6. Appendix

In this section, we supplement our paper by reporting additional information. First of all, we describe the implementation details of our networks in subsection 6.1. We then provide a discussion on the effects of the number of groups with qualitative results in subsection 6.2. Third, we qualitatively and quantitatively compare our model with the baseline models on CelebA dataset in subsection 6.3. Finally, we report extra results on CelebA dataset in subsection 6.4.

6.1. Implementation

Content encoder. The content encoders $\{E^c_{\mathcal{A}}, E^c_{\mathcal{B}}\}$ are composed of a few strided convolutional (conv) layers and four residual blocks. The size of the output activation map is in $\mathcal{R}^{256\times\frac{H}{4}\times\frac{W}{4}}$. Note that we use the instance normalization [4] along with the entire layers in E_c in order to flatten the content feature [2, 1].

Style encoder. The style encoders $\{E^s_{\mathcal{A}}, E^s_{\mathcal{B}}\}$ consist of several strided conv layers with the output size in $\mathcal{R}^{256\times\frac{H}{16}\times\frac{W}{16}}$. After the global average pooling, the style feature s is forwarded into the MLP_{CT} and MLP_{μ}. We use the group normalization [6] in E_s to match the structure of s with MLP_{CT} by grouping the highly correlated channels in advance.

Multi layer perceptron. Each of $\{\text{MLP}^{\text{CT}}_{\mathcal{A}}, \text{MLP}^{\text{CT}}_{\mathcal{B}}\}$ and $\{\text{MLP}^{\mu}_{\mathcal{A}}, \text{MLP}^{\mu}_{\mathcal{B}}\}$ is composed of several linear layers. The input dimension of MLP_{CT} depends on the number of group. Specifically, the partial style feature in $\mathcal{R}^{\frac{C}{G}}$ is forwarded as the input feature and the output size is the square of the input dimension. On the other hand, both of the input and output dimension of MLP_{μ} is thesame with the number of channels, 256.

Generator. The generators $\{G_{\mathcal{A}}, G_{\mathcal{B}}\}$ are made of four residual blocks and several sequence of upsampling layer with strided conv layer. Note that GDWCT is applied in the process of forwaring G.

Discriminator. The discriminators $\{D_{\mathcal{A}}, D_{\mathcal{B}}\}$ are in the form of multi-scale discriminators [5]. The size of the output activations are in $\mathcal{R}^{4\times4}, \mathcal{R}^{8\times8} and \mathcal{R}^{16\times16}$.

Training details. We use the Adam optimizer [3] with $\beta_1 = 0.5, \beta_2 = 0.999$ with a learning rate of 0.0001 for all generators and discriminators. Other settings are chosen differently based on the experimented dataset. In CelebA, we apply a batch size of eight with the image size of (216×216). The original image size (178×218) is resized to (216×264.5), followed by the center-crop to be (216×216)). Corresponding models are trained for 500,000 iterations with a decaying rate of 0.5 applied from the 100,000th iteration in every 50,000 iterations. In all other datasets, We train the model with the batch size of two and the image size of (256×256) (we first resize each image up to 286, then perform a random cropping). We set 200,000 iterations for the training and apply the decaying rate of 0.5 from 100,000th iterations in every 10,000 iterations. All the experiments are trained using a single NVIDIA TITAN Xp GPU for three days with the group size of eight.

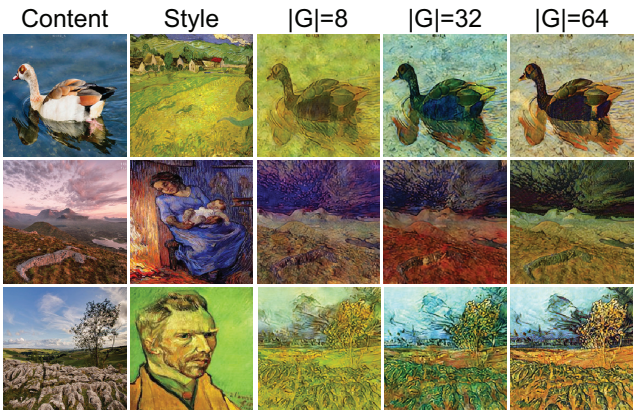

Figure 1. Effects of the number of groups.

6.2. Effects of Different Number of Groups

We discuss and conduct additional experiments on the effects of different group sizes.

First of all, the number of groups, $|G|$, is closely related to the number of model parameters to represent the style statistics of a given exemplar. Specifically, the number of model parameters is equivalent to $C^2/|G|$, where C and $|G|$ represent the numbers of channels and groups, respectively, as discussed in Section 3.3. Thus, increasing $|G|$ has the effect of reducing the model size, i.e., the number of parameters.

We also conduct a qualitative experiment to show the effects of $|G|$ on the final output. As shown in Fig. 1, a small value of $|G|$ tends to focus on the low-level information of the style. For example, those results with $|G| = 8$ in the third column mainly reflect the colors of the exemplar, while those results with $|G| = 64$ in the rightmost column do not necessarily maintain the color tone. We additionally observe that the style becomes more distinguishable across different objects in a resulting image as $|G|$ increases, such that the color of the duck in the first row becomes more distinct from the background as $|G|$ gets larger. We believe it is ascribed to the fact that larger $|G|$ shows the better capability in capturing contrast between objects in the image.

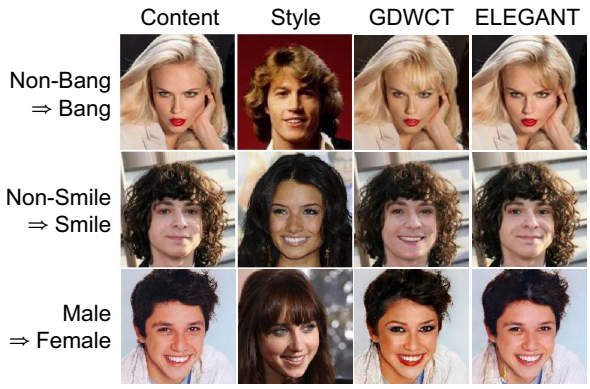

Figure 2. Qualitative comparison on attribute translation. Tested with image size of 216×216.

Although we attempted to rigorously figure out the effects of |G| on our method, however, through several experiments, |G| sometimes shows inconsistent patterns, so that the generalization of the effects of |G| is vague and difficult. Thus, as a future work, it is required to explore in-depth the influences the number of groups gives rise to.

6.3. Additional Comparison Results

In order to further validate the performances of our method, we additionally compare our method against ELEGANT [7], a recently proposed approach that focuses on facial attribute translation and exploits the adversarial loss. As shown in Fig. 2, qualitative results show that our method performs better than ELEGANT in terms of intended attribute translation. For instance, in the first row, our method generates more luxuriant bangs than the baseline method when translating from 'Non-Bang' to 'Bang'. Better results are also found in the Smile attribute, which shows the results closer to the given style. The person in the last row is translated to a female of a high quality with regard to the eyes. ELEGANT encodes all target attributes in a disentangled manner in a single latent space and substitutes a particular part of the feature from one image to another. Since ELEGANT neither decomposes a given image into the content and the style nor matches the statistics of the style to that of the content, it shows worse performances in properly reflecting the style image than our proposed model.

Furthermore, we also show the outstanding performance of our method in a quantitative manner, as illustrated in Table 1. In all cases, our model achieves a higher classification accuracy by a large margin.

6.4. Extra Results

Finally, we present the extra results of our model in Fig. 3, 4, 5, each translated attribute is written on the top of the macro column. All of the outputs in those figures are generated by the unseen data. Through the results, we verify a superior performance of our model.

	Gender	Bangs	Smile	Avg.
ELEGANT	77.15	61.73	70.88	69.92
Ours	**92.65**	**76.05**	**92.85**	**87.18**

Table 1. Classification accuracy in percentages.

References

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

[2] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[3] D Kinga and J Ba Adam. A method for stochastic optimization. In *ICLR*, 2015.

[4] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The missing ingredient for fast stylization, 2016.

[5] 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.

[6] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018.

[7] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. ELEGANT: Exchanging latent encodings with gan for transferring multiple face attributes. In *ECCV*, 2018.

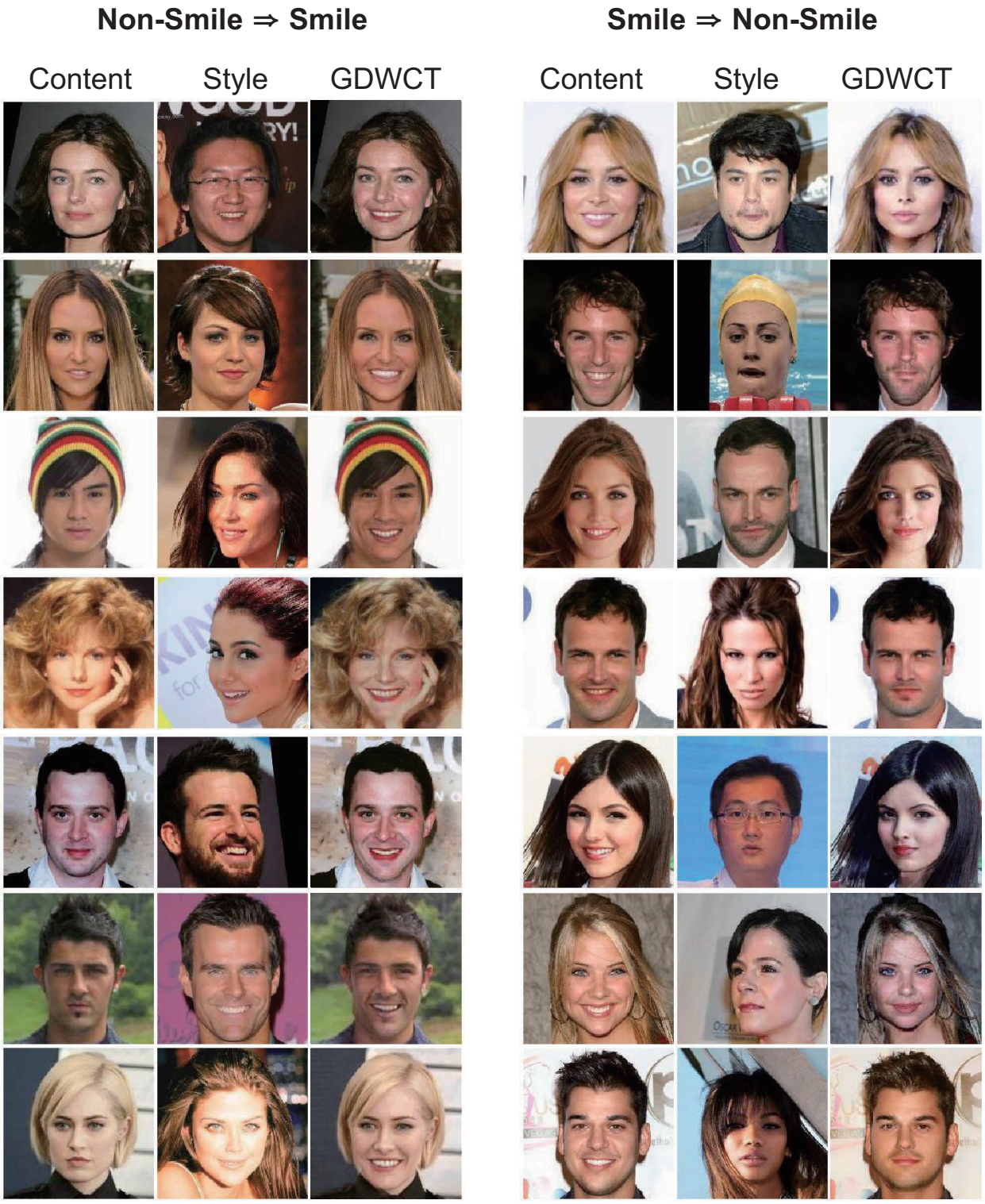

Figure 3. Extra results on CelebA dataset.

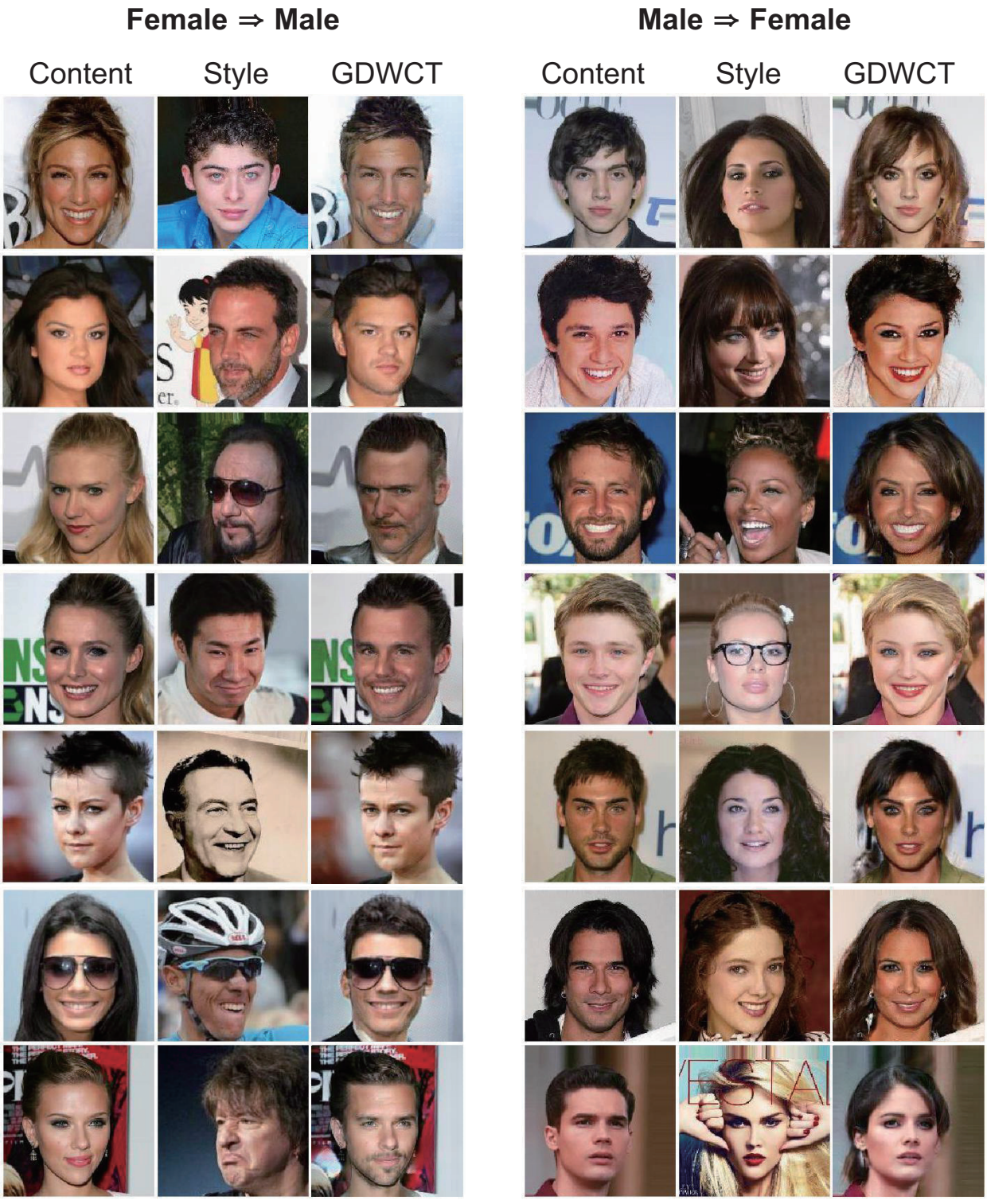

Figure 4. Extra results on CelebA dataset.

Figure 5. Extra results on CelebA dataset.