GAN을 이용한 동영상 스타일 생성 및 합성 네트워크 구축

최희조*, 박구만**. 김상준*, 이유진*, 상혜준**
*서울과학기술대학교 일반대학원 IT미디어공학과
**서울과학기술대학교 전자IT미디어공학과
heejo0624@seoultech.ac.kr, gmpark@seoultech.ac.kr

A Video Style Generation and Synthesis Network using GAN

Heejo Choi*, Gooman Park**
*Dept. of IT Media Engineering, Seoul National University of Science and Tech.
**Dept. of Electronic and IT Media Engineering , Seoul National University of Science and Tech.

요 약

이미지와 비디오 합성 기술에 대한 수요가 늘어남에 따라, 인간의 손에만 의존하여 이미지나 비디오를 합성하는데에는 시간과 자원이 한정적이며, 전문적인 지식을 요한다. 이러한 문제를 해결하기 위해 최근에는 스타일 변환 네트워크를 통해 이미지를 변환하고, 믹싱하여 생성하는 알고리즘이 등장하고 있다. 이에 본 논문에서는 GAN을 이용한 스타일 변환 네트워크를 통한 자연스러운 스타일 믹싱에 대해 연구했다. 먼저 애니메이션 토이 스토리의 등장인물에 대한 데이터를 구축하고, 모델을 학습하고 두 개의 모델을 블렌딩하는 일련의 과정을 거쳐 모델을 준비한다. 그 다음에 블렌딩된 모델을 통해 타겟 이미지에 대하여 스타일 믹싱을 진행하며, 이 때 이미지 해상도와 projection 반복 값으로 스타일 변환 정도를 조절한다. 최종적으로 스타일 믹싱한 결과 이미지들을 바탕으로 하여 스타일 변형, 스타일 합성이 된 인물에 대한 동영상을 생성한다.

1. 서론

오늘날 이미지와 비디오 합성 기술에 대한 수요가 늘어남에 따라 컴퓨터를 통한 자연스러운 이미지 생성에 대한 수요와 중요성도 함께 증가하고 있다. 하지만, 수작업에만 의존하여 이미지나 비디오를 합성하는데에는 시간과 자원이 한정적이며, 특히 비디오에 대한 합성은 더욱 전문적인 지식을 요한다.
이러한 문제점을 해결하기 위해 StyleGAN 등의 GAN 네트워크를 통해 이미지를 생성하는 네트워크를 이용하여 타겟 이미지를 사용자가 원하는 이미지와 유사한 이미지로 생성하는 기술을 이용한다.

특히, StyleGAN2를 이용한 스타일 변환 네트워크를 비디오 속의 사용자의 얼굴의 스타일을 믹싱하여 자연스러운 비디오 생성에 대해 연구한다. 본 시스템은 먼저 애니메이션 토이스토리에 나오는 등장인물로 사용자의 스타일을 변환하여 표출한다. 영화 토이스토리 4편의 등장인물에 대한 이미지를 수집하여 데이터셋을 만들고, 이 데이터 셋과 FFHQ 데이터셋 (Flickr-Faces-HQ Dataset)을 딥러닝 스타일 변환 네트워크를 통해서 학습한다. 학습을 마친 두 모델에 대하여 블렌딩 처리를 거쳐 새로운 모델을 만들고, 이를 바탕으로 스타일 믹싱을 진행한다. 이를 통해 사용자가 타겟 이미지에 대하여 얼굴 스타일 믹싱된 이미지를 생성 할 수 있게 하였고, 믹싱된 이미지를 바탕으로 비디오를 생성한다.

2. 관련 기술

1. StyleGAN2

스타일갠2는 기존의 스타일갠을 개선한 두번째 버전으로서 생성 이미지에서 나타나는 물방울 아티팩트와 phase artifact를 개선하였다. 스타일갠2에서는 가중치 복조(weight demodulation)를 통해서 AdaIN(적응형 인스턴스 정규화) 대신에 사용하여 물방울 아티팩트를 제거한다. 정규화된 잠재 벡터에 대하여 매핑 네트워크를 거친 W를 인코더와 디코더 역할을 해주는 변조(modulate)와 복조(demodulate) 작업을 통해서 이미지 데이터의 특징을 추출한다. 이를 통해서 그림 1과 같이 W를 스타일 A로 바꿔준 후에 c의 값이 입력하여 컨볼루션 연산을 해준다. c는 생성기의 i번째 중간 활성화 함수의 채널 수

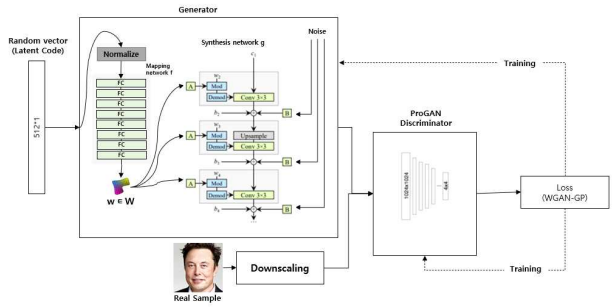

그림 1 Stylegan2 생성기와 판별기의 학습에 대한 전반적인 그림
(출처:https://upload.wikimedia.org/wikipedia/commons/8/85/Elon_Musk_Royal_Society_%28crop1%29.jpg)

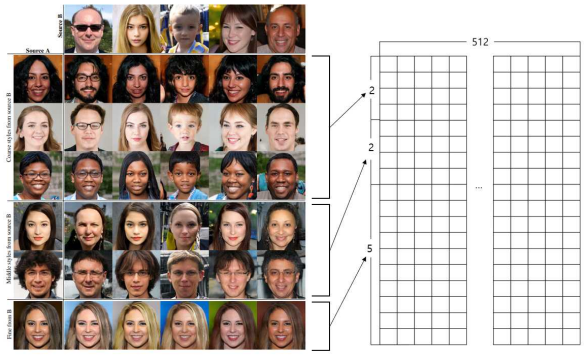

그림 2 모델 블렌딩을 위한 W 공간의 레이어 별 구성

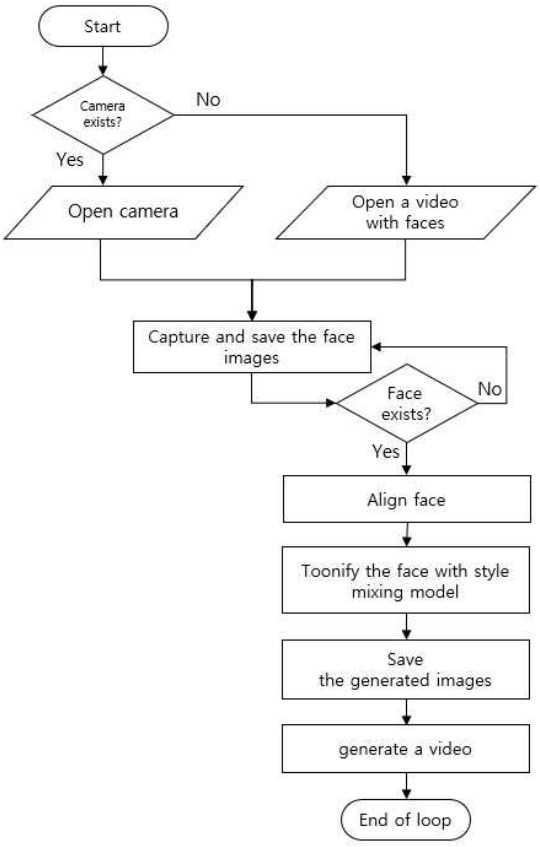

그림 3 제안하는 스타일 믹싱 비디오 생성 시스템의 전체적인 흐름도

이다. 여기에 바이어스 값과 노이즈를 추가하여 다음 스타일 블록의 입력으로 보낸다. 경로 길이 정규화(Path length regularization)는 자코비언 행렬(Jacobian matrix)을 이용해 w가 조금 변하면 생성 이미지에 대한 의미적인 부분도 그에 맞게 조금 변하도록 해주는 작업이다. 다시 말해, w가 변할 때 그로부터 생성되는 이미지 y도 변하는데 그 정도가 비슷하게 만들기 위해서 스타일갠2에서 추가로 적용된 내용이다. 이 논문에서는 스타일갠2의 config-f 구성을 사용하여 스타일갠의 네트워크를 확장한 네트워크를 사용한다.

1.1 Toonify[1](네트워크 블렌딩)

W 공간에 매핑된 데이터는 [그림 2]과 같은 형태이며, 아래와 같이 coarse, middle, fine style layer로 나뉜다. coarse 스타일 레이어에서는 세밀하진 않지만, 영상 내 구조 등의 전반적인 semantic feature들이 있고, middle style은 헤어 스타일, 눈을 뜨거나 감은 여부 등의 정보를 담는다. Fine style layer에서는 색상이나 미세한 구조와 같이 세밀한 부분을 담당한다. [그림 2]와 같이 거친 스타일 레이어(coarse style layer)는 4x4, 8x8에서 스타일 믹싱을 통해서 얼굴 자세나, 헤어스타일, 얼굴 모양 등의 특징을 믹싱할수있고, 중간 레이어 (middle layer)는 16x16,32x32 레이어를 통해서 얼굴의 특징, 좀 더 세밀한 헤어스타일, 눈감은 여부 등의 특징을 믹싱할 수 있다. 마지막으로, 좀더 세밀한 레이어(fine layer)는 64x64부터 1024x1024 레이어의 고해상도의 레이어에서의 스타일 믹싱을 통해서 변환할 수 있다. 전이 학습(Transfer learning)을 통해 타겟 이미지에 스타일 믹싱할 때 프로젝션을 적게 반복할수록 타겟 이미지의 특징에 가까워지고, 반복값 k를 늘릴수록 fine layer를 추출하여 texture 정보를 담고있는 모델의 스타일과 가까워지는 것을 확인할 수 있다.

III. 동영상 스타일 믹싱 시스템

1. 시스템 개요

제안하는 시스템의 전체적인 구조는 아래 [그림 3]과 같다. 먼저 웹 캠 등의 RGB 카메라를 열어 동영상을 입력받고, 카메라를 사용하지 않는 경우 영상을 선택하여 입력한다. 얻어진 영상에 대해서 프레임을 추출하며, 프레임마다 얼굴 검출 및 얼굴 정렬을 하여 얼굴이 있는 이미지만 저장한다. 얼굴을 검

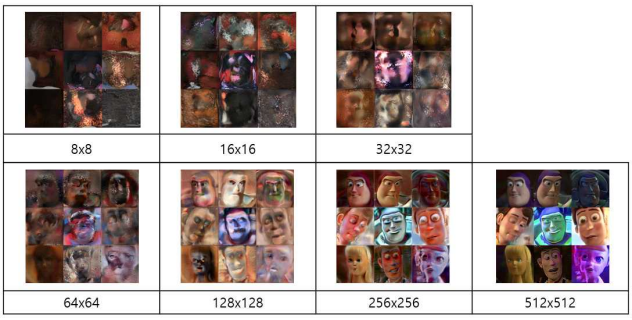

그림 4 5000번 학습한 모델과 미리 학습시킨 FFHQ 모델을 블렌딩한 모델에 대한 해상도별 결과 이미지

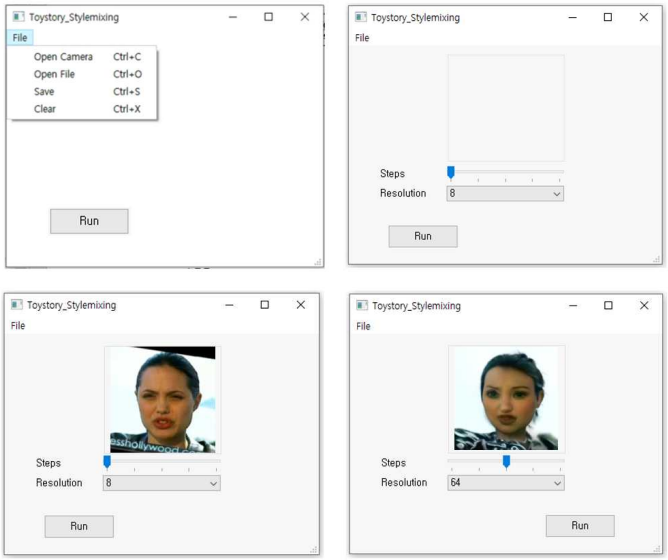

그림 5 인터페이스를 통해서 사용자로부터 인자를 입력받아 영상 표출

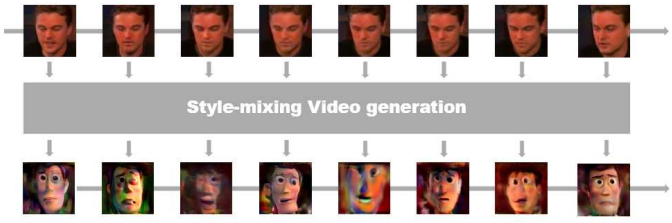

그림 6 YTF 데이터셋을 통해 얻은 영화배우에 대한 비디오의 프레임을 카툰 스타일로 스타일 믹싱

출한 이미지에 대하여 전처리를 해준 후, 미리 학습해놓은 딥러닝 모델에 입력하여 프로젝션하고, 이미지에 대한 스타일 변환을 진행한다. 그 후 생성 이미지를 저장하고, 이를 표출하며 최종적으로 동영상을 생성한다.

2. 스타일 변환 네트워크 학습

2.1 학습 데이터셋 구축

본 논문에서는 총 4편의 토이스토리 영화에 대하여 얼굴 검출 네트워크를 통해 등장인물의 얼굴 이미지를 3만여장 추출하고, 512x512의 동일한 사이즈로 이미지 내 얼굴을 정렬 (face alignment)하여 전처리한다. 그 후 tfrecords 파일로 만들어 학습 데이터셋을 준비하며, 이 데이터 셋은 모델 블렌딩 시 큰 구조를 맡는 coarse layer를 담당한다. 다음으로, 세밀한 디테일을 담당할 FFHQ 데이터 셋은 또한 얼굴 정렬하여 동일한 512x512 사이즈로 전처리한다.

2.2 딥러닝 모델 학습

만들어진 데이터 셋을 기반으로 config-f로 설정하여 GEFORCE RTX2080ti를 사용하여 5000번의 학습을 실시한다. Config a-f 중에서 config-f는 고해상도의 이미지 베이스라인 StyleGAN의 네트워크를 확장한 버전을 사용했으며, 연산량이 많은 점이 있다. 하지만, [4]에서 언급한 바와 같이 FID 점수와 Inception score가 가장 좋은 config-f를 사용했고, config-f으로 학습하기에 가장 적절한 512*512 사이즈의 이미지로 학습시켰다.

2.3 모델 블렌딩

토이스토리 데이터셋을 학습하여 만들어진 피클 파일을 통해서 ffhq 얼굴을 학습한 모델과 한다. 이 때, 저해상도 모델은 이미지 내에서 구조(coarse features)와 같이 이미지 내에서 다소 거시적이고 시맨틱하며, 결이 거친 특징을 담당하고, 고해상도의 모델은 텍스처(fine features)를 맡는 모델을 구분하여 모델을 블렌딩한다. 블렌딩한 이미지를 해상도 크기 별로 나타내면 아래의 [그림 4]과 같다.

2.4 스타일 믹싱 인터페이스

아래 [그림 5]은 만든 페이스 카툰화 비디오 생성 시스템을 위한 인터페이스이다. pyqt 라이브러리를 이용하여 사용자 인터페이스의 슬라이더와 콤보박스를 만들고, 프로젝션 반복 횟수와 해상도를 설정하여 프로젝션 할 때 이미지의 변환 정도를 조절한다.

3. 동영상 스타일 믹싱 시스템

[그림 6]는 Youtube Face 데이터[2](YTF dataset)으로부터 추출한 배우 레오나르도 디카프리오의 얼굴 이미지에 대하여 스타일 믹싱하여 생성한 동영상의 프레임 이미지이다. [그림 6]과 같이 GUI에서 슬라이더와 콤보박스를 만들어서 resolution과 iteration 값을 입력받으므로서 이미지 변환에 영향을 끼치는 인자를 사용자가 조절할 수 있게 한다.

프로젝션 해상도를 낮게 설정하면 블렌딩된 모델의 특징에 대한 토이스토리 데이터 셋의 특징이 두드러지며, 해상도가 512 사이즈에 가까워 질수록 FFHQ 모델에서의 특징이 더욱 두드러지는 것을 관찰할 수 있었다. 더불어, 프로젝션 반복 횟수를 늘릴수록 텍스처가 강조되며, 반복 값을 줄이면 거칠고 시맨틱한 특징들에 대한 표현이 강조되는 것을 관찰할 수 있었다. 더불어, 스타일 변환 시 값을 같게 하여 동영상이 더욱 자연스럽게 생성되도록 하였다. 64 크기의 해상도에 500번의 프로젝션이 제일 안정적이고 자연스럽게 합성한 결과 이미지를 생성해냈다. 또한, 학습 시 하나의 캐릭터에 대해 학습을 시키는 것보다 여러 캐릭터에 대해서 학습시켰을 때 더욱 영상에서 얼굴에 대한 표현의 자유도가 높아지며 더욱 자연스러운 합성 이미지를 생성하는 결과를 확인했다.

Ⅳ. 결론

학습 데이터에 대하여 5000 에포크의 학습을 돌린 결과를 가지고 블렌딩 및 스타일 믹싱을 진행해 본 결과, 512 사이즈의 영상을 학습하고, 생성한 이미지에 대하여 64 사이즈의 해상도로 500번 프로젝션 했을 때 가장 얼굴 영상 합성 결과물이 자연스러운 것을 확인하였다. 더불어, 얼굴이 정가운데로 올바르게 정렬되었을 때 가장 안정적으로 결과물을 내는 것을 확인할 수 있다. StyleGAN의 기본 네트워크에 가중치 복조과 경로 길이 정규화 등을 추가하여 확장한 config-f의 네트워크를 활용했을 때 가장 자연스럽고, 안정적으로 학습되었다. 반면, 학습된 데이터의 비중이 높아 서양인에 대한 데이터가 많아 아시아인을 프로젝션 시 구조나 자세등의 거시적인 특징들이 어색한 점이 있다. 앞으로 연구는 한국적인 데이터를 데이터 셋을 구축하고 이를 통하여 학습 및 응용해볼 예정이다. 또한, StyleGAN2 네트워크만을 사용하여서 학습 및 스타일 믹싱 시 조건을 줄 수 없던 점을 보완하기 위해 LSTM 등의 순환신경망을 통해서 비디오 생성 시 비디오 속 인물의 입모양, 표정, 시선 등의 부분도 조작할수 있도록 네트워크를 수정하여 학습할 예정이다. 더불어, 트랜스포머(Transformer)이나 GTP-3와 같은 큰 딥러닝 네트워크를 통해 학습시킨 모델을 통해 스타일 믹싱을 하여 더욱 성능을 향상 시킬 예정이며, 동영상에 대해 스타일 믹싱하는 프로세스 또한 더욱 경량화하여 개선해나갈 예정이다.

참고문헌

[1]A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation learning with Deep Convolutional GANs,” Int. Conf. Learn. Represent., 2016.

[2] Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” arXiv. 2017.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks”, arXiv. 2016.

[4] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, vol. 2019-June, doi: 10.1109/CVPR.2019.00453.

[5] unjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, Jaegul Choo, “StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation”,arXiv. 2017.

[6] Justus Thies1 Michael Zollhofer, Marc Stamminger1 Christian Theobalt Matthias Nießner,“Face2Face: Real-time Face Capture and Reenactment of RGB Videos”, CVPR. 2016

[7] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner“FaceForensics++:Learning to Detect Manipulated Facial Images”,ICCV. 2019

[8] Ivan Petrov, Daiheng Gao, Nikolay Chervoniy,Kunlin Liu, Sugasa Marangonda,Chris Ume´,Jian Jiang, Luis RP,Sheng Zhang,Pingyu Wu,Weiming Zhang,“DeepFaceLab: Integrated, flexible and extensible face-swapping framework”,arXiv. 2020

[9] T. Karras, S. Laine,5 M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” 2020, doi: 10.1109/CVPR42600.2020.00813.

[10] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” 2018.

[11] [Internet] Stylegan-network-blending https://www.justinpinkney.com/stylegan-network-blending/

[14][Internet] toonify-yourself, https://www.justinpinkney.com/toonify-yourself/

[15] [Internet] Leonardo Dicafrio, https://m.media-amazon.com/images/M/MV5BMjI0MTg3MzI0M15BMl5BanBnXkFtZTcwMzQyODU2Mw@@._V1_UY317_CR10,0,214,317_AL_.jpg

[16] [Internet] Chris Hemsworth, https://thumbor.forbes.com/thumbor/960x0/https%3A%2F%2Fspecials-images.forbesimg.com%2Fdam%2Fimageserve%2F968210608%2F960x0.jpg%3Ffit%3Dscale