Conceptual and Hierarchical Latent Space Decomposition for Face Editing -Supplementary Material-

Savas Ozkan Mete Ozay Tom Robinson
Samsung Research UK
savas.ozkan@samsung.com

1. Introduction

This supplementary document provides additional insights about our models and their evaluations, including comparisons with baselines.

2. Implementation Details

Please refer to our main manuscript (Section 4) for the details of transformer networks used in TLSE and GAN controller. The generator model used in TLSE is based on [2]. It contains 18 layers where modulated convolution [5] and Leaky ReLU [11] are sequentially utilized at each layer. Modulated convolutions have a filter size of 32 and a kernel size of 1×1. To reduce the computational complexity, the output resolution of the generator model is set to 64×64. No image upsampling is utilized in the intermediate layers so that the resolution of its Fourier block and coordinate embedding is set to 64×64.

Inference time to edit a GAN feature $\mathbf{x}$ on a single Nvidia RTX 3090 takes approximately 0.004 second using our GAN controller. Face synthesis time (the resolution of synthesized faces is 1024×1024) of the pre-trained StyleGAN2 model per a GAN feature $\mathbf{x}$ takes approximately 0.1 second on a single Nvidia RTX 3090. Note that reported inference time for a baseline [1] is 0.21 second on a single Nvidia Titan XP.

Edit	illum	pose	expr	all
RigNet [10]	0.967	0.961	0.979	0.939
Transformer (our)	0.981	0.986	0.983	0.963

Table 1: Face identity evaluation (cosine similarity) to compare different GAN controller methods using the same intermediate latent space $\mathcal{Q}$.

Edit	illum	pose	expr
RigNet [10]	0.43	0.16	3.61
Transformer (our)	0.41	0.11	3.49

Table 2: Face concept edit precision evaluation (mean absolute error) to compare different GAN controller methods using the same intermediate latent space $\mathcal{Q}$.

3. An Ablation Study of the Impact of Latent Space Editing with Transformers on the Performance

In this section, we compare our transformer-based GAN space controller with RigNet model used in [10]. Both models utilize the intermediate latent space $\mathcal{Q}$ in the optimization step. Tab. 1 presents the face identity scores (higher is better) calculated on the StyleFlow dataset with face embeddings [3]. The results demonstrate that our transformer-based controller outperforms RigNet model in term of preserving face identities under different face manipulation configurations. The reason is that the transformers can effectively capture face details at multiple abstraction levels by ensuring that identities remain unaffected during face manipulation. Furthermore, face concept editing precision is reported in Tab. 2. In particular, complex face concepts like pose and expression can be edited more precisely with transformers. Since these parameters are controlled by multiple abstraction levels in the GAN space (please refer to Fig. 10 in our main manuscript for details), disentangling them with RigNet model is not sufficient to capture all details.

4. An Ablation Study of the Impact of Coefficients and Multi-Task on the Performance

In this section, we analyse the impact of coefficients and multi-task learning on the training of our GAN controller. Results are reported in Tab. 3. Here, the base model represents our best model configuration where λ_{edit} and λ_{sparse} are set to $1e-2$ and $1e-4$, respectively, while multi-task learning is utilized. First, we test the impact of multi-task learning on identity and edit scores. Results show that our model overfits to pose and expression manipulation when multi-task learning is not employed. Second, we

	ID Score	Edit Score		
	all	illum	pose	expr
Base Model	0.963	0.41	0.11	3.49
w/o Multi-Task	0.965	0.47	0.10	3.51
$\lambda_{edit} = 1e^{-1}$	0.978	0.49	0.20	3.81
$\lambda_{edit} = 1e^{-3}$	0.943	0.40	0.10	3.45
$\lambda_{sparse} = 1e^{-3}$	0.971	0.45	0.16	3.65
$\lambda_{sparse} = 1e^{-5}$	0.955	0.44	0.13	3.57

Table 3: Identity and edit scores computed under different configurations in order to demonstrate the impact of coefficients (λ_{edit} and λ_{sparse}) and multi-task learning.

inspect the effect of the λ_{edit} coefficient. As expected, setting this value higher promotes keeping more face identities/attributes in the manipulation step. On the other hand, when it is low, face identities/attributes start to differ from the those in the input GAN features, leading to a decrease in the identity score. Later, we explore the contribution of the λ_{sparse} coefficient during the optimization step. Similarly, increasing the value of this coefficient limits the learning of a full GAN controller for face editing. However, when the value is reduced, both identity and edit scores are simultaneously affected, since unrelated abstraction levels contribute to the edited GAN features.

5. Additional Discussion for the Needs of Two Pairs in Optimization

The purpose is to unveil the unseen correlations of facial concepts in GAN space $\mathcal{X}$ using feature pairs $\mathbf{x_1}, \mathbf{x_2} \in \mathcal{X}$. For instance, $\mathbf{x_1}$ and $\mathbf{x_2}$ might represent a smiling face and a right-side illuminated face respectively. Our loss function enables to learn a new GAN feature $\mathbf{x_3}$ (edited feature) that jointly represents a smiling and right-side illuminated face. This ultimately improves the diversity and quality of edits for our controller.

6. Application on StyleGAN and Diffusion models

For StyleGAN, Ganspace [4] utilized residuals $\mathbf{x_{edit}} = \mathbf{x} + \Delta\mathbf{x_{edit}}$ (Eq. 5 in our main manuscript) for face manipulation. Since our model can optimize $\Delta\mathbf{x_{edit}}$ efficiently, our model can be employed to control GAN space $\mathcal{X}$ of StyleGAN models. As for diffusion-based models, a recent work [7] indicates that high-level hierarchical features can be estimated by latent variables. We consider that our latent codes can be employed to disentangle the feature space of latent diffusion models.

7. Results for Manipulation on Paintings

One of the advantages of manipulating GAN features is that the same GAN controllers can be used for different styles without the need of retraining. To be specific, our pre-trained pipeline can be used to manipulate face paintings by only replacing the original StyleGAN2 generator with a StyleGAN2 painting generator by $\mathbf{I_{xp}} = \sigma_p(\mathbf{x})$. Fig. 2 illustrates the visual results for editing paintings. Pose manipulation results are only reported, since no illumination or expression concept exists in the GAN space of paintings. At the end, the figure shows that high-quality and consistent results are achieved.

8. Results for Face Likeness Editing

Along with the facial expression θ, the scene illumination γ, and the head pose ($\mathbf{R}$, $\mathbf{t}$), the parameter space $\mathcal{P}$ (i.e., parameter space of 3D morphable face models) allows us to control the facial shape β and the skin texture ψ for face editing. To be specific, these parameters control facial features such as age, gender, and identity. For this task, we utilize our transformer-based GAN controller to edit these parameters using the same pipeline. First, we analyze the impact of changing these parameters onto StyleGAN2 features $\mathbf{x}$. The changes for different abstraction levels with mean square difference and variance are illustrated in Fig. 1. Similarly, our model sparsely edits the features by focusing on coarse and medium details in the GAN space. No result related to this experiment is reported in [10]. Indices determined in [1] for gender and age perfectly overlap with the indices estimated in our method.

To visualize the effectiveness of our method for face editing in terms of face likeness, we transfer shape and texture parameters from target images to source images. Results are shown in Fig. 3. The results demonstrate that age and gender attributes can be accurately transferred from target to source images using our method. Similarly, the background and other attributes observed in source images are largely preserved in edited images.

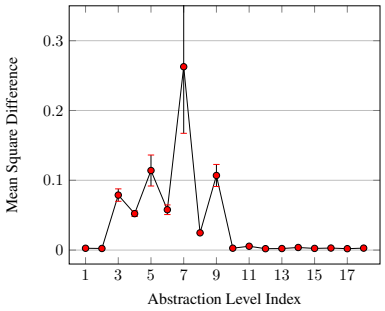

Figure 1: Change of StyleGAN2 features $\mathbf{x}$ at different abstraction levels after face likeness parameters (i.e., shape and texture parameters) are manipulated.

We also report results for face likeness editing on real faces in Fig. 4. Projections of [6] onto StyleGAN2 space are employed. High-quality results are similarly produced by manipulating age and gender attributes. Here, we also report the results for illumination manipulation.

9. Background Preservation

The undesired background changes might happen by setting a low value for λ_{edit} so that less details from source GAN feature $\mathbf{x}$ can be preserved (please see Tab. 3). Also, since image backgrounds are not explicitly represented as a separate face concept in $H(\cdot)$ model (i.e., the related category does not exist in 3DMMs), background information may be encoded with other concepts like gender and age.

10. Additional Comparative Analyses

As indicated in [1] and already shown in our results (please refer to Fig. 5 and Fig. 6 in our main manuscript for details), the baselines do not perform well when all face concepts are simultaneously applied. Therefore, we separately report comparative results in Fig. 5, Fig. 6 and Fig. 7 for pose, expression and illumination manipulation where severe artifacts can be observed for baselines. As baselines, InterfaceGAN (IG) [8], SeFa [9], StyleRig (SR) [10] and StyleFlow (SF) [1] are selected. Note that since there is no available IG model for illumination manipulation, we do not report any visual results on illumination manipulation for IG. Similar to the reported observations in our manuscript, for pose manipulation as illustrated in Fig. 5, SF modifies the other face attributes such as age during face manipulation (red boxes). On the other hand, SR can unintentionally change other concepts like expression and illumination (blue boxes). For SeFa, face identities are severely altered (orange boxes). IG yields consistent results with our results. However, when all face concepts are simultaneously applied (please refer to Fig. 6 in our main manuscript), the visual performance of IG significantly decreases.

The results for expression manipulation are demonstrated in Fig. 6. One of the common issues is that models have difficulty in editing faces beyond the limit set by the users (blue boxes). This indicates that the learning capacity of these models is very limited compared to our method. Another problem is that face identities can be altered during the expression manipulation (red boxes). Also, producing unrealistic expressions is severe for some models (orange boxes).

Fig. 7 illustrates the results for illumination manipulation. Similarly, our method performs significantly better than the baselines in preserving attributes and identities (red boxes), as well as producing consistent and satisfactory results (blue boxes).

11. Additional Visual Results

Fig. 8 shows our results when all face concepts are simultaneously transferred from target images to source images. We observe that our method can produce consistent results (each column) and successfully handle multiple concepts for face editing.

12. Limitations of Our Method

The main limitation of our method is that the 3D structures of faces are captured with 3DMMs. Hence, facial objects such as eyeglasses and hats are not represented in the parameter space $\mathcal{P}$. We found that the shadow patterns generated by these objects may remain on the manipulated faces under different illumination. Fig. 9 provides illustrative examples for this limitation.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. ACM Transactions on Graphics (ToG), 40(3):1–21, 2021. 1, 2, 3

[2] Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov. Image generators with conditionally-independent pixel synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14278–14287, 2021. 1

[3] Adam Geitgey. Github-face recognition, 2020. 1

[4] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. Advances in Neural Information Processing Systems, 33:9841–9850, 2020. 2

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8110–8119, 2020. 1

[6] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. ACM Transactions on Graphics (TOG), 42(1):1–13, 2022. 3

[7] Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In CVPR, pages 10684–10695, 2022. 2

[8] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9243–9252, 2020. 3

[9] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1532–1540, 2021. 3

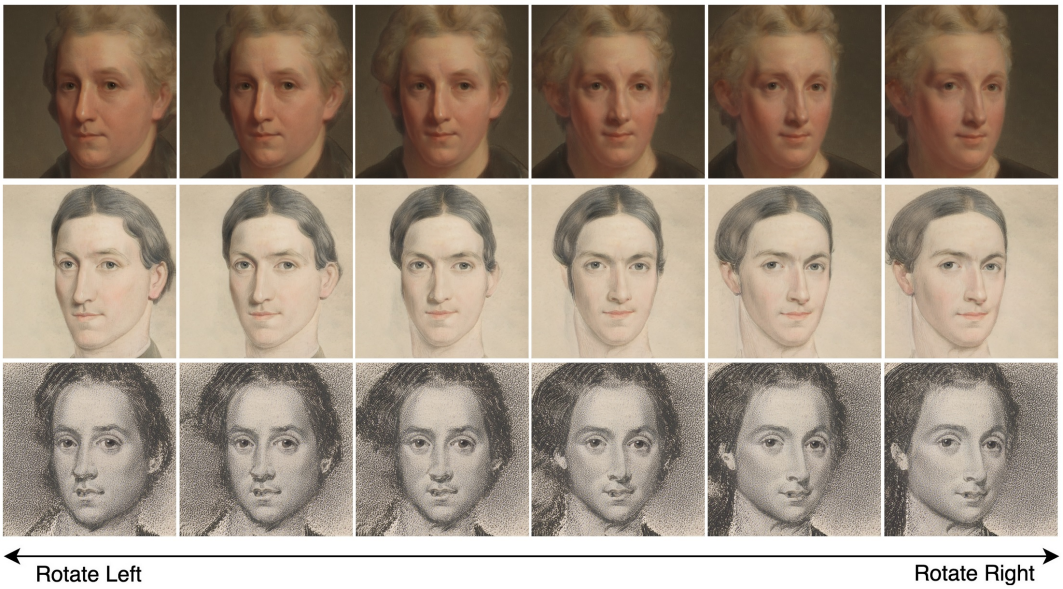

Figure 2: Results for face manipulation on paintings.

Figure 3: Sample results provided by our method for transferring face likeness (i.e., shape and texture parameters) from target to source images.

[10] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6142–6151, 2020. 1, 2, 3

[11] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853, 2015. 1

Figure 4: Sample results provided by our method for editing on real faces according to face likeness and illumination.

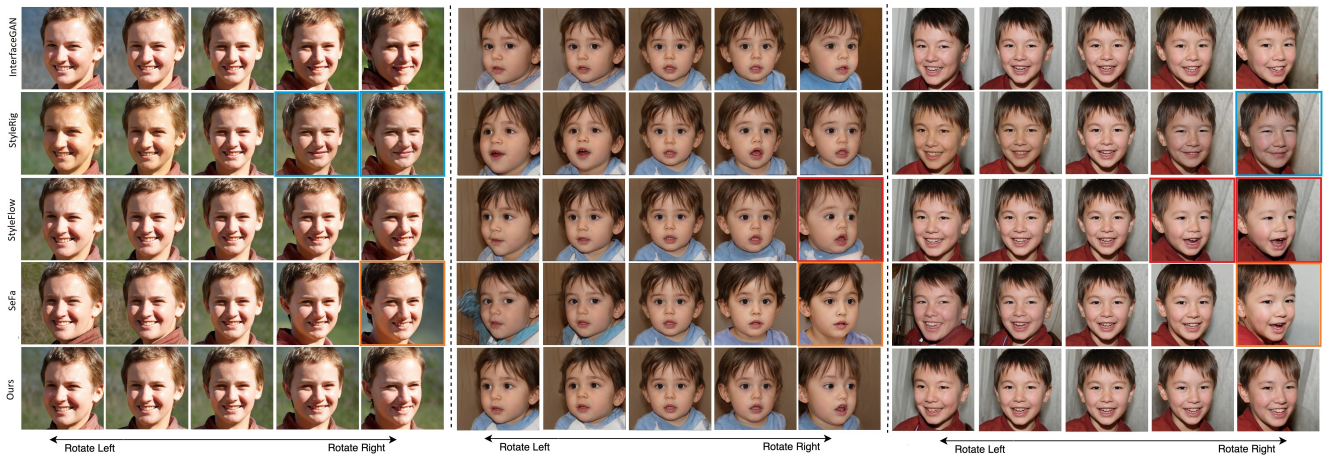

Figure 5: Comparative results for pose manipulation.

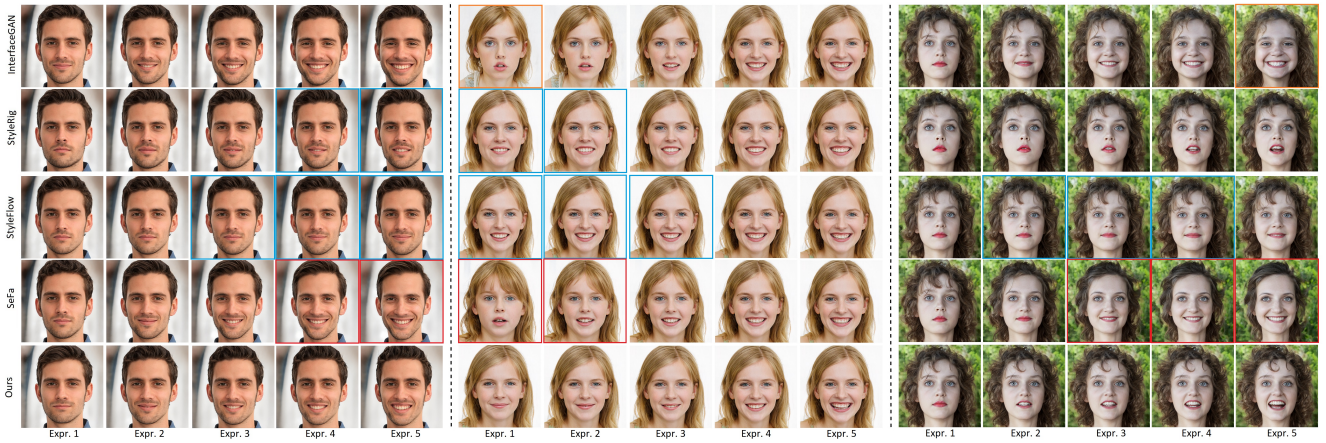

Figure 6: Comparative results for expression manipulation.

Figure 7: Comparative results for illumination manipulation.

Figure 8: Sample results provided by our method for transferring pose, expression and illumination simultaneously from target to source images.

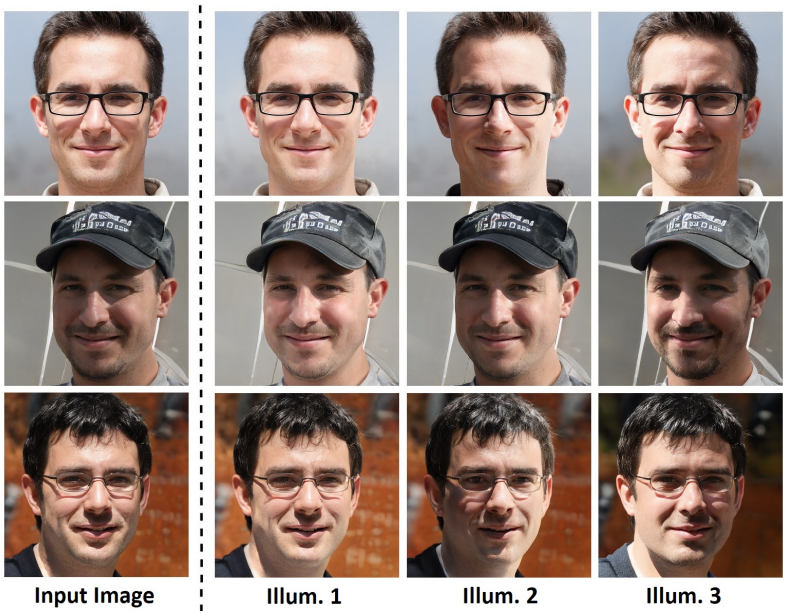

Figure 9: Limitation of our model under different illumination.