Analyzing and Improving the Image Quality of StyleGAN

**(Karras, T., Laine. S., Aittala, M., Hellsten, J., Lethinen, J. & Timo A.
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020)**

RAMI LAB.

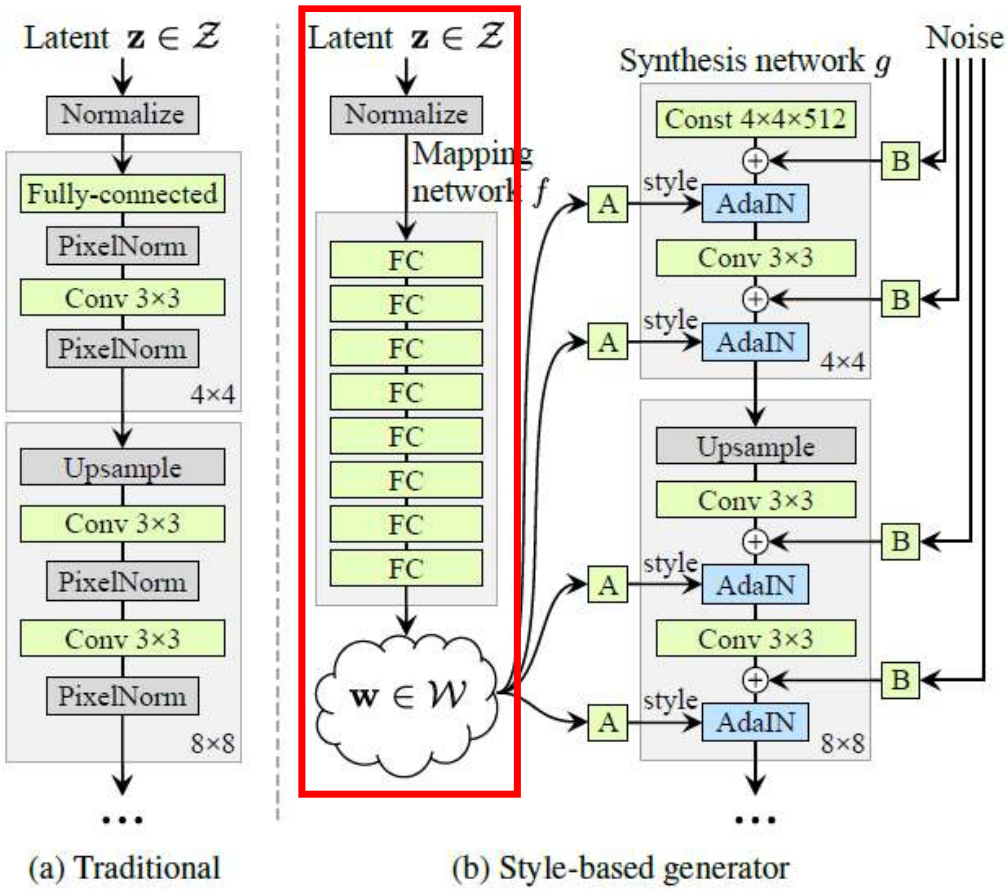
Sang-hyun Kim

Background

- **StyleGAN**
  - **Progressive GAN (PG-GAN) 을 기반으로 함.**
  - **여러 가지 component를 추가 하는 방식.**

	Method	CelebA-HQ	FFHQ
A	Baseline Progressive GAN [30]	7.79	8.04
B	+ Tuning (incl. bilinear up/down)	6.11	5.25
C	+ Add mapping and styles	5.34	4.85
D	+ Remove traditional input	5.07	4.88
E	+ Add noise inputs	**5.06**	4.42
F	+ Mixing regularization	5.17	**4.40**

Table 1. Fréchet inception distance (FID) for various generator designs (lower is better). In this paper we calculate the FIDs using 50,000 images drawn randomly from the training set, and report the lowest distance encountered over the course of training.



(a) Traditional (b) Style-based generator

StyleGAN2

- **StyleGAN2**
  - **StyleGAN 을 기반으로 함.**

Configuration	FFHQ, 1024×1024				LSUN Car, 512×384			
	FID ↓	Path length ↓	Precision ↑	Recall ↑	FID ↓	Path length ↓	Precision ↑	Recall ↑
A Baseline StyleGAN [24]	4.40	212.1	**0.721**	0.399	3.27	1484.5	**0.701**	0.435
B + Weight demodulation	4.39	175.4	0.702	0.425	3.04	862.4	0.685	0.488
C + Lazy regularization	4.38	158.0	0.719	0.427	2.83	981.6	0.688	0.493
D + Path length regularization	4.34	**122.5**	0.715	0.418	3.43	651.2	0.697	0.452
E + No growing, new G & D arch.	3.31	124.5	0.705	0.449	3.19	471.2	0.690	0.454
F + Large networks (StyleGAN2)	**2.84**	145.0	0.689	**0.492**	**2.32**	**415.5**	0.678	**0.514**
Config A with large networks	3.98	199.2	0.716	0.422	–	–	–	–

Table 1. Main results. For each training run, we selected the training snapshot with the lowest FID. We computed each metric 10 times with different random seeds and report their average. *Path length* corresponds to the PPL metric, computed based on path endpoints in $\mathcal{W}$ [24], without the central crop used by Karras et al. [24]. The FFHQ dataset contains 70k images, and the discriminator saw 25M images during training. For LSUN CAR the numbers were 893k and 57M. ↑ indicates that higher is better, and ↓ that lower is better.

Problem

- **Droplet artifact**
  - **최종 결과물에 물방울 무늬 같은 noise가 반복해서 나타나는 현상**

Figure 15. An example of the importance of the droplet artifact in StyleGAN generator. We compare two generated images, one successful and one severely corrupted. The corresponding feature maps were normalized to the viewable dynamic range using instance normalization. For the top image, the droplet artifact starts forming in 64^2 resolution, is clearly visible in 128^2, and increasingly dominates the feature maps in higher resolutions. For the bottom image, 64^2 is qualitatively similar to the top row, but the droplet does not materialize in 128^2. Consequently, the facial features are stronger in the normalized feature map. This leads to an overshoot in 256^2, followed by multiple spurious droplets forming in subsequent resolutions. Based on our experience, it is rare that the droplet is missing from StyleGAN images, and indeed the generator fully relies on its existence.

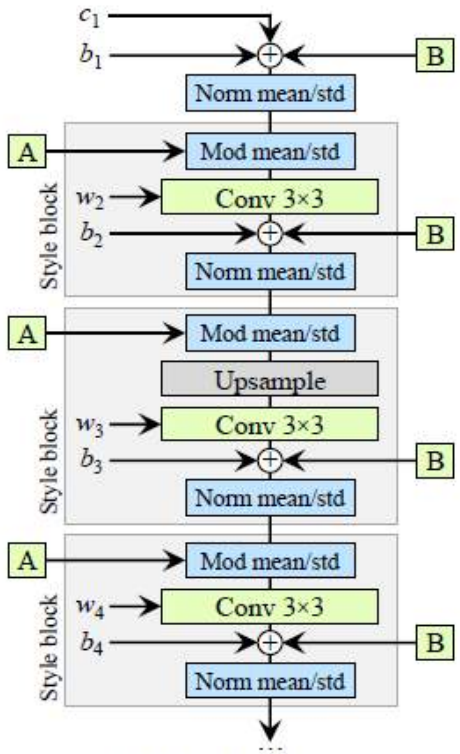
B. Weight demodulation

A	Baseline StyleGAN [24]
B	+ Weight demodulation
C	+ Lazy regularization
D	+ Path length regularization
E	+ No growing, new G & D arch.
F	+ Large networks (StyleGAN2)
	Config A with large networks

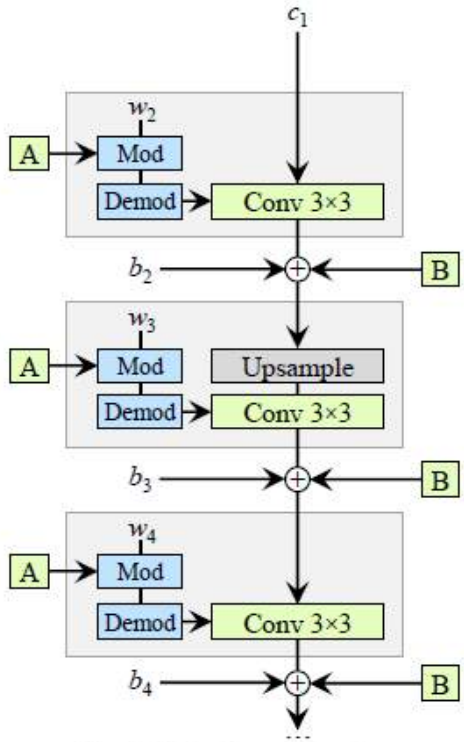
- **AdaIN (Adaptive Instance Normalization)**
  - **Conv layer와 Conv layer간의 사이의 관계를 방해.**

- **Modulation:** $w'_{ijk} = s_i \cdot w_{ijk},$

- **Demodulation:** $w''_{ijk} = w'_{ijk} \Big/ \sqrt{\sum_{i,k} w'^{\,2}_{ijk} + \epsilon}$



(b) StyleGAN (detailed)



(d) Weight demodulation

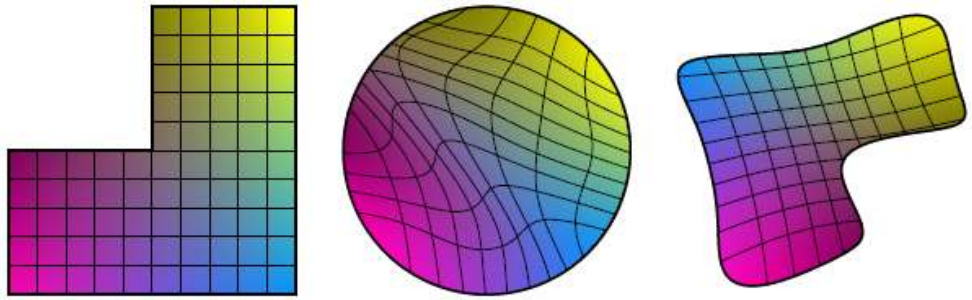
D. Perceptual Path Length

A	Baseline StyleGAN [24]
B	+ Weight demodulation
C	+ Lazy regularization
D	+ Path length regularization
E	+ No growing, new G & D arch.
F	+ Large networks (StyleGAN2)
	Config A with large networks

- **PPL (Perceptual Path Length)**
  - 이미지가 '지각적(perceptual)'으로 부드럽게 바뀌었는지 나타내는 지표.

(a) Low PPL scores (b) High PPL scores

Figure 4. Connection between perceptual path length and image quality using baseline StyleGAN (config A) with LSUN CAT. (a) Random examples with low PPL ($\leq 10^{\text{th}}$ percentile). (b) Examples with high PPL ($\geq 90^{\text{th}}$ percentile). There is a clear correlation between PPL scores and semantic consistency of the images.



(a) Distribution of features in training set (b) Mapping from $\mathcal{Z}$ to features (c) Mapping from $\mathcal{W}$ to features

D. Perceptual Path Length (Cont.)

A	Baseline StyleGAN [24]
B	+ Weight demodulation
C	+ Lazy regularization
D	+ Path length regularization
E	+ No growing, new G & D arch.
F	+ Large networks (StyleGAN2)
	Config A with large networks

- **PPL (Perceptual Path Length)**
  - **PPL은 낮을 수록 좋은 이미지.**
  - **PPL이 FID, Precision, recall에 비해서 인간의 관점에서 더 좋은 이미지를 detect하는데 유용하게 사용.**

Model 1: FID = 8.53, P = 0.64, R = 0.28, PPL = 924

Model 2: FID = 8.53, P = 0.62, R = 0.29, PPL = 387

(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

Reference: https://arxiv.org/pdf/1811.12231.pdf

D. Path Length Regularization & C. Lazy Regularization

A	Baseline StyleGAN [24]
B	+ Weight demodulation
C	+ Lazy regularization
D	+ Path length regularization
E	+ No growing, new G & D arch.
F	+ Large networks (StyleGAN2)
	Config A with large networks

- **Path Length Regularization**
  - J_w: Jacobian matrix, W space에서 weight의 derivation
  - y: random image
  - a: training 이 진행 됨에 따라 변하는 이동 평균

$$\mathbb{E}_{\mathbf{w},\mathbf{y}\sim\mathcal{N}(0,\mathbf{I})}\left(\left\|\mathbf{J}_{\mathbf{w}}^{T}\mathbf{y}\right\|_{2}-a\right)^{2}$$

- **Lazy Regularization**
  - 여러 regularization의 적용은 계산적으로 오버헤드가 있기 때문에, 16 배치마다 한 번씩 적용.

E. Phase artifact

- **Problem**
  - **이빨이나 눈동자 같은 부분에서 disentanglement가 자연스럽지 않은 현상.**

- **Solution**
  - **MSG-GAN architecture !!!**

A Baseline StyleGAN [24]
B + Weight demodulation
C + Lazy regularization
D + Path length regularization
E + No growing, new G & D arch.
F + Large networks (StyleGAN2)
Config A with large networks

Figure 6. Progressive growing leads to "phase" artifacts. In this example the teeth do not follow the pose but stay aligned to the camera, as indicated by the blue line.

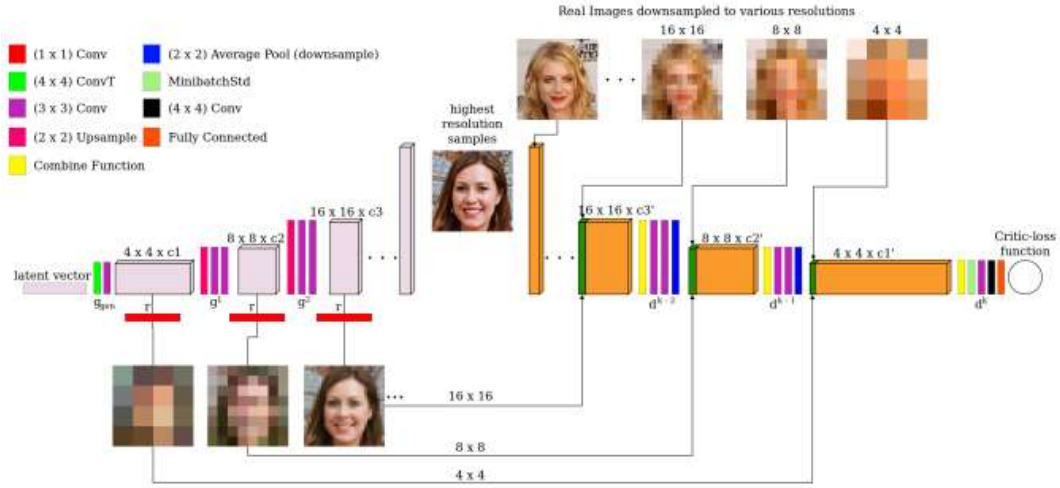


Figure 2: Architecture of MSG-GAN, shown here on the base model proposed in ProGANs [15]. Our architecture includes connections from the intermediate layers of the generator to the intermediate layers of the discriminator. Multi-scale images sent to the discriminator are concatenated with the corresponding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

Reference: https://arxiv.org/pdf/1903.06048.pdf

E. Phase artifact (cont.)

A Baseline StyleGAN [24]
B + Weight demodulation
C + Lazy regularization
D + Path length regularization
E + No growing, new G & D arch.
F + Large networks (StyleGAN2)
Config A with large networks

- **Problem**
  - **이빨이나 눈동자 같은 부분에서 disentanglement가 자연스럽지 않은 현상.**

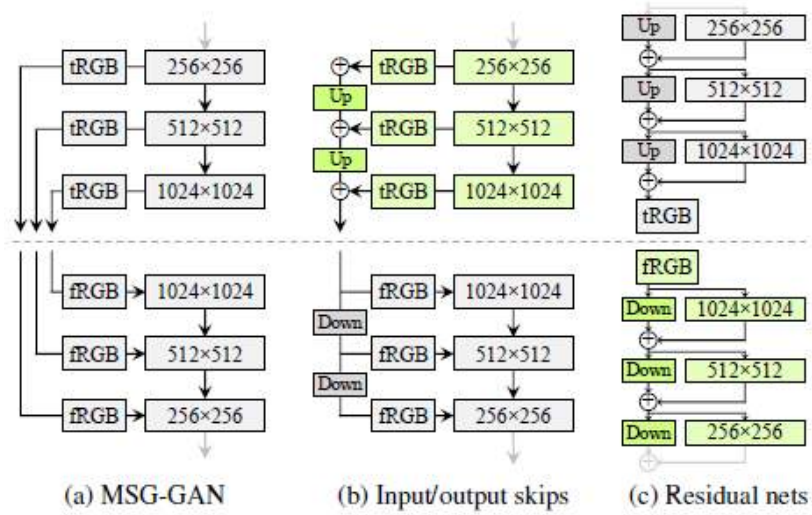


Figure 7. Three generator (above the dashed line) and discriminator architectures. Up and Down denote bilinear up and downsampling, respectively. In residual networks these also include 1×1 convolutions to adjust the number of feature maps. tRGB and fRGB convert between RGB and high-dimensional per-pixel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

FFHQ	D original		D input skips		D residual	
	FID	PPL	FID	PPL	FID	PPL
G original	4.32	265	4.18	235	3.58	269
G output skips	4.33	169	3.77	127	**3.31**	**125**
G residual	4.35	203	3.96	229	3.79	243

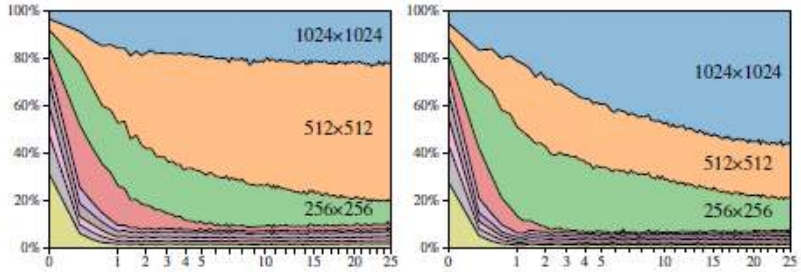
LSUN Car	D original		D input skips		D residual	
	FID	PPL	FID	PPL	FID	PPL
G original	3.75	905	3.23	758	3.25	802
G output skips	3.77	544	3.86	**316**	3.19	471
G residual	3.93	981	3.40	667	**2.66**	645

Table 2. Comparison of generator and discriminator architectures without progressive growing. The combination of generator with output skips and residual discriminator corresponds to configuration E in the main result table.

F. Large Networks

- **StyleGAN + MSG-GAN architecture**
  - **Contribution: 512 x 512 > 1024 x 1024**
- **Large network**
  - **Contribution: 512 x 512 < 1024 x 1024**

A Baseline StyleGAN [24]
B + Weight demodulation
C + Lazy regularization
D + Path length regularization
E + No growing, new G & D arch.
F + Large networks (StyleGAN2)
Config A with large networks



(a) StyleGAN-sized (config E) (b) Large networks (config F)

Figure 8. Contribution of each resolution to the output of the generator as a function of training time. The vertical axis shows a breakdown of the relative standard deviations of different resolutions, and the horizontal axis corresponds to training progress, measured in millions of training images shown to the discriminator. We can see that in the beginning the network focuses on low-resolution images and progressively shifts its focus on larger resolutions as training progresses. In (a) the generator basically outputs a 512^2 image with some minor sharpening for 1024^2, while in (b) the larger network focuses more on the high-resolution details.

Projection of images to latent space

- **Real Image or Generated Image**

A	Baseline StyleGAN [24]
B	+ Weight demodulation
C	+ Lazy regularization
D	+ Path length regularization
E	+ No growing, new G & D arch.
F	+ Large networks (StyleGAN2)
	Config A with large networks

StyleGAN — generated images | StyleGAN2 — generated images | StyleGAN2 — real images

Figure 9. Example images and their projected and re-synthesized counterparts. For each configuration, top row shows the target images and bottom row shows the synthesis of the corresponding projected latent vector and noise inputs. With the baseline StyleGAN, projection often finds a reasonably close match for generated images, but especially the backgrounds differ from the originals. The images generated using StyleGAN2 can be projected almost perfectly back into generator inputs, while projected real images (from the training set) show clear differences to the originals, as expected. All tests were done using the same projection method and hyperparameters.

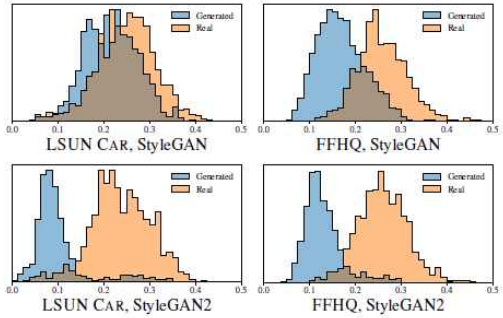


Figure 10. LPIPS distance histograms between original and projected images for generated (blue) and real images (orange). Despite the higher image quality of our improved generator, it is much easier to project the generated images into its latent space $\mathcal{W}$. The same projection method was used in all cases.

Conclusion

- **StyleGAN 개선**
- **Droplet artifact을 제거하기 위해 AdaIN 대신 Weight modulation 제시**
- **Progressive Growing을 제거**
- **Large Network**

THANK YOU!!!

Q & A