High-resolution Face Swapping via Latent Semantics Disentanglement – Supplementary Materials –

Yangyang Xu[1,3] Bailin Deng[2] Junle Wang[3] Yanqing Jing[3] Jia Pan[4] Shengfeng He[1*]
[1]South China University of Technology [2]Cardiff University
[3]Tencent [4]The University of Hong Kong

We provide more content in this supplementary, include the detail of architectures, more qualitative comparisons with MegaFS [3] and the quantitative comparisons of different variants.

1. Detail of Architectures

1.1. Landmark Encoder

We use a modification of pSp [2] as our landmark encoder. pSp inverts the input image to the $W+$ latent space with a feature pyramid through three levels: the coarse, medium, and fine, which controls different level of attributes. We discard pSp's higher layers and only use the coarse and medium layers to produce the semantic transfer direction $\overrightarrow{n}$.

1.2. Target Encoder

Our target encoder consists of 8 downsample blocks, each block contains a convolutional layer and a 'Blur' layer [1]. The target encoder produces the multi-resolution target features, which includes: $8 \times 8 \times 512$; $16 \times 16 \times 512$; $32 \times 32 \times 512$; $64 \times 64 \times 512$; $128 \times 128 \times 256$; $256 \times 256 \times 128$; $512 \times 512 \times 64$; $1024 \times 1024 \times 32$. Then they blend with the source features produced by StyleGAN generator with the corresponding resolution.

1.3. Decoder

The decoder has a mirror structure with target encoder, and we use the transpose convolution for upsampling. In the first upsampling block, we take the blending results of target and source features as input. And in each rest block, we concatenate the blended features with the output of last block as input.

*Corresponding author (hesfe@scut.edu.cn).

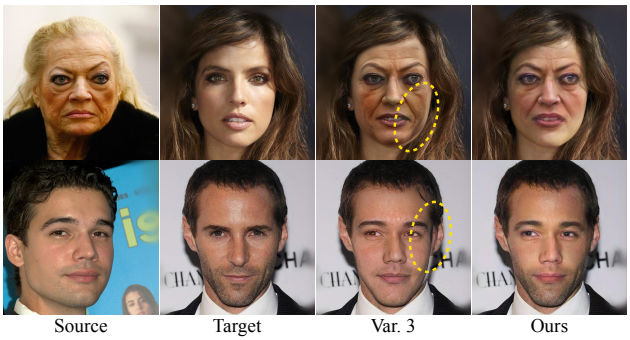

Figure 7. Qualitative results of Var.3. Problematic regions are circled out in yellow.

2. More Ablation Studies

2.1. Effectiveness of the landmark encoder.

To test the effectiveness of the landmark encoder, we implement a variant, named *Var.3*, that infers the structure transfer latent direction $\overrightarrow{n}$ from the face images directly without using the landmarks. We replace the landmark encoder with an image encoder, and feed the concatenation of the source and target face images to the encoder.

Its results are shown in Tab. 3 and Fig. 7. In Tab. 3, it has worse performance for the metrics *Pose Err.* and *Exp. Err.*, which indicates that the landmark encoder plays a vital role in pose and expression preservation. Besides, without the landmark guidance, Var.3 cannot effectively blend the target background with the source inner face using target mask; this leads to lower quality results than our method in Fig. 7.

2.2. Quantitative Comparisons of Different Variants

Tab. 3 shows quantitative evaluation results on CelebA-HQ. We can see that both Var.1 and Var.2 have a higher FID value than our approach, indicating worse quality of face images from both variants. This is because our appearance transfer can narrow the color gap while our background

Table 3. Quantitative of different variants on the CelebA-HQ dataset with four metrics. ↓ denotes the lower the better and vice versa, the best results are marked in **bold**.

Variants	ID Simi.↑	Pose Err.↓	Exp. Err.↓	FID↓
Baseline	0.5214	3.498	2.95	11.645
Var.1	0.5645	3.034	2.76	10.986
Var.2	0.5539	3.023	2.74	10.345
Var.3	0.5542	3.242	3.11	10.642
Ours	**0.5688**	**2.997**	**2.74**	**9.987**

transfer eliminates the blending artifacts, both helping to improve the quality of our results. On the other hand, Var.1 and Var.2 have similar performance as our approach on the other three metrics related to identity, pose and expression. This is because they also utilize the landmark encoder and the identity-preservation loss, which help to transfer the pose and expression while preserving the identity. The analysis of Var.3 can be seen in last subsection.

3. More Qualitative Comparisons

In Fig. 8, we present more of qualitative comparison on CelebA-HQ dataset with MegaFS [3], as shown in the main paper, our method can achieve favorably better face swapping results, it transfers the structure and appearance attributes as desired with identity-preserving.

References

[1] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 1

[2] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296, 2021. 1

[3] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *CVPR*, pages 4834–4844, 2021. 1, 2

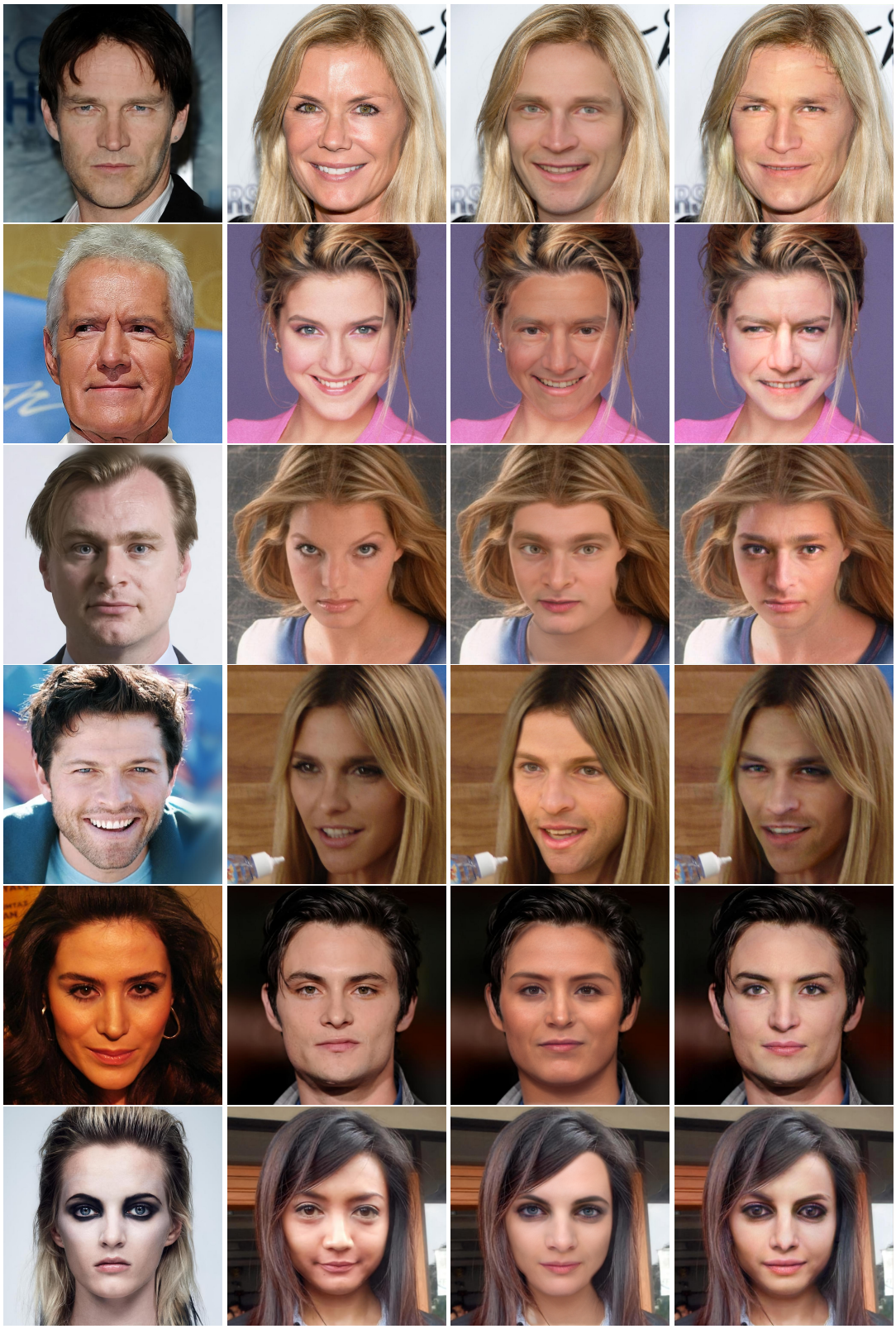

Figure 8. More qualitative comparison of face swapping on CelebA-HQ dataset.