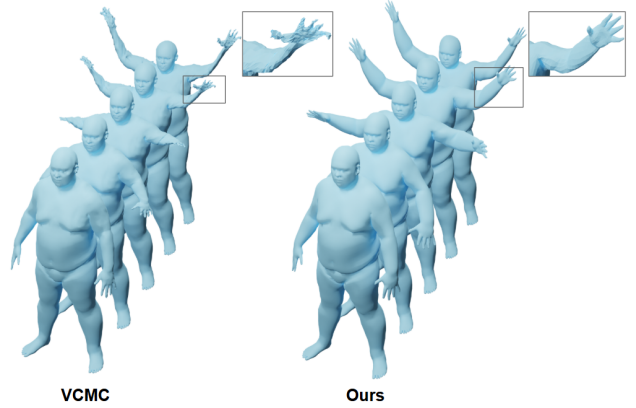


Figure 4. Qualitative comparison of latent interpolation. Our method shows more realistic interpolation trajectories, with limited distortion of the hand even with large non-linear deformations.

Partial deformations

Then we give another comparison of partial shape interpolation in Fig. 5.

Thanks to the different component of the framework, restricting deformations to specific is robust against remeshing as shown in Fig. 6.

Applications: Shape statistics

As mentioned in the main paper, PaNDaS can be used to build statistical models of non-rigid deformations, such as the mean of shape collections. We demonstrate the robustness of our method by showing in Fig. 8 the generalizability of our trained model. The estimated mean from similar poses using PaNDaS is consistent across different identities.

Additionally, we show results and comparison of mean estimation in Fig. 9.

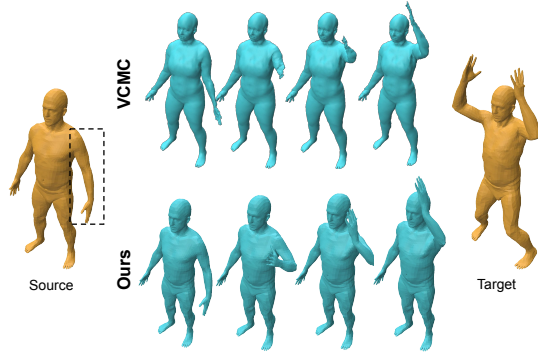


Figure 5. **Qualitative comparison of partial latent interpolation.** We computed a partial latent interpolation on an unseen identity between a *source* pose and a *target* pose displayed in yellow. The interpolated part is the left arm (dashed case). Previous comparable method VCMC fails to reconstruct the identity and exhibit important distortions of the selected part during the interpolation.

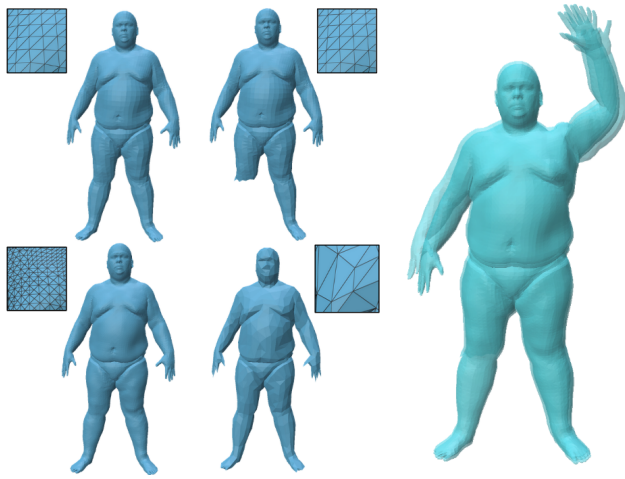


Figure 6. **Robustness of partial deformations.** From different meshing of the source mesh (on the left), we show that the resulting partial deformations restricted to the left arm stays similar by superposing the results (on the right).

Applications: Partial Expression transfer

We give examples of partial expression transfer in Fig. 10.

5. Animations

We attached to the supplementary material, several videos showing animations (on the left side of the videos) computed with partial deformations extracted from combinations of different input poses (on the right side of the videos). PaNDaS can generate complex, controllable motions from a limited number of input poses.

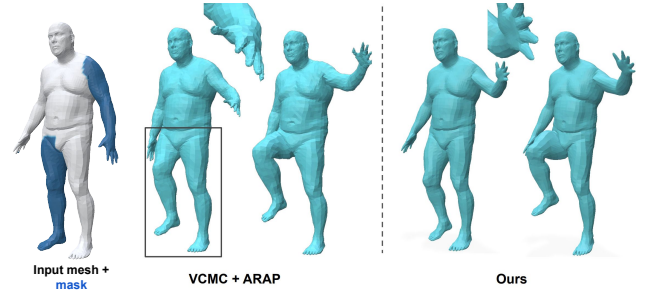


Figure 7. **Pose mixing using partial interpolations of a neutral pose with two other poses.** The first pose raises both arms, the second pose raises the right knee. We mix them and highlight how our method (on the right) computes smoother deformations and avoids limb shrinking (right leg on the left).

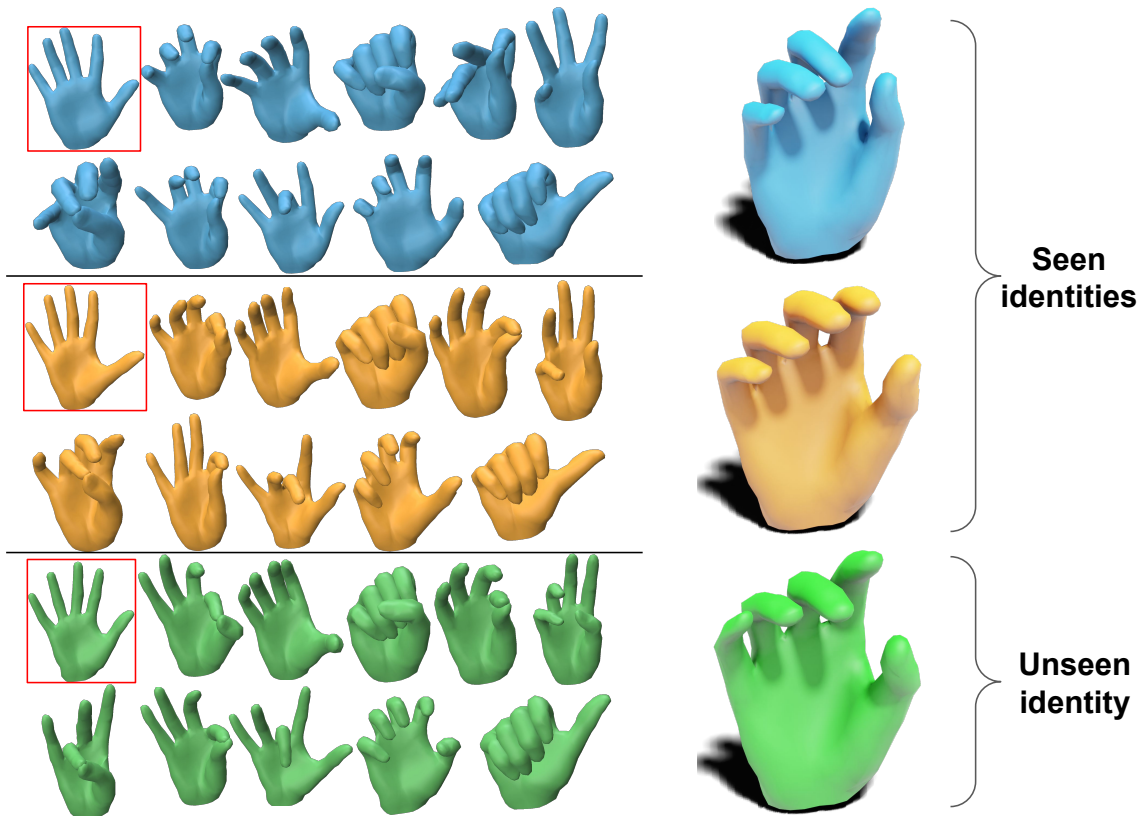


Figure 8. **Robustness of the mean shape estimation.** We perform mean shape estimation from a collection of 11 poses of hands. We predict the feature field associated with each pose and compute the Euclidean mean. The mean feature field is then used to compute the **mean pose**. In this figure, each color corresponds to a different identity (framed in red on the left). In particular, we highlight the robustness of our method across several hand identities, seen and unseen during training.

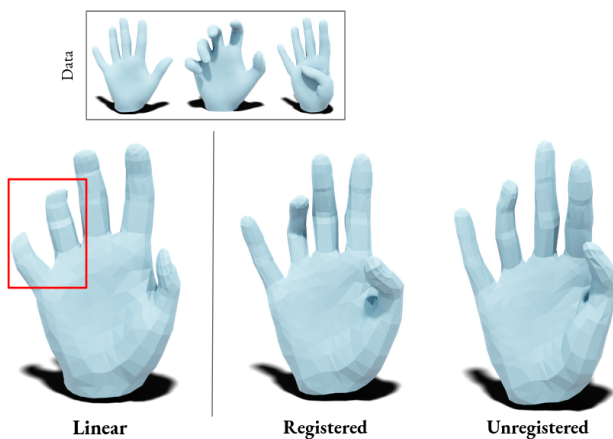


Figure 9. **Estimated mean from the same collection of hands.** From 10 different hand poses in MANO, we compute a linear mean (on the left) and the "latent mean" of the registered meshes along with the latent mean of the raw scans (unregistered).

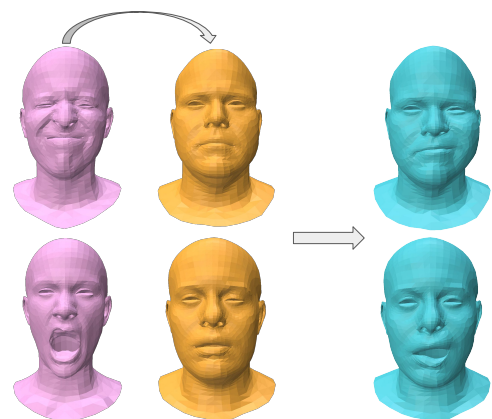


Figure 10. **Example of partial expression transfer.** In this example, the expression of the face in red is transferred to the yellow face restricted on the left side.

Acknowledgments

This work was partially supported by the project 4DSHAPE ANR-24-CE23-5907 of the French National Research Agency (ANR). This work was also supported by the ERC Consolidator Grant 101087347 (VEGA).

References

- [1] Noam Aigerman, Kunal Gupta, Vladimir G. Kim, Siddhartha Chaudhuri, Jun Saito, and Thibault Groueix. Neural jacobian fields: learning intrinsic mappings of arbitrary meshes. *ACM Trans. Graph.*, 41(4), 2022. [1](#), [2](#), [3](#)
- [2] Luca Cosmo, Antonio Norelli, Oshri Halimi, Ron Kimmel, and Emanuele Rodolà. Limp: Learning latent shape representations with metric preservation priors. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III*, page 19–35, Berlin, Heidelberg, 2020. Springer-Verlag. [2](#), [3](#)
- [3] Q. Huang, X. Huang, B. Sun, Z. Zhang, J. Jiang, and C. Bajaj. Arapreg: An as-rigid-as possible regularization loss for learning deformable shape generators. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5795–5805, Los Alamitos, CA, USA, 2021. IEEE Computer Society. [2](#), [3](#), [4](#)
- [4] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015. [2](#)
- [5] Yanir Kleiman and Maks Ovsjanikov. Robust structure-based shape correspondence. In *Computer Graphics Forum*, pages 7–20. Wiley Online Library, 2019. [2](#)
- [6] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*, page 109–116, Goslar, DEU, 2007. Eurographics Association. [1](#)
- [7] Yi Zhou, Chenglei Wu, Zimo Li, Chen Cao, Yuting Ye, Jason Saragih, Hao Li, and Yaser Sheikh. Fully convolutional mesh autoencoder using efficient spatially varying kernels. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. [1](#), [2](#), [3](#)