Supplementary for: Diverse Inpainting and Editing with GAN Inversion

Ahmet Burak Yildirim* Hamza Pehlivan* Bahri Batuhan Bilecen Aysegul Dundar
Bilkent University
{a.yildirim, hamza.pehlivan, batuhan.bilecen}@bilkent.edu.tr
adundar@cs.bilkent.edu.tr

In this supplementary document, we provide:

1. Architecture details of the two-stage framework we propose.
2. Visual comparison with pSp, HFGI, HyperStyle, CoModGAN, InvertFill and DualPath.
3. Visual inpainting results on AFHQ Cat and Dog datasets.
4. Semantic editing results on FFHQ dataset.

1. Architecture Details

The final architecture is given in Fig. 1. We follow a two-stage training pipeline. In the first stage, we train the Encoder and Mixing network. The architectures of them are as follows:

Encoder (E). We adopt the encoder architecture from pSp with minor modifications. First, we increase the first layer input channel number from 3 to 4 for taking the mask as an additional input. Then, we disable the normalization layers since we observe they decrease the performance of the model given that many input pixels may be 0 due to removal of them.

Mixing Network (Mi). We equip the mixing network with a neural network and gating mechanism as presented in the main paper. The dimensions of W^{enc} and W^{rand} are both 14×512.

In the second stage training, we set a new encoder which we refer to as Skip Encoder (S) in Fig. 1. In the second stage training, we set the first Encoder frozen. We are interested in learning high-rate features and feed them to StyleGAN generator to achieve better fidelity to input image. The architecture is as follows:

Skip Encoder (S). The Skip Encoder takes input from the final output of the first stage model. We additionally feed the mask and erased input image to the Skip Encoder (S). They are concatenated and are fed to the S. S starts with a convolution layer to increase the channel size from 7 to 32 with a filter size of 3×3 and padding of 1. The $32 \times 256 \times 256$ feature maps are fed into residual blocks. The residual blocks consist of three residual layers. Each residual layer consists of two convolution layers with batch normalization and parametric ReLu activation. Each residual block downsamples the input resolution to half in its first residual layer using max pooling layer. At each block the channel size increases. The Skip Encoder decreases the resolution to 32×32 at the end via 3 residual blocks. The channel size at each block are as follows 48, 64, 96 in the downsampling residual layers, respectively. After we extract the 32, 64, and 128 resolution feature maps, we pass them on 2 more convolution blocks to retrieve skip connection addition (G_{add}) and multiplication (G_{mult}) maps, whose channels are compatible with the StyleGAN at respecting resolution. We do not have an activation function for G_{add}, but we have a sigmoid function for extracting G_{mult}. Lastly, StyleGAN generator features (G_f) are changed as follows:

$$G_f = G_f + G_f * G_{mult} + G_{add} \tag{1}$$

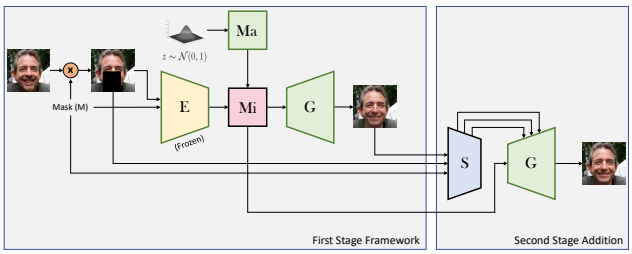

Figure 1. The overall architecture that is used in the second stage framework training. After learning the first stage model that includes E and Mi, we learn skip connections from skip network S to the generator G to achieve high-fidelity reconstructions and seamless transitions across the boundaries of the masks. The same W^{out} from the Mixing Network is used for both stages.

*Joint first authors, contributed equally.

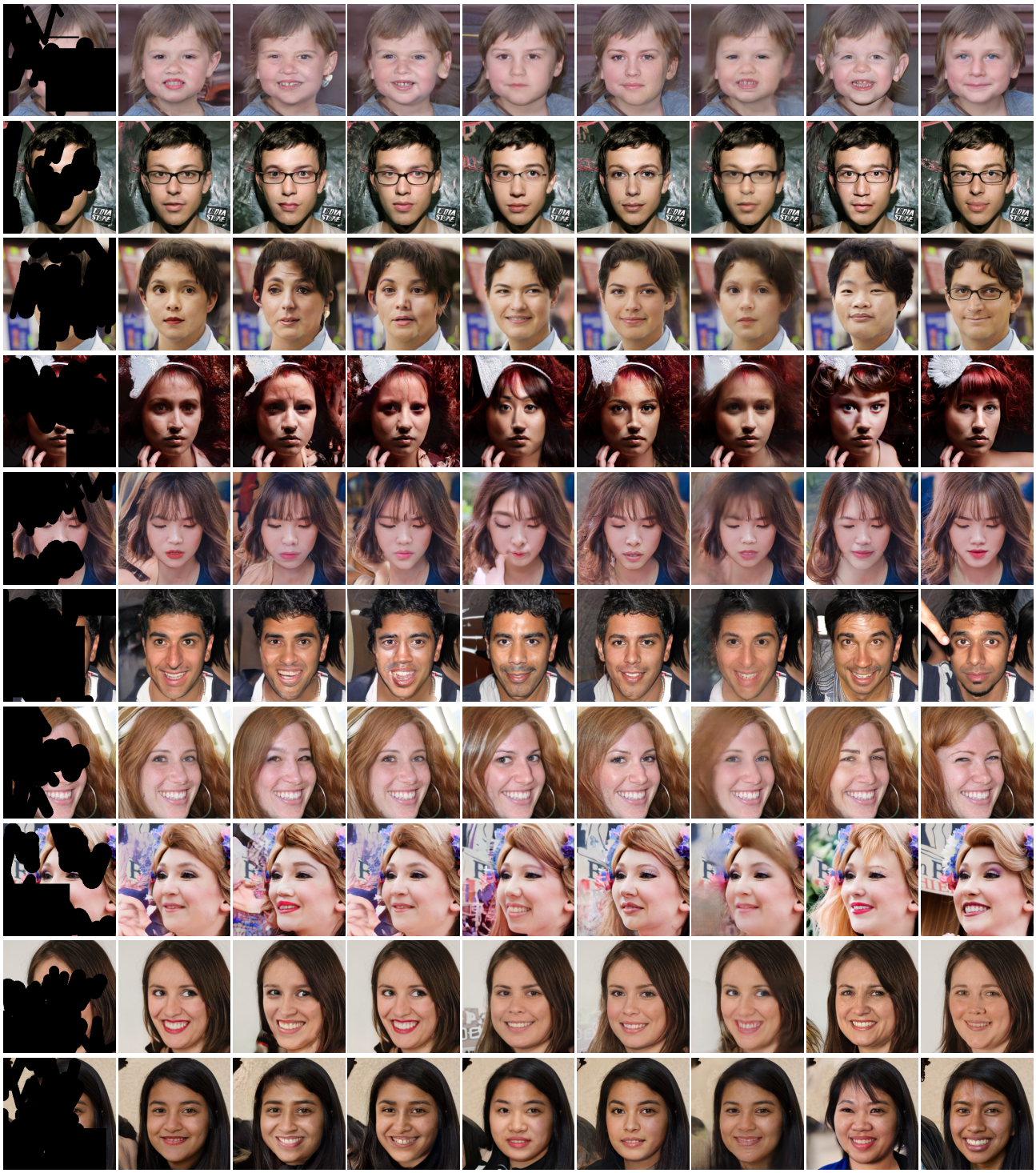

Figure 2. Qualitative results of our and competing methods on the FFHQ dataset. GT refers to original images.

Figure 3. Editing results of our method on the FFHQ dataset. Our method achieves diverse inpainting and editing under one framework.

Figure 4. Qualitative results of AFHQ-Cat and AFHQ-Dog dataset trainings.