KIPS Trans. Softw. and Data Eng.
Vol.11, No.11 pp.465~472
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.11.465

Style Synthesis of Speech Videos Through Generative Adversarial Neural Networks

Choi Hee Jo† · Park Goo Man††

ABSTRACT

In this paper, the style synthesis network is trained to generate style-synthesized video through the style synthesis through training Stylegan and the video synthesis network for video synthesis. In order to improve the point that the gaze or expression does not transfer stably, 3D face restoration technology is applied to control important features such as the pose, gaze, and expression of the head using 3D face information. In addition, by training the discriminators for the dynamics, mouth shape, image, and gaze of the Head2head network, it is possible to create a stable style synthesis video that maintains more probabilities and consistency. Using the FaceForensic dataset and the MetFace dataset, it was confirmed that the performance was increased by converting one video into another video while maintaining the consistent movement of the target face, and generating natural data through video synthesis using 3D face information from the source video's face.

Keywords : Generative Adversarial Network, Video Generation, Style Transfer, Style Synthesis Network, Video Synthesis Network

적대적 생성 신경망을 통한 얼굴 비디오 스타일 합성 연구

최 희 조† · 박 구 만††

요 약

본 연구에서는 기존의 동영상 합성 네트워크에 스타일 합성 네트워크를 접목시켜 동영상에 대한 스타일 합성의 한계점을 극복하고자 한다. 본 논문의 네트워크에서는 동영상 합성을 위해 스타일갠 학습을 통한 스타일 합성과 동영상 합성 네트워크를 통해 스타일 합성된 비디오를 생성하기 위해 네트워크를 학습시킨다. 인물의 시선이나 표정 등이 안정적으로 전이되기 어려운 점을 개선하기 위해 3차원 얼굴 복원기술을 적용하여 3차원 얼굴 정보를 이용하여 머리의 포즈와 시선, 표정 등의 중요한 특징을 제어한다. 더불어, 헤드투헤드++ 네트워크의 역동성, 입 모양, 이미지, 시선 처리에 대한 판별기를 각각 학습시켜 개연성과 일관성이 더욱 유지되는 안정적인 스타일 합성 비디오를 생성할 수 있다. 페이스 포렌식 데이터셋과 메트로폴리탄 얼굴 데이터셋을 이용하여 대상 얼굴의 일관된 움직임을 유지하면서 대상 비디오로 변환하여, 자기 얼굴에 대한 3차원 얼굴 정보를 이용한 비디오 합성을 통해 자연스러운 데이터를 생성하여 성능을 증가시킴을 확인했다.

키워드 : 적대적 생성 네트워크, 비디오 생성, 스타일 변환, 스타일 합성 네트워크, 동영상 합성 네트워크

1. 서 론

컴퓨터 비전에 대한 딥러닝 연구가 발전함에 따라 이미지 및 비디오 합성에 대한 관심이 높아지고 있다. 기존의 얼굴 스타일 변환 네트워크는 이미지 변환 위주의 연구가 활발히 진행되고 있다. 이미지 변환 기술이 고도화 및 안정화 됨에 따라 최근 다양한 도메인 간의 스타일 변환에 대한 연구가 시도되고 있다. 딥러닝 기반 생성 모델의 발전에 따라 여러 네트워크가 고안되기 시작하였으나, 동영상을 합성할 때 시퀀스를 추가하게 되면서 부자연스러운 동영상이 생성되어 안정적인 동영상에 대한 스타일 합성은 달성하지는 못하였다. 또한, 스타일 합성 네트워크인 스타일갠[1,2] 잠재공간 내에서의 프로젝션을 통한 생성이기 때문에 데이터에 치중한 표정들의 표현이 부자연스럽게 생성되며, 특히 얼굴에 대한 데이터를 생성할 때 여러 가지 앨리어스와 아티팩트 때문에 질적으로 떨어지는 경향이 있다. [3,22]는 표정 등의 합성에 대한 결과물에서 포즈를 재현하는 데에는 성공했지만, 시선, 입 모양, 표정 등에 대한 디테일을 표현하기에는 한계가 있다. 얼굴 비디오에 최적화된 스타일 변환 및 합성 시스템을 통해서

※ 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2017-0-00217, 투명도와 레이어 가변형 실감 사이니지 기술 연구).
※ 이 논문은 2021년 한국정보처리학회 ACK 2021의 우수논문으로 "GAN을 이용한 동영상 스타일 생성 및 합성 네트워크 구축"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 서울과학기술대학교 IT미디어공학과 석사과정
†† 비 회 원 : 서울과학기술대학교 IT전자미디어공학과 교수
Manuscript Received : December 29, 2021
First Revision : April 26, 2022
Accepted : May 2, 2022
* Corresponding Author : Park Goo Man(gmpark@seoultech.ac.kr)

얻을 수 있는 장점은 다음과 같다.

첫 번째, 3차원 얼굴 모델[4]을 통해 2차원의 프레임으로부터 추출된 3차원 얼굴 정보를 비선형적으로 얼굴의 각도와 시선에 대한 제어한다. 이를 통하여 인물의 스피치 비디오를 입력받아 자연스럽게 스타일을 변환[5,23]하여 스타일 합성된 비디오를 출력한다.

두 번째로, 개연성과 일관성이 유지되는 안정적인 스타일 합성 비디오를 생성한다. 페이스 포렌식++ 데이터셋[6]과 같은 인물의 스피치 동영상 데이터셋을 통한 비디오를 학습하여 새로운 비디오를 만들 때 입 모양이나 시선 처리 등의 디테일에 대한 오차로 인해 부자연스러운 동영상이 생성되는 것을 소스 인물의 표정과 포즈를 통하여 개연성과 스타일에 대한 일관성을 유지한다.

본 논문의 네트워크는 기존의 동영상 합성 네트워크인 헤드투헤드++[7,8]를 기반으로 하여 동영상에 대한 스타일 합성 네트워크의 한계점을 극복하기 위해서 생성 네트워크를 확장하고자 한다. 더불어, 다양한 환경에서 얼굴 영역에 좀 더 집중하여 학습을 시켰던 부분은 동영상 얼굴 합성에서의 자기 얼굴 재연에서는 자연스러운 데이터를 생성하여 성능을 증가시키고자 한다.

2. 관련 연구

2.1 딥러닝 기반의 스타일 변환 기술

1) 스타일갠

스타일갠의 네트워크의 생성기는 Fig. 1과 같다. a)는 스타일갠의 베이스라인이 되는 PGGAN[16]의 네트워크에 대한 그림이고, 그림1의 (b)는 스타일갠의 네트워크에 적응형 인스턴스 정규화를 사용하여 스타일을 변환한다. 네트워크는 맵핑 네트워크와 합성 네트워크로 나누어 진행된다. 맵핑 네트워크의 완전 연결 네트워크를 통하여 w 공간에 맵핑시킨다. 그 후 이를 합성 네트워크 g의 적응형 인스턴스 정규화에

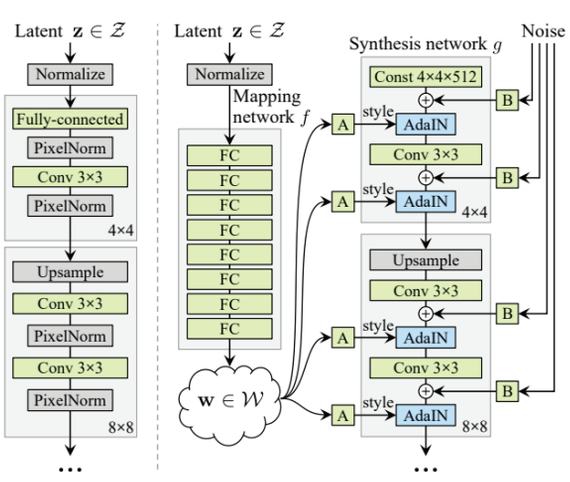

Fig. 1. The Structure of StyleGAN Generator

스타일로 입력한다. 맵핑 네트워크에서는 4x4x512 이미지에 대하여 확률론적 변이가 반영될 수 있도록 노이즈 B를 추가한다. 확률론적 변이[17]란, 모공, 수염 자국, 주근깨, 여드름, 머리카락의 흐트러짐 등의 미세한 노이즈를 자연스럽게 추가하기 위하여 컨볼루션 연산한 결과에 요소별로 노이즈 값을 추가하여 세세한 부분들에 변화를 주기 위해 사용된다. 적응형 인스턴스 정규화를 통해서 스타일을 변형을 담당한다. 그리고, 컨볼루션 3x3에 노이즈 B를 추가하고 적응형 인스턴스 정규화에 스타일 A를 추가하여 수행한다. 이를 다음 연산에서는 8x8, 16x16으로 1024까지 점차 크기를 키워 업샘플링하여 이미지 크기를 늘려가면서 학습을 진행한다. 이미지가 생성 네트워크를 거치면서 점점 고화질의 이미지가 생성된다. 이때 스타일과 노이즈가 반영될 수 있게 되면서 스타일의 더욱 선형적이고 스타일 얽힘 현상(entanglement)을 감소시킨다.

2) 스타일갠2(StyleGAN2)

스타일갠2는 스타일갠의 물방울 아티팩트, 위치 아티팩트 등 이미지의 자연스러운 생성을 저해하는 요소를 네트워크 구조를 변형하여 보완한다. 스타일갠2의 전체적인 네트워크는 그림2와 같다. 물방울 아티팩트는 스타일 갠의 적응형 인스턴스 정규화을 사용할때 물방울과 같은 아티팩트들이 생기는 경향이 있다. 가중치 복조로 대체하면 이미지 및 활성화 함수에서 특징적인 아티팩트가 제거된다. 위치 아티팩트는 스타일갠에서 레이어를 점차 키우는 PGGAN을 베이스라인으로 사용하여 생긴다. 위치 아티팩트는 치아가 얼굴 포즈를 따르지 않고, 파란색 선으로 표시된 것처럼 카메라에 정렬된 상태를 유지한다. 스타일갠2에서는 PGGAN을 스킵 연결으로 대체하여 각 해상도가 순간적으로 출력 해상도로 작용하여 최대 주파수 디테일을 생성하도록 한다. 그리고 학습된 네트워크가 중간층에서 지나치게 높은 주파수를 가지게 하여 이동 불변성을 손상시킨다는 문제를 해결한다.

자코비언 행렬을 통한 연산이 너무 무거워서 지연 정규화을 통해 16번에 한번 손실을 더할 때가 매 회 손실을 더하는 것보다 계산 비용과 메모리를 상당히 절감하여 스타일갠의 성능을 보완한다.

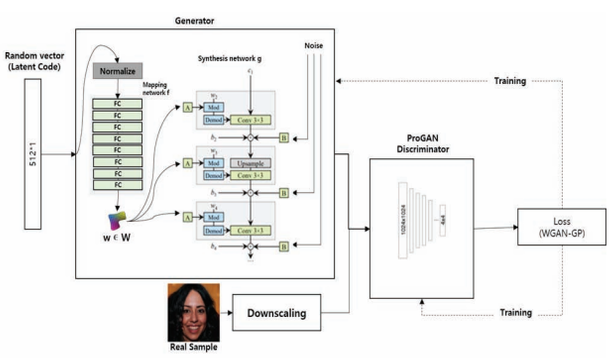

Fig. 2. StyleGAN2 Architecture for Training a Generator and a Discriminator