Improving Visual Quality of Image Synthesis by A Token-based Generator with Transformers

Anonymous Author(s)
Affiliation
Address
email

A More visual results

To evaluate the model performance on the synthesis of images with high-frequency details and complex scenes, we choose the commonly-used benchmark of unconditional image synthesis, i.e., FFHQ [1] and LSUN CHURCH [3] with different resolutions in the paper. As a supplement of the visual results (Figure 4) in the main paper, we show more synthesis results in Figure 1, 2, 3 and 4 for FFHQ-1024 [1], FFHQ-256 [1], LSUN CHURCH [3], respectively. As shown in the results, the token-based generator can synthesize high-frequency details in human faces with resolution 1024×1024, as well as various styles of complex scenes for the church.

In addition to the results in Figure 6 in the main paper, we show more results of image interpolation in Figure 4. The results show that the token-based generator can synthesize the images of a smooth transition of high-level attributes, e.g., the interpolated results from a front pose to a lateral pose in the first case, faces with and without hats in the third and the fourth cases, faces from a child to an adult in the fifth case, etc.

B Hyperparameters and training details

We build upon a Pytorch implementation of StyleGAN2 [2], which has reported the same performance with the reported results in the paper[1]. We inherit most of the training details from that implementation. In particular, we use the same discriminator architecture, minibatch standard deviation layer at the end of the discriminator, an exponential moving average of the generator, 512 channels for $\mathbf{z}$, eight fully connected layers as the mapping network. We up-sample the feature maps by first reshaping them to 2D feature maps, up-sampling features by bilinear filtering, and finally reshaping the features back to visual tokens. We initialize all the weights of the fully connected and affine transform layers and the constant input in the generator network using $\mathcal{N}(0, 1)$. We initialize the embedding of the keys for the styles to orthogonal matrices.

References

[1] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019.

[2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020.

[3] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[1] https://github.com/rosinality/stylegan2-pytorch

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Figure 1: Images produced by the token-based generator (config E) with the FFHQ-1024 [1] dataset at resolution 1024×1024. As shown in the results, the token-based generator is able to synthesize human faces with various styles, i.e., hairstyle, glasses, gender, poses, facial expressions, etc..

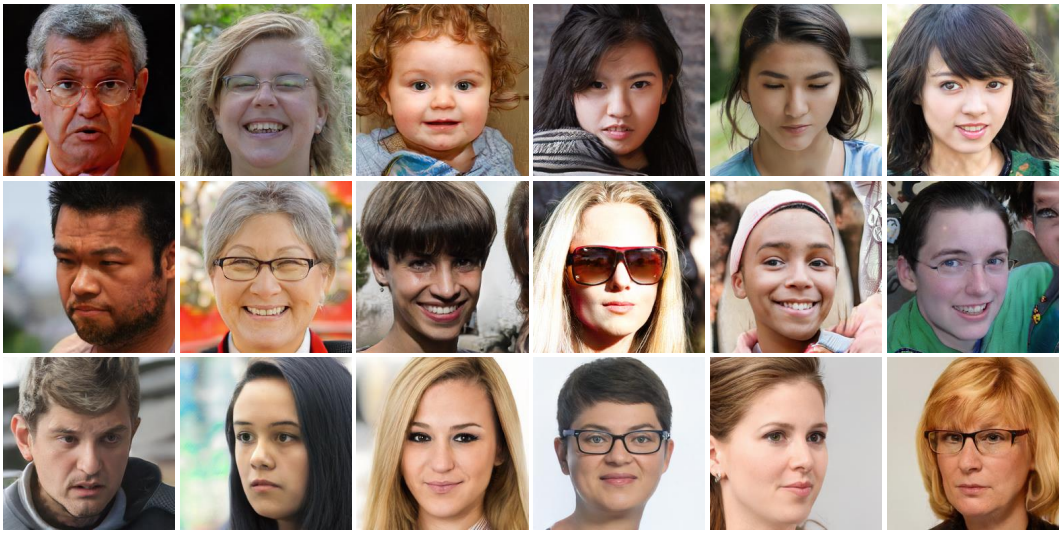

Figure 2: Images produced by the token-based generator (config E) with the FFHQ-256 [1] dataset at resolution 256×256. As shown in the results, the token-based generator is able to synthesize high-frequency details in human faces, i.e., beards, wrinkles, hairs, teeth, etc..

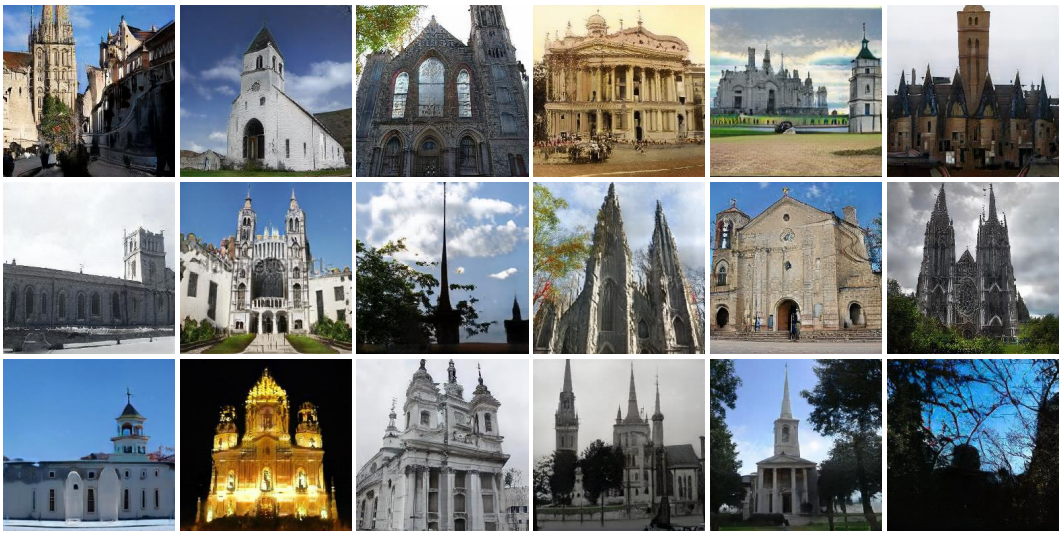

Figure 3: Images produced by the token-based generator (config E) with the LSUN CHURCH [3] dataset at resolution 256×256. As shown in the results, the token-based generator is able to synthesize the complex scenes of the church with various styles. The complex scene is usually composed of many different elements, e.g., different styles of buildings, trees, skies, etc..

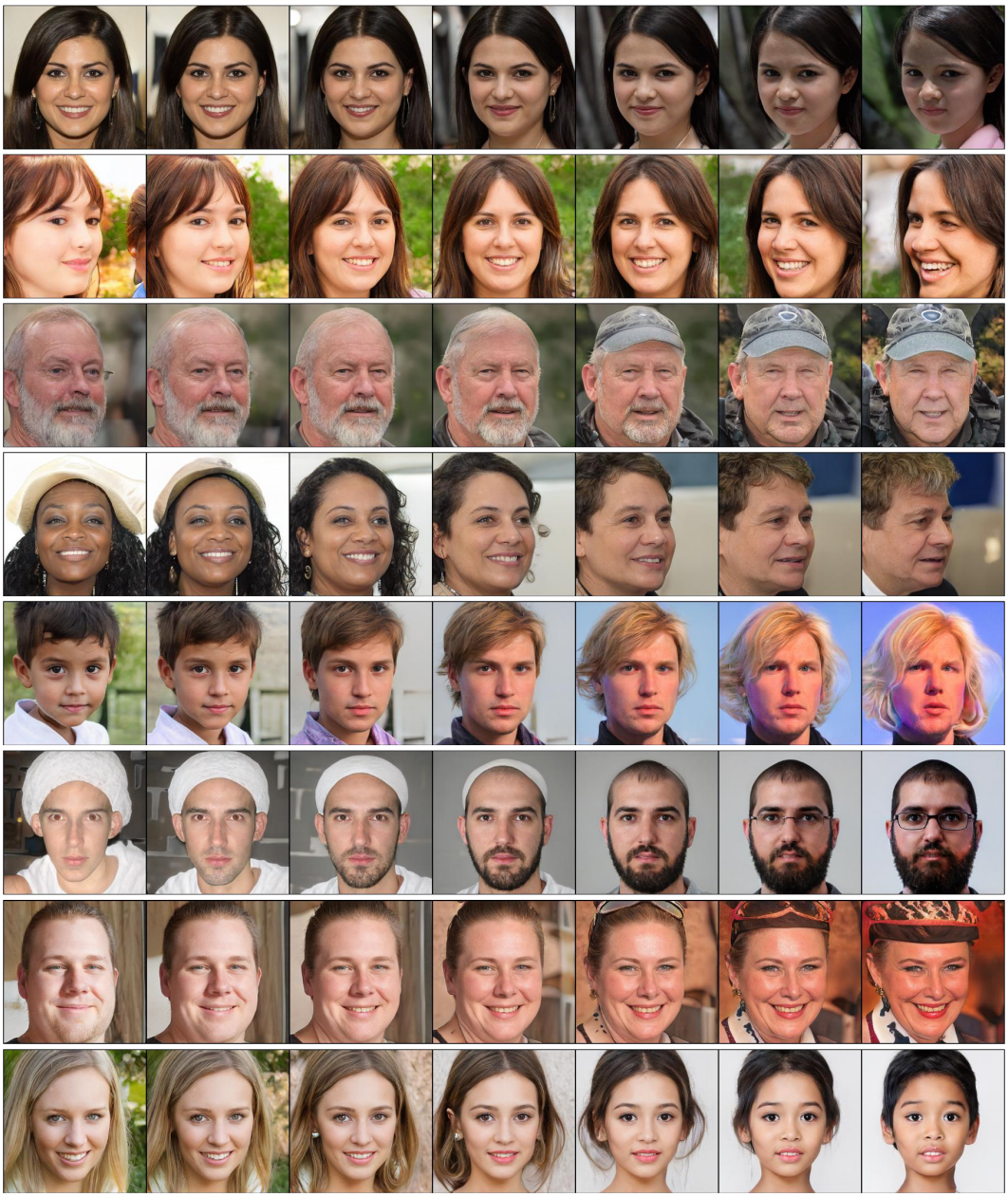

Figure 4: Results of image interpolation produced by the token-based generator. In each row, the first column and the last column are two samples randomly generated by the token-based generator, and the interpolated results of these two samples are shown between them (from 2nd to 6th column). As shown in the results, the token-based generator can synthesize a smooth transition of the high-level attributes, e.g., from children to adults, from long-hair to short-hair, from male to female, etc..