ToonerGAN: Reinforcing GANs for Obfuscating Automated Facial Indexing

Supplementary Material

1. Implementation Details for Reproducibility

We discuss the architectural details of different components of the proposed approach below:

- **Identity Encoder ($G(.;\theta_e)$):** Identity encoder consists of one convolution block and three downsampling blocks, progressively reducing the resolution from 256x256 to 128x128, 64x64, and finally to 32x32.
- **Style Decoder ($G(.;\phi_d^s)$):** Four upsampling blocks of style decoder are designed to incorporate inputs from identity encoder downsampling blocks via skip-connections. It takes bottleneck input of size 512x16x16 and generates a toon image of 3x256x256.
- **Anonymization Decoder ($G(.;\phi_d^a)$):** For this, we employ the last six blocks of StyleGAN2 weights $G(.;\phi_d^a)$, starting from the bottleneck output of size 512x16x16. We then feed this output through a series of upsampling blocks, progressively increasing the resolution from the lowest 16x16 to the highest 1024x1024. To expedite the training process, we begin with pre-trained StyleGAN2 weights on the FFHQ dataset, as it is easier to transition from the face distribution to the toon distribution rather than starting from a random point.
- **PSP Style Encoder ($\epsilon(.;\xi^s)$):** A PSP-encoder pre-trained on the face-images is utilized to distill the style features into the mapping network of the StyleGAN2.
- **Face Parsing**: To help the network better identify facial attributes, we compute the face parsing map using BiSeNet [2] and embed them with the real face image as input in Steps 1 and 3.
- Additionally, skip-connections are established between all the blocks of the encoder to both decoders: Identity encoder $G(.;\theta_e)$ to style decoder $D(.;\phi_d^s)$ and identity decoder $D(.;\phi_d^a)$.

2. Limitations

We notice the following limitations of the proposed algorithm and propose their corresponding fixes:

- In some cases, for the face images with sunglasses, the model generates eyes on top of the sunglasses.
- The proposed algorithm is not completely robust against facial occlusion.

The above-mentioned issues can be visualized in Figure 1. We hypothesize that the identified issues occur due to the lack of face images featuring closed eyes and facial occlusions in the FFHQ dataset. These challenges could be effectively resolved by incorporating these types of images into the training set.

Figure 1. Sample images to illustrate some failure cases. The real face images are borrowed from the public FFHQ dataset [1] and blurred for display purposes here to maintain anonymity.

References

[1] 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.

[2] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.