고해상도 GAN 영상 편집을 위한 중간 층 미리 보기

김성태 [O1], 강경국 [2], 조성현 [1,2]

[1] 포항공과대학교 인공지능대학원, [2] 컴퓨터공학과

seongtae0205@postech.ac.kr, kkang831@postech.ac.kr, s.cho@postech.ac.kr

요 약

사람 얼굴 영상은 한 사람을 잘 표현하는 대표적인 수단 중 하나이며 이를 편집하는 소프트웨어들이 많이 각광받고 있다. 이러한 관심에 힘입어 생성적 적대 신경망을 이용한 영상 편집 기법이 많이 연구되었다. 생성자의 잠재 공간을 분석하면 얼굴의 나이, 성별 등의 의미론적 표현에 대한 잠재 벡터를 알 수 있으며, 실제 영상을 잠재 공간에 투영시키는 인버전을 통하여 실제 영상의 의미론적 표현을 편집 할 수 있게 된다. 하지만 인버전 기법을 이용한 영상 편집은 최적화 과정이 필요하기에 고해상도 영상에서 많은 시간이 요구된다는 문제점을 가지고 있다. 따라서 본 논문에서는 신경망의 중간 해상도 영상을 사용자에게 미리 제공하는 방법론을 제안하고자 한다. 중간 해상도 영상을 얻기 위해서는 생성자의 중간 층 출력 결과를 시각화 하여야 한다. 이를 위해 본 논문에서는 파인 튜닝과 적응형 인스턴스 정규화 기법을 이용한 방법을 제안하며, 결과적으로 중간 해상도의 미리 보기를 통해 사용자와의 보다 빠른 상호작용을 기대할 수 있다.

그림 1. FFHQ 얼굴 데이터셋으로 학습된 StyleGAN[2]에서 합성된 1024x1024 얼굴 영상

1. 서론

페이스북, 인스타그램과 같은 SNS 의 발전으로 자신의 얼굴을 개성 있게 표현할 수 있도록 하는 시각콘텐츠의 편집 기술이 많이 연구되어 왔다. 하지만 얼굴 영상이 가지고 있는 나이, 성별, 각도 등의 의미론적 표현(Semantic Representation)을 편집하는 데는 많은 어려움이 있었다. 이를 해결하기 위해 최근 생성적 적대 신경망(GAN, Generative Adversarial Network)[1]을 이용한 사실적인 영상 편집 기법이 연구되었다.

GAN 은 생성자(Generator)와 구별자(Discriminator)가 경쟁적으로 학습을 진행하는 구조를 가지고 있다. 생성자가 임의의 잠재 벡터(latent vector)로부터 영상을 합성하면 구별자는 해당 영상이 실제 영상인지 합성된 영상인지 구별하게 된다. 생성자는 구별자가 영상을 잘 구별하지 못하도록 학습되며 구별자는 합성된 영상과 실제 영상을 잘 구별하도록 학습된다. 학습이 완료되면 생성자와 구별자는 내쉬 평형을 이루게 되고, 생성자는 실제 영상과 구별하기 힘든 정도의 합성 영상을 생성할 수 있게 된다.

GAN 모델 중 하나인 StyleGAN[2]은 낮은 해상도에서 시작해 점차적으로 높은 해상도의 영상을 생성하는 구조를 가지고 있다. 또한 잠재 벡터가 생성자의 각 층에 입력으로 주어져 영상을 컨트롤하는 구조를 가지기 때문에 높은 수준의 사실적인 고화질 얼굴 영상 합성이 가능하다 (그림 1).

생성자에게 입력으로 주어지는 잠재 벡터는 합성된 영상의 의미론적 정보가 인코딩(Encoding)되어 있다. 따라서 잠재 벡터를 조작하게 되면 합성되는 영상의 의미론적 표현을 편집할 수 있게 된다[5, 6]. 이를 실제 영상에 적용하기 위해서는 영상을 StyleGAN 잠재 공간(latent space)에 투영(projection)시키는 과정인 GAN 인버전(inversion)이 필요하다[3, 4].

[4]는 인버전을 효과적으로 하기 위하여 실제 영상을 잠재 공간으로 직접 임베딩(Embedding)시키는 인코더(Encoder)를 도입하였다. 하지만 그럼에도 불구하고 높은 수준의 인버전을 위해서는 임베딩 된

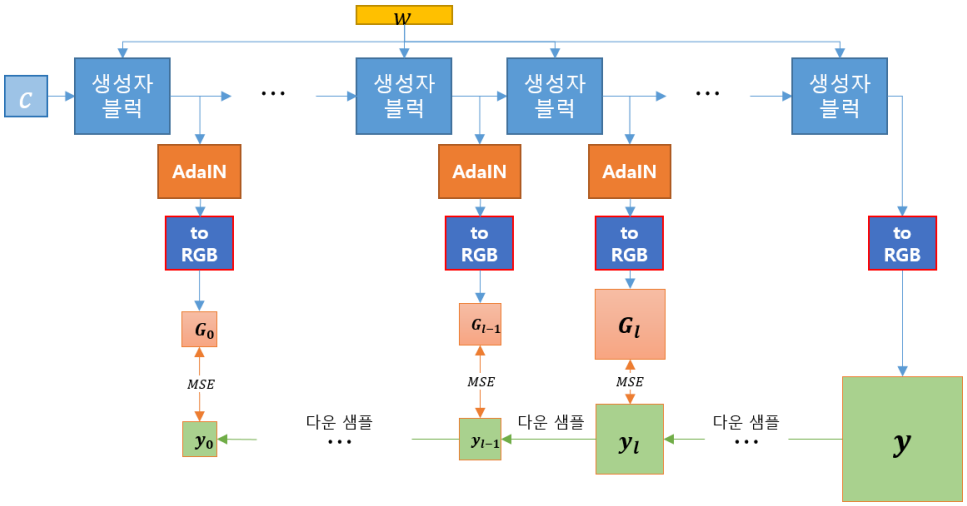

그림 2. 중간 층 시각화를 위한 모델 모식도. 중간 toRGB 층은 출력이 다운 샘플된 원본 영상과 유사해지도록 학습된다. toRGB 층 학습 시에는 AdaIN 층 파라미터가 각각 1 과 0 으로 고정된다. 이후 샘플에 대하여 중간 해상도 영상을 생성 할 때, 부족한 색감을 보완해주기 위하여 AdaIN 파라미터 최적화 과정을 거친다.

벡터의 추가적인 최적화 과정이 필수적으로 요구된다.

StyleGAN 을 실제 영상 편집에 활용하려면 사용자와의 빠른 상호작용이 필요하다. 하지만 현재의 StyleGAN 인버전은 많은 연구가 진행되었음에도 최적화 과정으로 인한 많은 시간이 소요된다. 이는 인터렉티브한 상호작용에 병목 현상으로 연결되고, 병목 현상은 고해상도 영상을 인버전 시킬 수록 더욱 두드러지기 때문에, GAN 영상 편집을 실제 어플리케이션에 적용하는데 어려움을 겪고 있다.

본 논문에서는 고해상도 영상의 인버전 병목 현상을 해결하기 위하여, StyleGAN 생성자의 중간 층 출력을 시각화해 사용자에게 미리 제공하는 방법을 제안하고자 한다. 중간 층 출력을 사용하면, 중간 해상도에서의 인버전을 이용하여 보다 빠른 속도로 사용자와 상호작용이 가능해진다. 또한 사용자가 편집을 완료하면 중간 해상도에서 찾은 잠재 벡터를 고화질 합성 영상에 이용할 수 있기 때문에 보다 높은 수준의 영상 편집을 제공할 수 있다.

중간 층 출력을 시각화 하기 위해 StyleGAN 에 존재하는 toRGB 층을 활용하기로 한다. toRGB 층은 최종 특성맵을 RGB 영상으로 출력해주는 1x1 합성곱층(Convolution layer)이다. 이는 StyleGAN 학습 시에 사용되며 학습이 완료되면 중간 toRGB 층은 본래의 기능을 잃게 된다. 본 논문에서는 미리 학습된 StyleGAN 의 toRGB 층을 파인 튜닝(fine-tuning)하여 중간 층 결과를 시각화 할 수 있도록 하였다. 추가적으로, 합성된 중간 층 영상이 원본과 비슷한 색감을 가지도록 적응형 인스턴스 정규화(AdaIN, Adaptive Instance Normalization)[7] 층을 이용한 최적화 과정을 제안하였다.

그림 3. InterFaceGAN[6]을 이용한 얼굴 나이 편집

2. 생성자의 중간 층 결과 시각화

그림 (2)는 전체적인 학습 모델을 표현한 모식도이며 StyleGAN 에서 toRGB 층 전에 적응형 인스턴스 정규화 층을 가진 구조를 가지고 있다. 적응형 인스턴스 정규화 층은 toRGB 층 학습 시에 출력 영상의 대비로 인한 문제를 해결해주며, 샘플 출력 시에는 원본 영상에서 부족한 색감을 보완해주는 역할을 한다. 2.1 장에서는 toRGB 층 학습에 대해 설명하며, 2.2 장에서는 적응형 인스턴스 정규화에 대해 설명하고자 한다.

2.1 toRGB 층 학습

StyleGAN 은 PGGAN [9]의 구조를 기반으로 하고 있다. StyleGAN 은 낮은 해상도에서의 학습을 완료시킨 후 높은 스케일을 담당하는 층을 추가하여 신경망 전체를 다시 학습시키는 과정을 반복한다. 이때 toRGB 층은 최종 특성맵(feature map)을 RGB 영상으로 출력해주도록 학습된다. 특정 스케일에서의 학습이 완료되면 toRGB 층은 해당 스케일에서의 최종 출력을 담당하게 된다. 하지만 이후 높은 스케일에 대해 신경망이 다시 학습되게 되면 이전 스케일에 대한 특성맵 분포가 바뀌기 때문에 중간 toRGB 층

(a) 목표 영상 (b) StyleGAN[2] (c) (b) +인스턴스 정규화 (d) (c)+toRGB 층 파인 튜닝 (e) 본 방법

그림 4. 1024x1024 영상을 합성하도록 학습된 StyleGAN 에서 256x256 중간 해상도의 단계별 출력 결과. (a)는 1024x1024 합성 영상을 256x256 크기로 다운 샘플 한 결과. (b)는 toRGB 층 학습 전 StyleGAN 의 256x256 스케일 출력 결과. (c)는 StyleGAN toRGB 층 전에 인스턴스 정규화를 적용한 결과. (d)는 toRGB 층 학습을 완료하였을 때의 중간 해상도 출력 결과. (e)는 (d)결과에 적응형 인스턴스 정규화 층 최적화를 적용하였을 때의 결과

출력 결과는 본래의 기능을 잃게 된다. 본래의 기능을 잃은 toRGB 층이 중간 해상도의 온전한 영상 출력하게 하기 위해서, 학습이 완료된 StyleGAN 으로부터 toRGB 층의 추가적인 학습과정을 추가하였다.

보다 온전한 중간 해상도 영상 출력을 위해서는 각 스케일 출력을 담당하는 toRGB 층이 다운샘플(down-sampling)된 목표 영상과 같은 출력을 가지도록 파인 튜닝(fine-tuning)하여야 한다.

$$Loss = \sum_l \|G_l - y_l\|^2 \qquad (1)$$

toRGB 층은 식 (1)의 손실 함수를 최소화하는 방향으로 학습이 이루어진다. l은 해상도 스케일을 의미하며, G_l과 y_l는 각각 l 스케일에서의 toRGB 출력과 l 스케일로 다운 샘플된 목표 영상을 의미한다.

2.2 적응형 인스턴스 정규화 층

적응형 인스턴스 정규화(AdaIN) 층은 toRGB 층의 학습 과정에서 활용되며, 샘플 합성 시 원본 영상과의 색감을 맞추는 역할을 하게 된다.

$$\mathrm{AdaIN}(x_i, s_i, b_i) = s_i \frac{x_i - \mu(x_i)}{\sigma(x_i)} + b_i \qquad (2)$$

i는 특성맵 채널 인덱스를 의미하며 x_i는 i번째 특성맵(feature map) 채널을 의미한다. s_i 는 스케일 파라미터, b_i는 바이어스 파라미터로 영상의 색감과 같은 스타일(style)을 컨트롤한다[7].

미리 학습된 StyleGAN 의 중간 toRGB 층 출력을 확인하면 매우 높은 대비로 인한 색상 왜곡이 발생된다 (그림 4 (b)). 이는 원본 영상과 많은 차이를 불러 일으키기에 toRGB 층 학습을 더욱 어렵게 한다. 이를 해결하기 위해 학습과정에 인스턴스 정규화(Instance Normalization)[8]를 적용하였다. 인스턴스 정규화는 영상의 대비를 정규화하는 기능을 한다[8]. 인스턴스 정규화를 위해 식 (2)에서 s_i와 b_i를 각각 1과 0으로 설정하였으며 이는 toRGB 층 학습과정에서 고정된다.

그림 4 (d)를 보면 toRGB 층의 학습만으로는 중간 해상도 영상이 목표 영상과 동일한 색감을 가지지 않는 것을 확인 할 수 있다. 목표 영상과의 색감을 맞추기 위해 영상 합성 시 적응형 인스턴스 정규화 층 파라미터를 최적화 하는 과정을 추가하였다. 학습해야하는 파라미터의 차원은 s 와 b 모두 특성맵 채널 수 이기 때문에, 상대적으로 적은 계산량을 요구하게 된다. 적응형 인스턴스 정규화 층의 최적화 또한 식 (1)를 최소화 하도록 설정하였다.

3. 실험 결과 및 분석

본 논문에서는 FFHQ 데이터셋으로 미리 학습된 StyleGAN 을 이용하였다. toRGB 층 파인 튜닝에는 FFHQ 데이터셋을 사용하였다.

toRGB 층에 특성(feature)이 입력 되기 전 인스턴스 정규화 과정을 거치게 되면 대비로 인해 왜곡되었던 색감이 많이 복구 되는 것을 확인 할 수 있다 (그림 4 (c)). 인스턴스 정규화를 기반으로 toRGB 층을 파인 튜닝하게 되면 목표에 더욱 근접한 영상이 생성되는 것을 확인 하였다 (그림 4(d)). 그 후 샘플에 대한 적응형 인스턴스 정규화 층 최적화를 거치게 되면 목표 영상과 비슷한 색감이 추가 복원된 중간 해상도 영상을 확보 할 수 있다(그림 4(e)).

4. 결론

영상의 의미론적 표현을 편집하기 위해서는, 영상의 잠재 벡터를 찾는 인버전 과정이 필수적이다. 하지만, 고해상도 영상을 인버전하는 데는 많은 시간이 요구된다. 인버전을 통해 GAN 으로 편집한 영상을 보다 빨리 사용자에게 제공하기 위하여 본 논문에서는 StyleGAN 의 중간 층 출력을 시각화하는 방법론을 제시하였다. 중간 층 출력을 담당하는 toRGB 층을 파인 튜닝하였으며, 부족한 색감을 복원하기 위해 적응형 인스턴스 정규화 층의 최적화 기법을 적용하였다.

하지만 본 방법은 아직 원본 영상과 동일한 수준의 출력을 가지지 못한다. 이는 생성자의 중간 층 출력은 높은 층에 입력으로 들어가는 잠재 벡터의 정보를 가지지 못한 채로 영상을 합성하기 때문이다. 따라서 향후에는 중간 층 출력 영상이 가지고 있는 한계를 극복하는 추가 연구를 진행할 계획이다.

감사의 글

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2019-0-01906, 인공지능대학원지원(포항공과대학교))과 한국연구재단의 지원(No. 2020R1C1C1014863)을 받아 수행된 연구임.

참고문헌

[1] Lan et al., Generative Adversarial Nets, NIPS 2014

[2] Karras et al., A Style-Based Generator Architecture for Generative Adversarial Networks, CVPR 2019

[3] Rameen et al., Image2stylegan: How to embed images into the stylegan latent space?, CVPR 2019

[4] Zhu et al., In-Domain GAN Inversion for Real Image Editing, ECCV 2020

[5] Tewari et al., PIE: Portrait Image Embedding for Semantic Control, SIGGRAPH 2020

[6] Yujun Shen et al., Interpreting the Latent Space of GANs for Semantic Face Editing, CVPR 2020

[7] X. Huang and S. J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. CoRR, abs/1703.06868, 2017

[8] Ulyanov et al., Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022, 2016.

[9] Karras et al., Progressive Growing of GANs for Improved Quality, Stability, and Variation, ICLR 2018