Face Generation

2023년도 하계 세미나

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

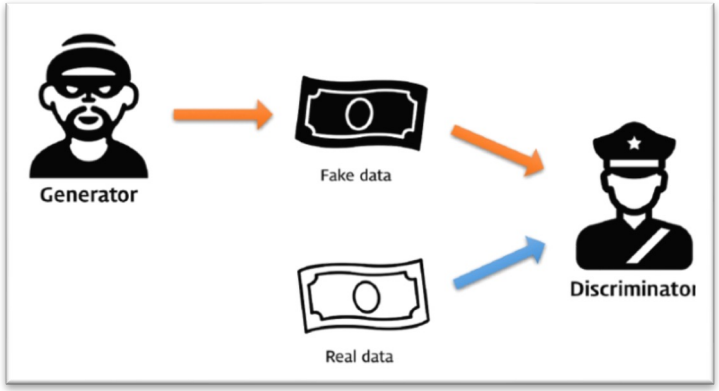
Presented By
Seo Young Oh

Outline

- Background
  - Generative Adversarial Network(GAN)
  - Swin Transformer
- A Style-Based Generator Architecture for GANs
  - CVPR 2019
- StyleSwin: Transformer-based GAN for High-resolution Image Generation
  - CVPR 2022

Background

- GAN
  - Generator와 Discriminator를 적대적으로 학습시키며 train data와 비슷한 이미지를 생성



GAN Architecture

$$\min_G \max_D V(D,G)$$

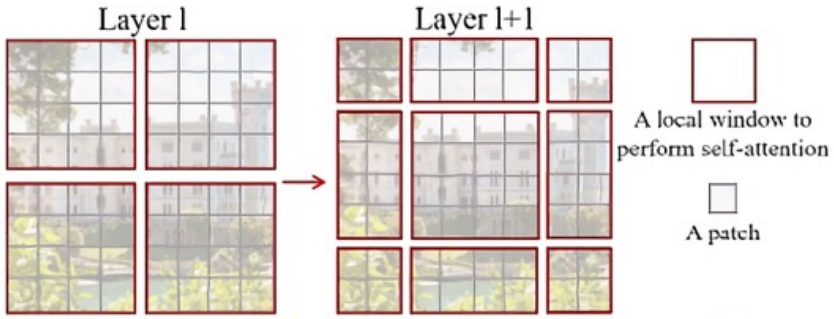
$$V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z))]$$

무조건 1로 판별해야 하는 값 — 무조건 1로 판별해야 하는 값

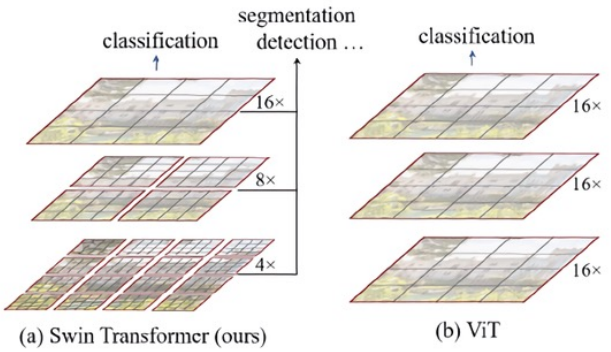
[MinMax Loss]

Background

- Swin Transformer
  - 기존 Vision 분야에서 Transformer based approach가 가진 한계
    - Computational complexity on high-resolution images
    - 이미지의 해상도가 커질 수록 모든 patch의 조합에 대해 self-attention을 수행하는 것이 불가능해짐
  - 이를 Shifted Windows와 Hierarchical을 통해 해결
    - (Shifted Windows) Layer l의 분할 위치에서 $(\lfloor \frac{M}{2}, \frac{M}{2} \rfloor)$칸 떨어진 위치에서 Layer l의 window 분할
    - 이전 Layer의 window와 현재 Layer의 window 사이를 이어주며 모델의 성능을 효과적으로 향상
    - (Hierarchical) 기존의 VIT가 이미지를 작은 patch들로 쪼개는 방향으로 진행된다면, Swin Transformer는 더 작은 단위의 patch부터 시작해서 점점 merge해 나감



[Shifted Windows]



[Hierarchical]

- **A Style-Based Generator Architecture for Generative Adversarial Networks**
  - StyleGAN
  - CVPR 2019

Introduction

- PGGAN 포함 기존의 이미지 합성 기술이 가진 한계
  - 내부 과정이 Black box임
    - 합성되는 이미지의 attribute를 조절하기 어려움
    - 생성된 이미지의 퀄리티가 불안정함

[PGGAN으로 생성된 이미지 예시]

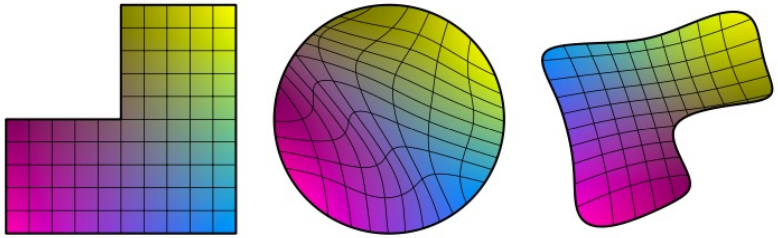
Introduction

- StyleGAN은 Style Transfer에 기반한 새로운 Generator 구조를 제안함
  - 이미지를 style의 조합으로 보고, generator의 각 layer마다 style 정보를 입히는 방식으로 이미지를 합성
  - 추가되는 style은 이미지의 coarse feature부터 fine detail까지 각기 다른 레벨을 조절
  - 훨씬 안정적이고 높은 퀄리티의 이미지를 생성

[StyleGAN으로 생성된 이미지 예시]

Style-based generator

- Mapping Network
  - 기존의 방식은 Input vector Z로부터 이미지를 직접 생성함
  - StyleGAN은 mapping network를 거쳐 intermediate vector W로 먼저 변환한 후 이미지를 생성함
    - Entangle은 여러 feature가 섞여 있어 구분하기 어려운 상태를 말함
    - 반대로 Disentangle은 각 feature가 잘 구분되어 있는 상태임
      - StyleGAN은 mapping network를 통해 entangle한 latent space를 disentangle하게 만들 수 있음



(a) Distribution of features in training set

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

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

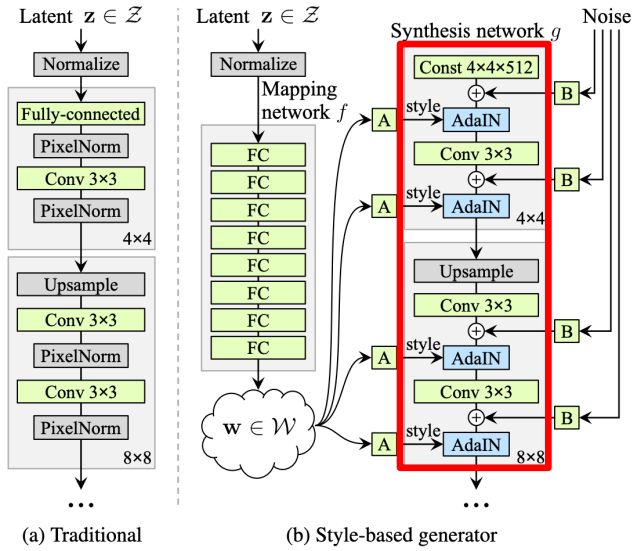
[Mapping Network 예시]

Style-based generator

- AdaIN(Adaptive Instance Normalization)
  - AdaIN을 활용해 각 Layer에 style 정보를 추가함
    - 파라미터 학습이 필요하지 않으며 style transfer 네트워크에서 좋은 성능을 보임
    - Feature map에 대해 instance normalization을 수행한 뒤 style y의 scalar components를 통해 스케일링됨

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

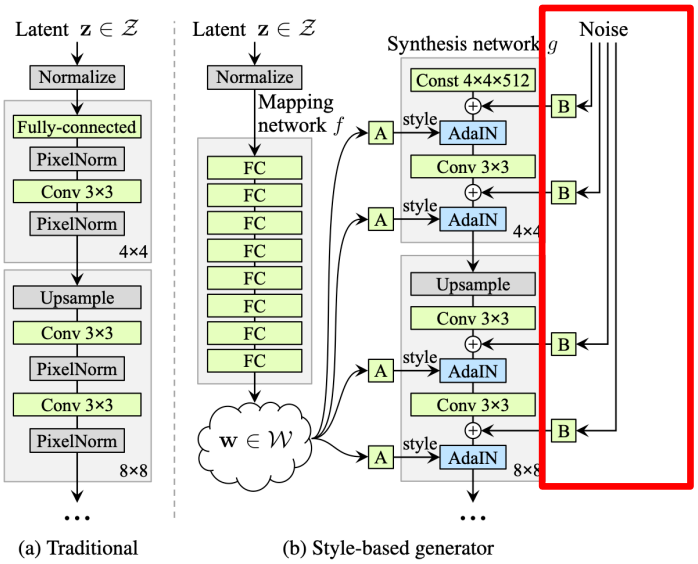
[AdaIN 수식]



[AdaIN 구조]

Style-based generator

- Stochastic Variation
  - 이미지에 포함되는 다양한 stochastic aspects를 노이즈로 조절
    - 기존의 네트워크에서는 초기값에서 다양한 random vector를 생성하는 방식만 가능
    - StyleGAN은 Synthesis network의 각 layer마다 random noise를 추가함으로써 이를 해결
      - 이를 통해 더욱 사실적인 이미지를 생성
      - 전체적인 구조나 대상은 바뀌지 않고 확률적 요소만 변화하게 됨



[Noise 구조]

Style-based generator

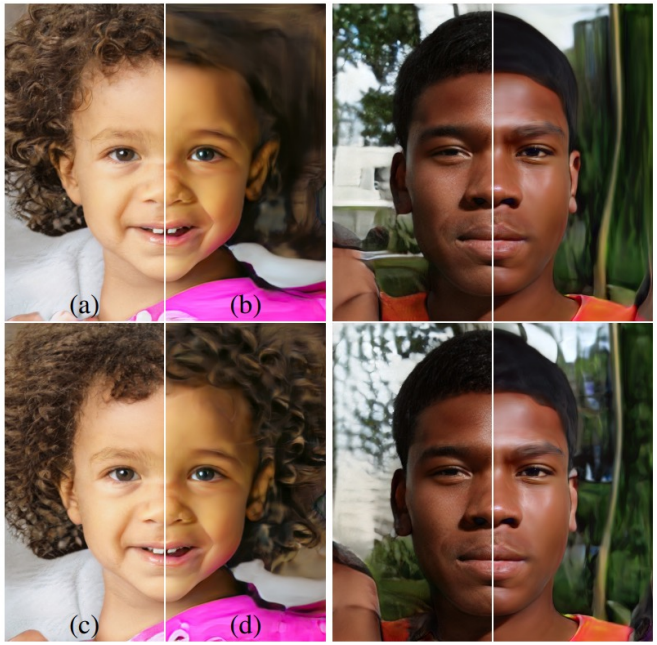
- Stochastic Variation
  - 이미지에 포함되는 다양한 stochastic aspects를 noise로 조절
    - 기존의 네트워크에서는 초기값에서 다양한 random vector를 생성하는 방식만 가능
    - StyleGAN은 Synthesis network의 각 layer마다 random noise를 추가함으로써 이를 해결
      - 이를 통해 더욱 사실적인 이미지를 생성
      - 전체적인 구조나 대상은 바뀌지 않고 확률적 요소만 변화하게 됨

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

[Noise 적용 예시]

Style-based generator

- Stochastic Variation
  - 이미지에 포함되는 다양한 stochastic aspects를 noise로 조절
    - 기존의 네트워크에서는 초기값에서 다양한 random vector를 생성하는 방식만 가능
    - StyleGAN은 Synthesis network의 각 layer마다 random noise를 추가함으로써 이를 해결
      - 이를 통해 더욱 사실적인 이미지를 생성
      - 전체적인 구조나 대상은 바뀌지 않고 확률적 요소만 변화하게 됨



(a) All layers
(b) No noise
(c) Fine layers
(d) Coarse layers

[Noise 적용 예시]

Experiments

	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.

	Method	Path length		Separability
		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

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		Separability
			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**

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.

- **StyleSwin: Transformer-based GAN for High-resolution Image Generation**
  - StyleSwin
  - CVPR 2022

Introduction

- 기존 GAN 연구
  - 초기 연구는 GAN 학습을 안정화하는 것에 집중
    - 적절한 regularization이나 adversarial loss design을 활용
  - 최근 연구는 architectural modification을 통해 성능을 개선
    - Self-attention, style-based generators, ect.
    - Discriminator에 transformer를 적용한 연구에 힘입어 generator에도 transformer를 적용하려는 시도

- StyleSwin 구조를 제안
  - Basic building block으로 Swin transformer를 활용하여 computational cost를 줄임
  - 모든 image scale에 적용 가능한 expressivity를 보임
    - 이미지 합성이 높은 해상도에서도 scalable함 i.e. 1024×1024
  - Local attention에서 locality inductive bias를 도입하여 재학습할 필요가 없음

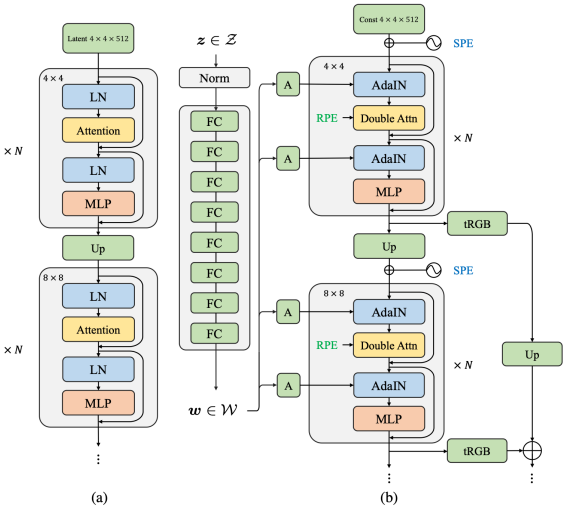
Introduction

- Three instrumental architectural adaptions
  - Style-based architecture에 local attention을 적용
  - Double attention을 통해 receptive field를 키움
    - Computational overhead 없이 generator capacity를 향상
  - sinusoidal positional encoding을 통해 이미지의 absolute position을 유지

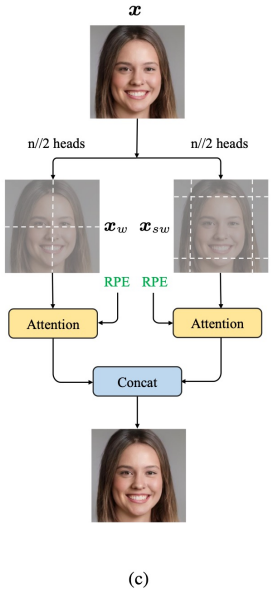
- 고해상도 이미지 합성 과정에서 blocking artifacts가 관찰됨
  - Wavelet discriminator를 통해 이를 개선

Method

- Overall architecture
  - Swin transformer
    - Basic building block으로 활용
    - multi-head self-attention을 non-overlapping windows로 계산
  - Style injection
    - Mapping network로 만든 latent code를 주입하여 스타일 적용
  - Double attention
    - An enlarged receptive field를 얻기 위해 사용



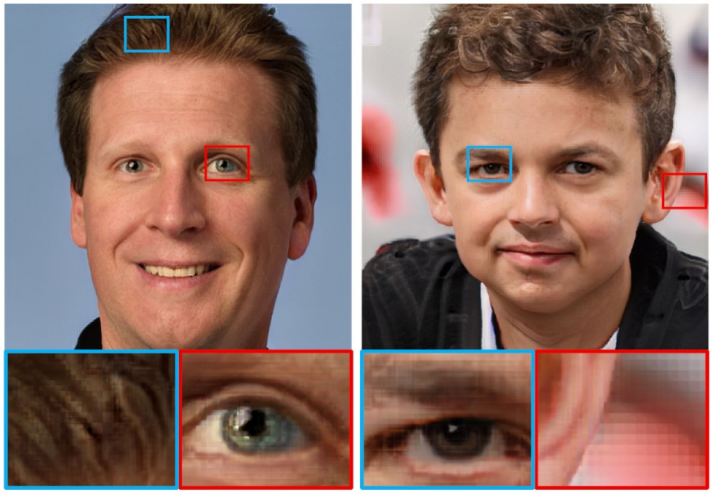
[Transformer based generative model]



[Double attention]

Method

- Blocking artifact in high resolution synthesis



[Blocking artifact 예시]

- Transformer 구조에서 발생한 것으로 추측
  - Local attetntion의 window size와 강한 correlation
    - Window-wise processing이 spatial coherency를 파괴

Method

- Artifact suppression
  - Artifact-free generator
    - Token sharing
      - HaloNet과 같이 windows간의 shared tokens 생성
      - 그러나 artifact는 여전히 발생
    - Sliding window attention
      - 이론적으로 artifact-free result여야 함
      - 그러나 지나치게 cost가 많이 들기 때문에 inference시에만 사용함
    - Reduce to MLPs on fine scales
      - Self-attention 구조를 삭제하고 point-wise MLP를 사용
      - 그러나 high-frequency details는 포기하여야 함
  - Artifact-suppression discriminator
    - Wavelet discriminator
      - 주기적인 artifact pattern은 spectral dom$ain에서 발견

Experiments

Methods	FFHQ	CelebA-HQ	LSUN Church
StyleGAN2 [35]	3.62*	-	3.86
PG-GAN [32]	-	8.03	6.42
U-Net GAN [55]	7.63	-	-
INR-GAN [56]	9.57	-	5.09
MSG-GAN [31]	-	-	5.20
CIPS [1]	4.38	-	**2.92**
TransGAN [29]	-	9.60*	8.94
VQGAN [14]	11.40	10.70	-
HiT-B [72]	2.95*	3.39*	-
StyleSwin	**2.81***	**3.25***	2.95

Table 3. Comparison of state-of-the-art unconditional image generation methods on FFHQ, CelebA-HQ and LSUN Church of 256 × 256 resolution in terms of FID score (lower is better). The subscript (∗) indicates that bCR is applied during training.

Methods	FFHQ-1024	CelebA-HQ 1024
StyleGAN[1] [35] [34]	**4.41**	5.06
COCO-GAN	-	9.49
PG-GAN [32]	-	7.30
MSG-GAN [31]	5.80	6.37
INR-GAN [56]	16.32	-
CIPS [1]	10.07	-
HiT-B [72]	6.37	8.83
StyleSwin	5.07	**4.43**

Table 4. Comparison of state-of-the-art unconditional image generation methods on FFHQ and CelebA-HQ of resolution 1024 × 1024 in terms of FID score (lower is better). [1]We report the FID score of StyleGAN2 on FFHQ-1024 and that of StyleGAN on CelebA-HQ 1024. For fair comparison, we report results of StyleGAN2 without style-mixing and path regularization.

Main Results

(a) (b)

Figure 7. Image samples generated by our StyleSwin on (a) FFHQ 1024 × 1024 and (b) CelebA-HQ 1024 × 1024.

감사합니다.