Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis

Bingchen Liu, Yizhe Zhu, Kunpeng Song, Ahmed Elgammal

Playform - Artrendex, Rutgers University

ICLR 2021

Presented by Minho Park

Why Few-shot Image Synthesis?

- The available samples to train GAN can be minimal.
  - E.g., medical images of a rare disease, art creation applications, etc.
- Transfer learning with a pre-trained model need to find a compatible pre-training dataset.
  - Also fine-tuning probably leads to even worse performance.
- In this paper, a GAN model converges from scratch with just **few hours of training on a single RTX-2080 GPU** even with less than **100 training samples on 1024×1024 resolution.**

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.

Contribution

- **Train GAN on High Resolution (1024×1024).**
  - Problem: Increased model parameters lead to a **rigid gradient flow** to optimize G.
    The target distribution on 1024×1024 is super **sparse**.
  - Prior works: Multi-scale GAN structures (e.g., pix2pixHD, MSG-GAN).
    However, increase computational cost.
  - FastGAN: Skip-Layer channel-wise Excitation (SLE) module.

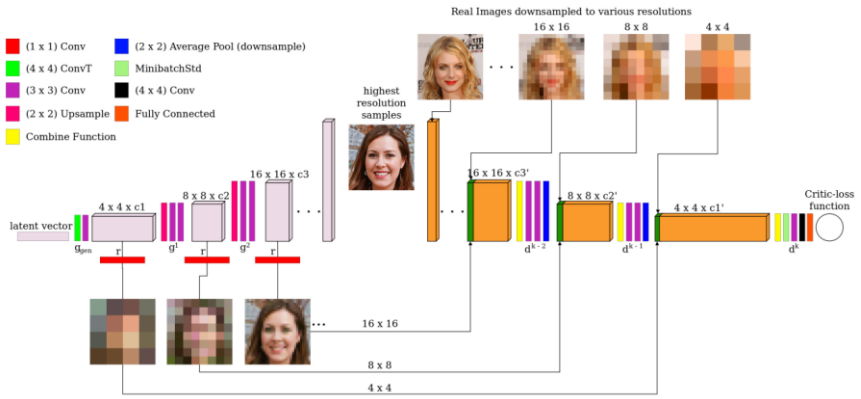


Figure 2: Architecture of MSG-GAN, shown here on the base model proposed in ProGANs [15]. Our architecture includes connections from the intermediate layers of the generator to the intermediate layers of the discriminator. Multi-scale images sent to the discriminator are concatenated with the corresponding activation volumes obtained from the main path of convolutional layers followed by a combine function (shown in yellow).

Multi-scale GAN (MSG-GAN)

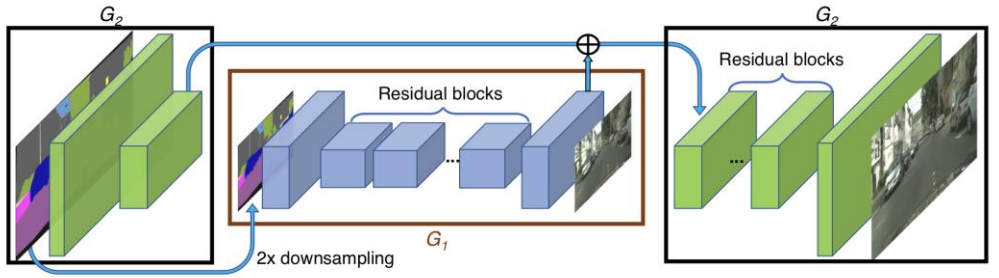


Figure 3: Network architecture of our generator. We first train a residual network G_1 on lower resolution images. Then, another residual network G_2 is appended to G_1 and the two networks are trained jointly on high resolution images. Specifically, the input to the residual blocks in G_2 is the element-wise sum of the feature map from G_2 and the last feature map from G_1.

Multi-scale conditional generator (pix2pixHD)

Wang, Ting-Chun, et al. "High-resolution image synthesis and semantic manipulation with conditional gans." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
Karnewar, Animesh, and Oliver Wang. "Msg-gan: Multi-scale gradients for generative adversarial networks." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

Contribution

- **Stabilize the GAN training.**
  - Problem: **Mode-collapse** on G, because of the **overfitted D**.
  - Prior works: Different objectives, regularizing the gradients, augmenting the training data, etc.
    However, these methods degrades fast when the training batch-size is limited.
  - FastGAN: A self supervised discriminator D trained as a feature-encoder with an extra decoder.

Skip-Layer channel-wise Excitation (SLE)

- Skip-connection in ResBlock requires the spatial dimensions of the activations to be the same.
- SLE enables skip-connection to a much longer range without an extra computation burden.

$$y = \mathcal{F}(x_{low}, \{W_i\}) \cdot x_{high}$$

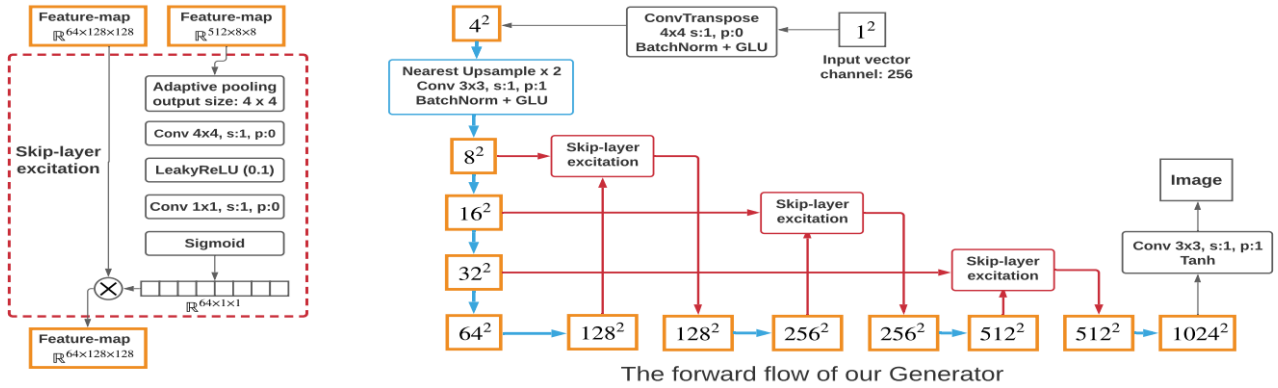


Figure 3: The structure of the skip-layer excitation module and the Generator. Yellow boxes represent feature-maps (we show the spatial size and omit the channel number), blue box and blue arrows represent the same up-sampling structure, red box contains the SLE module as illustrated on the left.

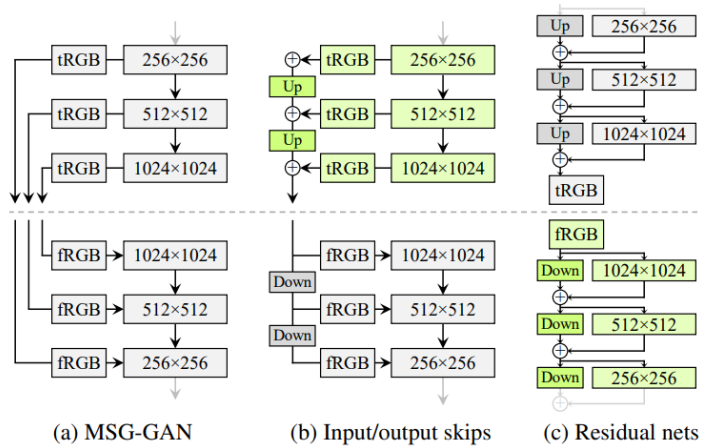


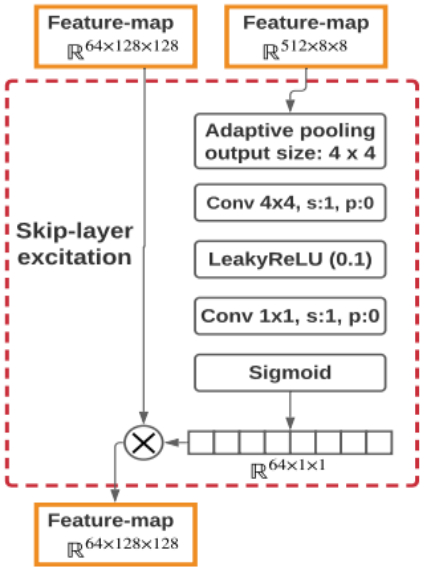
Figure 7. Three generator (above the dashed line) and discriminator architectures. Up and Down denote bilinear up and down-sampling, respectively. In residual networks these also include 1×1 convolutions to adjust the number of feature maps. tRGB and fRGB convert between RGB and high-dimensional per-pixel data. Architectures used in configs E and F are shown in green.

StyleGAN2 architecture

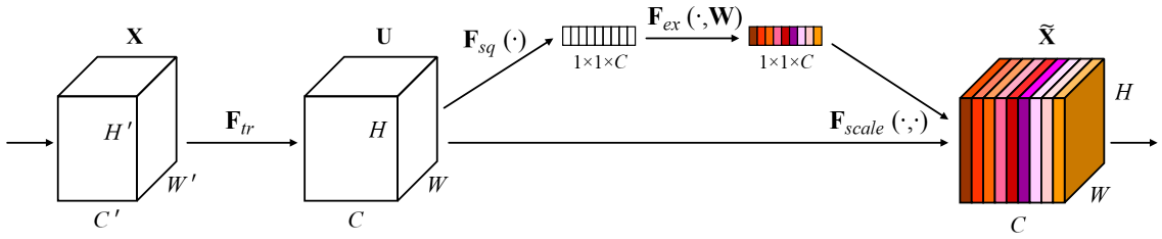
Karras, Tero, et al. "Analyzing and improving the image quality of stylegan." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

SLE vs. SE

- Both blocks activate channel-wise like Instance Normalization, which is widely used in style transfer.
- However, SE operates within one feature map, SLE performs between feature-maps that are far away from each other.



SLE block



SE block

Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Self-Supervised Discriminator

- Self-supervision strategy for regularizing D in form of an auto-encoder (AE).

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

- Such reconstructive training makes sure that D extracts a more comprehensive representation from the inputs.

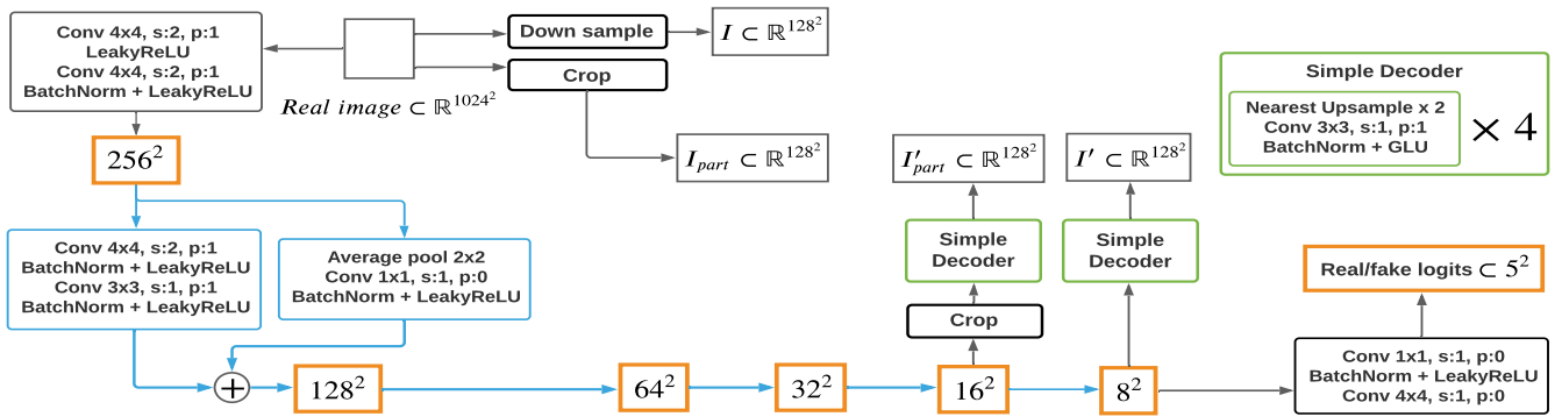


Figure 4: The structure and the forward flow of the Discriminator. Blue box and arrows represent the same residual down-sampling structure, green boxes mean the same decoder structure.

Experiments

- The hinge version of adversarial loss:

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

- Datasets: To show the model's robustness on multiple high-fidelity datasets.
  - Animal-Face Dog and Cat, 100-Shot-Obama, Panda, Flickr-Face-HQ, Oxford-flowers, etc.
- Metrics: FID, LPIPS (using GAN inversion).
- Compared Models: StyleGAN2, baseline (StyleGAN2 for the best training on few-samples datasets, e.g., configs, data augmentation).

Table 1: Computational cost comparison of the models.

		StyleGAN2@0.25	StyleGAN2@0.5	StyleGAN2	Baseline	Ours
Resolution: 256^2 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 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

Quantitative Results

- Few-sample datasets with $256^2, 1024^2$ resolution.

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**

Quantitative Results

- With more images.

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

- Testing mode collapse with back-tracking / Self-supervision methods

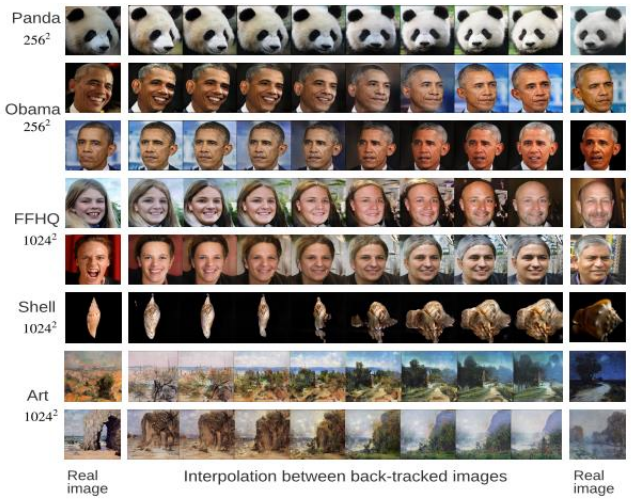


Figure 5: Latent space back-tracking and interpolation.

Table 5: LPIPS of back-tracking with G

	Cat	Dog	FFHQ	Art
Resolution	256		1024	
Baseline @ 20k iter	2.113	2.073	2.589	2.916
Baseline @ 40k iter	2.513	2.171	2.583	2.812
Ours @ 40k iter	**1.821**	**1.918**	2.425	2.624
Ours @ 80k iter	1.897	1.986	**2.342**	**2.601**

Table 6: FID of self-supervisions for D

	Art paintings	Nature photos
a. contrastive loss	47.14	57.04
b. predict aspect ratio	49.21	59.22
c. auto-encoding	**42.53**	**43.65**
d. a+b	46.02	54.23
e. a+b+c	44.21	47.65

Qualitative Results

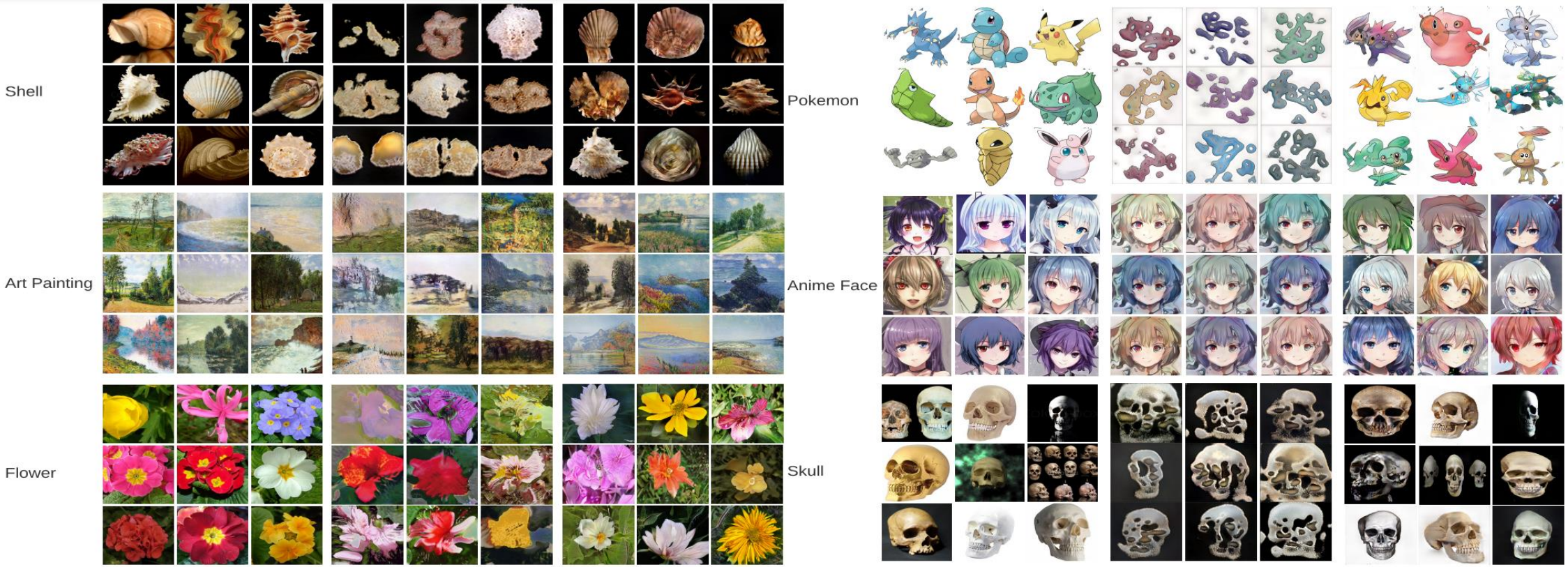
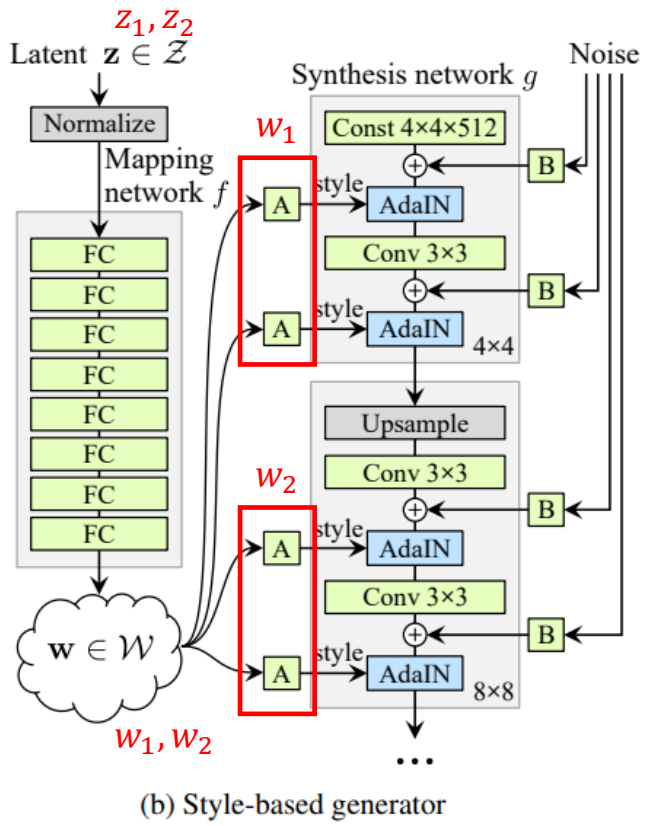


Figure 6: **Qualitative comparison between our model and StyleGAN2** on 1024^2 resolution datasets. The left-most panel shows the training images, and the right two panels show the uncurated samples from StyleGAN2 and our model. Both models are trained from scratch for 10 hours with a batch-size of 8. The samples are generated from the checkpoint with the lowest FID.

Style Mixing Like StyleGAN

- What is the style mixing?



Style mixing in StyleGAN

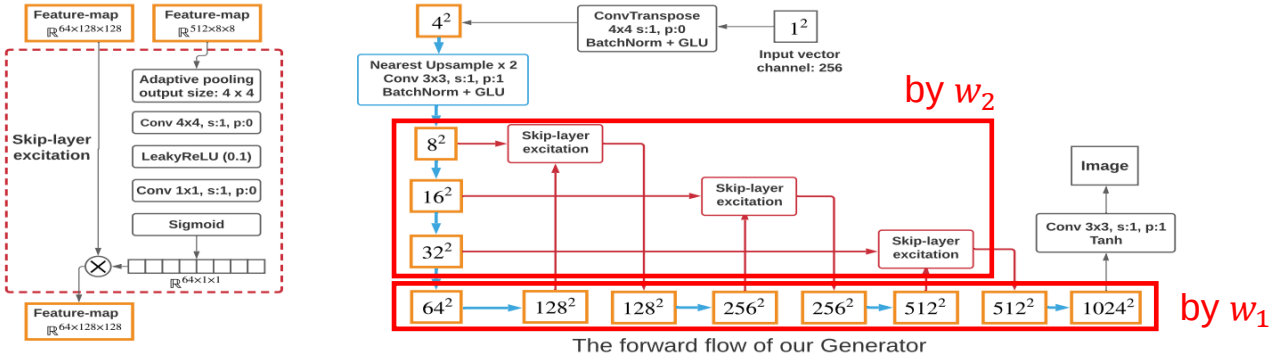


Figure 3: The structure of the skip-layer excitation module and the Generator. Yellow boxes represent feature-maps (we show the spatial size and omit the channel number), blue box and blue arrows represent the same up-sampling structure, red box contains the SLE module as illustrated on the left.

Style mixing in FastGAN

Style Mixing Like StyleGAN

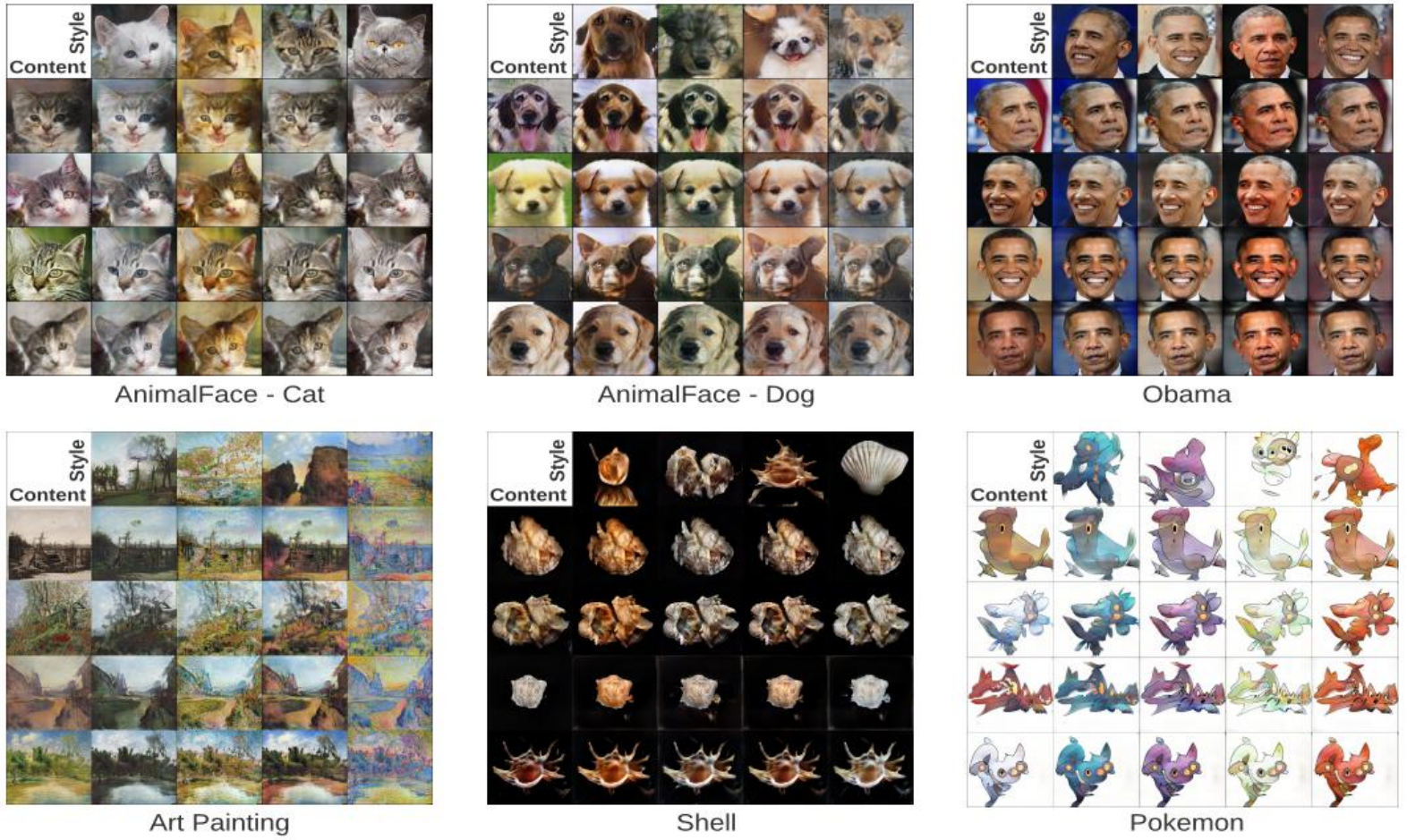


Figure 7: **Style-mixing results** from our model trained for only 5 hours on single GPU.