FastSwap: A Lightweight One-Stage Framework for Real-Time Face Swapping (Supplementary Material)

Sahng-Min Yoo[1,2], Tae-Min Choi[2], Jae-Woo Choi[2], Jong-Hwan Kim[2]
[1]KLleon AI Research
[2]RIT Lab., KAIST
sahngmin.yoo@klleon.io, {smyoo, tmchoi, jwchoi, johkim}@rit.kaist.ac.kr

1. Appendix

1.1. Training Objectives

1.1.1 Adversarial Loss

We use a multi-scale discriminator [3], and each discriminator has a PatchGAN [1] formulation of the adversarial learning. Specifically, the discriminator outputs a matrix of realism scores in which each element represents the realism score $\mathbf{r}$ of each corresponding patch of the input image.

The objective of the generator (L_{adv}^{G}) in adversarial learning can be written as follows:

$$\frac{1}{H_r W_r} \sum_{h,w} \max\left(0, 1 + \mathbf{r}_{h,w}(G.T.)\right) + \max(0, 1 - \hat{\mathbf{r}}_{h,w}(\hat{Y})) \tag{1}$$

where $r(G.T.)$ and $\hat{r}(\hat{Y})$ are realism scores of ground truth image and generated image, respectively, and $H_r \times W_r$ is a spatial size of a realism score matrix.

The objective of the discriminator (L_{adv}^{D}) in adversarial learning can be written as follows:

$$\frac{1}{H_r W_r} \sum_{h,w} \max\left(0, 1 - \mathbf{r}_{h,w}(G.T.)\right) + \max(0, 1 + \hat{\mathbf{r}}_{h,w}(\hat{Y})) \tag{2}$$

1.2. Additional Experiments

1.2.1 Quantitative Results of Ablation Study

We additionally evaluated the ablation models under the same experimental setting as the main paper. Table 1 shows the quantitative results of the ablation models with metric *ID*, *Pose*, and *FID*. In the first row of the table, sPose and dID refer to the shallow pose network model (*64*64 Pose*) and the deep identity encoder model (*1*1 ID*), respectively.

The quantitative evaluation results are consistent with the tendency expected from the qualitative results (Figure. 9, 10, and 11), suggesting that our proposed TAN block and data augmentation are valuable.

	Ours	w/o I	w/o P	w/o C	w/o D.A.	sPose	dID
ID (↑)	0.54	0.46	0.60	0.63	0.58	0.32	0.64
Pose (↓)	0.61	0.65	0.72	0.82	0.83	0.43	0.85
FID (↓)	60.1	67.4	67.3	68.6	64.5	60.0	61.5

Table 1: Quantitative evaluation results of the ablation models in the main paper. ↑ indicates that the higher the value, the better performance, and the ↓ indicates the opposite.

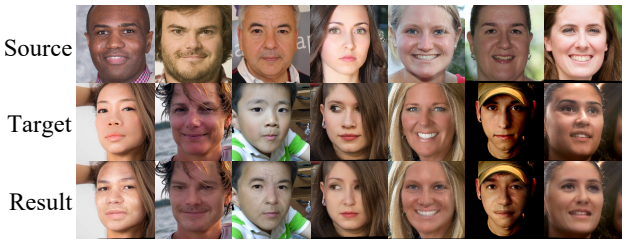

Figure 1: FastSwap results when using FFHQ [2] facial images as inputs.

1.3. Validation on FFHQ Dataset

We validated FastSwap on the other face benchmark FFHQ [2] without any additional training (see Figure 1). The results indicate that FastSwap also generates high-fidelity results in an additional dataset.

For the GPU memory usage, 2.1GB of memory is used, supported by Intel Core i7-7700K CPU when running our FastSwap with batch size 1 input. In consequence, FastSwap is a lightweight framework for real-time face swapping that can be used in a low-spec GPU with a minimum of 2.1GB of memory.

1.3.1 Source Faces

Figure 2 shows ten source faces used for the quantitative comparison experiments in Section 4.2 of the main paper. The faces are chosen to be evenly distributed according to

Figure 2: Ten source faces used for the quantitative comparison (Section 4.2 in the main paper).

gender and race.

1.3.2 Network Structures of Ablation Models

For ablation models in Section 4.5.3 of the main paper, the detailed structures of *1*1 ID* and *64*64 Pose* are depicted in Figures 3 and 4.

1.3.3 Additional Switch-Test Strategy Results

Additional results for switch-test strategy is shown in Figure 5.

References

[1] 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.

[2] 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.

[3] 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.

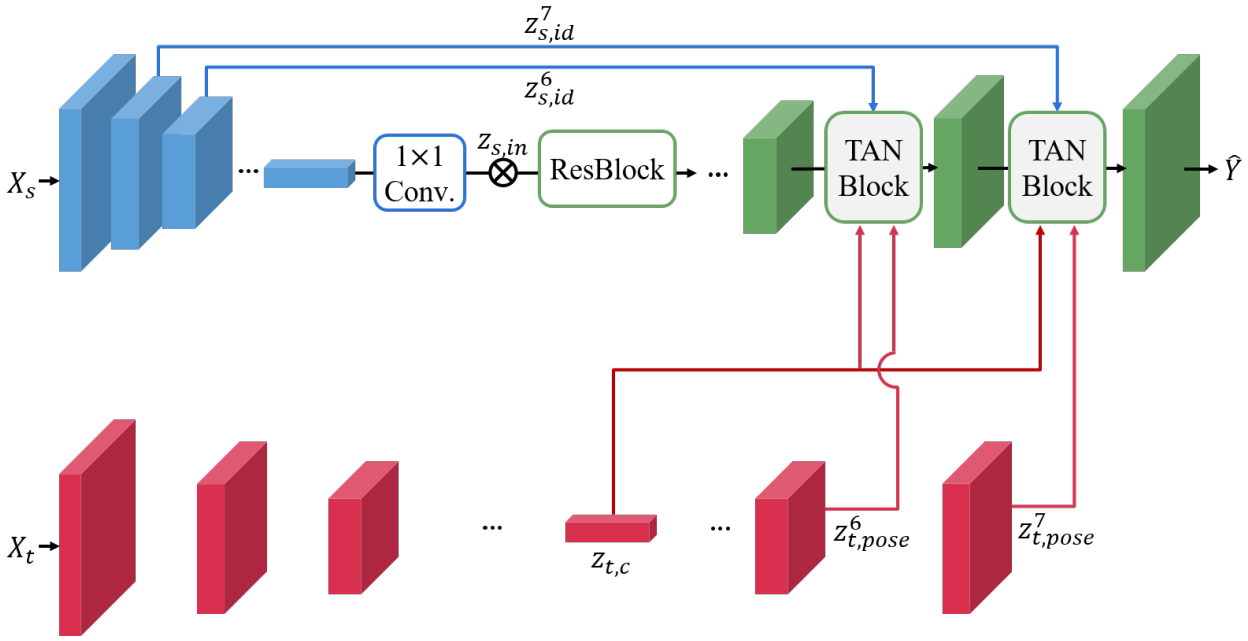

Figure 3: Network structure of *1*1 ID* in Section 4.5.3 of the main paper.

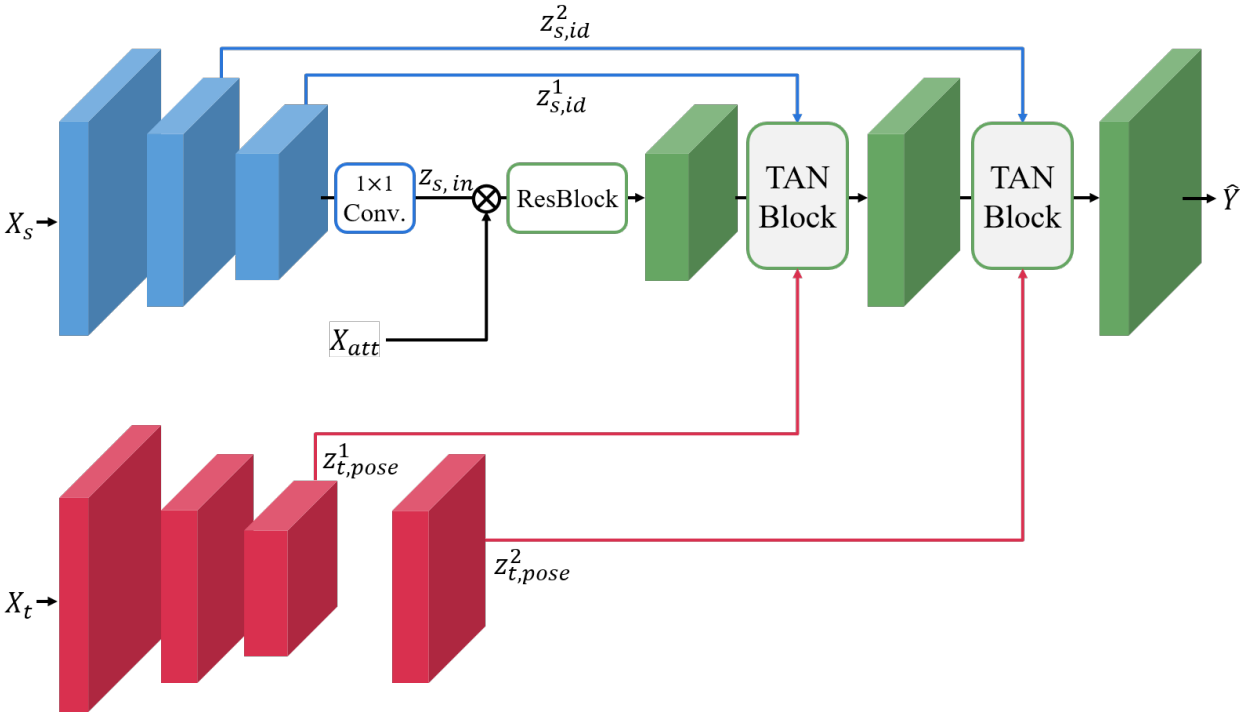

Figure 4: Network structure of *64*64 Pose* in Section 4.5.3 of the main paper.

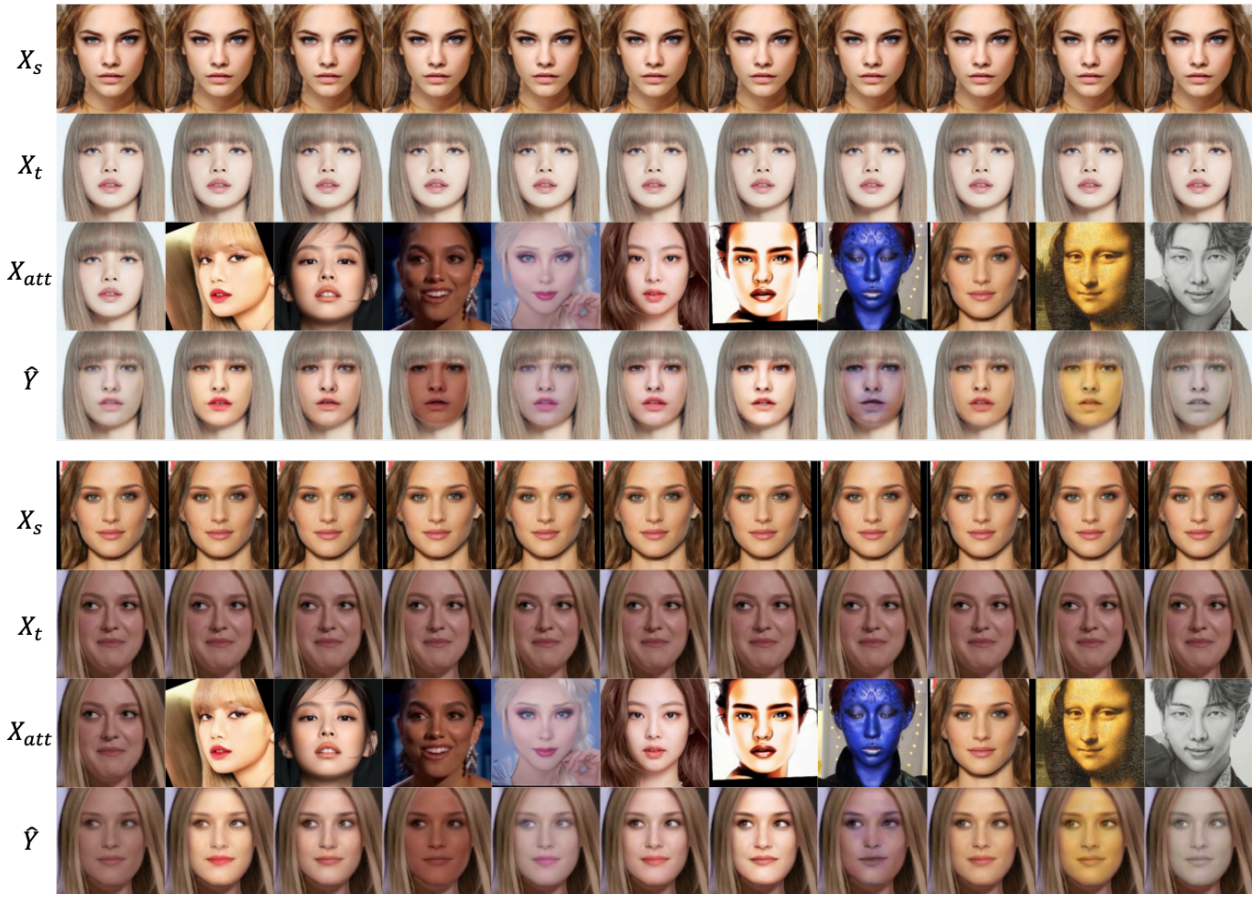

Figure 5: Controllable attribute editing examples by using a switch-test strategy.