[논문리뷰]A Style-Based Generator Architecture for Generative Adversarial Networks

1. Introduction

이 논문에서는 generator 에 다음 3가지를 새로이 디자인하여 이미지의 합성을 이뤄낸다.

1. a learning constant input 으로 시작
2. 서로 다른 크기에 따른 conv 다음에 style 을 추가, 합성
3. noise 를 주입 (stochastic variation)

Input latent code 를 바로 사용하는 것이 아니라 중간에 MLP를 통해 새로운 latent code를 구성함으로써 latent space 가 entanglement 되는 것을 피할 수 있다.

2가지 자동화된 metric 소개 : generator 가 linear 하고 덜 entangle 되었다는 것을 보여주는 지표

1. perceptual path length
2. linear separability

새로운 고해상도 사람 얼굴 데이터 세트 : FFHQ 제시

2. Style-based generator

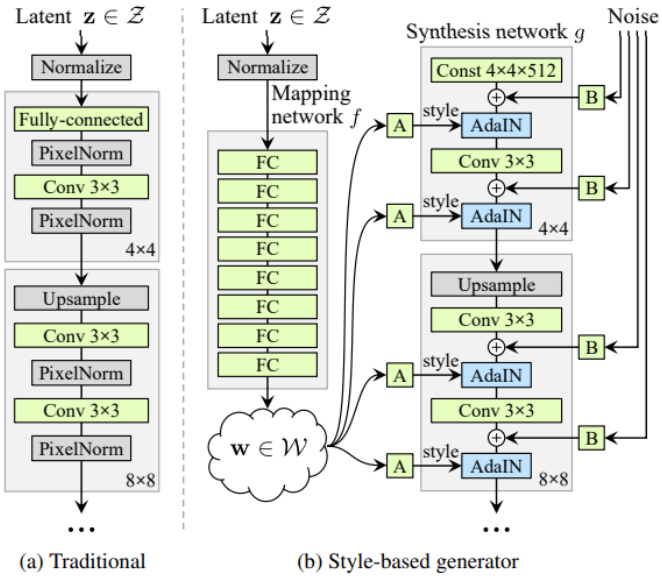


Figure 1.(a) 에서 알 수 있듯 전통적인 방법은 latent code를 바로 input으로 넣는 방식을 이용한다. 그러나 논문에서 제시한 방식인 (b) 는 latent code z 가 non-linear mapping f : Z → W 를 통해 W 를 생성한다. 이때 Mapping function 은 8개의 MLP layer로 구성되고 결과 차원은 512 이다.

이러한 방식은 entanglement 되는 것을 피할 수 있다.

생성된 intermediate latent vector W 는 이후 affine tranformation 을 진행하여 shape 를 맞춘다. 해당 결과와 합성 네트워크의 결과값을 기반으로 AdaIN 을 적용함으로써 style을 입힌다.

즉 각 feature 정보의 scaling 과 biaas 를 변화시켜 입력 스타일로 분포를 바꿀 수 있게 해준다.

마지막으로 noise input 을 새롭게 만들어 합성 네트워크의 layer에 추가한다.

	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**

위 표가 결과인데 점진적인 추가에 따라 FID 가 성능이 향상됨을 확인할 수 있다.

3. Properties of the style-based generator

3.1 Style mixing

Style 에 대한 localize 를 위해 mixing regularization 을 수행한다. 기존에 하나의 latent code 만을 사용하는 것과 달리 여기서는 2개의 랜덤 latent code를 사용한다.

1. Latent code z_1, z_2 에 대해 mapping network 를 진행하고 이에 상응하는 w_1, w_2 를 생성한다.
2. Cross point 이전에는 w_1 을 적용하고 그 이후에는 w_2 를 적용한다.

이를 통해 인접한 style들끼리 correlate 되는 것을 방지하고 regularization 효과를 보인다. 또한 각 layer 와 block 이 특정 부분의 스타일에 대한 구분되도록 진행된다.

Figure 3. Two sets of images were generated from their respective latent codes (sources A and B); the rest of the images were generated by copying a specified subset of styles from source B and taking the rest from source A. Copying the styles corresponding to coarse spatial resolutions ($4^2 - 8^2$) brings high-level aspects such as pose, general hair style, face shape, and eyeglasses from source B, while all colors (eyes, hair, lighting) and finer facial features resemble A. If we instead copy the styles of middle resolutions ($16^2 - 32^2$) from B, we inherit smaller scale facial features, hair style, eyes open/closed from B, while the pose, general face shape, and eyeglasses from A are preserved. Finally, copying the fine styles ($64^2 - 1024^2$) from B brings mainly the color scheme and microstructure.

(1) Coarse Style : pose, hair, face shape

(2) Middle Style : facial features, eyes

(3) Fine Style : color scheme

3.2 Stochastic variation

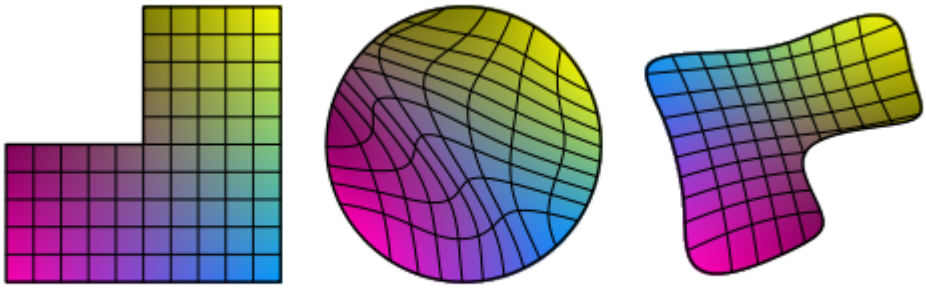
머리결, 주근깨, 모공 등 인지에는 크게 영향을 끼치며 랜덤하게 생성되는 부분을 stochastic 으로 간주한다.

본 논문에서는 stochastic variation 을 위해 각 conv 다음 per-pixel noise 를 추가한다.

(a) Generated image (b) Stochastic variation (c) Standard deviation

Figure 4. Examples of stochastic variation. (a) Two generated images. (b) Zoom-in with different realizations of input noise. While the overall appearance is almost identical, individual hairs are placed very differently. (c) Standard deviation of each pixel over 100 different realizations, highlighting which parts of the images are affected by the noise. The main areas are the hair, silhouettes, and parts of background, but there is also interesting stochastic variation in the eye reflections. Global aspects such as identity and pose are unaffected by stochastic variation.

4. Disentanglement studies



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

Disentanglement 란 하나의 변동 요소를 제어할 수 있는 선형 부분 공간들로 그성된 latent space를 의미한다. 그러나 제약된 분포에서 훈련 데이터 세트에서의 분포를 바로 따라야 하는 일반적인 mapping에서는 disentanglement 가 어렵다.

그러나 제안된 mapping function 을 통해 생성한 intermediate latent space W 는 고정되니 분포로부터 sampling 을 바로 하지 않기 때문에 선형적인 분포를 생성할 수 있다.

4.1 Perceptual path length

latent space vector 에서 interpolation 의 linearity 가 없다면 이는 entangle된 것을 의미한다. 즉 두 지점에 대해서 중간 지점에 급격한 변화가 있거나 존재하지 않는다면 entangle된 것이다.

4.2 Linear seperability

Latent space point 들이 linear hyperplane 을 통해 서로 다른 두 집합으로 잘 분리되는지를 정량화하는 지표를 제안한다.

Discriminator 와 같은 구조의 classifier 를 사용하여 40개의 속성이 보유된 CelebA-HQ 를 이용하여 40개의 분류 모델을 학습한다.

이때 20만개의 이미지를 생성 후 분류 모델을 통해 분류한다.

confidence 가 낮은 절반을 제거하고 10만개의 label된 latent-space vector를 생성한다.

각 속성마다 linear SVM 을 학습하고 각 속성별로 conditional entropy를 구한 separability score 를 구한다.

	Method	Path length full	Path length end	Separability
B	Traditional generator $\mathcal{Z}$	412.0	415.3	10.78
D	Style-based generator $\mathcal{W}$	446.2	376.6	3.61
E	+ Add noise inputs $\mathcal{W}$	**200.5**	**160.6**	3.54
	+ Mixing 50% $\mathcal{W}$	231.5	182.1	**3.51**
F	+ Mixing 90% $\mathcal{W}$	234.0	195.9	3.79

Table 3. Perceptual path lengths and separability scores for various generator architectures in FFHQ (lower is better). We perform the measurements in $\mathcal{Z}$ for the traditional network, and in $\mathcal{W}$ for style-based ones. Making the network resistant to style mixing appears to distort the intermediate latent space $\mathcal{W}$ somewhat. We hypothesize that mixing makes it more difficult for $\mathcal{W}$ to efficiently encode factors of variation that span multiple scales.

	Method	FID	Path length full	Path length end	Separability
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**

Table 4. The effect of a mapping network in FFHQ. The number in method name indicates the depth of the mapping network. We see that FID, separability, and path length all benefit from having a mapping network, and this holds for both style-based and traditional generator architectures. Furthermore, a deeper mapping network generally performs better than a shallow one.

Table 3는 Z 가 W보다 더 선형적이고 disentanglement 하다는 것을 알 수 있다. (Path length, Separability 낮음)

Table 4 는 Mapping Network 가 disentanglement 하다.