ICLR 2021

# TOWARDS FASTER AND STABILIZED GAN TRAINING FOR HIGH-FIDELITY FEW-SHOT IMAGE SYNTHESIS

Bingchen Liu, Yizhe Zhu, Kunpeng Song, Ahmed Elgammal

Chaerin Kong

# What's New

- 1024 x 1024 Image Generation – Few Hours Training on Single RTX-2080 GPU
- Comparable Performance with SOTA (StyleGAN2)
- Enables GAN Training in low-data regime

Figure 1: **Synthetic results on** $1024^2$ **resolution** of our model, trained from scratch on single RTX 2080-Ti GPU, with only 1000 images. Left: 20 hours on Nature photos; Right: 10 hours on FFHQ.

# Background

Difficulty in GAN Training

- Lack of Training Data (Medical Image, Art Work ⋯ ) → Overfitting, Mode Collapse
- High Computation Cost (StyleGAN, BigGAN ⋯ )

“Light-weight, Robust GAN Model Quickly Trainable on Small(~1000) Dataset”

# Contributions

1. Skip-Layer Channel-wise Excitation (SLE) Module

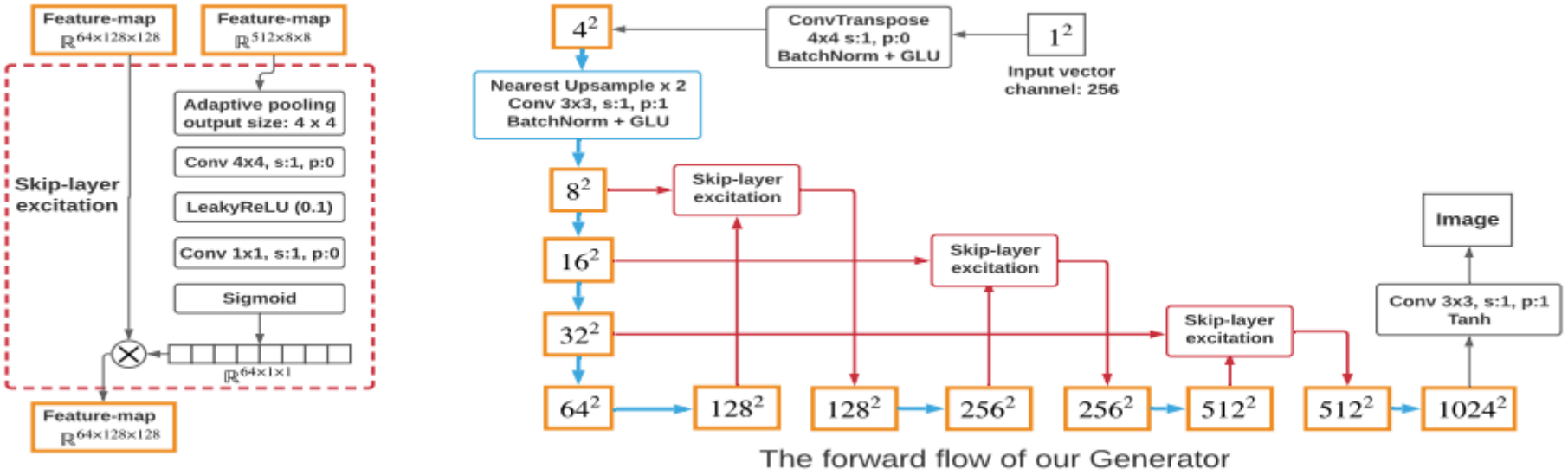
   → Gradient Flow + Style/Content Disentangle

2. Self-Supervised Discriminator trained as a Feature Encoder

   → Mitigate Overfitting + Provide Better Signal

3. Computationally Efficient GAN with High Fidelity Output

# SLE Module



The forward flow of our Generator

$$\mathbf{y} = \mathcal{F}(\mathbf{x}_{low}, \{\mathbf{W}_i\}) \cdot \mathbf{x}_{high}$$

- SLE = Skip Connection + Channel Attention
- $x_{low}$ (style) + $x_{high}$ (content) Disentanglement

# SLE Module

Unique from Residual Blocks in that

1) Apply channel-wise multiplication instead of element-wise addition

→ Can be applied when spatial dimensions are different

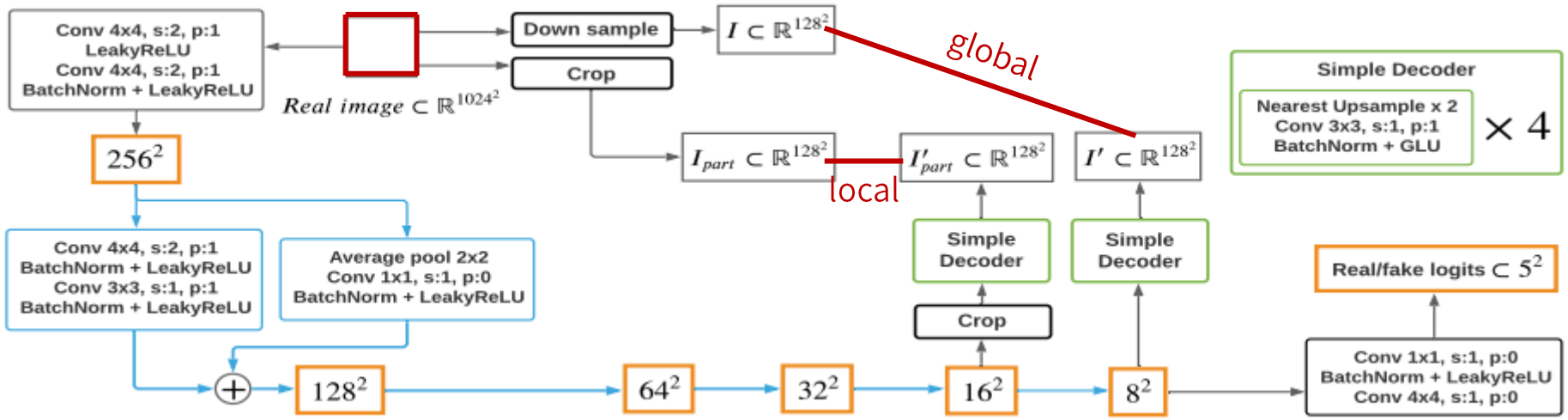
→ Relieves computation

2) Perform skip-connection between resolutions with much longer range

→ Use single conv layer on each resolution

→ Lighter computation

# Self-supervised Discriminator




$$\mathcal{L}_{recons} = \mathbb{E}_{\mathbf{f}\sim D_{encode}(x),\ x\sim I_{real}}[||\mathcal{G}(\mathbf{f}) - \mathcal{T}(x)||]$$

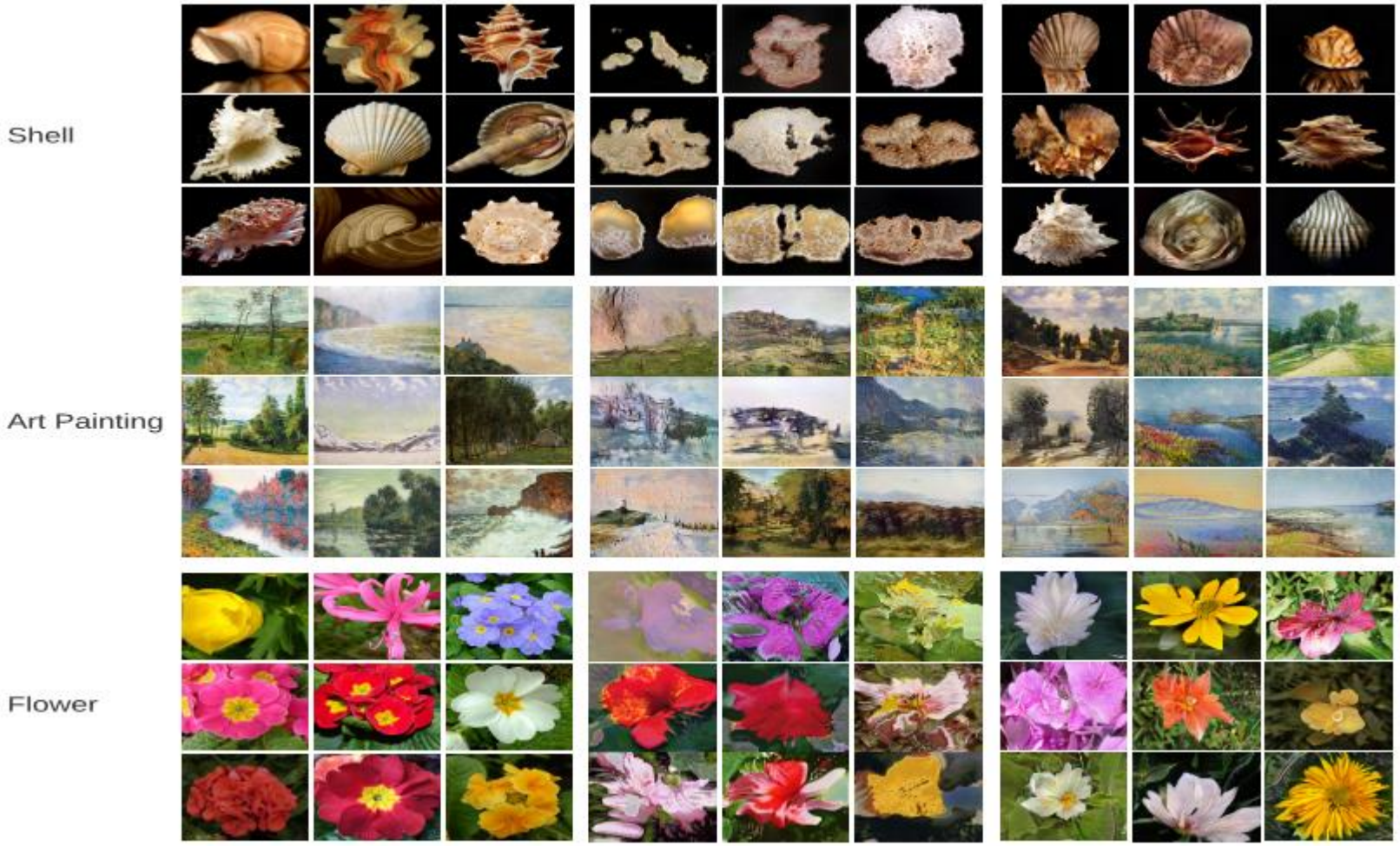
- D extracts more comprehensive representation from input covering both overall compositions and detailed textures
- Combats D overfitting by forbidding D to memorize and rely on small local patterns

# Final Objective

$$\mathcal{L}_D = -\mathbb{E}_{x \sim I_{real}}[min(0, -1 + D(x))] - \mathbb{E}_{\hat{x} \sim G(z)}[min(0, -1 - D(\hat{x})] + \mathcal{L}_{recons}$$
$$\mathcal{L}_G = -\mathbb{E}_{z \sim \mathcal{N}}[D(G(z))]$$

- Different GAN losses make little performance difference → Hinge Loss (Fastest)

# Experiments – (0) Qualitative Comparison




- Trained for 10 hours with batch size=8

# Experiments – (0) Qualitative Comparison


- Trained for 10 hours with batch size=8

# Experiments – (1) Computationally Efficient

Table 1: Computational cost comparison of the models.

| | | StyleGAN2@0.25 | StyleGAN2@0.5 | StyleGAN2 | Baseline | Ours |
|---|---|---|---|---|---|---|
| Resolution: $256^2$<br>Batch-size: 8 | Training time (hour / 10k iter) | 1 | 1.8 | 3.8 | 0.7 | 1 |
| | Training vram (GB) | 7 | 16 | 18 | 5 | 6.5 |
| | Model parameters (million) | 27.557 | 45.029 | 108.843 | 44.359 | 47.363 |
| Resolution: $1024^2$<br>Batch-size: 8 | Training time (hour / 10k iter) | 3.6 | 5 | 7 | 1.3 | 1.7 |
| | Training vram (GB) | 12 | 23 | 36 | 9 | 10 |
| | Model parameters (million) | 27.591 | 45.15 | 109.229 | 44.377 | 47.413 |

- StyleGAN2@0.25 cannot converge on 1024 x 1024 → Use StyleGAN2@0.5
- Nvidia RTX 2080-Ti GPU, Implemented in Pytorch

# Experiments – (2) High Quality Synthesis

Table 2: FID comparison at $256^2$ resolution on few-sample datasets.

| | | | Animal Face - Dog | Animal Face - Cat | Obama | Panda | Grumpy-cat |
|---|---|---|---|---|---|---|---|
| Image number | | | 389 | 160 | 100 | 100 | 100 |
| Training time on one RTX 2080-Ti | 20 hour | StyleGAN2 | 58.85 | 42.44 | 46.87 | 12.06 | 27.08 |
| | | StyleGAN2 finetune | 61.03 | 46.07 | **35.75** | 14.5 | 29.34 |
| | 5 hour | Baseline | 108.19 | 150.3 | 62.74 | 15.4 | 42.13 |
| | | Baseline+Skip | 94.21 | 72.97 | 52.50 | 14.39 | 38.17 |
| | | Baseline+decode | 56.25 | 36.74 | 44.34 | 10.12 | 29.38 |
| | | Ours (B+Skip+decode) | **50.66** | **35.11** | 41.05 | **10.03** | **26.65** |

Table 3: FID comparison at $1024^2$ resolution on few-sample datasets.

| | | | Art Paintings | FFHQ | Flower | Pokemon | Anime Face | Skull | Shell |
|---|---|---|---|---|---|---|---|---|---|
| Image number | | | 1000 | 1000 | 1000 | 800 | 120 | 100 | 60 |
| Training time on one RTX TITAN | 24 hour | StyleGAN2 | 74.56 | 25.66 | 45.23 | 190.23 | 152.73 | 127.98 | 241.37 |
| | | StyleGAN2 finetune | N/A | N/A | 36.72 | 60.12 | 61.23 | **107.68** | 220.45 |
| | 8 hour | Baseline | 62.27 | 38.35 | 42.25 | 67.86 | 101.23 | 186.45 | 202.32 |
| | | Ours | **45.08** | **24.45** | **25.66** | **57.19** | **59.38** | 130.05 | **155.47** |

# Experiments – (3) With More Data Samples

Table 4: FID comparison at $1024^2$ resolution on datasets with more images.

| Model | Dataset | Art Paintings | | | FFHQ | | | | Nature Photograph | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Image number | 2k | 5k | 10k | 2k | 5k | 10k | 70k | 2k | 5k | 10k |
| StyleGAN2 | | 70.02 | 48.36 | **41.23** | **18.38** | **10.45** | **7.86** | **4.4** | 67.12 | **41.47** | **39.05** |
| Baseline | | 60.02 | 51.23 | 49.38 | 36.45 | 27.86 | 25.12 | 17.62 | 71.47 | 66.05 | 62.28 |
| Ours | | **44.57** | **43.27** | 42.53 | 19.01 | 17.93 | 16.45 | 12.38 | **52.47** | 45.07 | 43.65 |

- StyleGAN2@0.5 does better with more image samples

- (1) Capacity of *Ours* not big enough
- (2) StyleGAN2@0.5 is still to big for few shot (~1000)

# Experiments – (4) Robust to Overfitting - D

Table 7: LPIPS on $D$'s feature-extracting performance

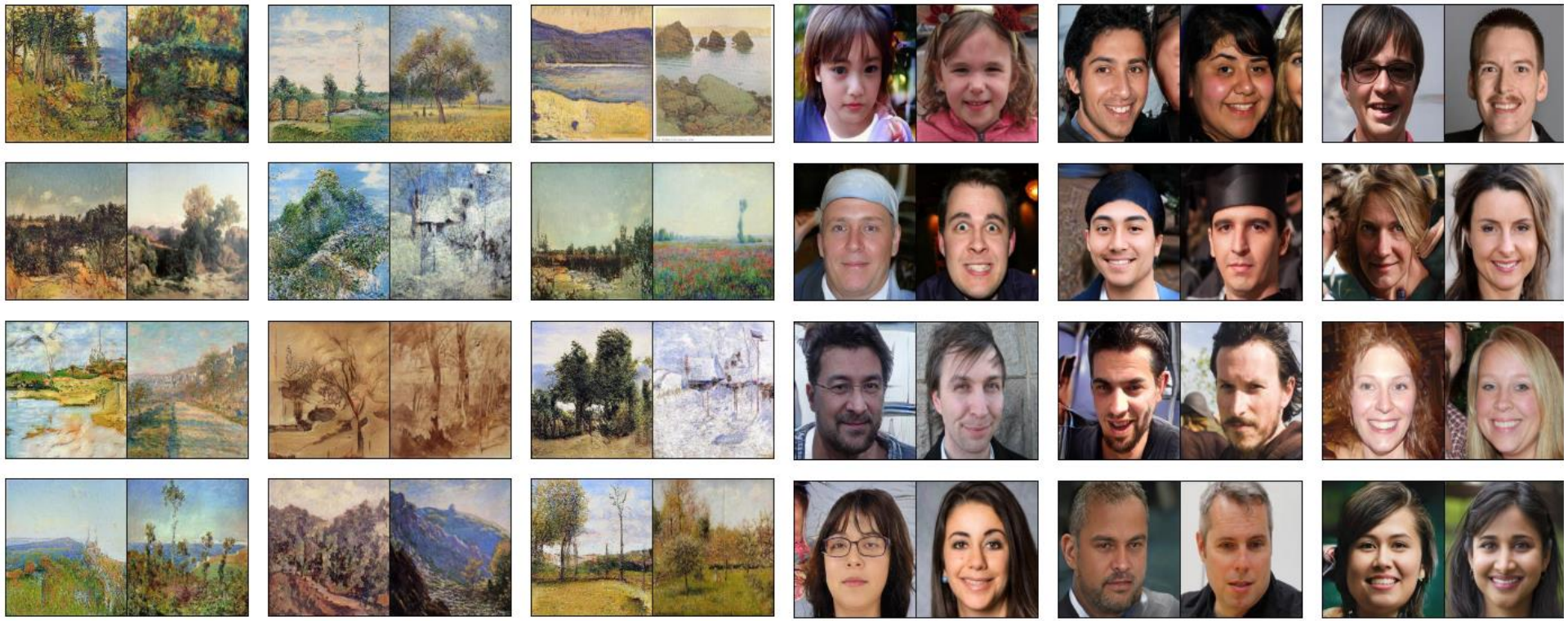
| | Grumpy Cat | Obama | FFHQ | | | Art | |
|---|---|---|---|---|---|---|---|
| Image number | 100 | 100 | 1k | 70k | 0 | 1k | 0 |
| StyleGAN2 | 0.914 | 0.652 | 3.177 | 2.43 | 2.289 | 3.051 | 2.761 |
| Baseline | 1.632 | 0.733 | 2.421 | N/A | 1.943 | 2.677 | 2.421 |
| Baseline + Contrastive | 1.251 | 0.647 | 1.821 | N/A | 1.943 | 2.124 | 2.421 |
| Baseline + AE | **0.725** | **0.405** | **1.075** | N/A | 1.943 | **1.806** | 2.421 |
| Baseline + AE + Contrastive | 1.156 | 0.578 | 1.345 | N/A | 1.943 | 1.927 | 2.421 |

- If D pays attention to all the regions of an image and encode it with minimum information loss
  → Easier for decoder to reconstruct

- Vanilla GAN training makes D worse as feature encoder – focus on rather local discriminative feature

- All SSL helps feature extraction, but Auto Encoding works best (in their setting)

# Experiments – (4) Robust to Overfitting - G



(a) Art-paintings 1k

(b) FFHQ 1k

- Nearest Real Images / Synthesized Images (Left-Right) by LPIPS

# Experiments – (5) Ablation – SLE

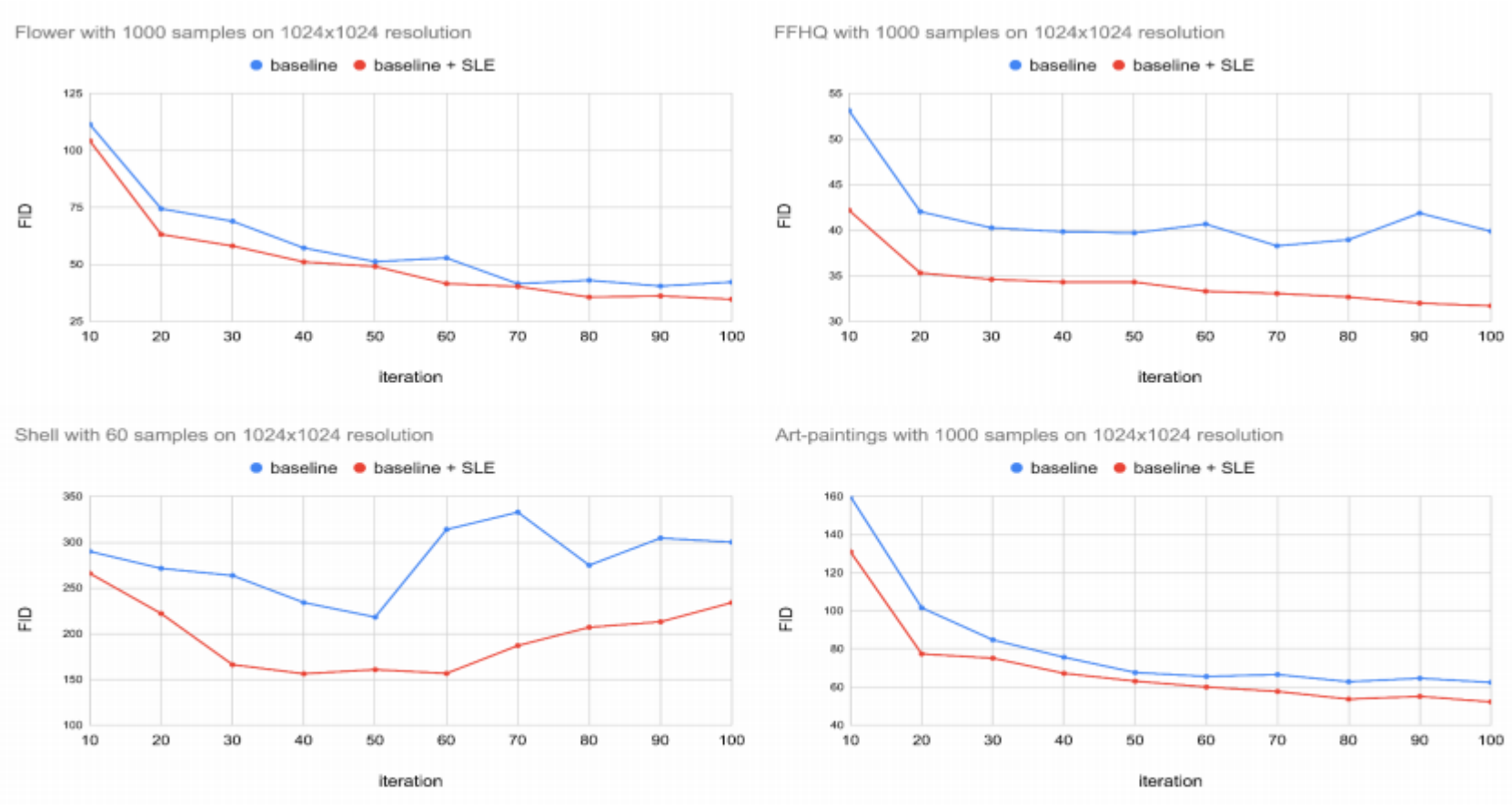



Figure 8: **Ablation study for SLE module** on $1024 \times 1024$ resolution datasets. Each unit on the x-axis represents 1000 training iterations, and y-axis represents the FID score.

# Experiments – (6) Style Mixing

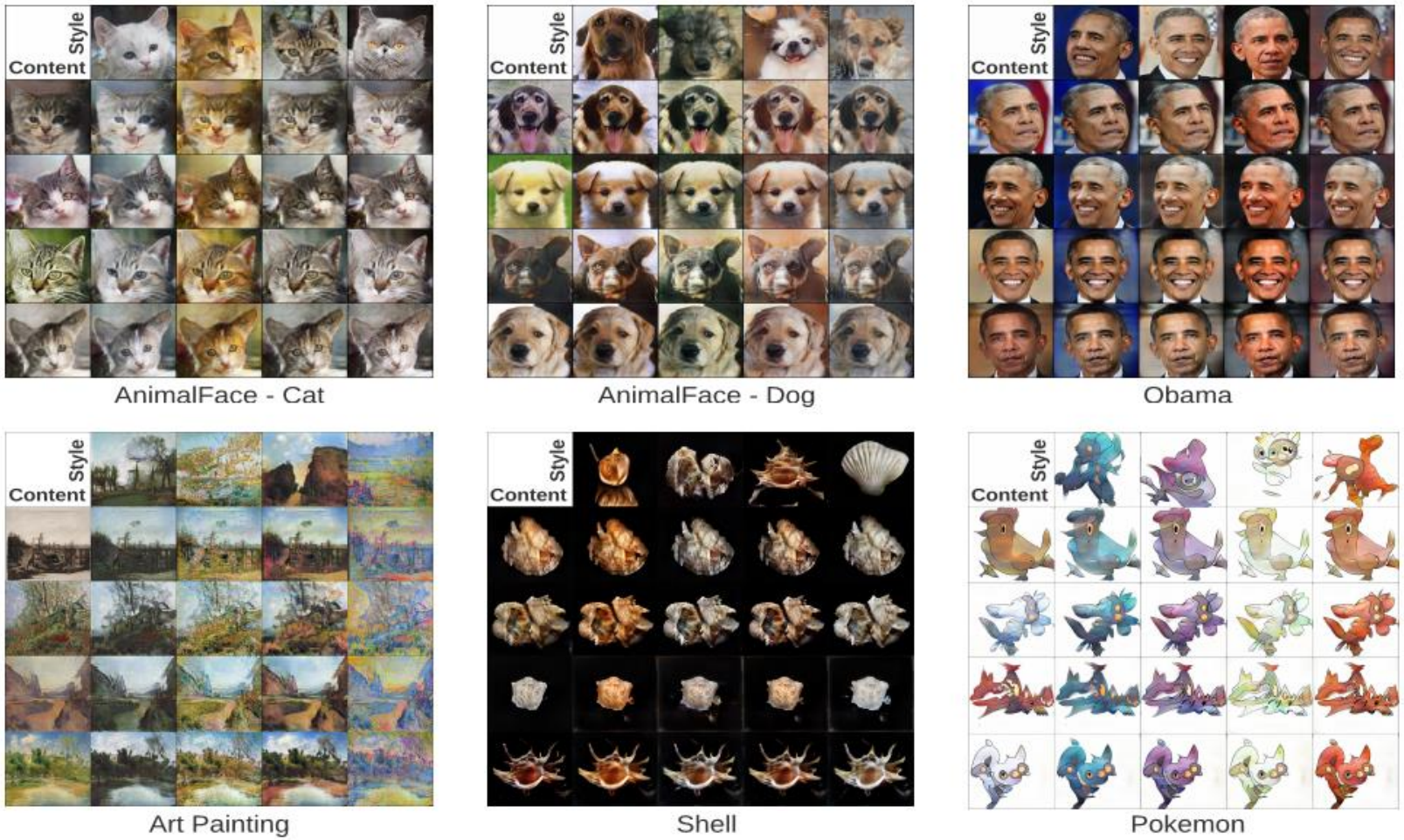



Figure 7: **Style-mixing results** from our model trained for only 5 hours on single GPU.