ShapeFusion: A 3D diffusion model for localized shape editing
Supplementary Material

Rolandos Alexandros Potamias, Michail Tarasiou, Stylianos Ploumpis, Stefanos Zafeiriou
Imperial College London, United Kingdom

1. Ablation Study

The proposed diffusion model incorporates three critical components, which contribute significantly to its effectiveness: the *relative features message passing scheme*, the *vertex-index positional embeddings* and the *hierarchical message passing scheme* (see lines 215-271, including Fig. 3, of the main manuscript). To systematically assess the contribution of one of the key elements to the overall performance, we perform a series of ablation experiments.

Relative Features. As has been shown in the literature [3, 11, 14], a key technique to increase the expressivity of a graph neural network is the use of relative features on the message passing scheme to enable propagation of relative information to each of the nodes. Without the relative features, the vertex update function relies only on the current vertex features along with each neighbor which limits the expressivity of the layer. This not only results in degradation of the performance, as shown in Tab. 2, but also in poor manipulation performance from anchor points. Specifically, as can be easily observed in Fig. 1 (first row), the manipulations result in discontinuous spikes around the anchor vertices. In contract, the proposed method, that utilizes a relative feature message passing scheme, achieves smooth manipulations that follow the anchor points without any artifacts.

Vertex-Index Embedding. As described in the main paper, training a network directly on the partially noisy shape space, is considerably challenging. The noise added on the vertices translates to permutation equivariance for the model, which results into learning point distributions without any semantic topological information. To break this equivariance, we introduced a vertex-index positional encoding $\mathbf{p}$ to enable the preservation of the topology on the generated meshes [2, 8]. As shown in Fig. 1 (middle row), the network trained without the vertex positional embeddings fails to generate results that preserve the topology of the input, resulting in meshes with permuted triangles. This can be quantified in Tab. 2, where the model trained without the vertex positional embeddings results in significant performance drop under both FID and ID metrics.

Hierarchical Message Passing. Inspired from [9], a key novelty of the proposed framework is the hierarchical message passing scheme that enables the aggregation of low and high level features across vertices (see Eq. 3 on the main paper). Using the proposed hierarchical message passing, each node features contain both global and local topological information, which is extremely important in tasks such as region sampling and manipulation. As the proposed model utilizes masked inputs, the generation of masked regions is strongly contingent upon the context of the unmasked regions. As shown in Fig. 1 (bottom row), without the use of hierarchical features, several artifacts arise around the region boundaries. In contrast using the proposed hierarchical message passing scheme, the generated regions are smooth and preserve the topology of the input mesh since its vertex contains contextual information of the global shape.

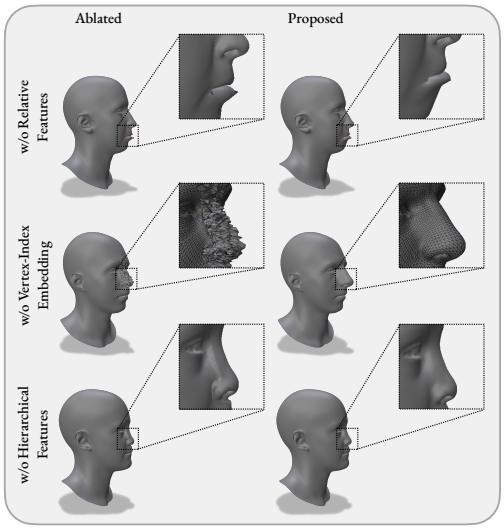

Figure 1. **Ablation Study.** Qualitative Evaluation of the different components of the proposed method.

Different GCN layers. In addition to the Spiral convolution layer, we have also compared our work with some recently proposed graph convolutional layers [1, 4, 7]. As can be seen in Tab. 2, the adaptive sprial convolution method achieves the best performance. In this paper, we utilized

Table 1. **Ablation Study.** Quantitative comparison between different ablated components on MimicMe dataset. All figures are measured in mm.

Method	FID (↓)	ID (↓)
w/o Relative Features	1.12	0.85
w/o Vertex-Index Emb.	14.1	2.19
w/o Hierarchical Features	1.32	0.12
Proposed	**0.30**	**0.05**

original spiral convolution layer for two reasons. First, as it was utilized by the baseline methods [5, 6], we can easily showcase the performance gain of the proposed framework. Second, during our experimentation we observed that the recently proposed adaptive spiral layer [1] despite achieving the best performance, its training time and model size were substantially larger.

Table 2. **Ablation Study.** Quantitative comparison between different mesh convolution layers on MimicMe dataset. All figures are measured in mm.

Method	DIV (↑)	FID (↓)	ID (↓)
Proposed	0.34	0.30	0.05
Proposed w. LSA-Conv [7]	0.38	0.25	0.04
Proposed w. GEM [4]	0.32	0.34	0.06
Proposed w. Adaptive Spiral [1]	**0.41**	**0.22**	**0.04**

2. Region Sampling

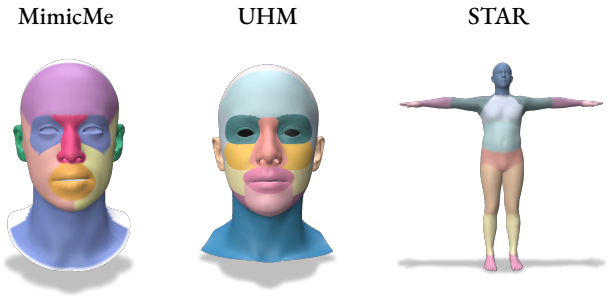

Figure 2. Regions used for experiment 4.1 for each dataset. Each color denotes a different region.

In the experiment of Sec. 4.1 of the main paper, we use the term *region sampling* to refer to the process of impainting a specific part of a shape. For this experiment, a 3D artist cropped the 3D face and body in several non-overlapping anatomical regions, as depicted in Fig. 2, that act as manipulation regions. In the supplementary video, we included additional qualitative results on region sampling for different face regions. As can be easily observed, compared to the M-VAE method, the proposed method achieves diverse and realistic samples that are fully disentangled from the rest of the shape.

Table 3. Quantitative evaluation of specificity metric between the proposed and the baseline methods for global sampling. Figures in mm.

Method	PCA(%90)	PCA(%95)	PCA(%99)	SD	LED	**Proposed**
Specificity (↓)	2.24	2.31	2.44	3.27	2.97	**2.11**

3. Global Sampling

As mentioned in Sec. 4.3 of the main paper, apart from region sampling, the proposed model can be used to sample full heads directly by simply masking the whole shape. In this section, we qualitatively and quantitatively evaluate the proposed model as a 3D morphable model. More specifically, to measure the generative abilities of the network we used *specificity* metric, which is commonly used for the assessment of parametric models. Specificity measures the realism of the generated faces and their similarity to the respective training samples. We compared our model with the VAE baselines SD [5] and LED [6] along with a PCA model trained on the same training set. To evaluate specificity metric, for each method, we generated 10,000 faces and measured their per-vertex distance from their closest sample on the ground-truth datasets. In Tab. 3, we report the specificity performance of the proposed and the baseline models. Following [10], for the PCA model, we report three different results based on the percentage of retained variance. The proposed method can generate diverse faces that follow the training distribution and manages to outperform PCA models with considerably less parameters. Fig. 3 illustrates samples generated from the proposed model. As can be easily observed the proposed model can generate diverse shapes with distinct characteristics that result to realistic faces.

4. Runtime Performance

Despite the considerable improvements that diffusion models have demonstrated on generative tasks, they suffer from the inherent drawback of prolonged inference times. In Tab. 4 we report the average inference performance on 100 runs of the proposed and the baseline methods on a single Nvidia RTX 4090. As expected the diffusion model attains slower inference times compared to VAE-based methods.

However, it is important to note that in contrast to the baseline methods, the proposed diffusion based model does not require any optimization for manipulation and fitting tasks. In this sense, the manipulation of a region translates to a single diffusion denoising pass, which requires the same inference time as region sampling. As shown in Sec 4.2 of the main paper, this can effectively reduce the

Figure 3. Faces sampled from the proposed method.

Table 4. Runtime performance of the proposed and the baseline methods across datasets. Times measured in seconds.

Method	MimicMe	UHM	STAR
SD [5]	**0.04**	**0.12**	**0.08**
LED [5]	0.06	0.14	0.09
Proposed	2.92	3.17	2.98

inference time by 10 $\times$ ($\sim$ 3.2sec) compared to the baseline methods ($\sim$ 22sec) for manipulation and fitting tasks.

5. Comparison with ARAP

We compared our method on shape editing against popular As-Rigid-As-Possible (ARAP) method [12]. Given that ARAP uses an alternating minimization strategy it is significantly slower on large meshes compared to the proposed method (10min for an edit vs 3.4sec for ours) and *does not have any shape prior model* which results in non feasible deformations (e.g. the pointing nose-tip and the curved eye) and artifacts. It is also important to note that ARAP method, similar to any non-learnable model, can not handle part swapping or generate and complete face regions.

6. Failure Cases

Editing shapes can cause failure cases when the anchors exceed the range of the statistical distribution. As shown in Fig. 5, when the anchor point (red) is dragged far away from the statistical plausible nose tips, the rest of the manipulated region fails to follow, resulting in a vertex-pick. This can also be quantified with the displacement plot, where the displacement of a masked vertex (green) in the neighborhood of the anchor gradually reach a plateau (green line) compared to the displacement of the anchor point that continues to increase as we linearly drag it (red line).

7. Effectiveness of Evaluation Metrics

To evaluate our model's generative performance we devised two heuristic measures (FID, ID) that leverage the PCA's latent space as a powerful and expressive prior. To measure the FID/ID losses we project the manipulated and the original samples and measure their distances in the latent space. Thus, the distances between the PCA projections of the manipulated and the original shapes would align with the model's manipulation performance. Intuitively, the smaller the edited region within a shape, the closer the distance be-

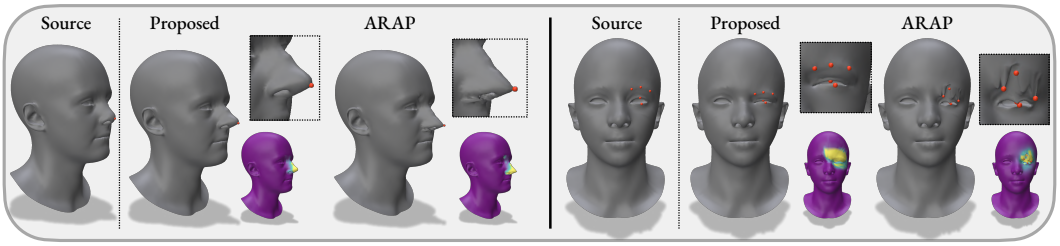

Figure 4. Qualitative and Quantitative comparison between the proposed and the ARAP [12] methods on face editing.

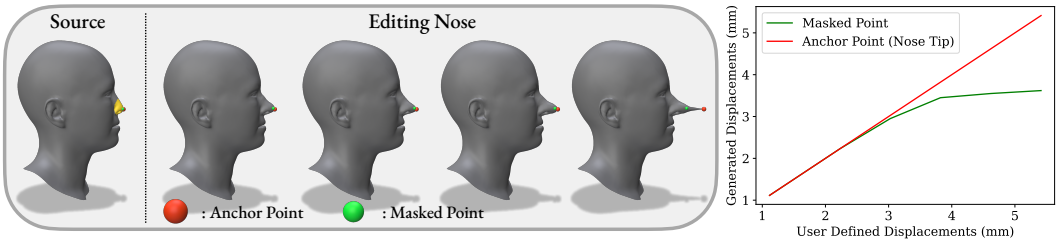

Figure 5. Failure cases of the proposed model.

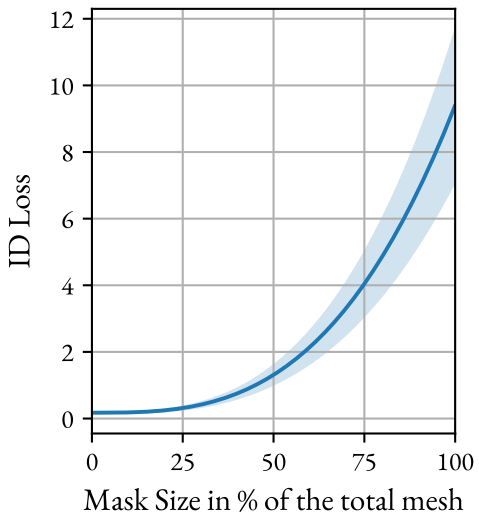

Figure 6. Effectiveness of the ID loss. As the manipulated region expands, it's expected that the ID loss will increase, resulting in more noticeable changes to the identity of the object.

tween the original and manipulated latent codes. This is also validated in Fig. 6, where the ID loss rises in correlation with the expansion of the manipulated area.

8. Discussion

Our work focuses on a critical task of 3D editing with numerous applications in 3D shape modeling. Despite extensive research in this field, the problem has not been effectively tackled. Our contribution is a simple, fast and powerful framework to effectively solve this issue. Previous methods [5, 6, 13], which enforce disentanglement via latent space partitioning, are *fundamentally flawed*. They compromise reconstruction quality for disentanglement and fail to produce localized edits. Additionally, these methods are limited to predefined regions and require slow optimization to edit shapes. In contrast, our approach provides an intuitive and interpretable solution, that is deeply embedded with the task, that can *guarantee localized edits* and allows flexible definitions of manipulated regions on the fly. In addition, this is the first study that has proposed a single versatile model to edit, manipulate, swap and generate 3D shapes and expressions *fully defined by the user*.

9. Limitations and Ethical Concerns

The development of a powerful 3D shape manipulation method undoubtedly raises concerns regarding the authorization of manipulations and the potential of using AI-generated content for harmful purposes. Additionally, since the model may have been trained on a dataset that does not accurately represent the world's population demographics, it is important to note racial biases that may occur and their possible impacts. Moreover, perpetuating unrealistic beauty standards through the alteration of facial and body features may contribute to body image issues and societal pressures. While the potential ethical concerns of the proposed method are limited, careful consideration of the ethical implications is essential in the development and deployment of these technologies.

Regarding the limitations of the proposed method, similar to previous approaches in 3D shape manipulation, our denoising model relies on a fixed template. This fixed topol-

ogy serves as a constraint for PCA and all explicit 3DMMs methods in the literature to achieve high-fidelity performance. Undoubtedly, a major reason behind the success of fixed topology models is their ability to leverage intrinsic correspondence across data, enabling the learning of high-frequency details using compact latent spaces. To mitigate the limitations of fixed topology, one could potentially explore the use of implicit models that do not operate within a fixed topology setting. However, to date, such models have severe limitations and cannot achieve the controllability and expressivity of 3DMMs.

References

[1] Francesca Babiloni, Matteo Maggioni, Thomas Tanay, Jiankang Deng, Ales Leonardis, and Stefanos Zafeiriou. Adaptive spiral layers for efficient 3d representation learning on meshes. *ICCV*, 2023. 1, 2

[2] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1985–1995, 2024. 1

[3] Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. *Advances in Neural Information Processing Systems*, 33:13260–13271, 2020. 1

[4] Pim De Haan, Maurice Weiler, Taco Cohen, and Max Welling. Gauge equivariant mesh cnns: Anisotropic convolutions on geometric graphs. *arXiv preprint arXiv:2003.05425*, 2020. 1, 2

[5] Simone Foti, Bongjin Koo, Danail Stoyanov, and Matthew J Clarkson. 3d shape variational autoencoder latent disentanglement via mini-batch feature swapping for bodies and faces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18730–18739, 2022. 2, 3, 4

[6] 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. 2, 4

[7] Zhongpai Gao, Junchi Yan, Guangtao Zhai, Juyong Zhang, Yiyan Yang, and Xiaokang Yang. Learning local neighboring structure for robust 3d shape representation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1397–1405, 2021. 1, 2

[8] 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. 1

[9] Rolandos Alexandros Potamias, Stylianos Ploumpis, and Stefanos Zafeiriou. Neural mesh simplification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18583–18592, 2022. 1

[10] Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, and Stefanos Zafeiriou. Handy: Towards a high fidelity 3d hand shape and appearance model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[11] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1

[12] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, pages 109–116. Citeseer, 2007. 3, 4

[13] 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. 4

[14] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. 1