



[StyleGAN3의 entanglement에 대한 연구]

이름: 김대호
 학번: 20190865
 연구지도교수: 조성현

1 연구 목적

최근 가장 주목받고 있는 분야 중 하나인 딥러닝은 다양한 학습방법이 존재한다. 그러나 이전까지 딥러닝에서 다루는 task는 classification과 regression에 집중되어 있었다. 이는 labeling된 data를 기반으로 하기 때문에 매우 직관적이며, 좋은 결과를 가져오기 때문이다. 그러나, Image Synthesis 등의 분야에서는 정답을 특정 label 하나로 표현하기 어렵기 때문에 이러한 학습 방법이 어울리지 않았으며 새로운 방법이 요구되었다. 이러한 문제를 해결하기 위해서는 labeling되어있지 않은 data들 사이의 pattern을 찾는 Unsupervised Learning을 사용해야 한다. 이는 기존에 정답이 있는 data를 활용한 Supervised Learning에 비해 정답이 없는 data들을 이용해야 했기 때문에 훨씬 어려운 방법이었다. 이는 2014년 GAN(Generative Adversarial Networks)이 나오며 이러한 부분들이 해결되기 시작했으며 폭발적인 성장이 시작되었다.

GAN은 Generator와 Discriminator라는 두 개의 model을 학습시키기 때문에, 그 결과가 수렴하기 쉽지 않다. 하지만 Computer Vision 분야에서는 model이 CNN기반으로 어느 정도 완성되어있기 때문에 다른 분야에 비해 비교적 다양한 시도를 해볼 수 있었다. 이로 인해 GAN을 이용한 이미지 생성의 resolution과 quality는 비약적인 상승이 이루어지고 있다. [2, 1, 3, 4] 이들은 image editing [7, 8], domain translation [12, 6], video generation [10] 등의 다양한 분야에서 활용되고 있다.

그럼에도 이미지 quality의 측면에서는 발전의 여지가 남아있다. 이미지의 구도를 달리 하였을 때, 육안으로 구분 가능할 정도로 어색한 부분들이 존재하는 texture sticking 문제가 남아있기 때문이다. 이는 특정 feature들이 pixel에 고착화되어있어 나타나는 문제였으며, 이러한 문제점을 해결하고자 다양한 방법들이 시도되었다.

StyleGAN3의 저자들은 해당 문제의 원인이 aliasing으로부터 비롯된다고 발표하였으며, 이는 단순히 StyleGAN 뿐만 아니라 딥러닝의 전반적인 부분들에서 발생한다고 보고하였다.[5] 그에 따라 이들은 equivariance한 모델을 통해 feature들이 hierarchical하게 학습하게 만들어 Figure 1에서 볼 수 있는 것처럼 texture sticking을 해결하였다.

하지만, StyleGAN3에는 새로운 문제가 생겼다. PGGAN과 StyleGAN의 가장 큰 차이점 중 하나는 disentanglement한 latent vector을 통해 원하는 feature을 가진 이미지를 만들 수 있다는 점이었다. StyleGAN3에서는 latent vector들이 상당히 entanglement하였다. 다시 말해 어색한 부분 등 이미지 자체의 quality는 높아졌지만, 원하는 이미지를 정확하게 만들 수 없게 되었다. 이러한 이유에 대해서는 아직 밝혀진 바가 없다.

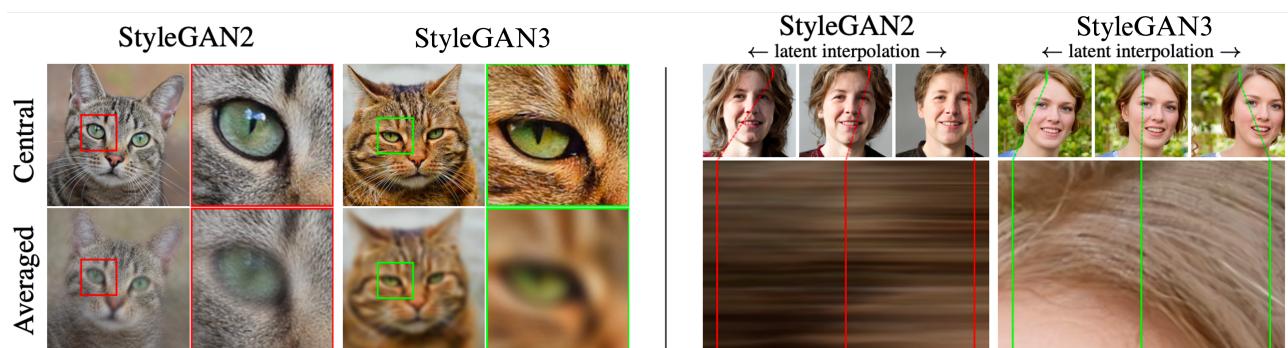


Figure 1: Result of StyleGAN3

Configuration	FID↓	EQ-T↑	EQ-R↑	Parameter	FID↓	EQ-T↑	EQ-R↑	Time	Mem.
A StyleGAN2	5.14	–	–	Filter size $n = 4$	4.72	57.49	39.70	0.84 ×	0.99 ×
B + Fourier features	4.79	16.23	10.81	* Filter size $n = 6$	4.50	66.65	40.48	1.00×	1.00×
C + No noise inputs	4.54	15.81	10.84	Filter size $n = 8$	4.66	65.57	42.09	1.18×	1.01×
D + Simplified generator	5.21	19.47	10.41	Upsampling $m = 1$	4.38	39.96	36.42	0.65 ×	0.87 ×
E + Boundaries & upsampling	6.02	24.62	10.97	* Upsampling $m = 2$	4.50	66.65	40.48	1.00×	1.00×
F + Filtered nonlinearities	6.35	30.60	10.81	Upsampling $m = 4$	4.57	74.21	40.97	2.31×	1.62×
G + Non-critical sampling	4.78	43.90	10.84	Stopband $f_{t,0} = 2^{1.5}$	4.62	51.10	29.14	0.86 ×	0.90 ×
H + Transformed Fourier features	4.64	45.20	10.61	* Stopband $f_{t,0} = 2^{2.1}$	4.50	66.65	40.48	1.00×	1.00%
T + Flexible layers (StyleGAN3-T)	4.62	63.01	13.12	Stopband $f_{t,0} = 2^{3.1}$	4.68	73.13	41.63	1.36×	1.25%
R + Rotation equiv. (StyleGAN3-R)	4.50	66.65	40.48						

Figure 2: StyleGAN3 results for FFHQ-U. **Left:** Training configurations. EQ-T and EQ-R are StyleGAN3’s equivariance metrics in decibels (dB); higher is better. **Right:** Parameter ablations StyleGAN3’s final configuration (R) for the filter’s support, magnification around nonlinearities, and the minimum stopband frequency at the first layer.

본 연구에서는 기존 StyleGAN2에서 StyleGAN3로 넘어오며 발생한 entanglement 문제의 이유를 밝히고자 한다. StyleGAN3에서는 Figure 2에서 볼 수 있는 것처럼 다양한 configuration들을 시도하고 추가하였다. 이러한 configuration들에 집중하여 어떠한 부분에서 latent vector들이 entanglement하게 되었는지 확인하는 것을 목표로 한다.

2 연구 배경

실제 물체들은 서로 다른 feature들이 hierarchical하게 움직이는 경향이 있다. 예를 들어, 머리를 움직이면 입이 움직이며 입술, 이빨, 수염 하나하나까지 모두 움직이게 된다. 그러나 이와 다르게 기존 StyleGAN2까지의 결과는 이러한 hierarchical한 이미지를 만들어내지 못했다.[5] 각각의 feature들의 resolution은 향상되었지만, 이들이 더 큰 크기의 feature이 아니라 픽셀에 고착화되어있는 것을 확인할 수 있다. 이러한 texture sticking 문제는 StyleGAN3에서 해결되었다.

Texture sticking의 문제는 크게 image boarder, per-pixel noise inputs, positional encoding, aliasing 4가지로 추려졌다. Image boarder가 spatial한 정보를 줘서 generator가 texture sticking이 되도록 하고 있었으며 이는 image 를 크게 잡고 나중에 crop하는 방식으로 해결하였다. 또한, 기존 StyleGAN에서는 pixel에 independant한 gaussian noise가 들어가는데 이를 제거하여 transform에 따라 다른 이미지가 생성되도록 하였다. Positional encoding 문제를 해결하기 위해서는 Fourier feature을 사용하였다.[9, 11] 마지막으로 aliasing과 같은 경우에는 generator에서 upsampling하는 과정에서 low-path filtering을 하지 않아, 원치 않은 high-frequency들이 계속 더해져 일어나거나, ReLU와 같이 non-linearity function을 지날 때, 값이 확 트여 일어난다. 이러한 aliasing은 yquist-Shannon sampling theorem로 해결할 수 있다. 저자들은 StyleGAN2의 generator를 신호학적으로 분석하여 이를 해결하였다. Figure 2에서 볼 수 있다시피 꽤나 유의미한 결과를 도출할 수 있었다.

하지만, 앞서 서술한대로 StyleGAN3은 GAN Inversion을 통한 latent vector들의 disentanglement가 좋지 않았다. 결국, 각 feature들이 얹혀있어 이를 조합하기가 어려워졌다. 이러한 문제를 해결하여, 원하는 feature들을 가지면서 높은 resolution과 quality를 보장할 수 있는 이미지와 영상을 만들 수 있는 길을 개척하고자 이번 연구를 진행하게 되었다.

3 연구 방법

기본적인 연구 방법은 StyleGAN2를 기반으로 StyleGAN3에서 바꾼 요소들 중 큰 요인으로 보이는 부분들에 대해 영향을 주는지 확인하고 그 이유를 알아내는 것이다.

StyleGAN3에서 바뀐 요인에 대해 자세한 내용은 Figure 2의 Configuration부분을 보면 알 수 있다. 각각의 요소에 대해 disentanglement를 판단하는 평가 지표를 이용하여 어느 부분에서 entanglement해지는지 확인한다.

Disentanglement를 판단하는 척도로는 크게 PPL(Perceptual path length)와 Linear seperability가 있다.[3] PPL과 같은 경우에는 z 값이 조금 변했을 때, 큰 차이가 없어야 disentanglement하다고 할 수 있다. Linear Seperability는 latent space가 disentanglement하다면 feature을 구분하는 정확한 방향 vector를 찾을 수 있어야 한다는 개념을 이용한 평가 지표이다. Linear hyperplane으로 latent space를 나눌 수 있다면 disentanglement하다고 할 수 있으므로 training 결과를 linear SVM(Support Vector Machine)을 이용해 분류한다. 분류가 잘 된다면 이는

lineary hyperplane으로 feature을 잡아낼 수 있다는 뜻이다.

이렇게 두 가지 평가지표를 활용하여 disentanglement를 판단하고 configuration을 분류한다. 최종적으로 가장 큰 영향을 주는 configuration을 찾고 그 이유를 알아낸다.

4 기대 효과

기본적으로 StyleGAN3의 문제점을 파악하고 이를 해결하는데 도움을 줄 수 있다. 이렇게 문제가 해결된다면 원하는 이미지를 높은 resolution과 quality로 만들 수 있을 것이다. 이는 실생활에서 온라인 모델, 인공지능 웹툰 등 다양한 분야에 활용될 수 있다.

5 연구 추진 일정

연구는 다음과 같이 진행할 예정이다.

3/11 과제연구 proposal 제출 및 연구 진행

4/22 연구 진행 상황 정리 및 보고서 제출

4/29 연구 진행 상황 발표

6/1 연구 결과 정리 및 보고서 제출

6/2 최종 발표

연구 진행 보고서 제출일 전까지 StyleGAN3에서 바뀐 부분에서 latent vector에 영향을 준 부분이 어떠한 요소인지 확인할 것이다. 이후, 확인된 요소가 entanglement에 영향을 주는 이유를 분석하고, 이를 사이에 어떠한 상관관계가 있는지 확인할 예정이다. 마지막으로 이를 어떤 방식으로 해결할 수 있는지 생학해보고, 정리한 결과를 발표할 예정이다.

참고문헌

- [1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. [1](#)
- [2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [1](#)
- [3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. [1](#), [2](#)
- [4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. [1](#)
- [5] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021. [1](#), [2](#)
- [6] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017. [1](#)
- [7] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021. [1](#)

- [8] Ryohei Suzuki, Masanori Koyama, Takeru Miyato, Taizan Yonetsuji, and Huachun Zhu. Spatially controllable image synthesis with internal representation collaging. *arXiv preprint arXiv:1811.10153*, 2018. 1
- [9] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singh, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 2
- [10] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 1
- [11] Rui Xu, Xintao Wang, Kai Chen, Bolei Zhou, and Chen Change Loy. Positional encoding as spatial inductive bias in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13569–13578, 2021. 2
- [12] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1