2021

인공지능세미나

정보찬

AILab.

2021.07.10.

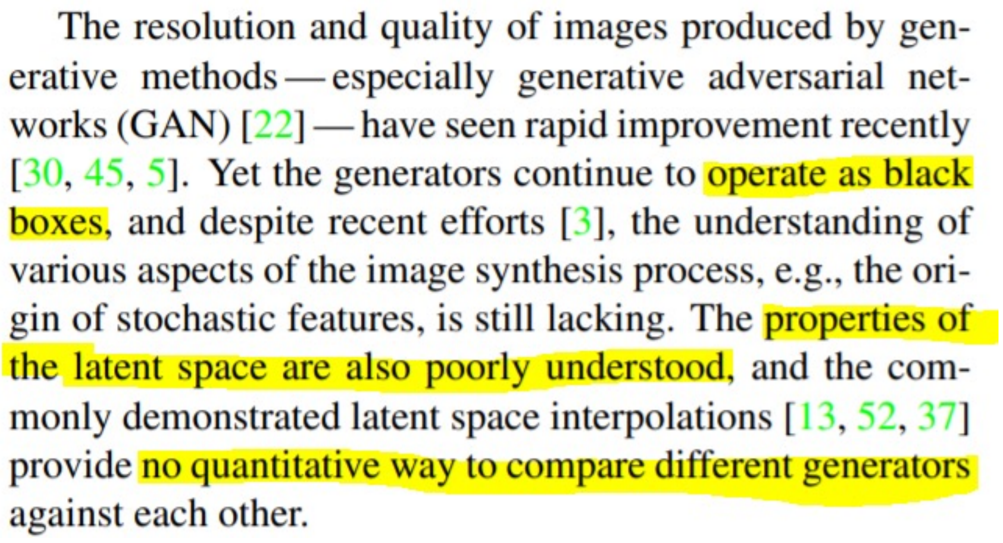
주식회사 바스젠바이오

Style GAN

1. Introduction
2. Style-based generator
3. Properties of the style-based generator
4. Disentanglement studies
5. Appendix

1 Introduction

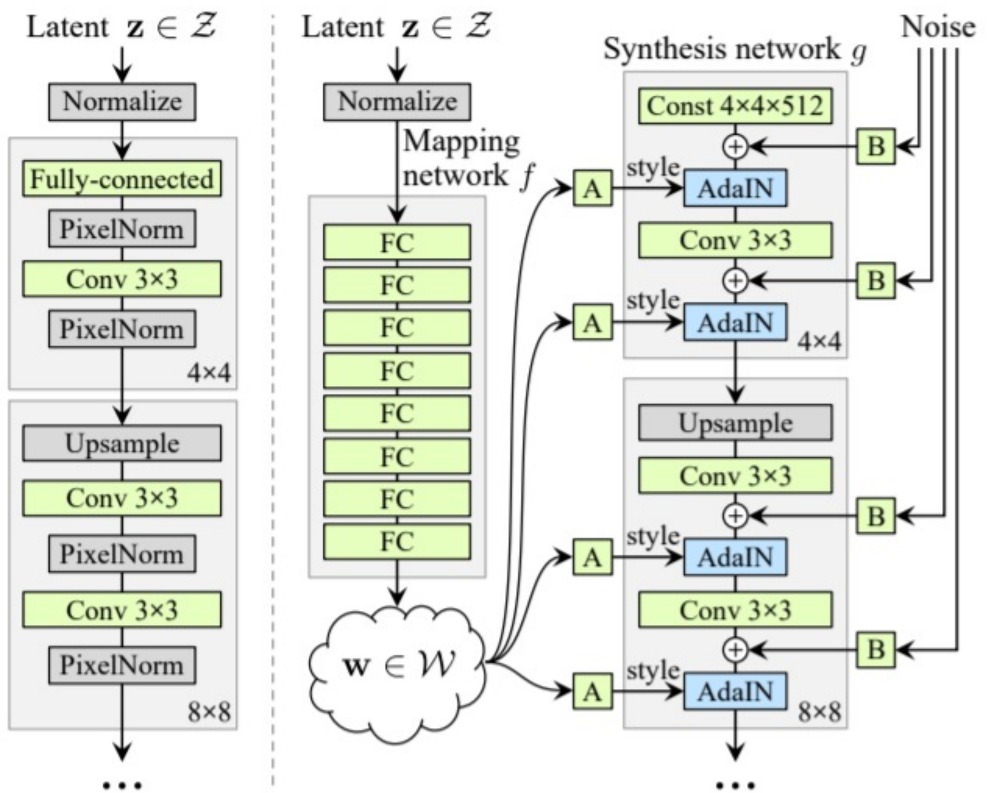
- 기존 GAN의 단점



The resolution and quality of images produced by generative methods — especially generative adversarial networks (GAN) [22] — have seen rapid improvement recently [30, 45, 5]. Yet the generators continue to operate as black boxes, and despite recent efforts [3], the understanding of various aspects of the image synthesis process, e.g., the origin of stochastic features, is still lacking. The properties of the latent space are also poorly understood, and the commonly demonstrated latent space interpolations [13, 52, 37] provide no quantitative way to compare different generators against each other.

- generator가 black box
- latent space의 성질에 대한 이해가 부족
- latent space interpolations에 대한 정량적인 비교 방법이 없음

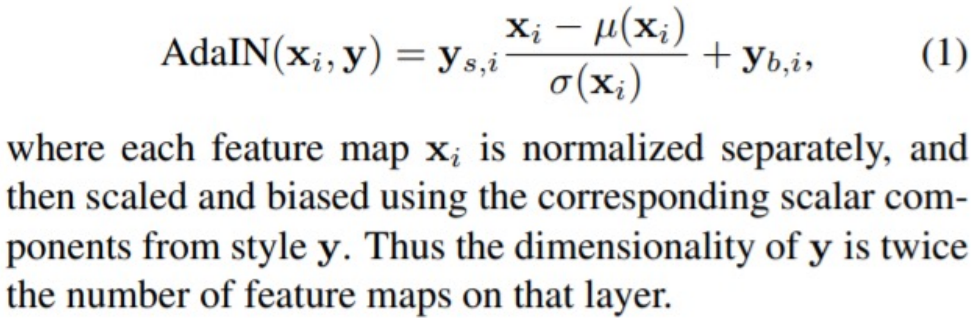
2 Style-based generator



(a) Traditional (b) Style-based generator

- latent tensor가 Mapping network를 통해 intermediate latent space를 거침
- learned constant tensor로부터 생성
- AdaIN 사용
- bilinear up/down sampling 사용
- Gaussian noise를 추가
- mixing regularization

2 Style-based generator



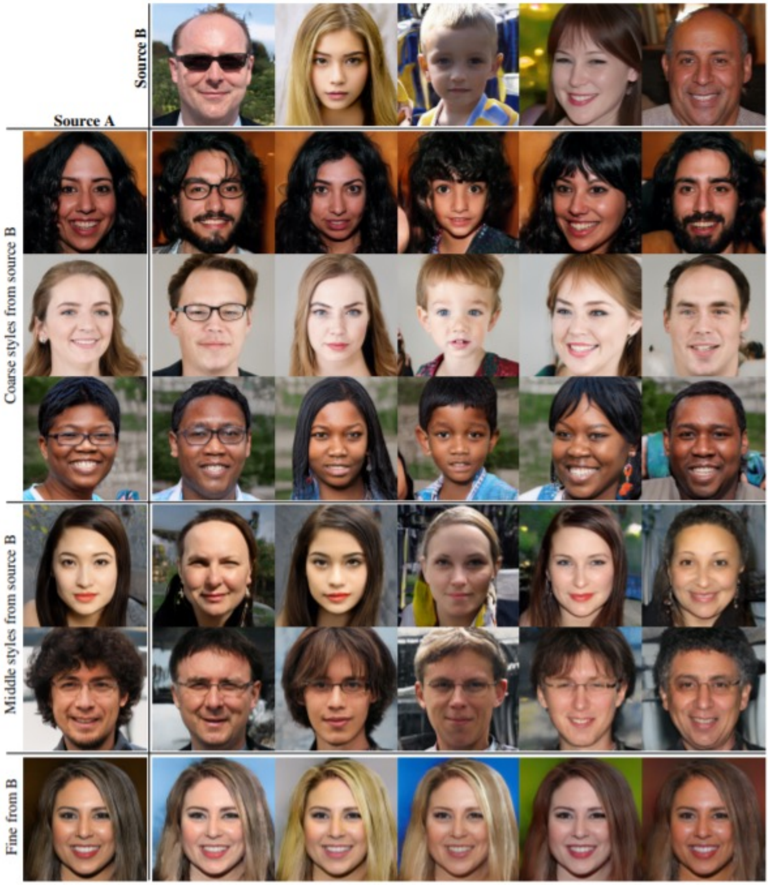
$$\mathrm{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i}\frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i}, \qquad (1)$$

where each feature map $\mathbf{x}_i$ is normalized separately, and then scaled and biased using the corresponding scalar components from style $\mathbf{y}$. Thus the dimensionality of $\mathbf{y}$ is twice the number of feature maps on that layer.

- 이미지마다 normalization
- x의 통계량 변화를 통해 스타일을 줌
- style transfer에 효과적이라 알려져 있음

3 Properties of the style-based generator

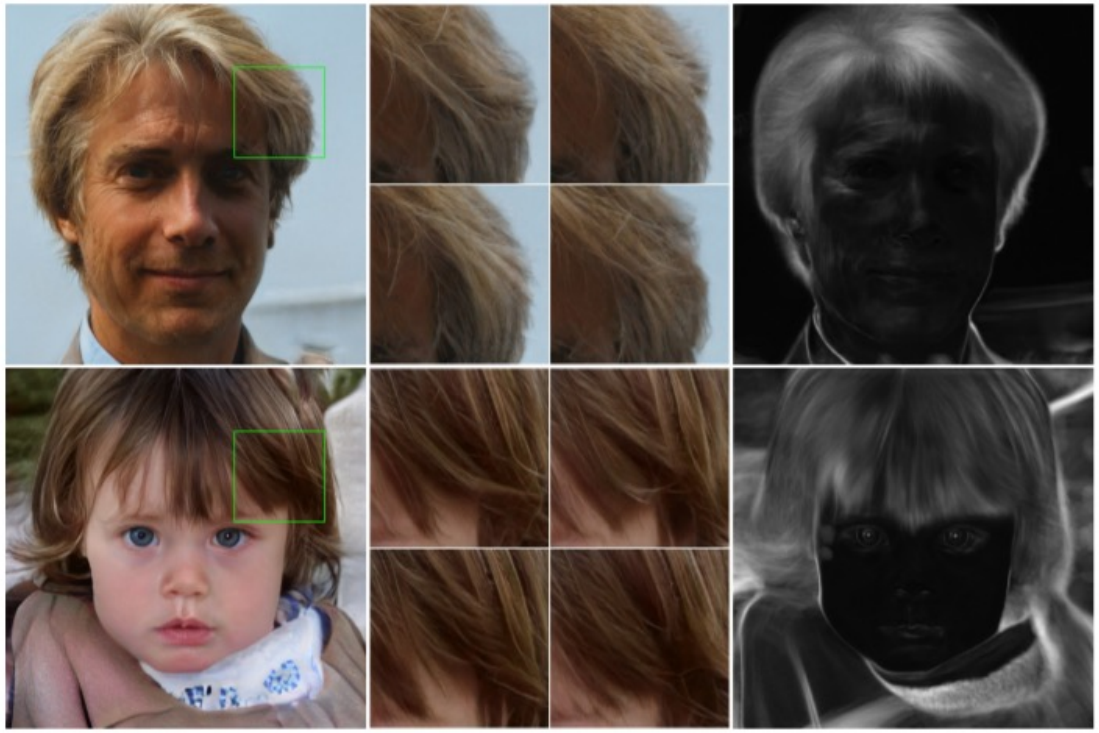
- style mixing



- 두 스타일 텐서w1, w2로부터 일부 레이어엔 w1을, 나머지엔 w2를 사용
- 인접한 스타일끼리의 correlated를 막음
- style을 localize함

3 Properties of the style-based generator

- stochastic variation

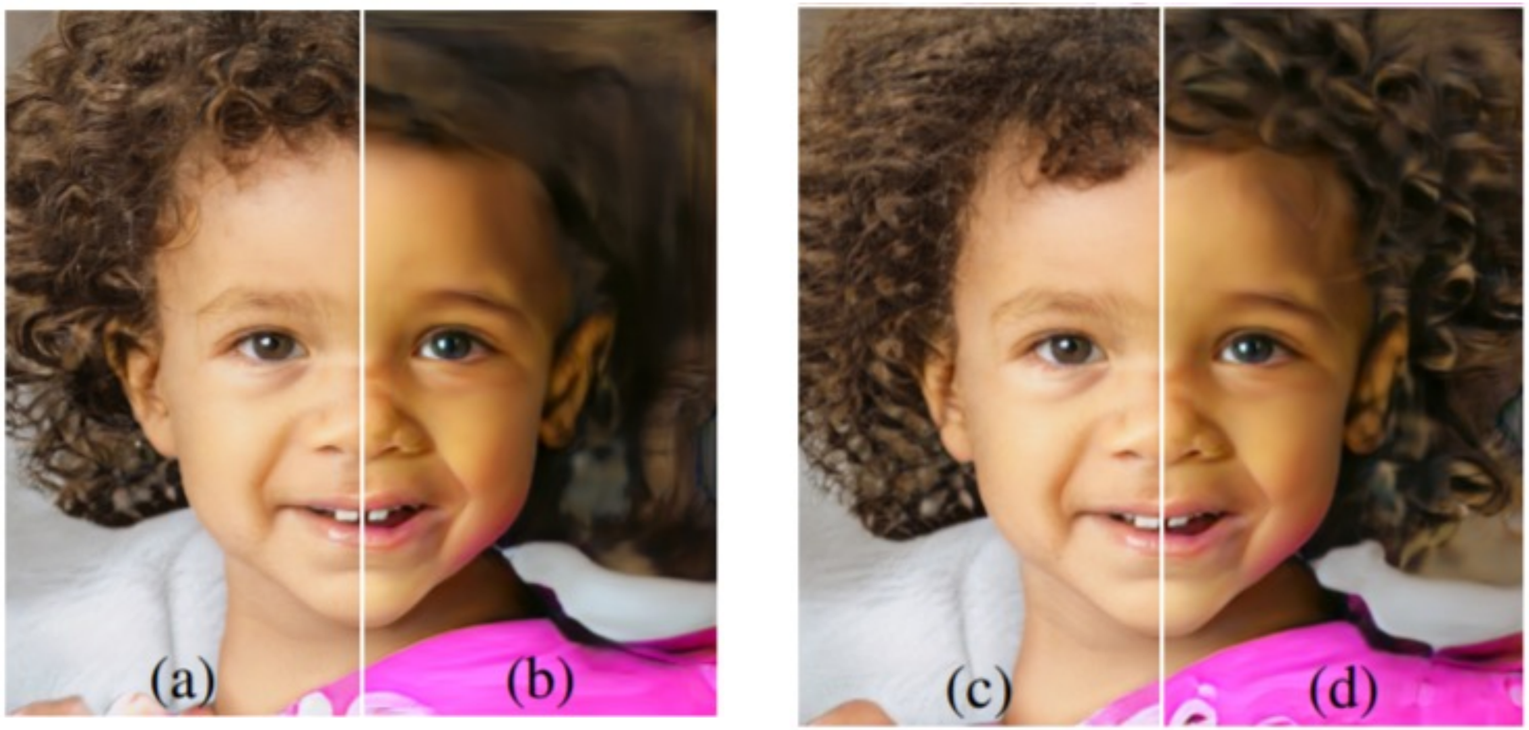


(a) Generated image (b) Stochastic variation (c) Standard deviation

- Noise를 통한 stochastic variation
- 머리카락, 주근깨 등 그날의 상태에 따라서 다를 수 있는 것을 표현

3 Properties of the style-based generator

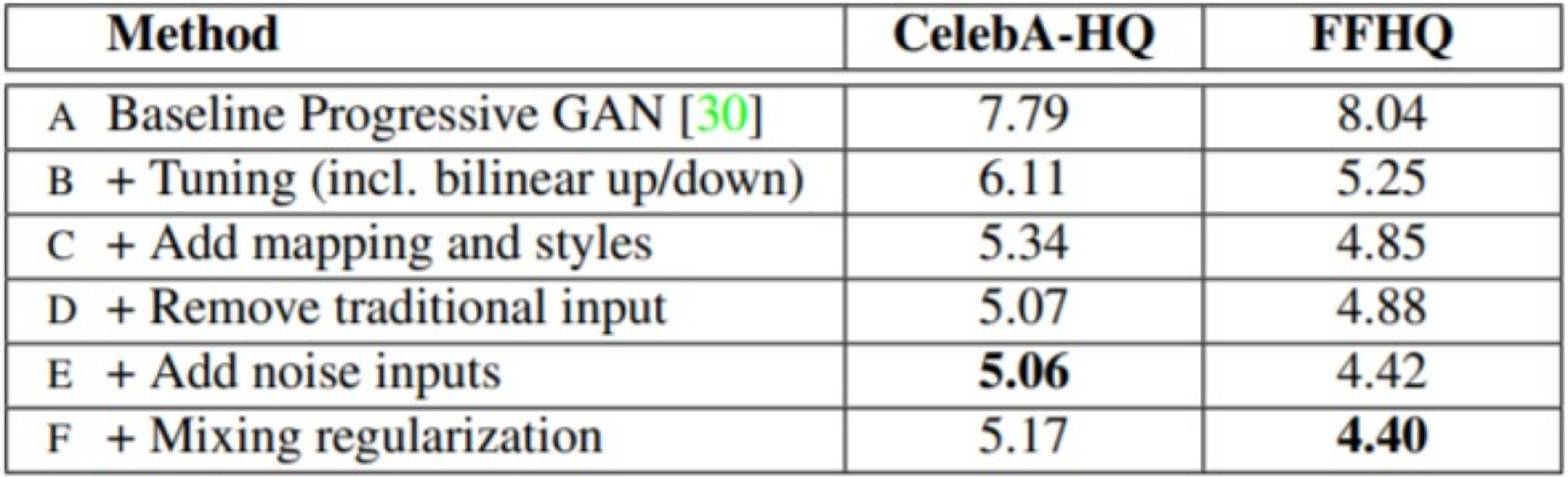
- stochastic variation



a : 모든 레이어에 noise, b : noise 사용 안함,

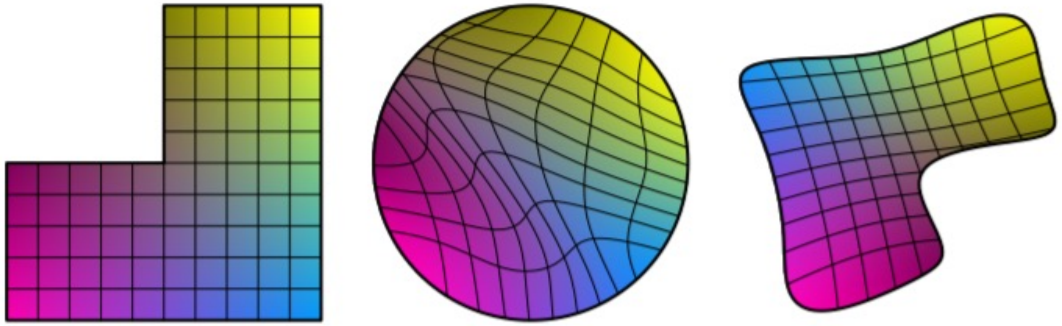
c : 64~1024(finer)레이어에 noise, d : 4~32(coarser)레이어에 noise

3 Properties of the style-based generator



	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**

4 Disentanglement studies



(a) Distribution of features in training set

(b) Mapping from $\mathcal{Z}$ to features

(c) Mapping from $\mathcal{W}$ to features

Figure 6. Illustrative example with two factors of variation (image features, e.g., masculinity and hair length). (a) An example training set where some combination (e.g., long haired males) is missing. (b) This forces the mapping from $\mathcal{Z}$ to image features to become curved so that the forbidden combination disappears in $\mathcal{Z}$ to prevent the sampling of invalid combinations. (c) The learned mapping from $\mathcal{Z}$ to $\mathcal{W}$ is able to "undo" much of the warping.

- disentanglement <=> latent space consists of linear subspaces
- 기존 방법(b)은 특정 분포(가우시안)을 따르기 때문에 학습 데이터의 분포와 다르지만, 새로운 방법(c)는 특정 분포를 따를 필요가 없음

4 Disentanglement studies

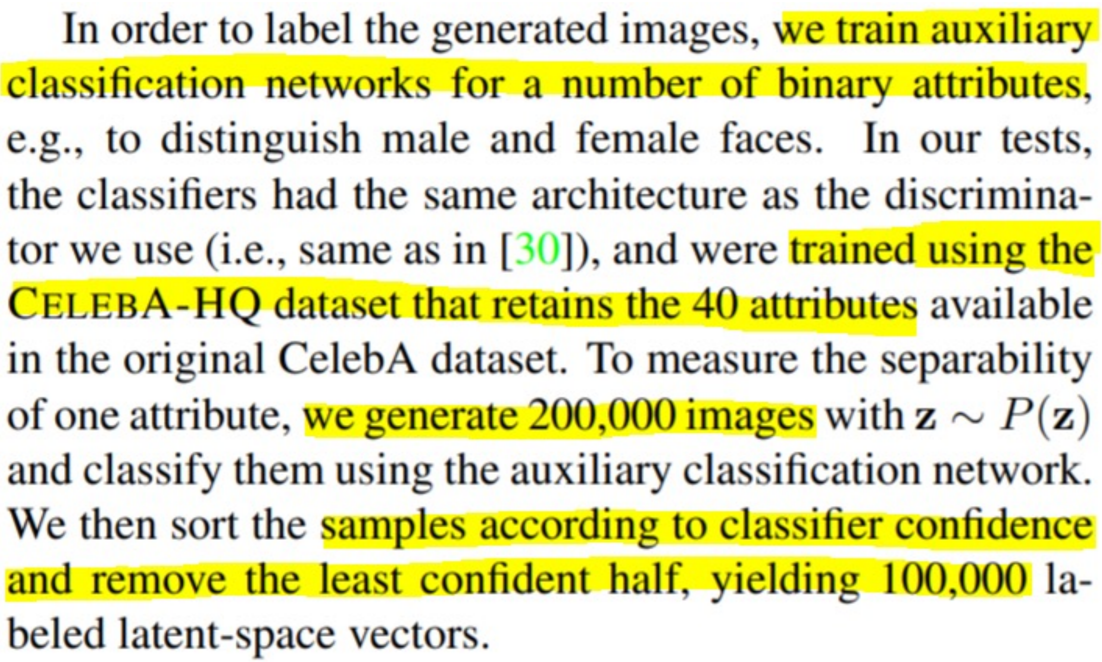
- Perceptual path length

$$
\begin{aligned}
l_{\mathcal{Z}} = \mathbb{E}\Big[\frac{1}{\epsilon^2} d\big(&G(\mathrm{slerp}(\mathbf{z}_1, \mathbf{z}_2; t)),\\
&G(\mathrm{slerp}(\mathbf{z}_1, \mathbf{z}_2; t+\epsilon))\big)\Big],
\end{aligned}
$$

$$
\begin{aligned}
l_{\mathcal{W}} = \mathbb{E}\Big[\frac{1}{\epsilon^2} d\big(&g(\mathrm{lerp}(f(\mathbf{z}_1), f(\mathbf{z}_2); t)),\\
&g(\mathrm{lerp}(f(\mathbf{z}_1), f(\mathbf{z}_2); t+\epsilon))\big)\Big],
\end{aligned}
$$

4 Disentanglement studies

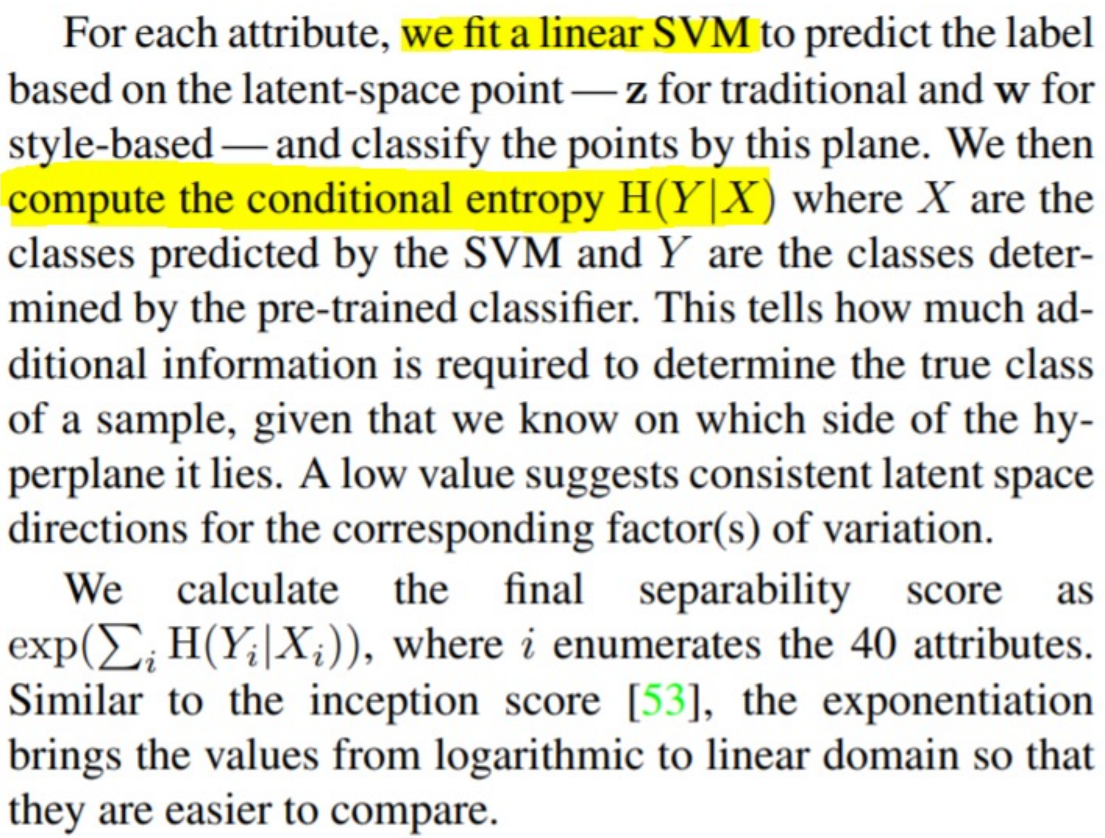
- Linear separability



In order to label the generated images, we train auxiliary classification networks for a number of binary attributes, e.g., to distinguish male and female faces. In our tests, the classifiers had the same architecture as the discriminator we use (i.e., same as in [30]), and were trained using the CELEBA-HQ dataset that retains the 40 attributes available in the original CelebA dataset. To measure the separability of one attribute, we generate 200,000 images with $\mathbf{z} \sim P(\mathbf{z})$ and classify them using the auxiliary classification network. We then sort the samples according to classifier confidence and remove the least confident half, yielding 100,000 labeled latent-space vectors.

4 Disentanglement studies

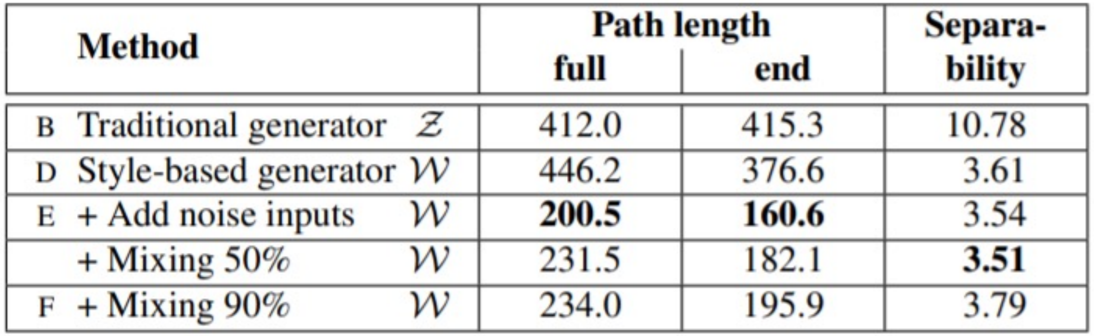
- Linear separability



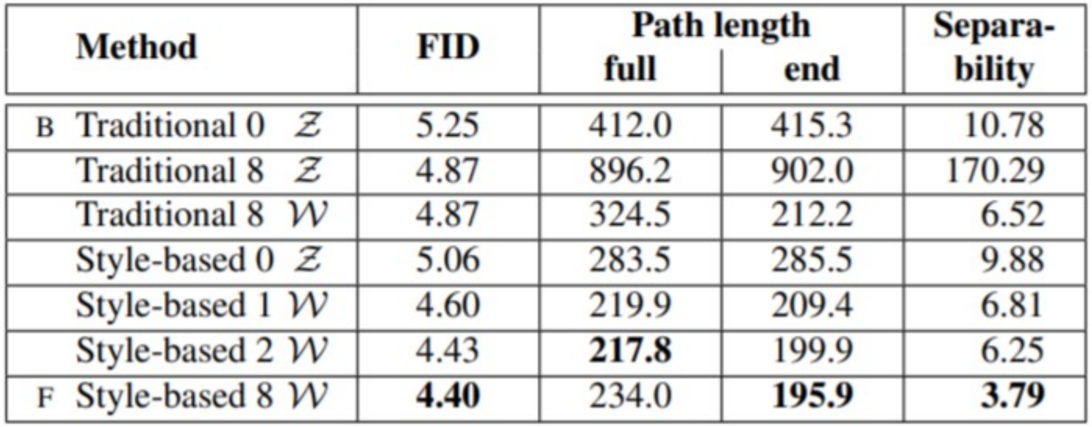
For each attribute, we fit a linear SVM to predict the label based on the latent-space point — **z** for traditional and **w** for style-based — and classify the points by this plane. We then compute the conditional entropy $\mathrm{H}(Y|X)$ where X are the classes predicted by the SVM and Y are the classes determined by the pre-trained classifier. This tells how much additional information is required to determine the true class of a sample, given that we know on which side of the hyperplane it lies. A low value suggests consistent latent space directions for the corresponding factor(s) of variation.

We calculate the final separability score as $\exp(\sum_i \mathrm{H}(Y_i|X_i))$, where i enumerates the 40 attributes. Similar to the inception score [53], the exponentiation brings the values from logarithmic to linear domain so that they are easier to compare.

4 Disentanglement studies

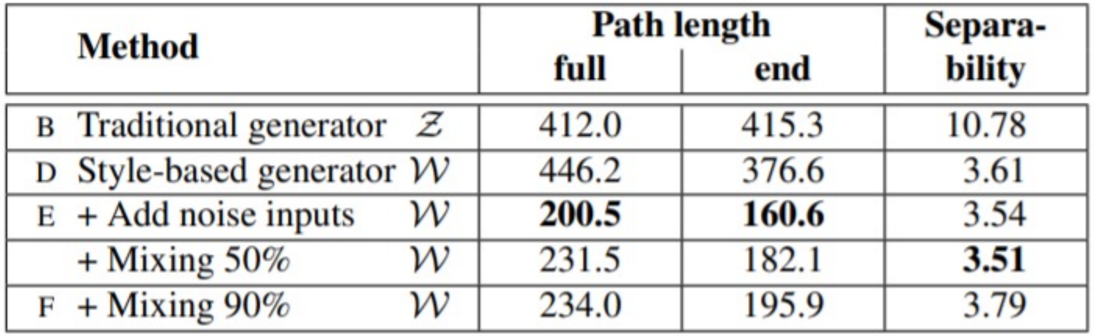


	Method		Path length		Separa-bility
			full	end	
B	Traditional generator	$\mathcal{Z}$	412.0	415.3	10.78
D	Style-based generator	$\mathcal{W}$	446.2	376.6	3.61
E	+ Add noise inputs	$\mathcal{W}$	**200.5**	**160.6**	3.54
	+ Mixing 50%	$\mathcal{W}$	231.5	182.1	**3.51**
F	+ Mixing 90%	$\mathcal{W}$	234.0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

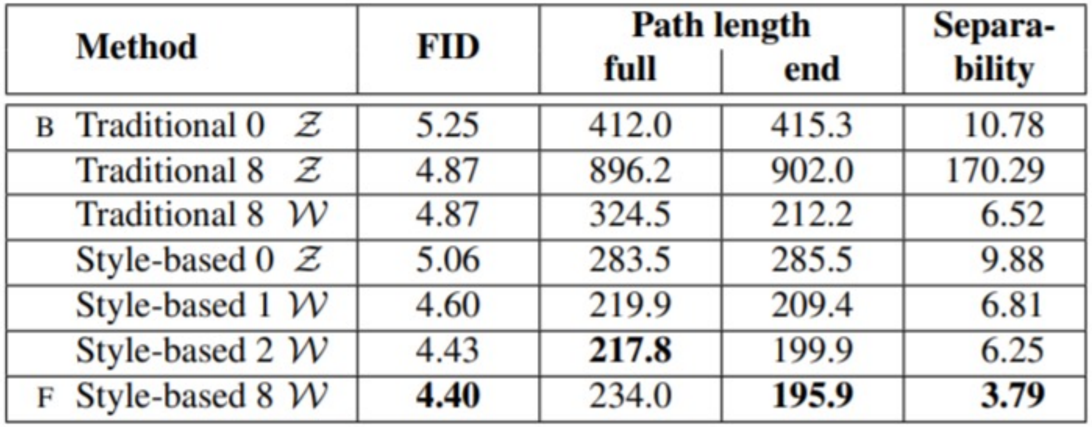


	Method		FID	Path length		Separa-bility
				full	end	
B	Traditional 0	$\mathcal{Z}$	5.25	412.0	415.3	10.78
	Traditional 8	$\mathcal{Z}$	4.87	896.2	902.0	170.29
	Traditional 8	$\mathcal{W}$	4.87	324.5	212.2	6.52
	Style-based 0	$\mathcal{Z}$	5.06	283.5	285.5	9.88
	Style-based 1	$\mathcal{W}$	4.60	219.9	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**

4 Disentanglement studies



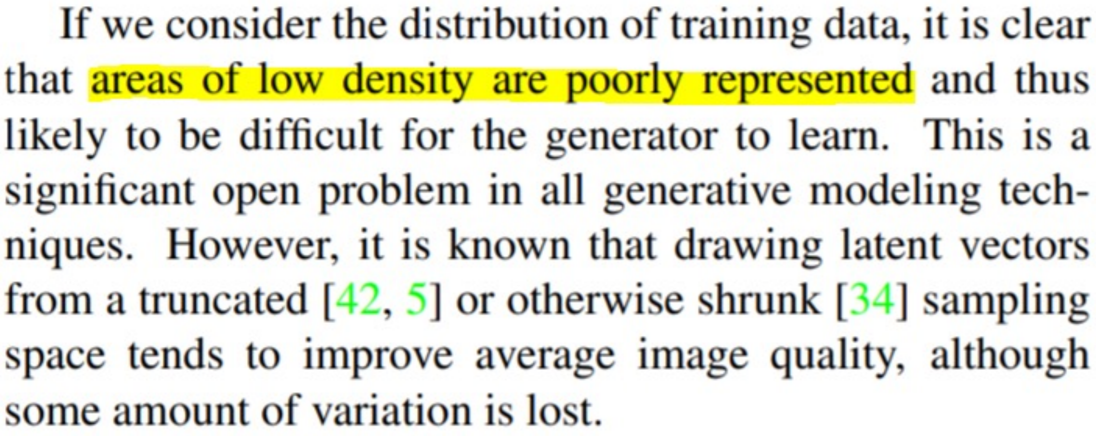
	Method		Path length		Separa-bility
			full	end	
B	Traditional generator	$\mathcal{Z}$	412.0	415.3	10.78
D	Style-based generator	$\mathcal{W}$	446.2	376.6	3.61
E	+ Add noise inputs	$\mathcal{W}$	**200.5**	**160.6**	3.54
	+ Mixing 50%	$\mathcal{W}$	231.5	182.1	**3.51**
F	+ Mixing 90%	$\mathcal{W}$	234.0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



	Method		FID	Path length		Separa-bility
				full	end	
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**

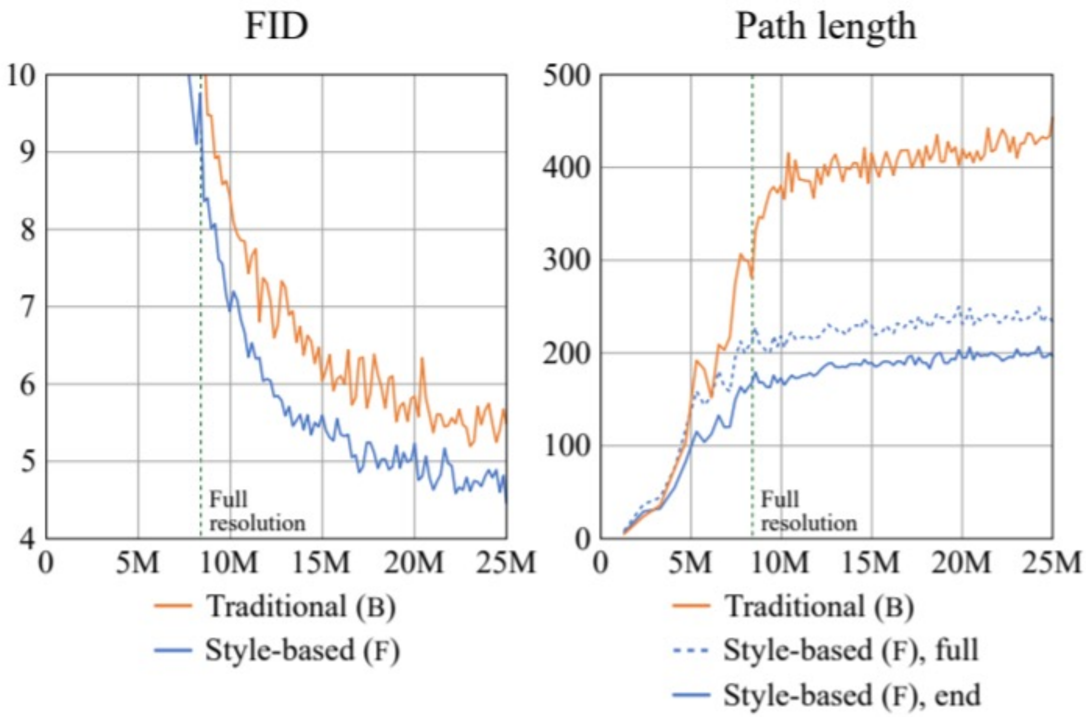
5 Appendix

- Truncation trick in W



If we consider the distribution of training data, it is clear that areas of low density are poorly represented and thus likely to be difficult for the generator to learn. This is a significant open problem in all generative modeling techniques. However, it is known that drawing latent vectors from a truncated [42, 5] or otherwise shrunk [34] sampling space tends to improve average image quality, although some amount of variation is lost.

- 평균에 가까운 애들을 고르면 좋은 이미지를 얻을 수 있다!



Full resolution이 되고 나서는 FID는 점점 줄어드는데, path length는 점점 늘어남.

FID와 entanglement의 trade off 발생

5 Appendix

- coarse => 카메라 구도
- middle => 가구 배치
- fine => 색상, 재질