MaskGAN: Towards Diverse and Interactive Facial Image Manipulation Supplementary Material

Cheng-Han Lee[1] Ziwei Liu[2] Lingyun Wu[1] Ping Luo[3]

[1]SenseTime Research [2]The Chinese University of Hong Kong [3]The University of Hong Kong

A. Additional Implementation Details

Our MaskGAN is composed of four key components: MaskVAE, Dense Mapping Network, Alpha Blender, and Discriminator. Specifically, Dense Mapping Network contains two elements: Image Generation Backbone, Spatial-Aware Style Encoder. More details about the architecture design of these components and training details are shown below.

MaskVAE. The architecture of MaskVAE is similar to UNet [6] without skip-connection. Detailed architectures of Enc_{VAE} and Dec_{VAE} are shown in Fig. 1 which uses BN for all layers.

Image Generation Backbone. We choose the architecture of Pix2PixHD [7] as Image Generation Backbone. The detailed architecture is as follow:
$c7s1{-}64, d128, d256, d512, d1024, R1024, R1024, R1024, R1024, u512, u256, u128, u64 - c7s1$.
We utilize AdaIN [2] for all residual blocks, other layers use IN. We do not further utilize a local enhancer because we conduct all experiments on images with a size of 512 × 512.

Spatial-Aware Style Encoder. As shown in Fig. 2, Spatial-Aware Style Encoder consists of two branches for receiving both style and spatial information. To fuse two different domains, we leverage SFT Layers in SFT-GAN [8]. The detailed architecture of SFT Layer is shown in Fig. 3 which does not use any normalization for all layers.

Alpha Blender. Alpha Blender also follows the desing of Pix2PixHD but only downsampling three times and using three residual blocks. The detailed architecture is as follow:
$c7s1{-}32, d64, d128, d256, R256, R256, R256, u128, u64, u32 - c7s1$ which uses IN for all layers.

Discriminator. Our design of discriminator also follows Pix2PixHD [7] which utilize PatchGAN [3]. We concatenate the masks and images as inputs to realize conditional GAN [5]. The detailed architecture is as follow:
$c64, c128, c256, c512$ which uses IN for all layers.

Training Details. Our Dense Mapping Network and MaskVAE are both updated with the Adam optimizer [4] ($\beta_1 = 0.5$, $\beta_2 = 0.999$, learning rate of $2e^{-4}$). For Editing Behavior Simulated Training, we reduce the learning rate to $5e^{-5}$. MaskVAE is trained with batch size of 16 and MaskGAN is trained with the batch size of 8.

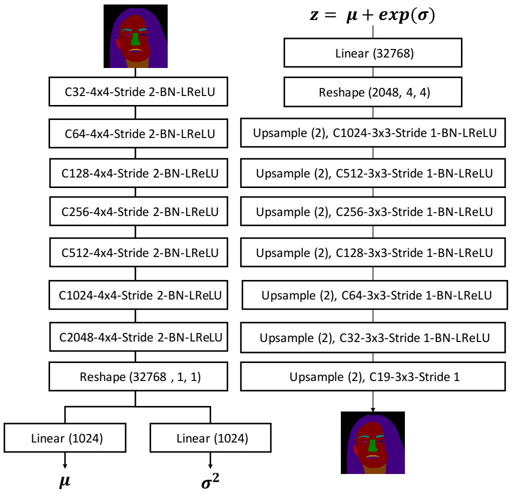
$z = \mu + exp(\sigma)$
Linear (32768)
Reshape (2048, 4, 4)
Upsample (2), C1024-3x3-Stride 1-BN-LReLU
Upsample (2), C512-3x3-Stride 1-BN-LReLU
Upsample (2), C256-3x3-Stride 1-BN-LReLU
Upsample (2), C128-3x3-Stride 1-BN-LReLU
Upsample (2), C64-3x3-Stride 1-BN-LReLU
Upsample (2), C32-3x3-Stride 1-BN-LReLU
Upsample (2), C19-3x3-Stride 1
C32-4x4-Stride 2-BN-LReLU
C64-4x4-Stride 2-BN-LReLU
C128-4x4-Stride 2-BN-LReLU
C256-4x4-Stride 2-BN-LReLU
C512-4x4-Stride 2-BN-LReLU
C1024-4x4-Stride 2-BN-LReLU
C2048-4x4-Stride 2-BN-LReLU
Reshape (32768 , 1, 1)
Linear (1024)
Linear (1024)
μ
σ^2

Figure 1: Architecture of MaskVAE.

B. Additional Ablation Study

A simple quantitative comparison is shown in Table. 1. SFT layers utilize more parameters to fuse to different domains together. As a result, it is reasonable that SFT layers have better effect than concatenation.

In Fig. 4, we show a visual comparison of style copy. The results with EBST have better color saturation and attribute keeping quality (heavy makeup).

C. Additional Visual Results

In Fig. 5, Fig. 6, Fig. 7, and Fig. 8, we show additional visual results of attribute transfer for a specific attribute: **Smiling**. We compare our MaskGAN with state-of-the art methods including Pix2PixHD [7] with modification,

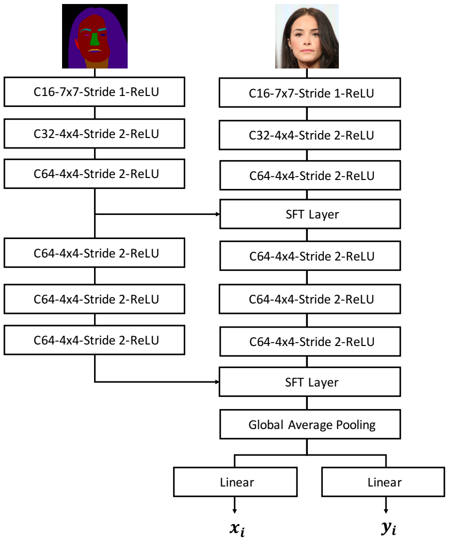

Figure 2: Architecture of Spatial-Aware Style Encoder.

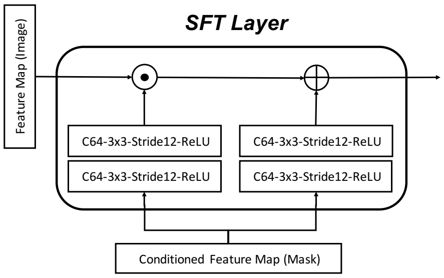

Figure 3: Architecture of Spatial Feature Transform Layer.

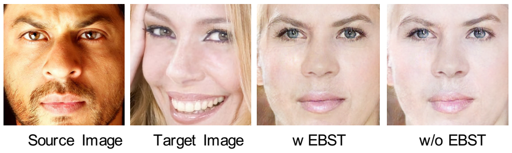

Figure 4: Visual comparisons of training with and without EBST.

ELEGANT [9], and StarGAN [1].

In Fig. 9, Fig. 10, Fig. 11 and Fig. 12, we show additional visual results of style. We compare our MaskGAN with state-of-the art methods including Pix2PixHD [7] with modification.

In the accompanying video, we demonstrate our interactive facial image manipulation interface. Users can edit the shape of facial components or add some accessories toward manipulating the semantic segmentation mask.

Metric	Attribute cls. acc(%)			Seg(%)	FID
MaskGAN-concat	63.1	61.3	84.8	90.8	27.13
MaskGAN-SFT	67.7	67.1	89.0	92.5	26.22
GT	96.9	88.1	95.4	93.4	-

Table 1: Ablation study on style copy. Attribute types in attribute classification accuracy from left to right are **Male**, **Heavy Makeup**, and **No Beard**. P.S. The train/test split here is different from the main paper.

References

[1] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 2

[2] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 1

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[5] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015. 1

[7] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 1, 2

[8] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 1

[9] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. *arXiv preprint arXiv:1803.10562*, 2018. 2

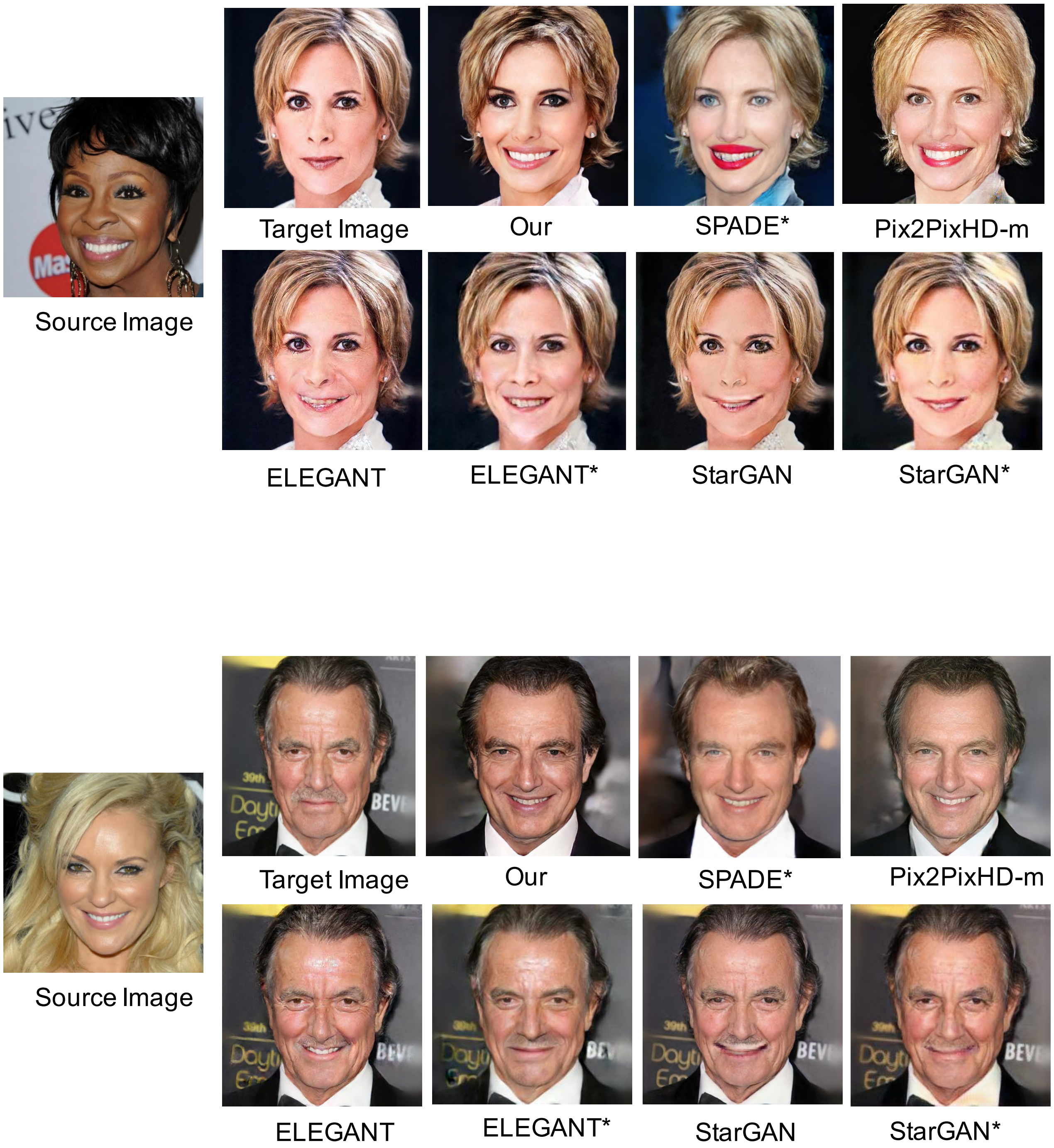

Figure 5: Visual results of attribute transfer for a specific attribute: **Smiling**. * means the model is trained with a size of 256 × 256.

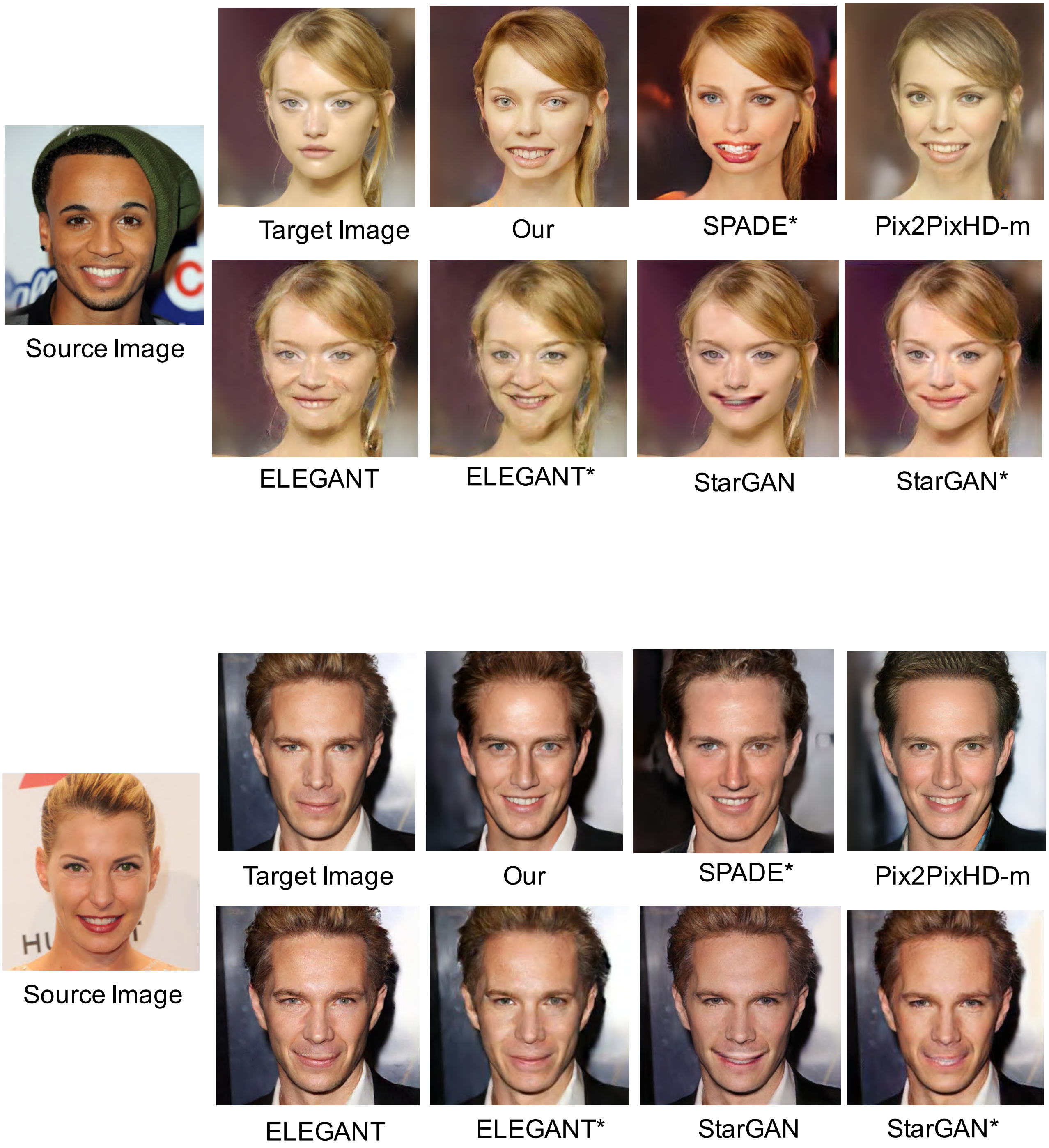

Figure 6: Visual results of attribute transfer for a specific attribute: **Smiling**. * means the model is trained with a size of 256 × 256.

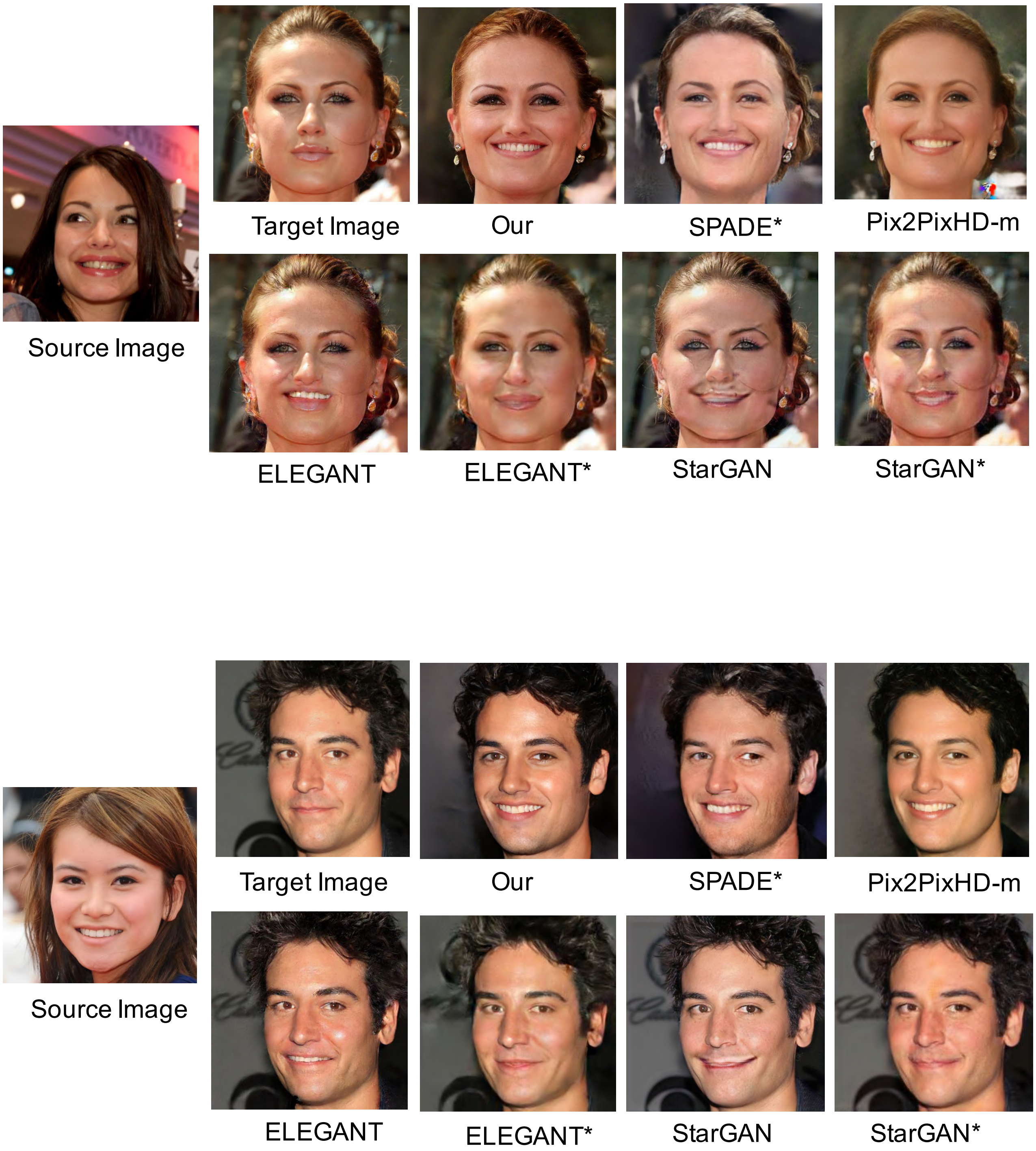

Figure 7: Visual results of attribute transfer for a specific attribute: **Smiling**. * means the model is trained with a size of 256 × 256.

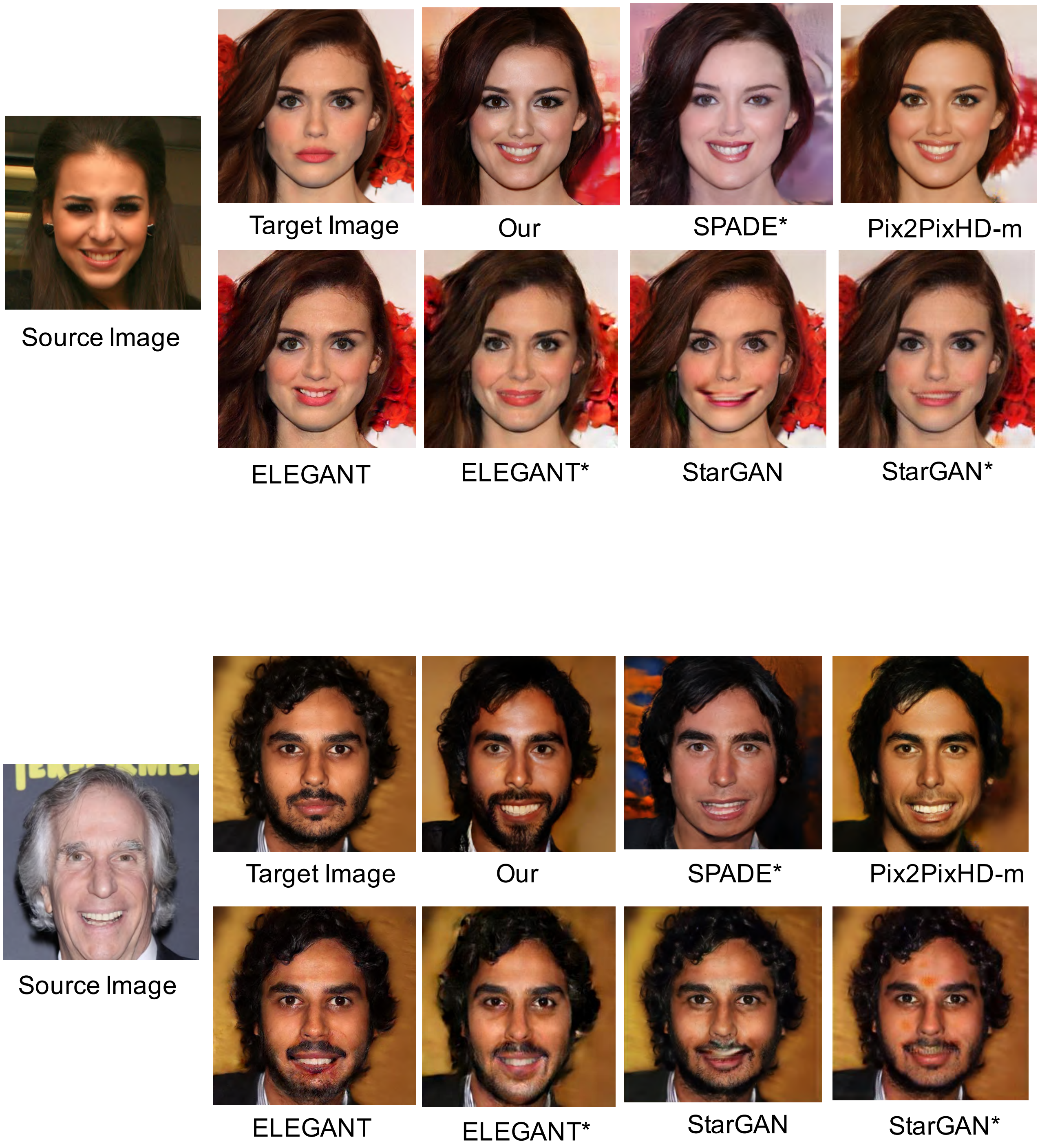

Figure 8: Visual results of attribute transfer for a specific attribute: **Smiling**. * means the model is trained with a size of 256 × 256.

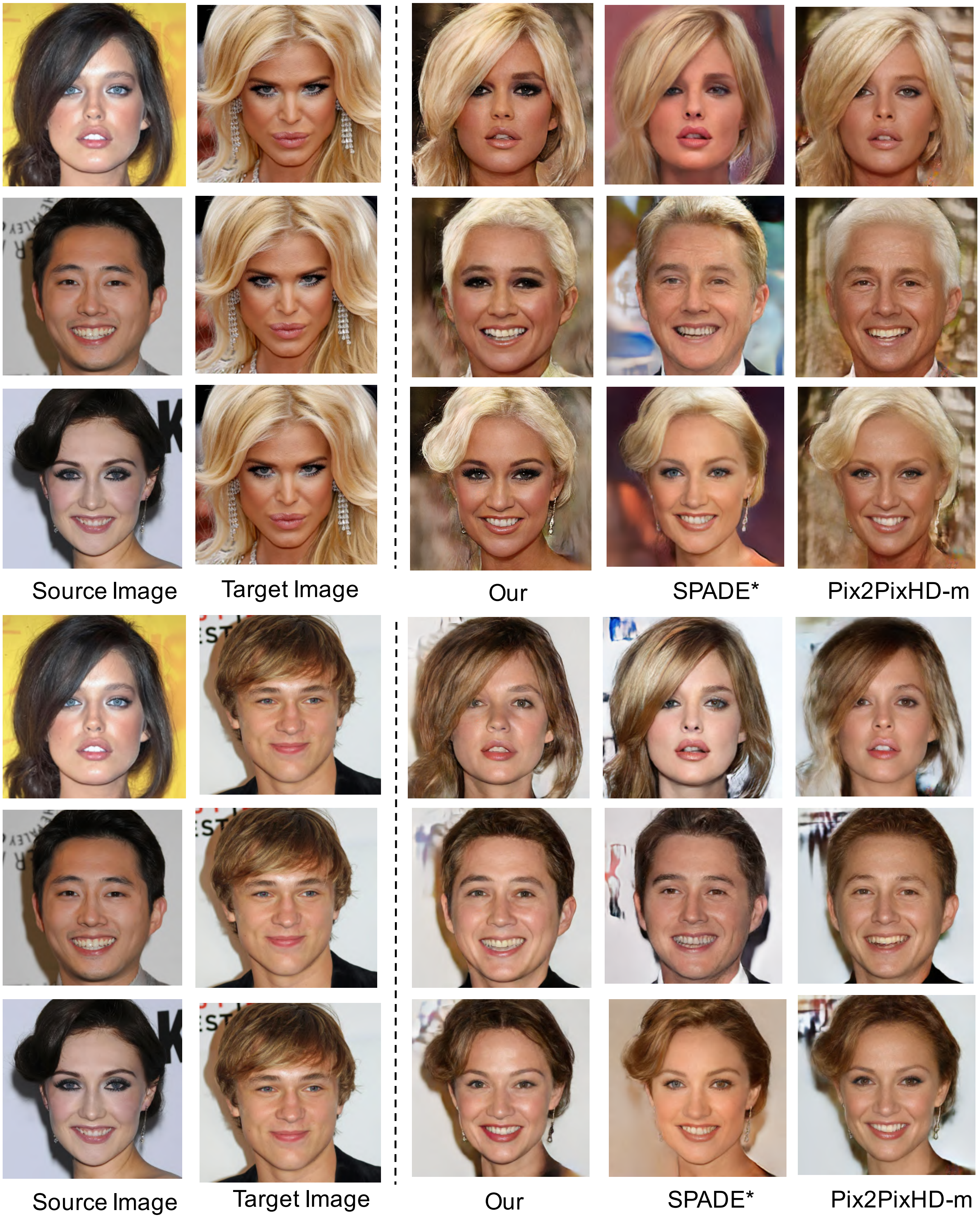

Figure 9: Visual results of style copy.

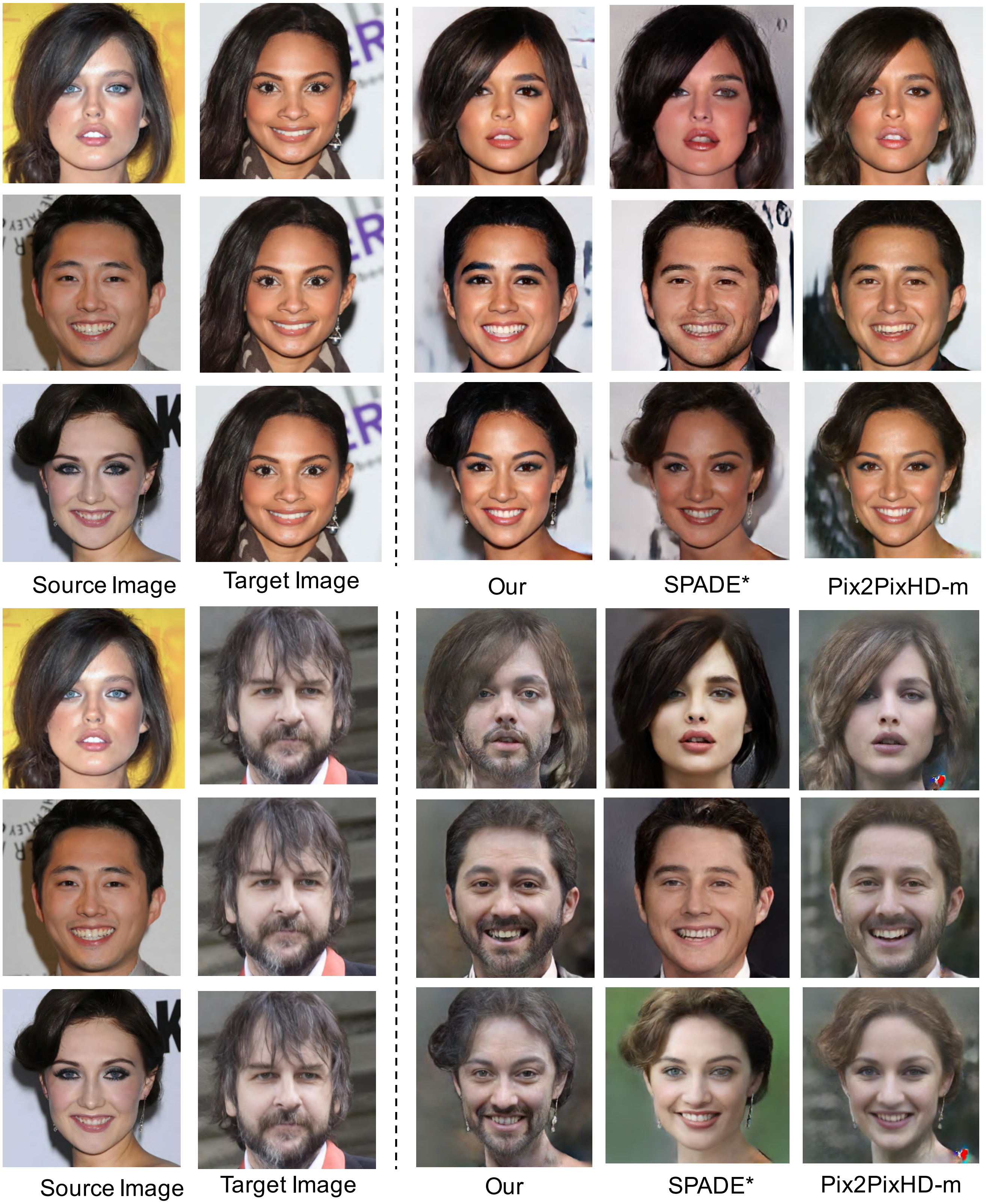

Figure 10: Visual results of style copy.

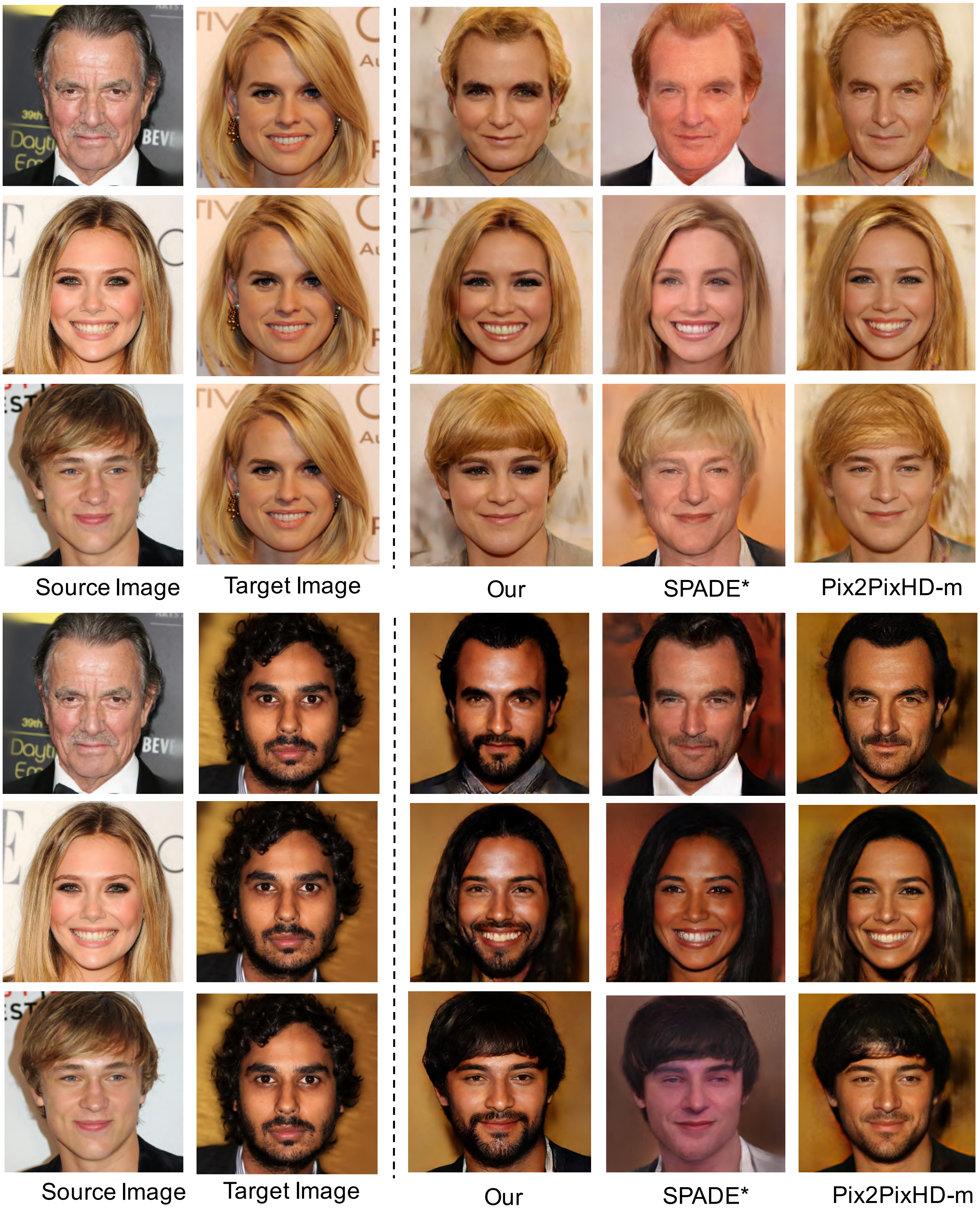

Figure 11: Visual results of style copy.

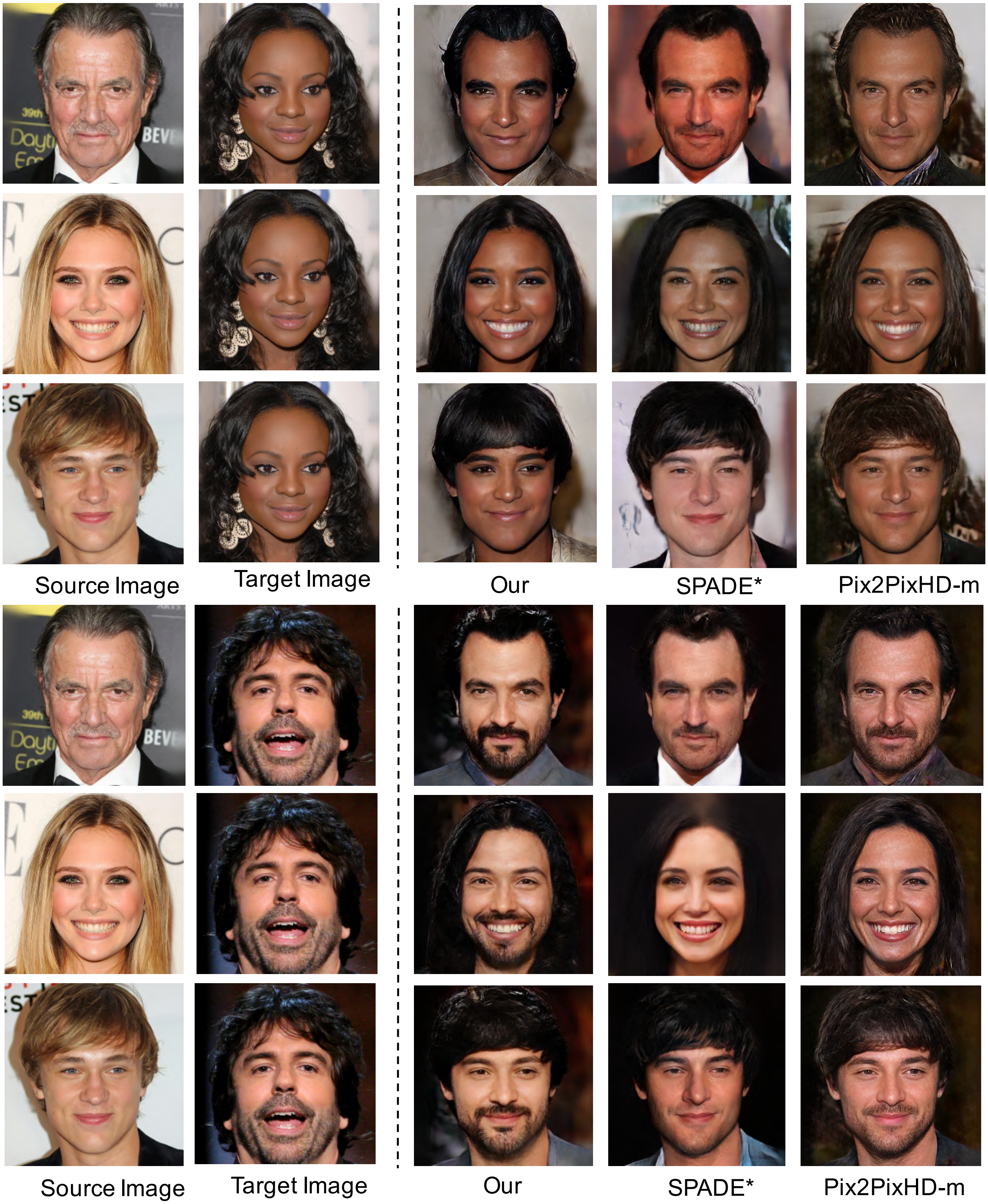

Figure 12: Visual results of style copy.