GRIT: GAN Residuals for Paired Image-to-Image Translation (Supplemental Material)

Contents

1. Implementation details

1.1. Network architecture

Our network architecture resembles a typical multi-modal I2I frameworks (*e.g.*, [7, 11]) with modifications to the decoder part of the generator to accommodate the proposed GAN Residuals. Our network consists of a U-Net [9] generator G and a style encoder E^S. The generator G consists of a content encoder E^C and a decoder with skip connections. Both the content and style encoders $\{E^C, E^S\}$ consist of 6 downsampling blocks, followed by a fully connected layer that generates a 512-dimensional latent code. Each downsampling block is a residual block borrowed from [1], with replacing average-pooling with blur-pooling. We use 64 feature maps at the first encoder layer and double this number after each downsampling block with a maximum of 512 feature maps. The decoder network consists of 6 upsampling blocks to form a U-Net with the content encoder. The architecture of each decoder block is similar to that of StyleGAN [3], with skip connection with the corresponding downsampling block from E^C. The decoder outputs 6 channels: 3 RGB channels for the low-frequency reconstruction component $\hat{I}^B_{\text{rec}}$ and 3 channels for the GAN Residual $\hat{I}^B_{\text{res}}$. Both $\hat{I}^B_{\text{rec}}, \hat{I}^B_{\text{res}}$ are summed to form the final combined output $\hat{I}^B$. During training, we also use a discriminator network whose architecture we adapt from from [4].

1.2. Training details

We train all of our experiments and the baselines on the CelebAMask-HQ dataset [6] for approximately 200 epochs. We follow [2] and use equalized learning rate in all of our networks. We use an Adam optimizer [5] with $\beta_1 = 0, \beta_2 = 0.999$, and a learning rate of 0.001 for all networks. We linearly decay the learning rate by a factor of 10 during the last epoch of training. Our training employs three losses. First, a conditional adversarial loss $\mathcal{L}_{\text{adv}}$ where the conditional input I^A is concatenated to either the real/fake images $I^B/\hat{I}^B$ and fed to a discriminator network. Second, we use a simple L_1 loss as our reconstruction loss $\mathcal{L}_{\text{rec}}$ between our output $\hat{I}^B$ and the ground truth I^B. We use a relative weight $\lambda_{\text{rec}} = 30$ as a relative weight between the two losses. While the value of λ_{rec} was selected to bring the two loss terms to a close value range, we found the training not sensitive to the setting of this hyper-parameter. Finally, we use an L_2 regularization term on the style latent code z^s to encourage a compact latent space. For more implementation details, please refer to our code, which will be released upon acceptance.

2. Standard Deviation of Spatial Noise

In addition to the examples shown in the main paper, we add more examples of the standard deviation computed over diverse generations for each output image. The results are shown in Figure 1. As can be seen in the third row, the highest variations occur in the high-frequency regions corresponding to hair, around eyes, lips, and nose.

3. VGG vs. L1 loss

We analyze the effect of using L1 vs. VGG loss for reconstruction in Table 1. L1 loss is consistent with low-freq reconstruction and suits our decomposition of reconstructed and GAN residual better. As can be seen from the table, L1 is better on the PSNR, SSIM and L1 metrics. This can be attributed to the fact that L1 enforces pixel-wise reconstruction and these metrics mostly focus on that. VGG on the other hand is a perceptual loss like LPIPS so it works better for the LPIPS metric. Also while FID is better for VGG, we visually observe it has more artifacts compared to L1.

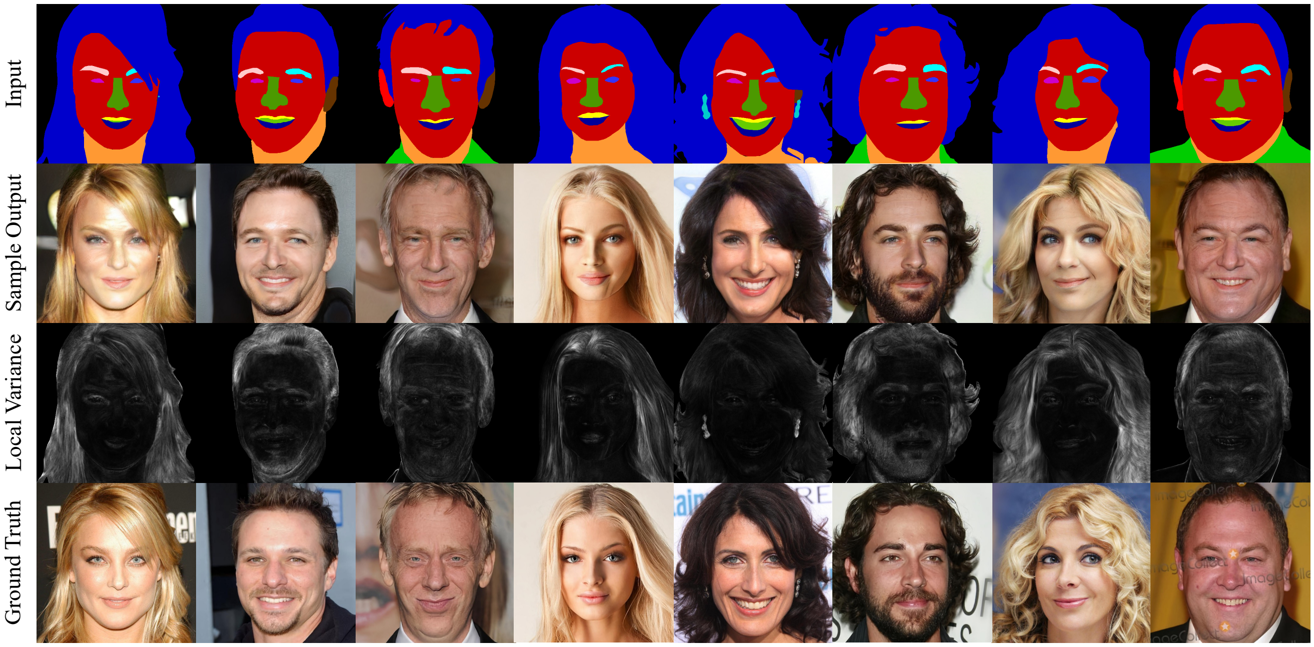

Figure 1. Examples of local stochastic variations. Top to bottom rows represent the input image, one sample output, the standard deviation of each pixel over 20 different outputs of the same input, and the ground truth image respectively.

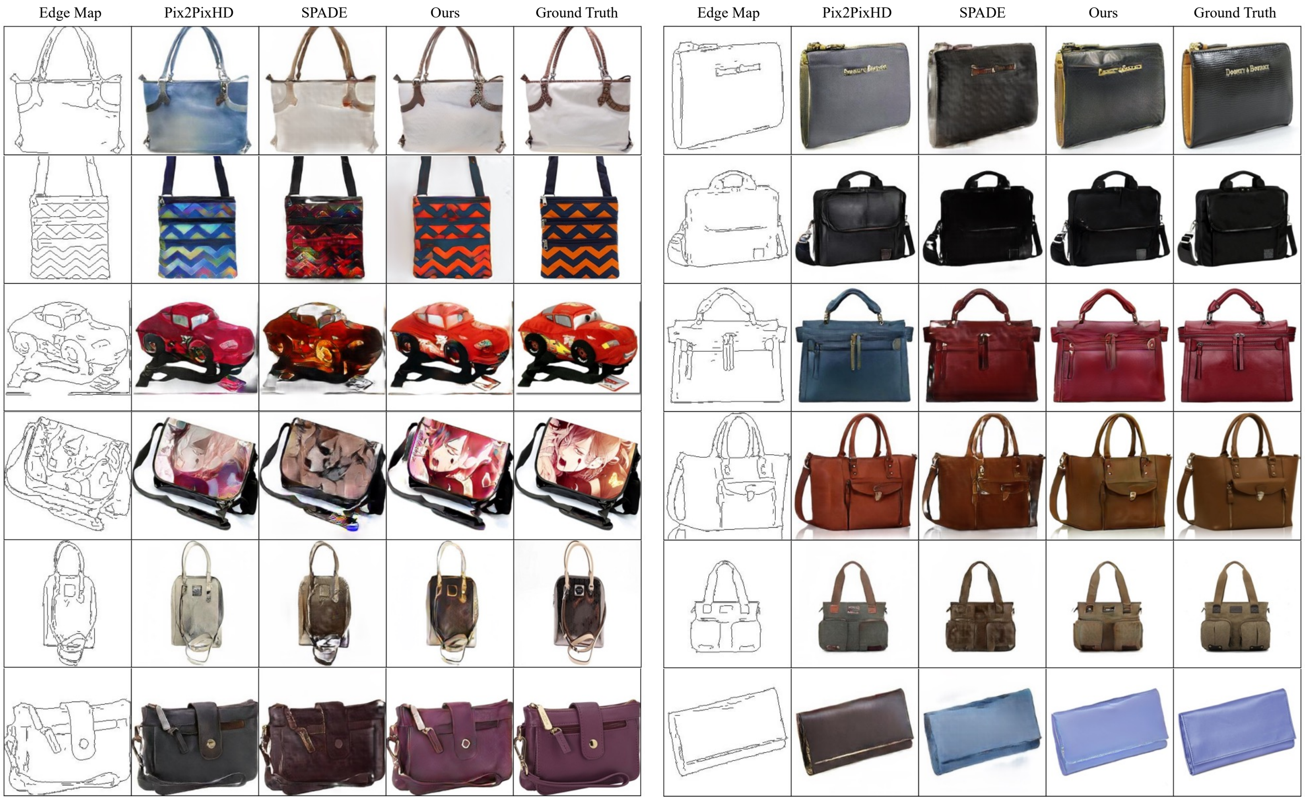

Figure 2. Qualitative comparison with baselines on Edges2Handbags dataset.

Loss	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
L1	18.02	19.85	0.520	0.25	19.82
VGG	22.44	17.96	0.499	0.24	18.89

Table 1. Comparison between using L1 vs. VGG losses for supervising the reconstruction component $\hat{I}^B_{\text{rec}}$.

4. Style transfer

In Figure 3, we show that the network is also capable of performing style transfer. To generate these samples we generate every possible pair of translated images for 10 subjects using the images and corresponding label maps. As shown in the figure, the network is able to use the style from one image and label maps from another to synthesize realistic output in most cases.

5. Qualitative results

In Figure 4 and Figure 5, we show results for qualitative comparison with [8, 10] at a 512×512 resolution on the CelebAHQ-Mask dataset. We chose these two baselines to compare to as they were the best performing baselines at 256×256 resolution. Also in Figure 2 we show comparisons with the same two baselines [8, 10] on the Edges2Handbags dataset which is at 256×256. It can be clearly seen that our method generates the most realistic outputs which better matches the ground truth.

6. GAN Residuals

In Figure 6, we show examples of the outputs generated by our method. As reported in the paper, the reconstructed image encodes most of the structure and content of the image while the GAN residual captures the high frequency details. Combining both of them gives a realistic translation output which is close in appearance to the ground truth.

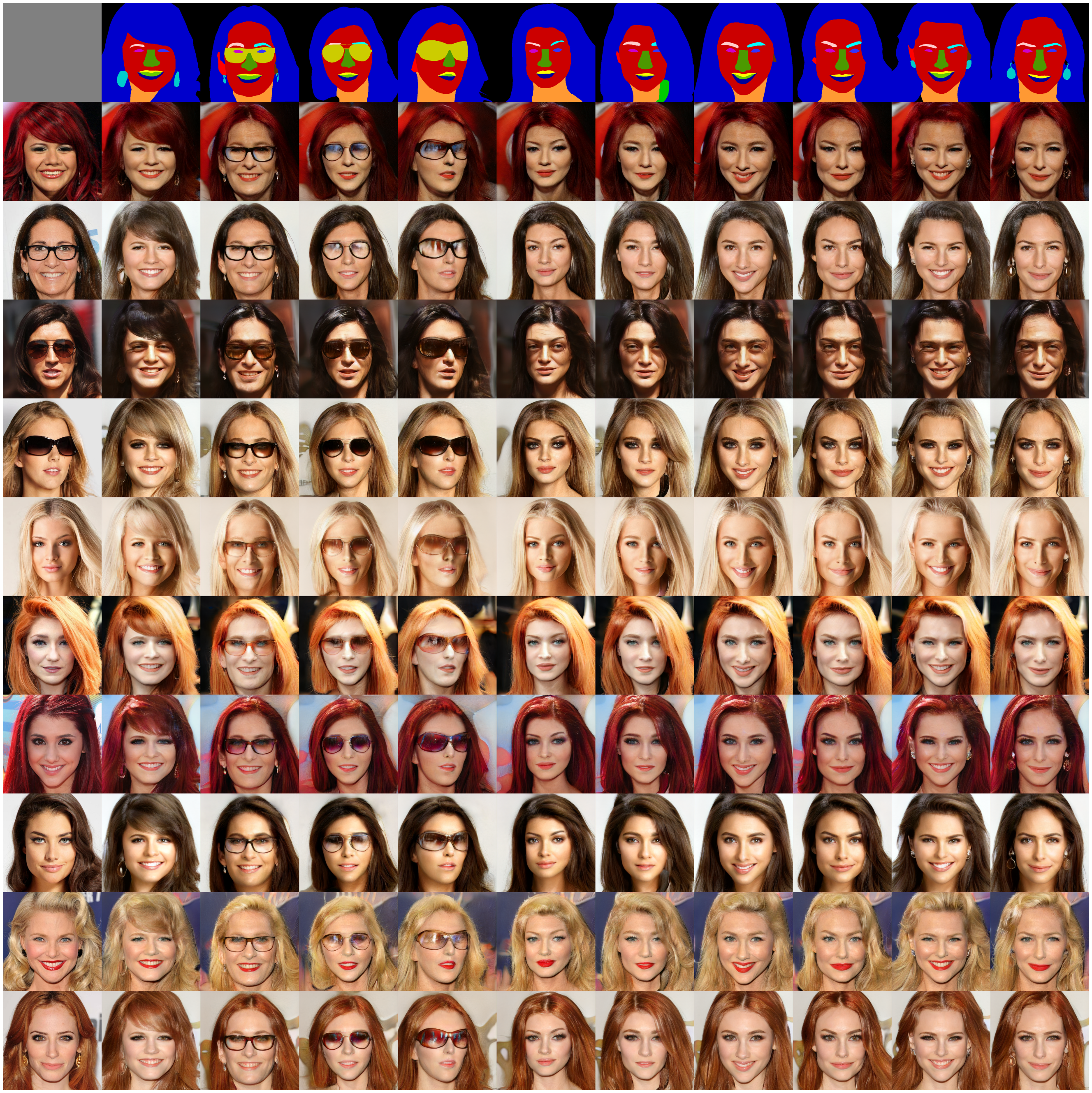

Figure 3. Examples of style transfer by using input label maps and style images from 10 different subjects.

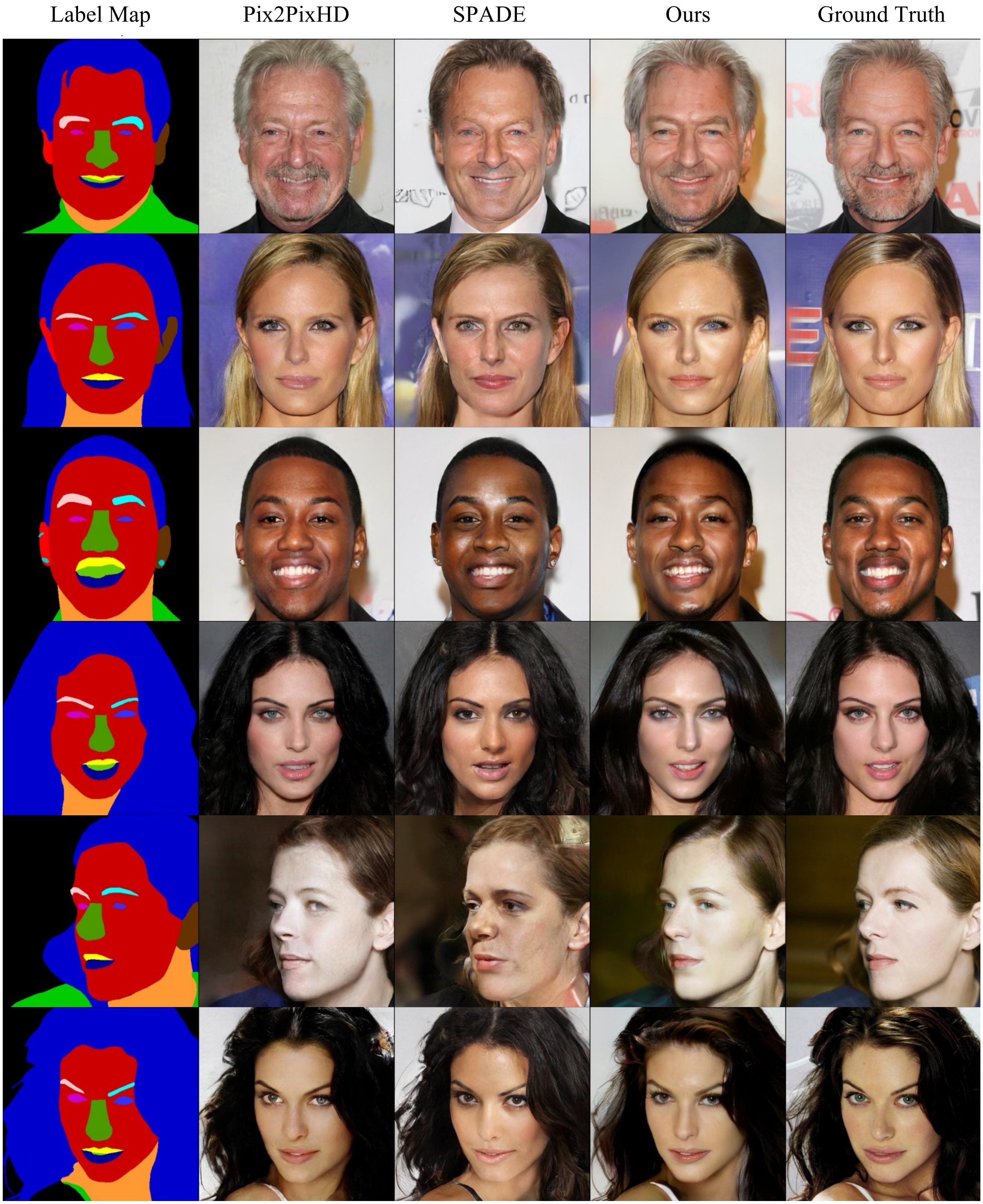

Figure 4. Qualitative comparisons with baselines at 512×512 resolution.

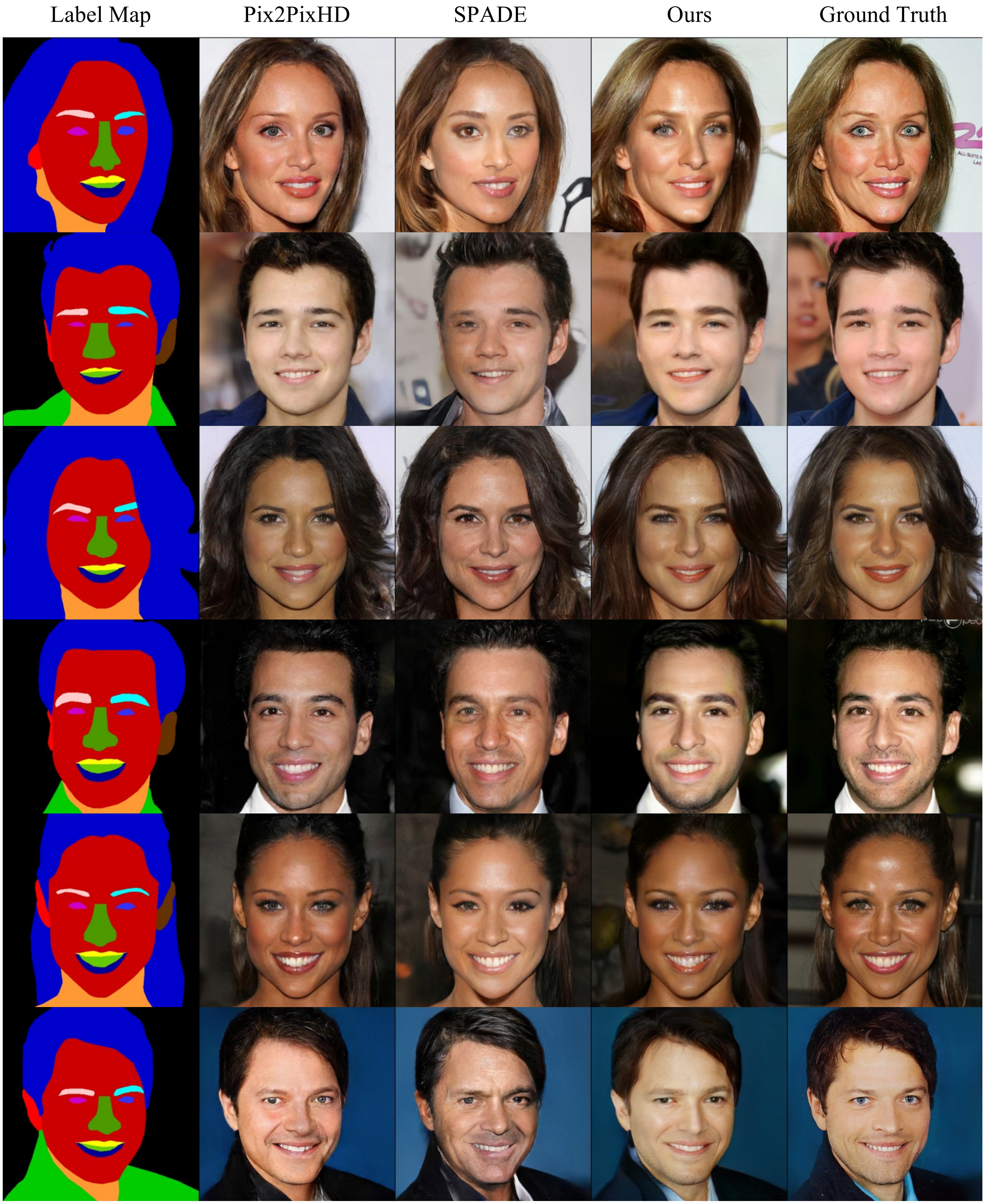

Figure 5. Qualitative comparisons with baselines at 512×512 resolution.

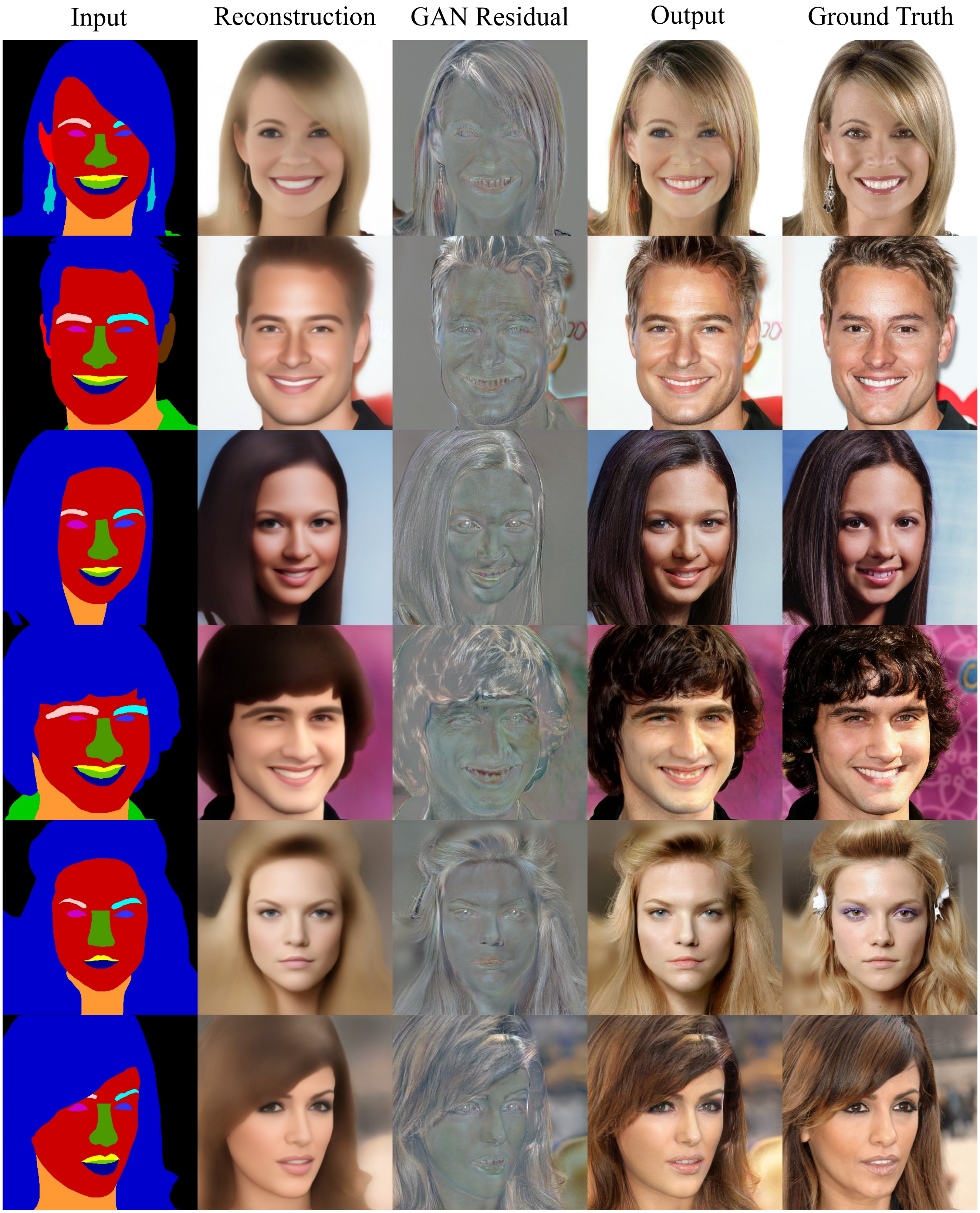

Figure 6. Examples of the different outputs of our method along with the input label map and ground truth image.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *Int. Conf. Learn. Represent.*, 2019.

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Int. Conf. Learn. Represent.*, 2018.

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

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

[6] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] Moustafa Meshry, Yixuan Ren, Larry S Davis, and Abhinav Shrivastava. Step: Style-based encoder pre-training for multi-modal image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3712–3721, 2021.

[8] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[9] 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*, 2015.

[10] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[11] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Adv. Neural Inform. Process. Syst.*, 2017.