ImHead: A Large-scale Implicit Morphable Model for Localized Head Modeling Suplementary Material

Rolandos Alexandros Potamias, Stathis Galanakis, Jiankang Deng
Athanasios Papaioannou, Stefanos Zafeiriou
Imperial College London
https://rolpotamias.github.io/imHead/

1. Ablation Study

To justify the technical choices we made and evaluate the contribution of each component we perform an ablation study.

Impact of global latent space. In particular, we divide the ablation into three major categories to capture all aspects of the proposed model. We first evaluate the contribution of the global latent code by modifying imHead latent space to a set of local latents, reported as *w. Local Lat.*. We follow NPHM and use 32 latent dimensions for each of the K=32 regions resulting in a total 1248 latent space, 4.87$\times$ increase compared to 256 that we use in imHead. In addition we report the performance of another variation that extends the local latents to include an additional global identity latent, following the architectural design of NPHM, reported as *w. Local and Global Lat.*. The total latent space of this model is 1344 (same as NPHM) which reflects to 5.25$\times$ increase in the latent size. Finally, to demonstrate the impact of a single global latent space, we report the results of a model trained with a local latent space where each region receives a local latent of size 8, resulting in a latent space size of 312. As can be easily observed in Tab. 1, utilizing a split latent space diminishes the reconstruction performance of the network. This significantly deteriorates when we use a latent space with the size of 312, where the model struggles to achieve reasonable performance. The reason behind this, as suggested in [4, 11, 12], is that global patterns of the shape are copied in each local latent which inevitably increase the size of the model. To enable a fully local latent space, whilst also achieving sufficient reconstruction performance, it is necessary to increase each latent sufficiently enough to encode both global and local information. An intermediate solution is to build a local-global latent space, similar to NPHM model. Although this approach achieves similar performance with imHead, it suffers from two main factors: a) a $5\times$ larger latent space which limits the shape compression and b) a highly constrained latent space that prohibits localized face editing as the latent codes are now extended with global information. imHead can successfully bridge both worlds by leveraging a compact latent space along with an intermediate localized representation that can facilitated disentangled manipulation.

Impact of FusionNet. To demonstrate the impact of the proposed structural blending network, we train a model that directly regress the local SDF from each local-part network without using an intermediate feature representation as in imHead. Despite being slightly lighter model, the performance of the the model drops significantly, as each of the local networks need to directly predict the global SDF. It is also important to note that the normal consistency of the reconstructions deteriorates due to non-smooth blending. In contrast, when using the proposed FusionNet, the local features are aggregated and the SDF values are regressed using an intermediate feature representation. This allows the model to learn more complex representations while achieving smooth reconstructions.

Impact of Local Canonical Space. We additionally report the effect of using a per-region canonical space (*w/o Local Canonical Space*). In particular, each local-part network uses a canonical space that is defined around its corresponding keypoint $\boldsymbol{k}_j$ as:

$$\mathbf{f}_x^j = \boldsymbol{g}_j(\boldsymbol{x} - \boldsymbol{k}_j, \boldsymbol{z}_{id}^j) \tag{1}$$

where $\mathbf{f}_x^j$ denotes the j-th feature embedding corresponding to point $\boldsymbol{x}$ and $\boldsymbol{k}_j$ represent the generated landmark keypoint corresponding to region j. This canonical space can effectively reduce the workload of each local part network and facilitate the training process. As can be seen in Tab. 1, apart from the training stability, the canonical space has a positive impact on the reconstruction performance of imHead, as we observe a significant performance improvement when using a canonical space for each local-part network (*imHead-Full*).

Method	NPHM			MimicMe		
	CD ↓	NC ↑	F@5mm ↑	CD ↓	NC ↑	F@5mm ↑
w. Local Lat. ($d = 312$)	0.876	0.915	0.689	0.874	0.914	0.721
w. Local Lat. ($d = 1248$)	0.775	0.948	0.743	0.767	0.939	0.788
w. Local and Global Lat. ($d = 1344$)	0.494	0.964	0.841	0.569	0.958	0.857
w/o FusionNet	0.595	0.954	0.808	0.674	0.947	0.812
w/o Local Canonical Space	0.723	0.934	0.723	0.884	0.946	0.732
imHead-Full	**0.459**	**0.988**	**0.898**	**0.533**	**0.986**	**0.873**

Table 1. **Ablation Study** of different key components of imHead.

2. Robustness to Noise

Given that the proposed model was trained on raw scans with a considerable amount of noise, it can achieve robust reconstructions even under noisy point cloud inputs. In particular, to evaluate the reconstruction performance of imHead under noise scenarios, we add Gaussian noise of different standard deviations to the input point clouds and measure the performance drop. As can be seen in Fig. 2, imHead can achieve reasonable reconstruction that retain the identity characteristics even with noise levels that correspond to 1.5 standard deviations.

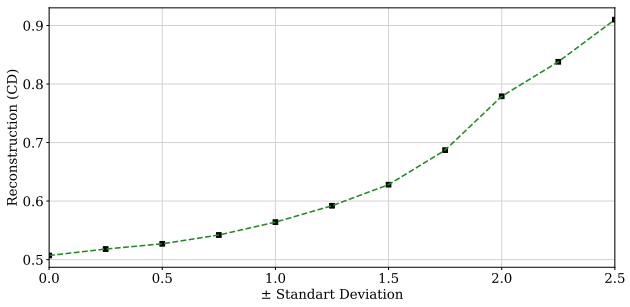

Figure 1. **Reconstruction Error under Noisy Inputs**. We measured the reconstruction error under different noise levels of the input point cloud.

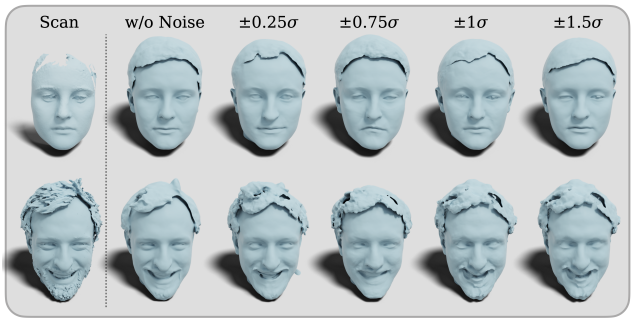

Figure 2. **Qualitative Evaluation of fitting under Noisy Inputs**. We insert Gaussian noise to the input point clouds and measure the reconstruction performance.

3. Limitations and Societal Impact

As stated in the main paper, although imHead makes a step towards full head modeling, it still suffers from some limitations. In particular, implicit models, in contrast to explicit 3DMMs suffer from slow inference times. To obtain a high resolution head it is required to sample and predict the SDF for a sufficient number of points which could significantly reduce the runtime of the method. It must be also noted that SDFs require an additional post-processing marching cubes step which can further reduce the inference speed of the method. In contrast, 3DMMs can leverage fast rendering techniques and may provide a more efficient method in tasks where runtime performance is key priority. Implicit surfaces are also known to struggle capturing fine-grained details and fail to accurately model thin surfaces such as the hair strands. In addition, although as we experimentally show, imHead preserves the face correspondences there is not an 1-1 mapping similar to the case of explicit models. Furthermore, as noted in the main paper, localized editing is constrained by the fixed number of anchors that define each region. The editing process can also be influenced by the contributions of nearby local-part networks, which are designed to ensure smooth and plausible surfaces, but will affect the accuracy of edits especially at the boundaries. Finally, despite curating a large-scale dataset, there are still race biases within the dataset. This also includes the hair regions which are directly adapted from the NPHM dataset, which has also limited diversity and cannot adequately represent all hair types. As an extend, imHead also shares the same demographic biases that should be taken into consideration when using imHead for downstreaming tasks. Despite the biases, as can be seen in 3, imHead can generalize well in out-of-distribution and non-Caucasian ethnicities.

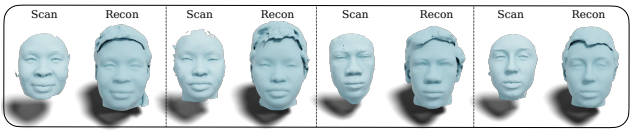

Figure 3. **Reconstruction performance on non-Caucasian ethnicities.** Despite the demographic biases, imHead can accurately reconstruct out-of-distribution samples.

4. Dataset Curation

To enable large-scale head modeling we utilized MimicMe datasets [8], which consists of 5,000 distinct subjects under different expressions. MimicMe dataset was collected using a 3dMD face capture system. The raw scans have a resolution of approximately 60,000 vertices. We filter the dataset to avoid noisy scans, resulting in a total of 4,000 distinct subjects being retained, with available metadata including gender (57% male, 43% female), age ($1 - 81$ years

old) and ethnicity (73% White, 13% Asian, 7% Mixed and 3% Black, 4% Other). Notably, the collected head scans demonstrate significant diversity across age, ethnicity, and height, marking progress toward a universal full head model. In comparison to previous implicit head models [5, 15], the curated dataset encompasses over 600 children under the age of 12, as well as more than 100 individuals aged over 60.

To bring the raw scans into dense correspondence, we utilized a multi-step pipeline. Initially, the scans were rendered from multiple views and 2D joint locations were detected using RetinaFace [3]. Subsequently, the 2D landmark locations were lifted to 3D by utilizing a linear triangulation and projected to the 3D surface. Using the 3D detected keypoints, we fit FLAME parametric model by optimizing the pose and expression parameters to align the template head to the exact pose, expression and shape of each raw scan. Specifically, we optimize the pose θ, expression ψ and shape β parameters using following loss function:

$$\mathcal{L} = \mathcal{L}_J + \mathcal{L}_{cd} + ||\beta||_2 + ||\psi||_2 + ||\theta||_2 \quad (2)$$

where $\mathcal{L}_J = ||J - \hat{J}||_2$ is a keypoint loss that enforces FLAME landmakrs $\hat{J}$ to match the detected keypoints $\hat{J}$ and $\mathcal{L}_{cd}$ is the chamfer distance loss that minimizes the scan to FLAME distance. The optimization process was performed using Adam optimizer with learning rate of $1e-3$. We complete the full head of the aligned scans by fitting NPHM model [5]. However, a lot of the identity details of the subject might have been diminished during the fitting process. To retrieve the identity details we perform a Non-rigid Iterative Closest Point algorithm (NICP) [1] between the fitted meshes and the 3D raw scans. The proposed fitting and registration process enables the capture of rich facial details while ensuring plausible head surfaces with minimal reconstruction error. As shown in Fig. 4, the non-rigid ICP step helps mitigate racial biases that may arise during the fitting process.

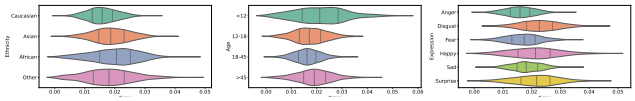

Figure 4. **Registration and Data Curation Errors.** We report reconstruction errors during the data curation process for different ethnicities and expressions.

5. Implementation Details

In this section we provide the implementation details of the different components of our network.

5.1. Identity Network

The identity network of the proposed imHead model is composed of three main modules: the *Decomposition network*, the *Local-Part Networks* and the *Fusion network*. Bellow we describe the implementation details for each one of them:

Decomposition network. The Decomposition network is responsible for the mapping of the global identity latent codes $\boldsymbol{z}_{id}$ to a set of localized embeddings $\{\boldsymbol{z}_{id}^j\}$ that span the 3D head. We define $\boldsymbol{z}_{id}$ using a simple Embedding layer that maps dataset instances to a 256-dimension latent code. Using a fully connected layer, we project the global latent $\boldsymbol{z}_{id}$ to K embeddings of 32 dimensions each. We follow NPHM [5] and select $K = 39$ keypoints that span the 3D head, resulting in a localized embedding with a total of 1248-dimensions. Using this simple yet efficient mapping we can achieve both compact global latent space, which can effectively improve the reconstruction capabilities of the network [12, 14] along with a localized intermediate representation that enables localized editing.

Landmark Regression. Following the latent embedding split, we use an MLP to regress the keypoints of the head, that will serve as the local coordinates for each region. In particular, we use a three-layer MLP that receives the set of local embeddings $\{\boldsymbol{z}_{id}^j \in \mathbb{R}^{32}\}$ as input and predicts the K=39 facial keypoints $\{\boldsymbol{k}^j \in \mathbb{R}^3\}$. We opted to use the intermediate local embedding representation to regress the facial landmarks as it can provide more robust estimations even after shape manipulations.

Local-Part Networks. Using a point sampled from the 3D space $\boldsymbol{x} \in \mathbb{R}^3$, we use an enseble of local-part networks to extract a point-specific feature $\mathbf{f}_j$ per region. To acquire the local part-specific feature $\mathbf{f}_j$, we feed point $\boldsymbol{x}$ along with the localized embeddings $\{\boldsymbol{z}_{id}^j\}$ to their corresponding local-part module. To better capture the high frequency details of the shapes [10], we use a set of positional embeddings as defined in [7]:

$$\begin{aligned}\gamma(\boldsymbol{x}) = \big(&\boldsymbol{x},\\ &\sin(2^0\pi\boldsymbol{x}), \cos(2^0\pi\boldsymbol{x}),\\ &\sin(2^1\pi\boldsymbol{x}), \cos(2^1\pi\boldsymbol{x}), \ldots,\\ &\sin(2^{L-1}\pi\boldsymbol{x}), \cos(2^{L-1}\pi\boldsymbol{x})\big)\end{aligned}$$

that map the points $\boldsymbol{x}$ to a high dimensionality. We use $L = 7$ frequency bands. Before feeding each point to the corresponding local-part network, we first normalize it according to the keypoint $\boldsymbol{k}_{id}^j$ associated with each part-network. This step is essential to normalize the coordinate system of each part network and not only achieve efficient and stable training but increase the expressivity of the network. We implement each local-part network using a small DeepSDF module with 4 layers and a hidden dimension [9] of 200. Following the implementations of [9] we use `softplus` activation function.

Fusion Network. The final step of our identity network is to fuse the extracted feature codes $\mathbf{f}_j$ from each part-

network j back to a single global feature that will be used to regress the final SDF of point $\boldsymbol{x}$. Although an obvious choice would be to directly regress the fused SDF from the local-part networks, as we experimentally show in the ablation study, this choice significantly reduces the reconstruction quality and limits the editing properties of the network. We obtain the fused global feature vector using:

$$\hat{\mathbf{f}}_x = \sum_j^K w(\boldsymbol{x}, \boldsymbol{k}_j)\mathbf{f}_x^j \tag{3}$$

where $w(\boldsymbol{x}, \boldsymbol{k}_j)$ scales the contribution of each feature embedding based on position of the point $\boldsymbol{x}$:

$$w(\boldsymbol{x}, \boldsymbol{k}_j) = \frac{e^{\frac{-||\boldsymbol{x}-\boldsymbol{k}_j||_2}{\sigma}}}{\sum_j^K e^{\frac{-||\boldsymbol{x}-\boldsymbol{k}_j||_2}{\sigma}}} \tag{4}$$

The final feature vector along with the correspond point $\boldsymbol{x}$ is then fed to the FusionNet to predict the final signed distance field y:

$$y = \mathcal{F}_\theta(\boldsymbol{x}, \hat{\mathbf{f}}_x) \in \mathbb{R} \tag{5}$$

We implement the fusion network as a small DeepSDF module [9] with 4 layers and 200 latent dimensions. Similar to the local-part networks, we use `softplus` activation function.

5.2. Expression Warping Module

Our expression module is responsible for backward-warping the sampled points from the observation space $\boldsymbol{x}_{obs} \in \mathbb{R}^3$ to the canonical space of the identity network. To enable fast integration to existing pipelines we define $\boldsymbol{z}_{exp}$ using the expression parameters of FLAME model [6] acquired during the fitting process of the dataset. The FLAME expressions are then fed to a higher dimensional latent space and used to condition the expression warping module. Given that imHead is conditioned on FLAME expression parameters, it can be easily adapted to existing pipelines and generalize to unseen expressions as shown in Fig. 5 Similar to the previous networks, we implement the expression module using a DeepSDF network with 8-layers with 128-hidden dimensions.

Figure 5. **Generalization to unseen expressions**. Given that imHead is relies on FLAME [6] expression space, it can easily generate out-of-distribution expressions.

6. Backward vs. Forward Warping

Backward warping has been widely used across implicit field [2, 13, 15] achieving robust results and offering several advancements over traditional forward deformation warping. Specifically, backward warping does not require any costly registration process to bring the scans in dense correspondence. In contrast, forward deformation methods such as NPM [5] and NPHM [5] require a registration step to non-rigidly aling the scans to calculate the target deformation fields. Additionally, forward deformation methods heavily rely on iterative root finding schemes, which apart from time consuming optimization processes introduced, can also affect the robustness of the parametric model. In particular, as shown in Fig. 6, forward deformation methods, can fail in cases of noisy scans where the inverse correspondences are not established correctly

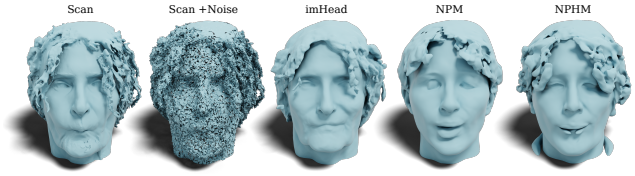

Figure 6. **Failure cases of forward deformation methods**. Given that forward warping methods rely on iterative root-finding schemes, inaccurate correspondences can significantly impact reconstruction performance.

References

[1] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 3

[2] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 20364–20373, 2022. 4

[3] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 3

[4] Simone Foti, Bongjin Koo, Danail Stoyanov, and Matthew J Clarkson. 3d generative model latent disentanglement via local eigenprojection. In *Computer Graphics Forum*. Wiley Online Library, 2023. 1

[5] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 4

[6] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 4

[7] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[8] Athanasios Papaioannou, Baris Gecer, Shiyang Cheng, Grigorios Chrysos, Jiankang Deng, Eftychia Fotiadou, Christos Kampouris, Dimitrios Kollias, Stylianos Moschoglou, Kritaphat Songsri-In, et al. Mimicme: A large scale diverse 4d database for facial expression analysis. In *European Conference on Computer Vision*, pages 467–484. Springer, 2022. 2

[9] 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. 3, 4

[10] Rolandos Alexandros Potamias, Alexandros Neofytou, Kyriaki Margarita Bintsi, and Stefanos Zafeiriou. Graphwalks: efficient shape agnostic geodesic shortest path estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2968–2977, 2022. 3

[11] Rolandos Alexandros Potamias, Michail Tarasiou, Stylianos Ploumpis, and Stefanos Zafeiriou. Shapefusion: A 3d diffusion model for localized shape editing. In *European Conference on Computer Vision*, pages 72–89. Springer, 2024. 1

[12] Michail Tarasiou, Rolandos Alexandros Potamias, Eimear O'Sullivan, Stylianos Ploumpis, and Stefanos Zafeiriou. Locally adaptive neural 3d morphable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1867–1876, 2024. 1, 3

[13] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7743–7753, 2022. 4

[14] Jiali Zheng, Rolandos Alexandros Potamias, and Stefanos Zafeiriou. Design2cloth: 3d cloth generation from 2d masks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1748–1758, 2024. 3

[15] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3d morphable face model with implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20343–20352, 2022. 3, 4