EGC: Image Generation and Classification via a Diffusion Energy-Based Model

Qiushan Guo[1], Chuofan Ma[1], Yi Jiang[2], Zehuan Yuan[2], Yizhou Yu[1], Ping Luo[1,3]
[1]The University of Hong Kong [2]ByteDance Inc. [3]Shanghai AI Laboratory

A. Appendix

A.1. Implementation Details

Algorithm 2 Sampling

```
Sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ do
    Sample noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\boldsymbol{\epsilon} = 0$
    $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t + (1 - \alpha_t)\nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t)) + \sqrt{\beta_t}\boldsymbol{\epsilon}$
end for
return $\mathbf{x}_0$
```

We adopt the UNet architecture used in LDM [2] and IDDPM [1], with the group normalization layers retained. To improve convergence speed, we do not include spectral normalization and weight normalization regularization. We adjust the channel width, multiplier, attention resolution, and depth compared to IDDPM and LDM, as shown in Tab. 5. We use the *conv_resample* instead of the *Res-Down/Up Block* to upsample and downsample features. To optimize our model, we set the learning rate to 0.0001, batch size to 128 and weight decay to 0 across all datasets except ImageNet, for which we use a batch size of 512. The Adam optimizer is used to update the model parameters. To save computational resources, the ImageNet and LSUN Church images are compressed to 32×32×4 latent features by the KL-autoencoder [2], while CelebA-HQ images are compressed to 64×64×3 latent features. To augment the data, we randomly flip the images for all the datasets except CIFAR, for which all the images for training are padded with 4 pixels on each side and a 32 × 32 crop is randomly sampled from the padded image or its horizontal flip, and cutout is used to avoid overfitting. The classification results at $t = 0$ are reported in Table 1-3. To balance the reconstruction and classification losses, we set $\gamma = 0.001$ for CIFAR and $\gamma = 0.005$ for ImageNet in Algorithm. 1.

We follow the sampling strategy used in DDPM and describe it in detail in Algo. 2. To conduct conditional sampling, we replace the unconditional score $\nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t)$ with the conditional score $\nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t|y)$.

	CIFAR	LSUN-Church	CelebA-HQ	ImageNet
Diffusion steps	1000	1000	1000	1000
Noise Schedule	cosine	linear	linear	linear
Channel	192	256	256	384
Depth	3	2	3	2
Channel Multiplier	1,2,2	1,2,2	1,2,3,4	1,2,4
Attention Resolution	16,8	32,16,8	16,8	32,16,8
Iteration	200k	200k	100k	700k
Batch Size	128	128	128	512

Table 5: The hyperparameters of the ECG models producing the results shown in Table 1-3.

B. Additional Samples

Fig. 9 and 10 show additional examples of image interpolation on CelebA-HQ 256^2 and LSUN-Church 256^2 datasets. Fig. 11, 12 and 13 show uncurated samples from the learned models on CIFAR-10, CelebA-HQ 256^2 and LSUN-Church 256^2 datasets.

References

[1] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

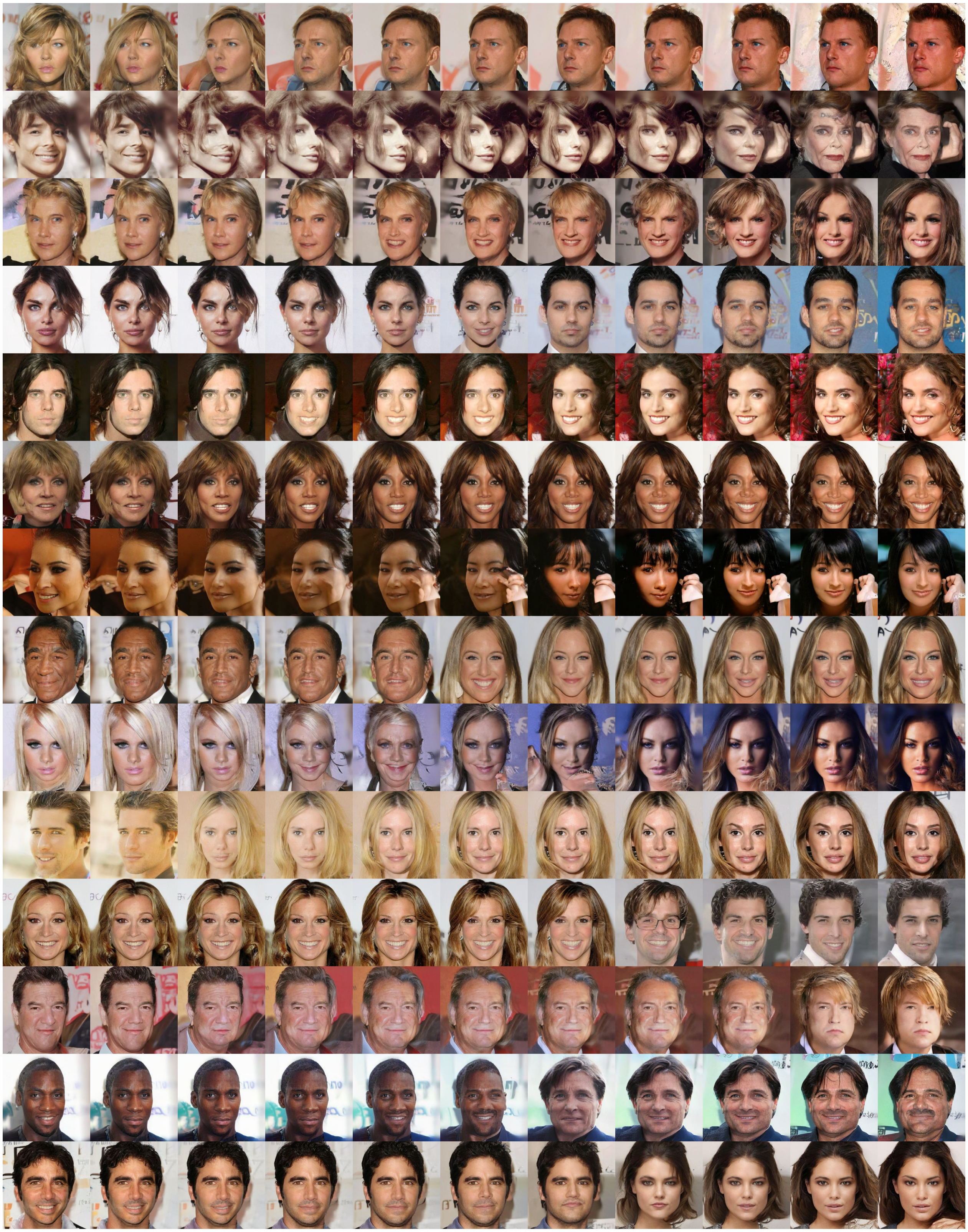

Figure 9: More interpolation results on CelebA-HQ dataset.

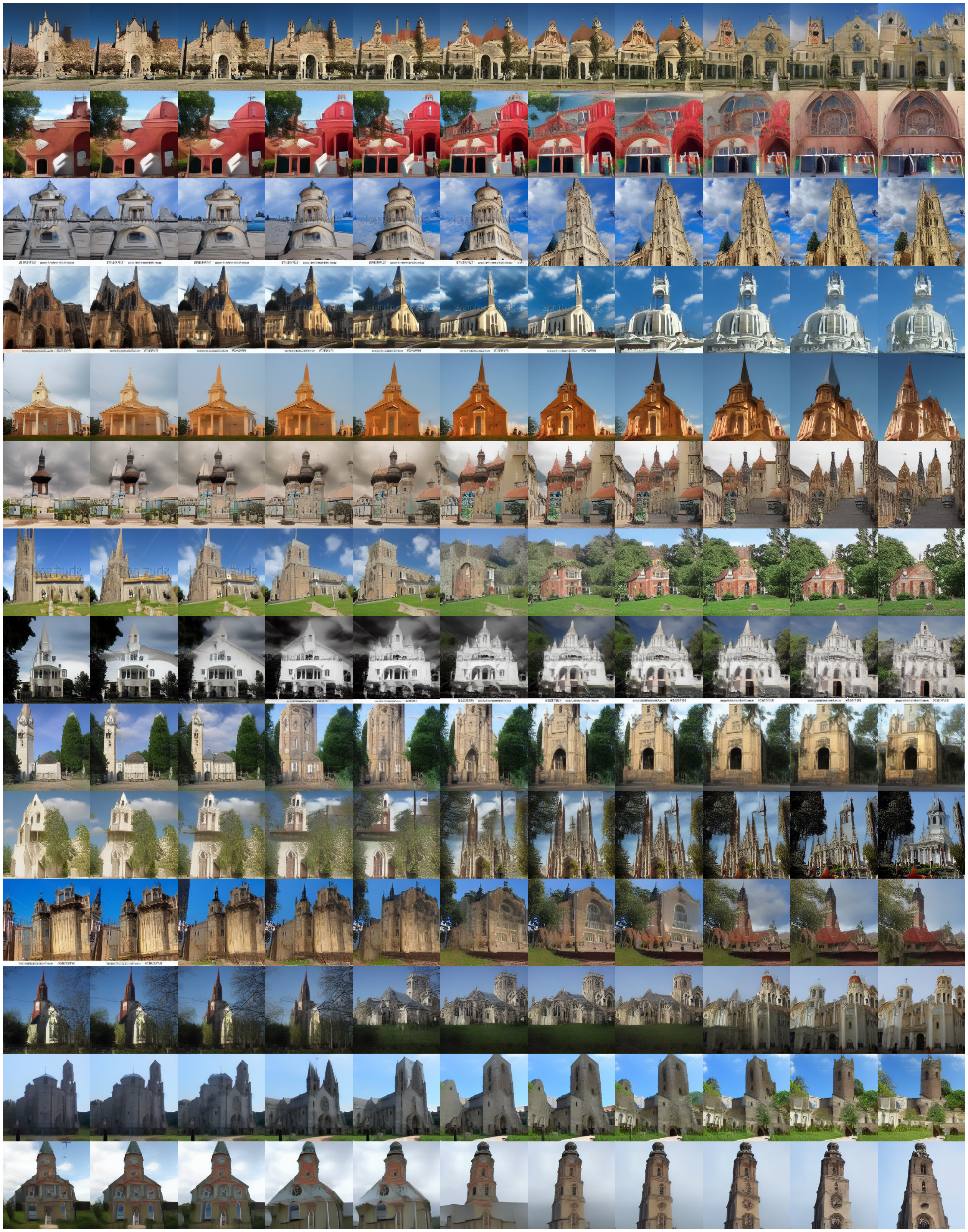

Figure 10: More interpolation results on LSUN-Church dataset.

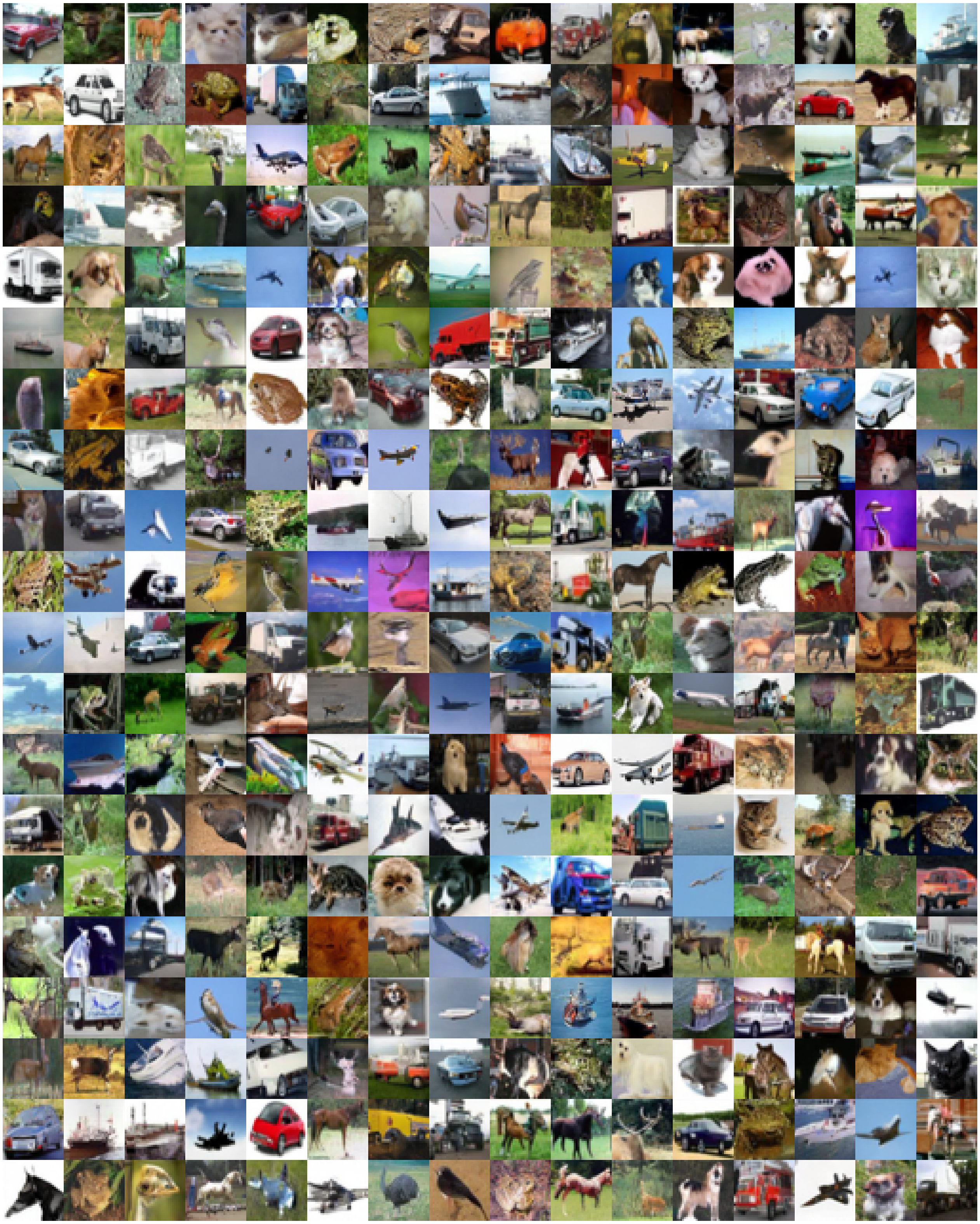

Figure 11: More results on CIFAR-10 dataset. FID=3.30.

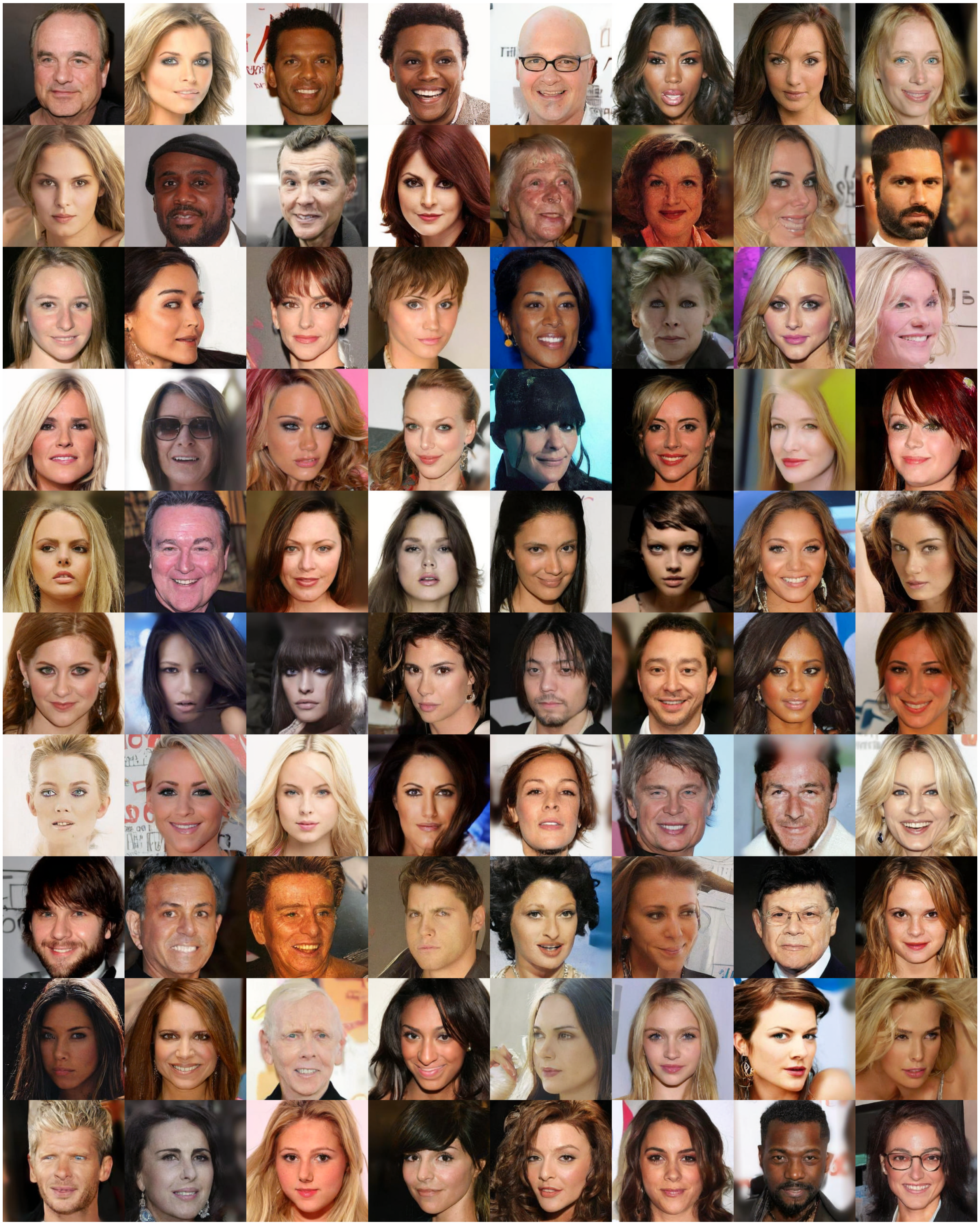

Figure 12: More results on CelebA-HQ 256×256 dataset. FID=7.75.

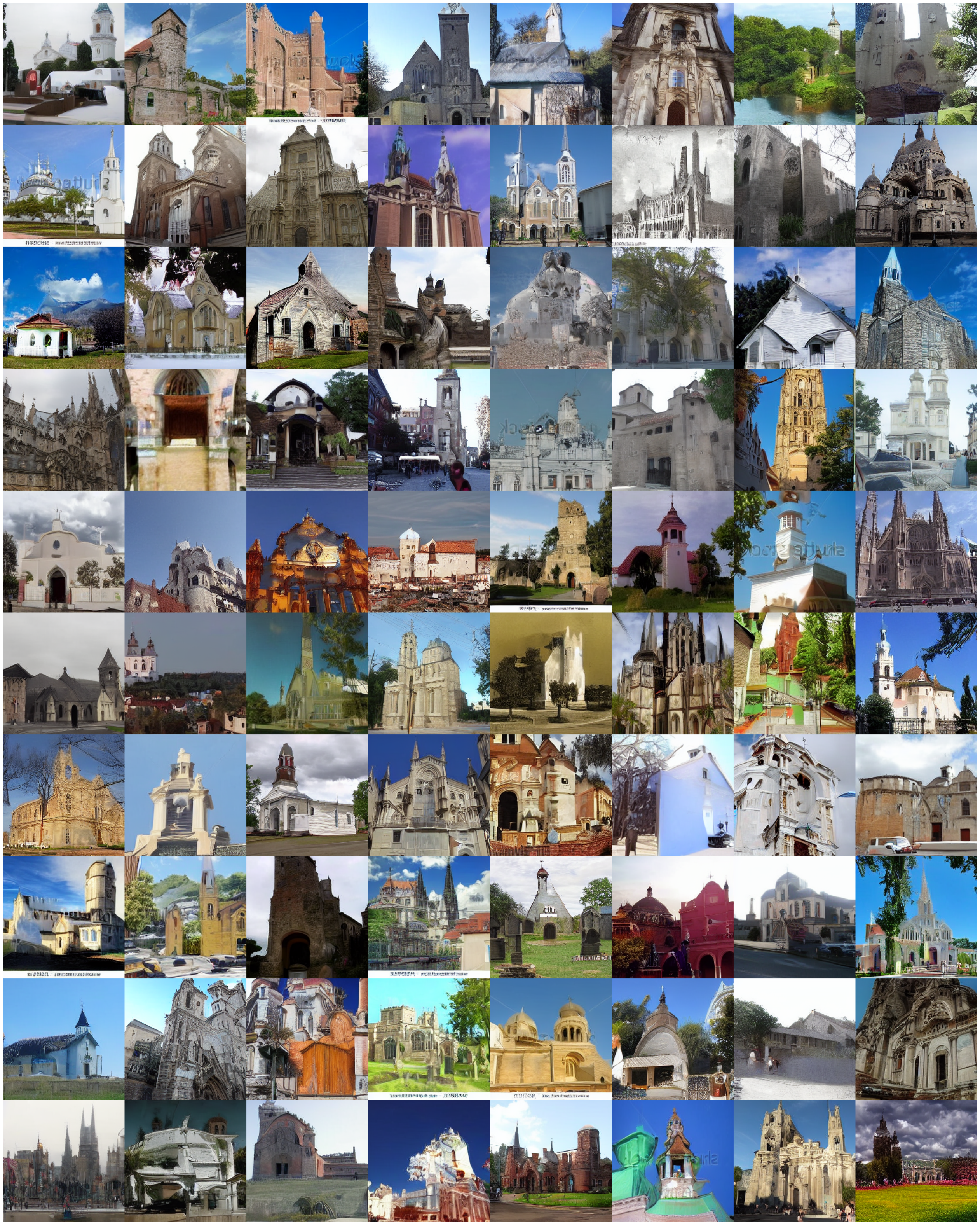

Figure 13: More results on LSUN-Church 256×256 dataset. FID=8.97.