A Style-Based Generator Architecture for Generative Adversarial Network (StyleGAN)

Author: Tero Karras, Samuli Laine, Timo Aila (NVIDIA).

Year: 2019

San Jose State University

Data 255: Assignment 1

Aryama Ray

Background of StyleGAN : Generative Adversarial Network (GAN)

- Proposed by Ian Goodfellow et al.(2014)
- Consists of two deep networks the **generator -** produces images from random noise , and the **discriminator -** learns to distinguish real from fake images
- GAN is limited to small image sizes because of model stability.
- **Progressive GAN**(T Kerras et al., 2018) introduced a novel approach to generate high-quality images by progressively growing resolutions in generator and discriminator layers.

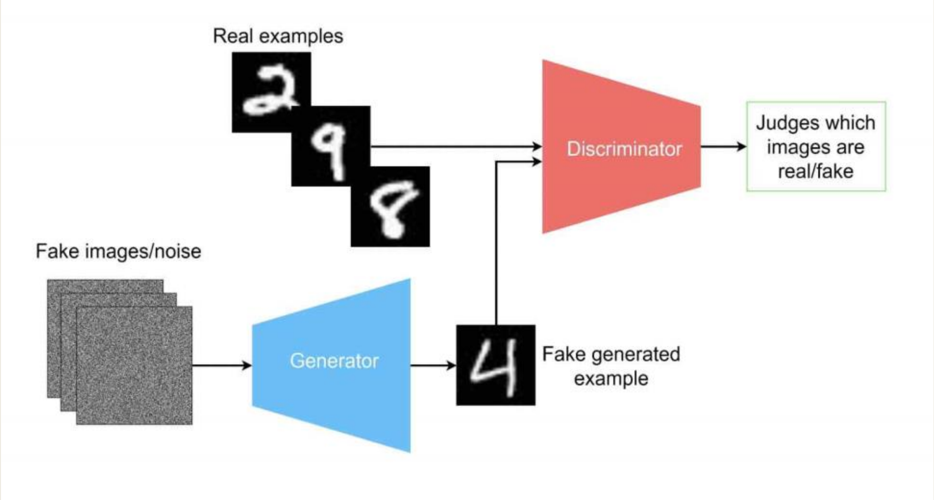


Image: https://developer.ibm.com/articles/generative-adversarial-networks-explained/

Motivation for StyleGAN

- Traditional GANs **lack control over image generation** - **Generators were treated as black box**, due to their highly entangled and opaque generation process, which stems from the direct use of latent code in the generator's input layer.
- There was **no quantitative way to assess disentanglement** or to modify specific attributes in isolation.
- In Traditional GANs, modifying the latent vector often leads to **simultaneous changes in multiple image attributes**—for example, altering hair color may also unintentionally change age or pose.

Main Contributions of StyleGAN : Summary

1. **Novel Generator Architecture : controls the image synthesis process**. Here generator starts with learned constant input and modulates each convolution layer's activations using "style" of the images based on the latent code.

2. **Mapping Network :** simplified traditional architecture by removing input layer. **Intermediate Latent Space W maps** input vectors Z via an 8-layers MLP (Multilayer Perceptron). This mapping is unconstrained by the training data distribution, allowing W to encode the information in a more disentangled way.

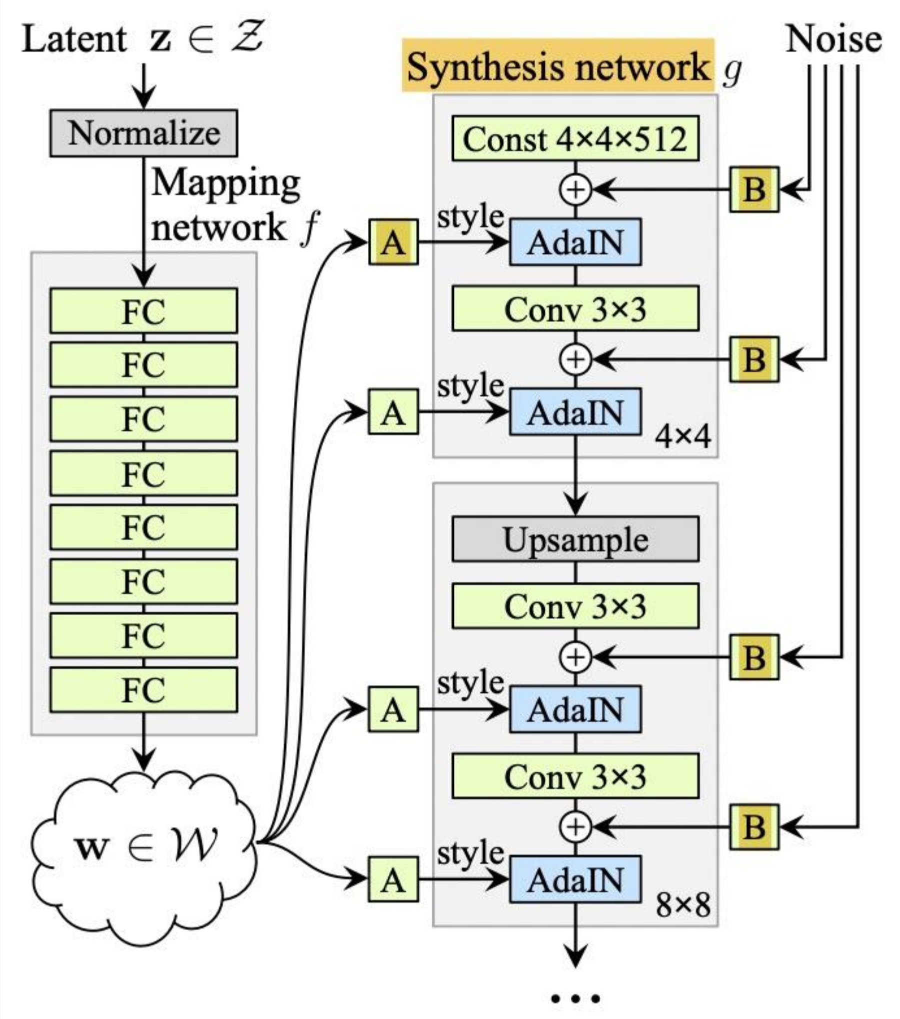
3. **Adaptive Instance Normalization (AdaIN) :** being used at each convolution layer of the generator. It allows the generate more control over the image generation and adjust the "style" of the images.

4. **Gaussian Noise Injection and Novel Mixing Regularization** : incorporates at each layer of the synthesis network to introduce stochastic variation in the generated images, thereby enhancing output quality. Novel mixing regularization method decorrelates adjacent styles, enabling finer-grained control over the generated imagery.

5. **New Evaluation Metrices** : **Perceptual Path Length & Linear Separability -** showcase that the proposed generator admits a more linear, less entangled representation of different factors of variation.

6. **New dataset - Flickr-Faces-HQ (FFHQ) :** consists of 70,000 high-quality face images at 1024^2 resolution. FFHQ has diverse data in terms of age, ethnicity, lighting, and accessories unlike CelebAHQ dataset.

StyleGAN Architecture



StyleGAN : Generator Architecture
Image : https://arxiv.org/abs/1812.04948

- **Baseline Progressive GAN** : start with small images, in this case, 4×4 images.
- **StyleGAN's generator** takes a latent vector Z and feeds it into the network
- **The proposed generator** is split into two parts: a **Mapping network f** and a **Synthesis network g**.
- **The mapping network (f)** (an 8-layer MLP) : transforms the input vector z into an intermediate latent space w. A learned affine transform(A) turns **w** vectors into **styles** which is fed to the synthesis network.
- **Synthesis Network (g) : s**tarts from learned constant 4×4×512 tensor, applies styles via AdaIN.
- The image is generated through a **sequence of convolutional** layers that progressively upsample the resolution (4×4 → 8×8 → … → 1024×1024) similar to Progressive GAN
- Each layer also has a **Noise Injection** added after each convolution to generate fine stochastic detail.

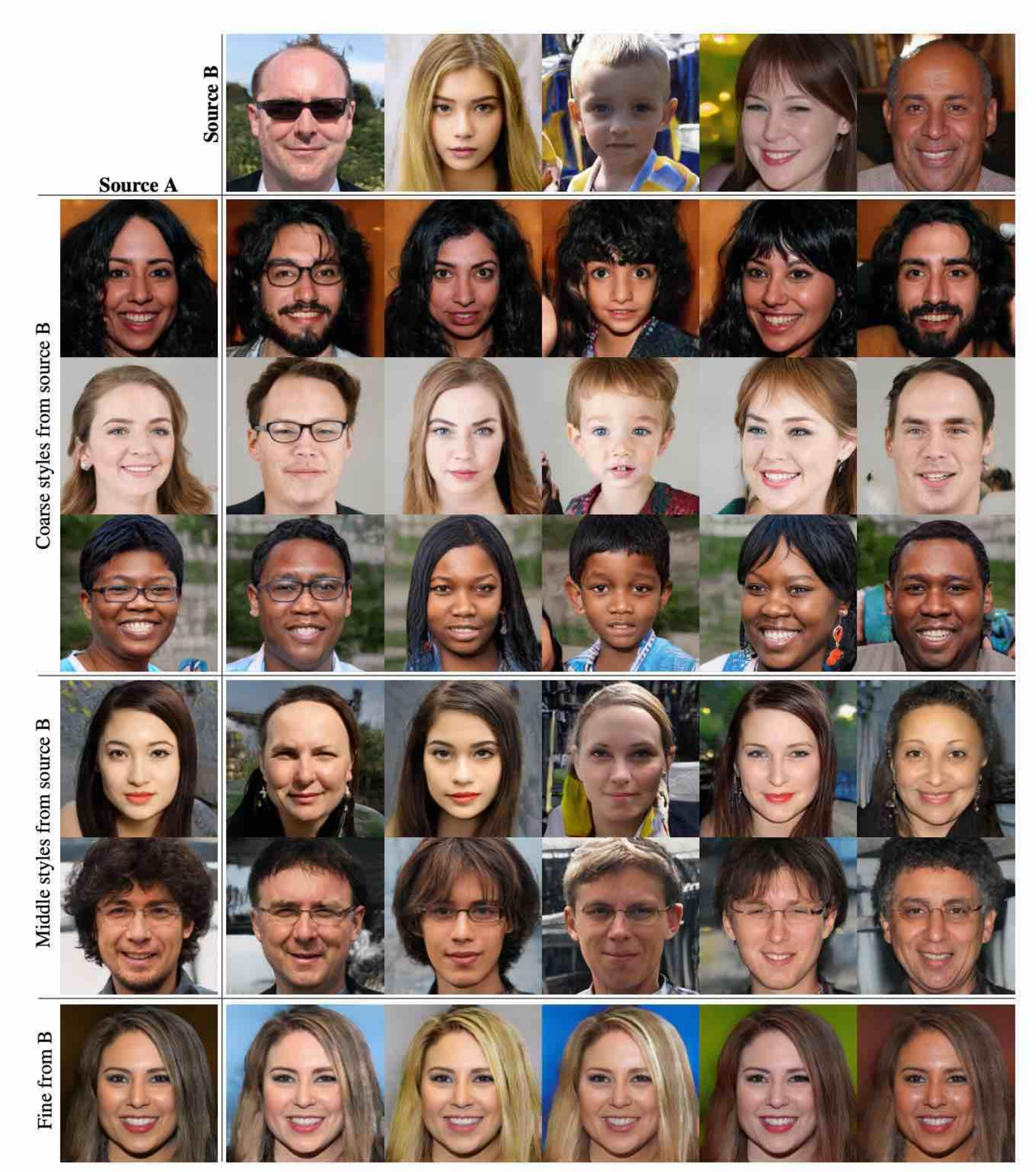
Flickr-Faces-HQ (FFHQ) Dataset

- New dataset was introduced.
- Consists of 70k images of human faces at 1024×1024 resolution, sourced from Flickr and automatically aligned/cropped.
- It spans a wide range of ages (babies to the elderly), ethnic backgrounds, various viewpoints and lighting conditions, and accessories like eyeglasses and hats
- FFHQ offers far more variation than the CelebA-HQ dataset, which was used in earlier GANs

Image : https://arxiv.org/abs/1812.04948

Training Details

Base Framework	Base model is similar to Progressive GANTraining starts at a low resolution and gradually grows to high resolution (1024×1024), stabilizing learning at each stage.
Loss Functions and Regularization	Early versions used **WGAN-GP** loss, but for better stability and quality, the later version switches to **non-saturating loss** with **R1 regularization** (gradient penalty on real images)
Upsampling and Downsampling	Traditional **nearest-neighbor sampling** is replaced with **bilinear sampling + low-pass filtering** for smoother transitions and reduced aliasing artifacts.
Training Schedule	Training starts from low resolutions and it progresses to **1024×1024**, with a **transition phase** at each resolution. Each stage trains on a large number of images — up to **25 million** total images seen by the discriminator.
Optimization and Hyperparameters	Uses **Adam optimizer** with settings from the original PGGAN. Uses **exponential moving average (EMA)** of the generator for inference and FID evaluation. The **learning rate** is typically 0.003, but reduced to **0.002** at higher resolutions to stabilize training. The **mapping network** (for z→w) has 8 layers, but uses a **much lower learning rate** to avoid instability.
Augmentation and Regularization	**Mirror augmentation** is used for datasets like CelebA-HQ and FFHQ. **No batch norm, dropout, spectral norm, attention, or pixelwise normalization** is used—favoring simplicity and stability.



Inference and Control of Image Generation

StyleGAN allows **unprecedented control** at inference.

Here two source images with their respective latent codes were shown.

- **Coarse resolutions [4x4–8x8]:** eyes/hair/skin color (**small-level aspects**))are copied from **A** whereas the pose, hairstyle, and face shape (**high-level aspects**) are copied from **B.**
- **Middle resolutions [16x16–32x32]:** high-level aspects are copied from **A** and small aspects are copied from **B**.
- **Fine resolutions [64x64–1024x1024]:** almost all the style is copied from **A** and only some color details are copied from **B**.

Image : https://arxiv.org/abs/1812.04948

Evaluation Metrics and Results

- **FID (Fréchet Inception Distance):** widely used metric to measure image quality and diversity – it compares the statistics of generated images to real images (lower FID is better)

Two new metrices were proposed.

- **Perceptual Path Length (PPL):** measures the degree of changes done on the image when performing interpolation. Smooth changes give better results(lower is better)

- **Linear Separability:** measures how well the latent space points corresponding to two image classes can be separated via a hyperplane.

	Method	FID	Path length full	Path length end	Separa-bility
B	Traditional 0 $\mathcal{Z}$	5.25	412.0	415.3	10.78
	Traditional 8 $\mathcal{Z}$	4.87	896.2	902.0	170.29
	Traditional 8 $\mathcal{W}$	4.87	324.5	212.2	6.52
	Style-based 0 $\mathcal{Z}$	5.06	283.5	285.5	9.88
	Style-based 1 $\mathcal{W}$	4.60	219.9	209.4	6.81
	Style-based 2 $\mathcal{W}$	4.43	**217.8**	199.9	6.25
F	Style-based 8 $\mathcal{W}$	**4.40**	234.0	**195.9**	**3.79**

- W-space combined with a style-based generator architecture gives the best FID (Frechet Inception Distance) score, perceptual path length, and separability

Table: https://arxiv.org/abs/1812.04948

Strengths of StyleGAN

- Style-based generator architecture that decouples the generation process into an affine style modulation at each layer, enabling unprecedented control over the generated image's properties
- Achieved new State-of-the-Art Image Quality
- Proposed intermediate latent space (W) yields *more linear, less entangled* representation of images
- Learned unsupervised separation of high-level vs. stochastic features in images (e.g., overall structure vs. micro-detail), and scale-specific image control automatically
- Contributed FFHQ dataset & code released publicly.
- Foundation for further improvements (StyleGAN2, StyleGAN3) and variations

Shortcomings of StyleGan

- Instance normalization causes water droplet -like artifacts in StyleGAN images. Blob-like artifacts (resembling water droplets)is visible at 64×64 resolution in all feature maps and it get worse in higher resolution.

Image : https://arxiv.org/pdf/1912.04958v2

- Image generation control offered by StyleGAN is inherently limited to the generator's learned distribution and can only be applied to images generated by StyleGAN itself.
- The architecture isn't explicitly spatially invariant – a limitation later addressed by StyleGAN3 to handle aliasing and rotation.
- Training requires high computational resource.

References

1. https://jonathan-hui.medium.com/gan-whats-generative-adversarial-networks-and-its-application-f39ed278ef09
2. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8107–8116. https://doi.org/10.1109/cvpr42600.2020.00813
3. https://medium.com/@arijzouaoui/stylegan-explained-3297b4bb813a
4. Divya Saxena and Jiannong Cao. 2021. Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions. ACM Comput. Surv. 54, 3, Article 63 (April 2022), 42 pages. https://doi.org/10.1145/3446374
5. Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. *International Conference on Learning Representations*. https://dblp.uni-trier.de/db/journals/corr/corr1710.html#abs-1710-10196
6. Bermano, A., Gal, R., Alaluf, Y., Mokady, R., Nitzan, Y., Tov, O., Patashnik, O., & Cohen-Or, D. (2022). State-of-the-Art in the architecture, Methods and applications of StyleGAN. *Computer Graphics Forum*, *41*(2), 591–611. https://doi.org/10.1111/cgf.14503
7. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-Free generative adversarial networks. *arXiv (Cornell University)*, *34*. https://arxiv.org/abs/2106.12423