LeGO: Leveraging a Surface Deformation Network for Animatable Stylized Face Generation with One Example

Supplementary Material

Overview

In this supplementary material, we present implementation details in Section A. Section B contains the details of the LeGO architecture. Section C covers additional experiments and their corresponding details. Section D is dedicated to visualizing example results of the applications of our method. Lastly, Section E presents additional results. For additional results, please visit the project page https://kwanyun.github.io/lego/.

A. Implementation Details

The implementation of LeGO including the surface deformation networks D_S and D_T, and MAGE on an Nvidia RTX 3090 GPU. The implementation details are sequentially presented, in the order of with D_S, D_T, MAGE, and the baselines for the experiments in the following paragraphs.

Training D_S To deform a template face to a person with a different identity and expression, we trained the source surface deformation network D_S using the FLAME [3] parameter Φ and its corresponding face mesh, as described in the main paper. We sampled 100k instances of Φ and their corresponding faces for self-supervised training, with Φ being sampled from a uniform distribution to learn diverse identities and expressions. We jointly trained $\mathcal{M}_{shape}$, $\mathcal{M}_{exp}$, and D_S for 400 epochs with a fixed learning rate of 1e-6. We adopted the SIMS approach to enable D_S to handle diverse topologies by training it using surface points instead of only sampling from mesh vertices.

Training D_T To modify the template face to incorporate styles while maintaining the same identity and expression as D_S, we trained the target surface deformation network D_T following the procedure outlined in the main paper. The training began with an initial learning rate of 3e-5, which gradually decreased to 1e-5 over 2,000 iterations. The balancing weights in Equation (9) in the main paper, λ_{vert},λ_{CLIP}, λ_{in}, λ_{across}, and λ_{style} were fixed at 80, 2e-3, 6e-3, 6e-3, and 4e-3, respectively.

During the training of D_T, we adopted a hierarchical rendering approach comprising three levels. The first level, featuring the most enlarged views, focused on rendering local facial parts such as the eyes, nose, and lips (illustrated in the blue box in Figure 1). The second level includes close-up views of faces from three directions, encompassing the front and sides of the face. The last level comprises full-face views from the same directions.

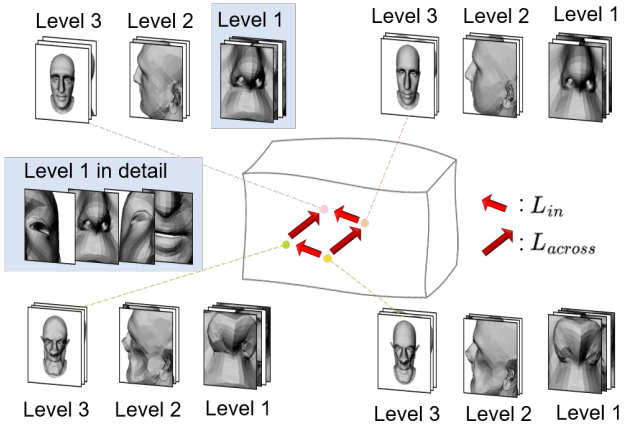

Figure 1. Detailed illustration of L_{in} and L_{across} with hierarchical rendering. The blue box shows an example of the predefined pivots of hierarchical rendering.

Training MAGE MAGE functions as an encoder that transforms faces with diverse topologies into the latent space of LeGO, specifically into the shared latent space of D_S and D_T. NFR [5] encoders were fixed during training, while ID2ID, exp2exp, and latent mapper were jointly trained. The model was trained with an initial learning rate of 3e-4, which gradually decreased to 5e-5 over 12,000 iterations.

Figure 2. First two rows are identity exemplar mesh and style exemplar mesh in our dataset, created for fine-tuning. Last row shows deformation target meshes for experiments

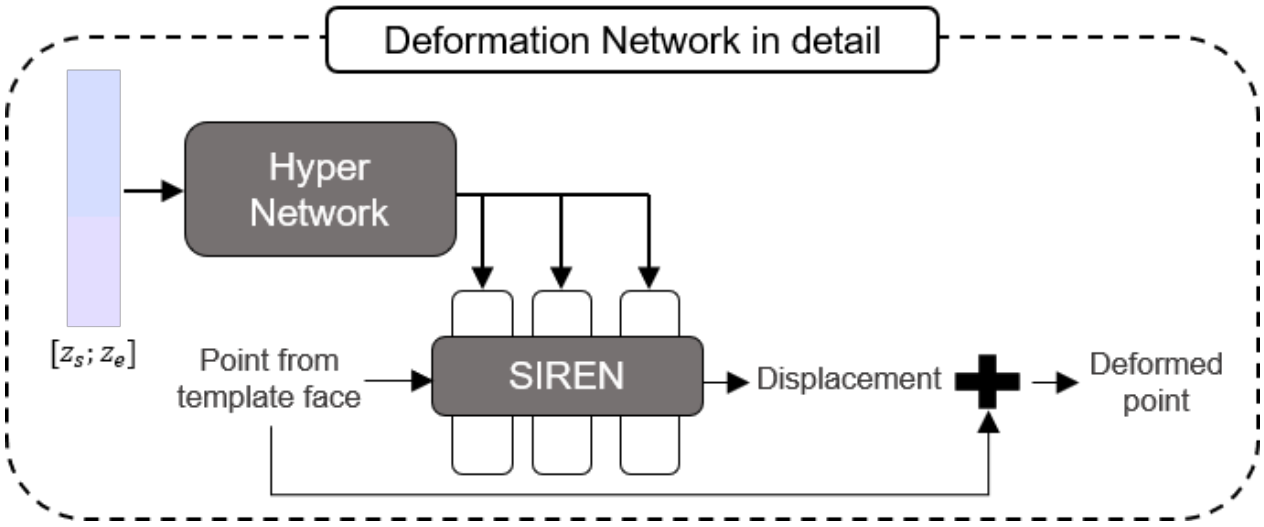

Figure 3. Visualization of deformation network in detail.

Baseline Methods All baseline methods were using their default settings as specified in their respective papers or the official code provided by the authors. As specified in the main paper, we utilized 8 different faces with corresponding manually crafted meshes as a dataset, which are shown in Figure 2. We also randomly sampled another 10 different identities without expressions from FLAME decoder as deformation targets for experiments. For the text-based methods [1, 4], we used the deformation target as the source template mesh and text to specify the target style. Eight text prompts that describe styles are as follows: "Bulldog makeup", "Disney Dwarf", "Exaggerated smile", "Musical Cats", "Orc", "Person with unicorn horn", "Person without nose", and "Pixar child".

B. LeGO Architecture

B.1 Deformation Network Architecure

The architectures of D_S and D_T are designed to compute the displacement of a point, either from a vertex or surface, and produce a deformed face output. The architectures are inspired by DD3C [2]. The architecture of the deformation network is illustrated in Figure 3. The latent code $[z_s;z_e]$ enters the hypernetwork, modifying the parameters of the SIREN MLP [6]. Subsequently, as the point from the template face traverses the network, the displacement is added to the point, determining the output position.

B.2 Rationale Behind Using the Deformation Network

The main reason to utilize a deformation network instead of simply adding an displacement lies in the inability of handling diverse inputs and outputs when the simple method is used. Another reason is to avoid severe artifacts that may occur when the identity of the deformation target and identity exemplar mesh are too different to directly transfer the displacement from one face to another. Examples of these artifacts are illustrated in Figure 4.

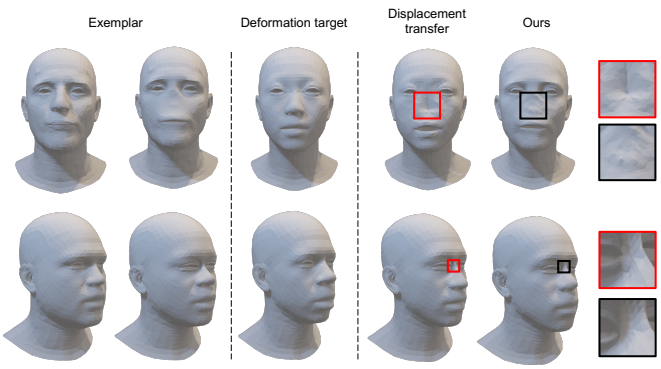

Figure 4. Artifacts occurred by simple displacement transfer. The red boxes show the close-up views. The boxes colored in red show the result of simple dispacement with artifacts(sunk nose and penetration on eye region) unlike Ours, colored in black.

C. Additional Experiments

C.1 Comparison with NFR

NFR [5] is a method that can transfer the expression of a target facial mesh to an unrigged identity mesh of arbitrary topology. Because NFR is specifically designed for expression transfer, it is difficult for the method to preserve both identity and style in the resulting mesh. However, because NFR is the backbone of MAGE, we compared it with our

Figure 5. Comparison with NFR and LEGO. Becaus NFR is designed for expression transfer, it's expression encoder cannot preserve identity.

method. As shown in Figure 5, although NFR could generate a stylized mesh, it failed to preserve the original identity, resulting in all similar outcomes that reflect the style exemplar.

C.2 Ablation on Direct Style Loss

As stated in the main paper, "Ours" and "Ours w. direct L_{style}" loss produced the least amount of surface artifacts. Here, the style loss that directly compared $D_T([{z_s}^{samp}; {z_e}^{samp}])$ and M_T, it forced the stylized face to have the same expression as M_T. In contrast, ours that compared $D_T([{z_s}^{samp}; {z_e}^{ref}])$ and M_T successfully maintained animatability. This is illustrated in Figure 6.

D. Applications

D.1 Visualization of Results Produced by the Applications.

We present additional results of style interpolation in Figure 7. These results demonstrate that our method can effectively construct the latent space for identities and styles, ensuring that even when mixing weights, the person's identity remains unchanged while the styles transition smoothly. This finding inspired the generation of new styles by blending existing ones. Additionally, we showcase further results on generating stylized 3D faces from 2D portraits, indicating that our method does not require a mesh as input; instead, any image can be used to create a stylized face with a specific identity, thereby broadening the practical application of our method.

D.2 Retargeting from Video

Retargeting is one of widely used applications in animation in which target follows the animation of the source. We performed an additional experiment on video driven stylized 3D face retargeting. Using a metrical photometric tracker [9], we can obtain the shape and expression parameters of FLAME from video, which can be directly adopted to LeGO. From these parameters, we achieved 3D stylized face retargeting as shown in Figure 8. Additional results are shown in the supplementary video.

E. Additional Results

We present additional stylization results produced by LeGO trained with a paired exemplar. Figures 9, 10 and 11 display the results of all eight styles and deformation targets with different topologies.

References

[1] William Gao, Noam Aigerman, Thibault Groueix, Vladimir G Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance. *arXiv preprint arXiv:2304.13348*, 2023. 2

[2] Yucheol Jung, Wonjong Jang, Soongjin Kim, Jiaolong Yang, Xin Tong, and Seungyong Lee. Deep deformable 3d caricatures with learned shape control. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

[3] 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. 1

[4] Yiwei Ma, Xiaoqing Zhang, Xiaoshuai Sun, Jiayi Ji, Haowei Wang, Guannan Jiang, Weilin Zhuang, and Rongrong Ji. X-mesh: Towards fast and accurate text-driven 3d stylization via dynamic textual guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2749–2760, 2023. 2

[5] Dafei Qin, Jun Saito, Noam Aigerman, Thibault Groueix, and Taku Komura. Neural face rigging for animating and retargeting facial meshes in the wild. *arXiv preprint arXiv:2305.08296*, 2023. 1, 2

Figure 6. Comparison with Ours and Ours w. direct L_{style}. Both methods generated the style well while Ours followed the expression from the deformation target better compared to Ours w. direct L_{style}.

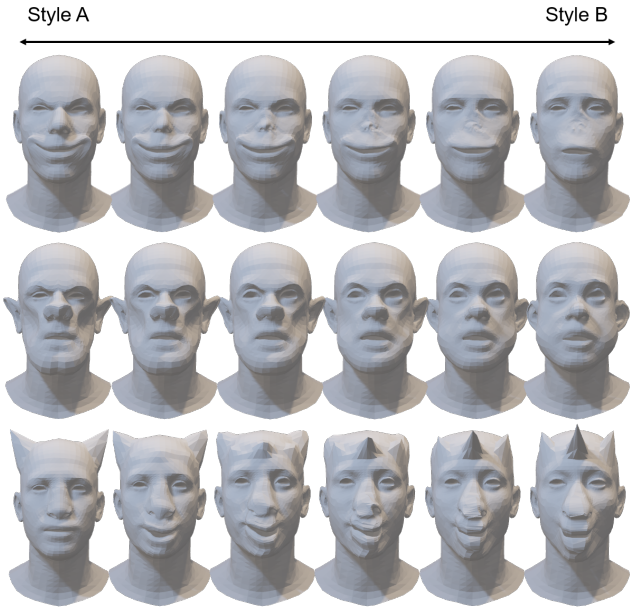

Figure 7. Additional results of style interpolation.

[6] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 2

[7] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 5

[8] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 5

[9] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 250–269. Springer, 2022. 3

Figure 8. Retargeting from video, two different styles are shown for each input. Left examples are from Talking-head-1KH [8] and right examples are from MEAD dataset [7].

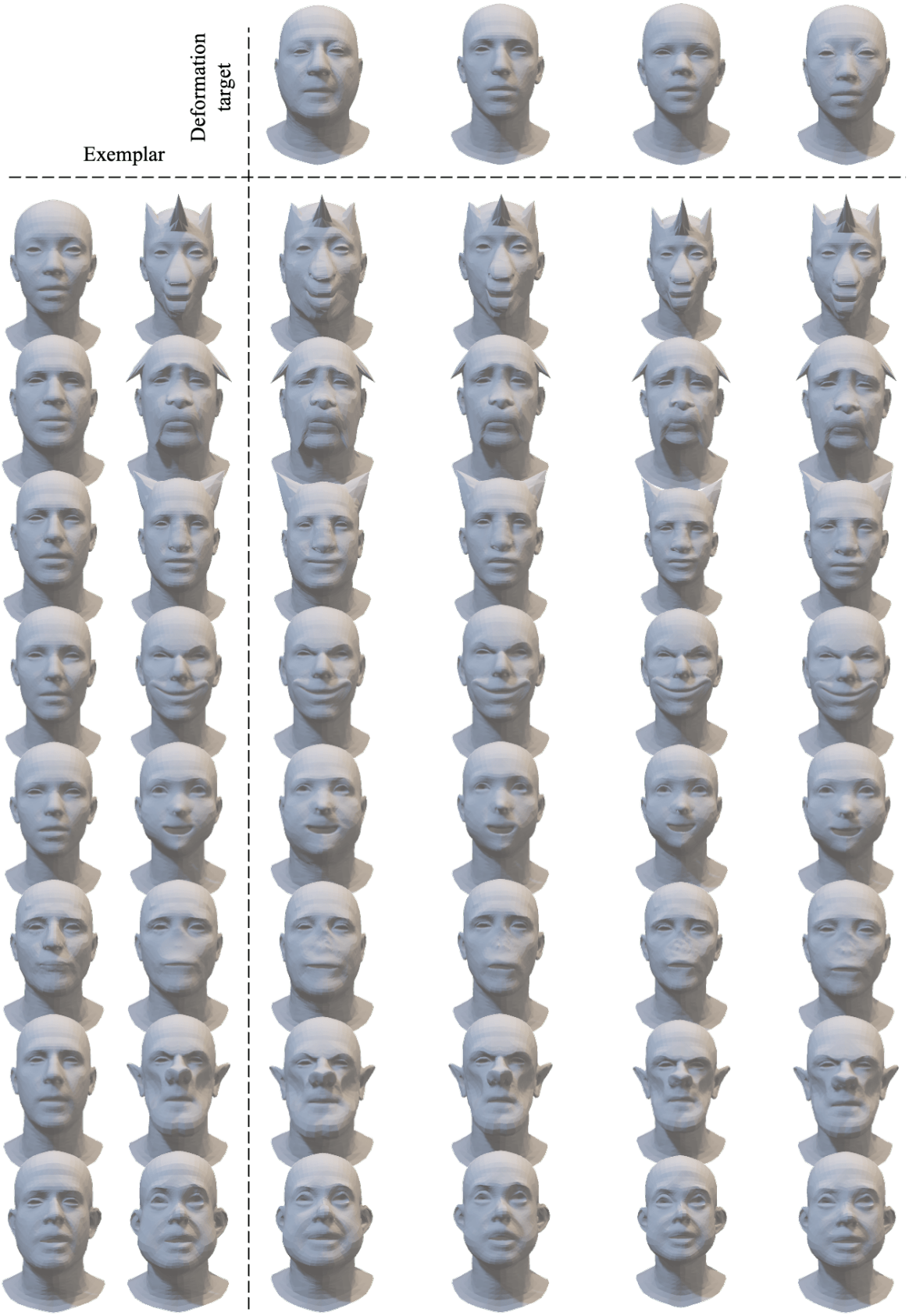

Figure 9. Additional results on all 8 different styles.

Figure 10. Additional results on all 8 different styles.

Figure 11. Additional results on all 8 different styles.