

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 스타일GAN 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 스타일GAN[1,2] 잠재공간 내에서의 프로젝션을 통한 생성이기 때문에 데이터에 치중된 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 레이어가 변형 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 것이다.

† 준회원 : 서울과학기술대학교 IT미디어공학과 석사과정

†† 비회원 : 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)

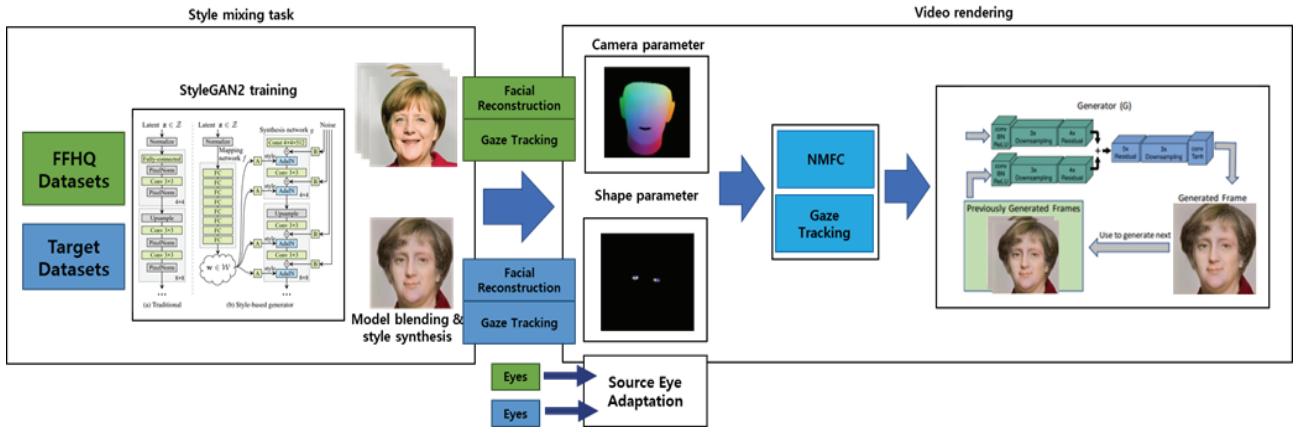


Fig. 3. The Overall Structure of the Proposed System

2.2 딥러닝 기반의 동영상 합성 기술

헤드투헤드++의 네트워크는 두 비디오에 대하여 얼굴의 3차원 기하학적 정보를 추출하고 이를 기반으로 3차원으로 얼굴을 재구성하여 동영상 내 원본 비디오와 대상 비디오의 자연스러운 합성이 되도록 학습한다.

1) 헤드투헤드++

3차원 얼굴 복원은 3차원 얼굴 모형을 통해 2차원의 프레임에서 3차원의 얼굴 정보를 추출한다. 3차원 얼굴 모형[4]은 사람 얼굴의 3차원 표현을 위한 생성 매개변수 모델이다. 헤드투헤드 네트워크는 3차원 얼굴 모형을 통해서 입력 시퀀스에 나타나는 얼굴을 3차원으로 재구성하고, 얼굴을 3차원으로 추적한다.

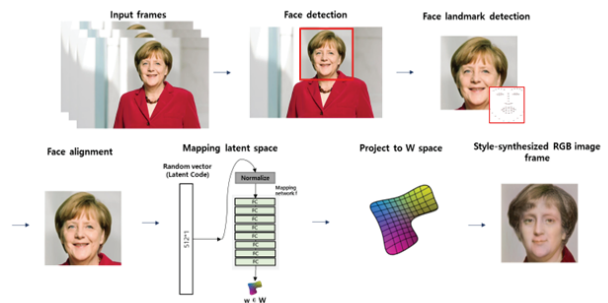


Fig. 4. The Process of Style Mixing Tasks

3. 얼굴 동영상 합성 네트워크

3.1 동영상 스타일 합성을 위한 네트워크 전체 구성

본 연구에서는 기존 동영상 스타일 변환의 개선점들을 반영하여 얼굴 동영상에 대하여 더욱 효율적으로 스타일을 합성할 수 있는 시스템을 제안한다. Fig. 3은 얼굴 동영상 합성을 위한 전체 시스템을 나타낸다. 이 시스템에서는 동영상에 대하여 일관적인 콘텐츠와 스타일을 생성하여 개연성 있는 동영상을 생성한다. 얼굴의 포즈와 같은 결과물은 자연스럽지만, 표정, 시선, 입 모양 등의 디테일한 얼굴의 특징에 대한 부분은 한계가 있다. Fig. 4와 같이 먼저 인물의 스피치 모습이 담긴 비디오가 입력되면, 프레임 내 얼굴이 있는지 확인하여 얼굴을 256 x 256의 크기로 얼굴 정렬 및 검출하여 크롭한다. 얼굴에 대한 전처리가 완료되면, 타겟 데이터셋 스타일 합성을 진행한다. 이 때, 얼굴 데이터셋을 학습한 스타일GAN2 네트워크와 메트로폴리탄 박물관의 명화 얼굴 데이터셋인 메트로폴리탄 얼굴 데이터셋을 학습한 스타일GAN2 네트워크를 서로 블렌딩하여 스타일 합성 모형을 만들고, 이를 통해 프레

임 합성을 진행한다. 프레임 속 얼굴에 대하여 3차원 재건을 하고 3차원 얼굴 랜드마크 검출 등의 카메라 파라미터와 모양 파라미터를 얻는다. Flownet2[25,26]를 통해서 합성된 이미지로부터 눈동자를 추출하여 시선을 검출한다. 얼굴 3차원으로 재건[4]하여 각 프레임 별로 스케일, 회전 등의 카메라 파라미터를 얻어온다. 이 단계에서 각 비디오에서의 평균 아이덴티티 계수를 얻어 저장한다. 3차원으로 얼굴 랜드마크를 검출하여 전처리 해준 후, 학습한다. 3개의 테스트를 하여 프레임 별로 출력하여 이를 비디오로 인코딩한다.

3.2 동영상 내 얼굴 검출 및 스타일 합성

1) 얼굴 검출 및 정렬

FFHQ[29] 데이터셋과 스타일 합성을 위한 데이터셋들의 모델을 블렌딩하여 사람의 얼굴에 스타일 변환을 진행할 수 있는 새로운 모델을 만든다. 페이스 포렌식[6] 데이터셋의 15명의 인물에 대한 비디오 데이터셋으로부터 총 344,700장의 RGB 이미지를 확보한다. 그리고 이미지 내의 얼굴들을 MTCNN[29]을 통해서 입력 이미지 내 얼굴을 검출하고 이를 얼굴을 정면의 모습으로 정렬하여 1024x1024 크기로 얼굴을 추출하여 얼굴에 대한 정확한 작업이 수행될 수 있도록 전처리 한다.

2) 얼굴 스타일 합성

얼굴 검출 네트워크 MTCNN의 네트워크를 통하여 얼굴을 검출하고, 얼굴 랜드마크와 비교하여 앞모습으로 정렬된 얼굴을 8개의 완전 연결층에 통과시켜 잠재공간 W에 맵핑한다. 맵핑된 얼굴 데이터를 미리 학습시켜놓은 모델을 통해 합성한다. 이때, 모델에 얼굴 이미지를 투사하여 500번의 반복으로 학습된 모델 중 특징값이 가장 비슷하도록 해야하기 때문에, 특징의 유클리드 거리를 가장 작은 값으로 만드는 얼굴을 찾아내어 이를 저장한다. 스타일 변형 시, 스타일 갠을 활용하여 학습된 합성 네트워크의 w 공간에서 프로젝션을 거쳐 가장 공간 내에서 가장 가까운 얼굴을 찾는 것이어서 표정, 포즈 등의 디테일이 합성된 모델의 분포를 자연스럽게 따라가게 된다.

3.3 동영상 프레임에 대한 3차원 전처리 및 렌더링 학습

1) 동영상 프레임에 대한 3차원 전처리

2차원의 얼굴 이미지에 대한 3차원 얼굴 정보 추출 및 얼굴 재건, 그리고 그 데이터의 학습 및 테스트를 위한 전반적인 흐름을 나타내는 구조도이다. Fig. 5에서 먼저 입력 이미지로부터 검출된 시선에 대한 정보와 3차원 얼굴 정보 추출 비디오 렌더링 학습을 진행한다. 모든 얼굴 이미지에 스타일 합성 처리를 하여 저장한 후, 데이터 셋 내에 3차원 얼굴 모델을 통해서 모든 비디오에 대한 3차원 얼굴 재건과 NMFC[37](Normalized Mean Face Coordinate) 이미지에 대한 연산을 수행한다. NMFC[37]는 얼굴의 3차원 좌표에 대한 평균을 정규화한 값이며, Fig. 6의 왼쪽 그림은 정규화된 평균 얼굴이고, 컬러맵으로 정규화된 얼굴과 함께 z-buffer에 의해 렌더링된 3차원 얼굴인 투사한 정규화 3차원 좌표[30]의 생성 이미지이다.

3차원 얼굴에 대한 재구성은 RGB 이미지로부터 Dense-FaceRec[37] 네트워크의 3차원 텐스 페이스 메쉬를 이용한다. Fig. 6는 3차원 얼굴 모형을 기반으로 생성된 데이터들에 대한 결과 이미지이다. 이미지 맵과 깊이 맵, 그리고 컬러 맵을 뽑아 카메라 파라미터를 추정하며, 3차원 얼굴 모형 베이스에 모양을 투영하여 3차원 얼굴 모형의 아이덴티티와 표정 계수를 생성한다. 이 때, 카메라 파라미터란 각 프레임에 대한 스케일, 회전, 변환과 같은 값을 의미하며, 3차원 아이덴티티와 표정 계수는 각 비디오에 대한 평균 아이덴티티 계수를 의미한다. 그리고, 3차원 얼굴을 스캔할때 사람의 시선 방향을 포착하기 어려운 점을 보완하기 위해 FlowNet2 [25, 26]을 활용한다. 추출한 랜드마크에 2개의 시선도 추가하여 총 70개의 랜드마크를 획득한다.

2) 동영상 렌더링 학습

학습 시, 비디오 전체 프레임의 2/3에 해당하는 프레임을 학습 데이터로 구성하고, 나머지는 테스트를 위한 데이터셋으로 구성하여 준비한다. 앞서 3차원 전처리를 통하여 구한 비디오 프레임과 RGB 비디오에서 검출한 입과 눈을 검출한

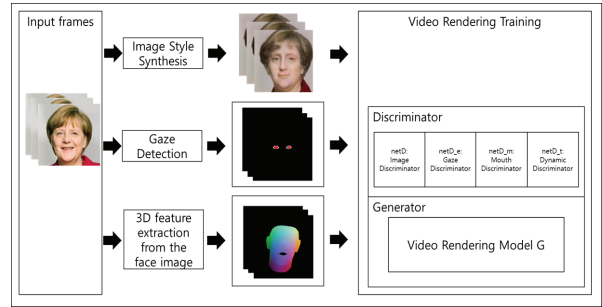


Fig. 5. The Structure of 3D Image Preprocessing and Video Rendering Training

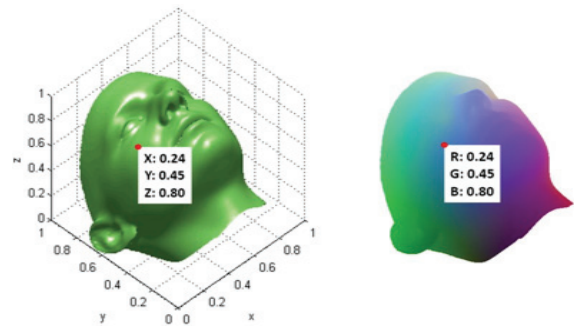


Fig. 6. Normalized Coordinate Code(NCC) Projected Normalized Coordinate Code(PNCC)

다. 이 때, 입 모양의 중간값과 눈의 중간값을 구하여 생성 모델에 입력한다. 이미지, 시선, 입 모양, 역동성에 대한 판별기는 모두 그림5와 같은 구성의 판별 모델을 사용하며, 멀티 스케일 판별기를 통해 입 모양 판별기와 시선 판별기의 손실함수를 구한다. 이때, 판별기는 합성 프레임과 실제 프레임을 구별하도록 학습되며, 실제 프레임과 합성 프레임에 대한 판별 후 입모양 판별기에 전달되어 입 모양이 생성된 입 모양 인지에 대한 진위여부를 판별한다. 비디오 렌더링 학습 네트워크는 그림의 생성모델과 같다. 역학 판별기를 통해서 실제와 생성된 시퀀스에 대해 임시로 만들어진 프레임과 눈동자에 대해 임시로 만들어진 프레임들을 손실함수를 통해 손실 값을 구하고, 이를 통해서 같은 인물인지 판별한다. 역학 판별기는 비디오에 대해 시간적인 역동성을 감지하도록 학습되는데, 연속적인 실제 프레임과 가짜 프레임을 입력으로 받아 옵티컬 플로우가 균일하게 생성되도록 판별한다.

4. 실 험

4.1 비디오 스타일 합성 실험

본 장에서는 비디오 생성 네트워크에 관하여 실험하였다. 실험 환경은 Table 1과 같다. Table 2는 학습 파라미터에 대한 내용이다. 학습에 사용된 데이터셋은 FFHQ, Face Forensic이며, 이를 스타일갠의 Stylegan2-config-f 모델을 이용하여 스타일 합성한다.

Table 1. Experiment Environment

		PC
H/W	CPU	i9-9700
	GPU	RTX-2080ti
	RAM	64GB
S/W	OS	Ubuntu 18.04
	CUDA	10.1
	python	3.6
	Tensorflow	1.14.0

Table 2. Training Parameters

		Training parameters
Dataset		FFHQ
		Cartoon Face
		Face Forensic++
Style mixing parameter	network	Stylegan2-config-f
	Style layer	128
	output size	256
Video rendering test		self-reenactment

이때, Stylegan2-config-f의 합성 레이어는 128을 사용하고, 이미지 사이즈는 256으로 출력한다. 그리고, 비디오 렌더링에 대한 학습은 self-reenactment를 사용한다. 페이스 포렌식 데이터셋의 총 15명의 인물에 대하여 스타일GAN2의 생성모델로 합성한 프레임과 3장에서 제안한 시스템을 통하여 자기 재구성에 대한 실험을 진행하였다.

1) 스타일 합성에 대한 정량적 성능 평가

제안하는 시스템에 대한 스타일 합성에 대한 성능을 평가 지표 PPL[10](Perceptual Path Length), FID[35](Fréchet Inception Distance)를 통해 측정한다. PPL과 FID는 데이터가 실제와 유사하게 생성되었는지를 측정하는 지표이다. Table 3은 정치인 마크롱의 스피치 비디오에 대한 자기 재구성의 정량적 성능을 측정하여 스타일GAN, 스타일GAN2의 성능과 비교한다.

2) 동영상 합성에 대한 정량적 성능 평가

제안하는 네트워크에 대한 동영상 합성에 대한 성능을 이미지의 품질과 다양성을 평가하는 지표인 LPIPS[10] ((Learned Perceptual Image path similarity)), FID의 평가 지표를 통해 측정한다. 이때, LPIPS의 지표는 높을수록, FID의 지표는 낮을수록 더욱 좋은 성능을 나타낸다. Table 4는 이미지 변환에 대한 정량적 성능을 측정하여 동영상 합성 네트워크들과 비교한다.

4.2 실험 결과

Fig. 7은 정치인 마크롱에 대한 스타일 자기 재구성을 한 결과 이미지이다. Fig. 7의 첫 번째 행은 입력 비디오를 프레

Table 3. Evaluation of Style Synthesis in Self-reconstruction

	PPL	FID
ours	152.4	64.79
StyleGAN	212.1	4.40
StyleGAN2	145.0	2.84

Table 4. Evaluation for Video Synthesis of Self-reconstruction

	LPIPS	FID
ours	0.344	64.79
DRIT++	0.201	63.8
CouncilGAN	0.430	39.1
StarGANv2	0.427	59.8
GNR	0.505	34.4
head2head	-	39.35

임별로 나열한 것이고, 두 번째 행의 그림은 스타일GAN2 네트워크만을 이용하여 매 프레임마다 디즈니 스타일로 합성한 프레임이다. 세 번째 행은 시선을 추출한 프레임이며, 네 번째는 정규화된 3차원 얼굴 좌표의 평균값에 대한 그림이다. 다섯 번째, 여섯 번째 행은 각각 이미지에 대한 히트맵과 얼굴에 대한 히트맵에 대한 그림이다. 마지막 행의 그림들은 자기 재구성 테스트를 통해 스타일을 합성한 프레임이다. Fig. 8은 정치인 메르켈, 마크롱, 오바마에 대한 스타일 자기 재구성하여 비디오 스타일 합성한 결과를 기존 이미지 합성 네트워크인 스타일GAN2의 결과 이미지와 비교해놓은 그림이다. 각 인물에 대해 첫 번째 줄은 원본(소스) 비디오에 대한 프레임들이고, 두 번째 줄은 세 번째 줄은 얼굴 비디오 스타일 합성을 진행한 결과 비디오의 프레임들이다. 그리고 세 번째 줄은 원본 비디오에 대하여 매 프레임 당 스타일 변환을 한 이미지이다.

비디오 렌더링 학습을 통하여 추출된 시선과 입모양 데이터를 통해서 원본 프레임의 표정, 시선, 입모양 등의 정보를 더욱 흡사하게 전이한다. 더불어, 기존의 방법처럼 매 프레임 스타일 변환을 했을 때 보다 더욱 일관적인 스타일 변환이 유지되는 것을 확인할 수 있다. 비디오 얼굴 합성을 통한 스타일 합성의 표정이나 포즈와 같은 디테일이 스타일GAN2 네트워크만을 이용한 스타일 합성의 결과보다 원본의 정보를 더욱 정확하고, 디테일하게 표현한다. 이 부분은 이미지 대 이미지 변환 네트워크를 기반으로 한 GNR 네트워크에서 비디오 합성 시 표정 등의 디테일을 옮겨오기 어려운 부분을 개선한다. 더불어, 헤드투헤드++ 네트워크의 움직임 판별기를 통해서 스타일 합성 시 스타일 변환으로 인해 콘텐츠 외적인 스타일 부분이 일정하지 못한 부분을 보완해준다. 이로 인해 좀 더 자연스럽게 안정적으로 비디오 내 인물의 스타일을 합성한다. Table 3과 Table 4는 Table 2에서 진행한 자기 재구성 테스트에 대하여 정량적 성능 평가에 대하여 정리한 표이다. Table 3에서는 본 연구에서 스타일 합성에 대한 성능을 측정하기 위해 PPL과 FID를 계산하여 기존의 스타일 합성 네트

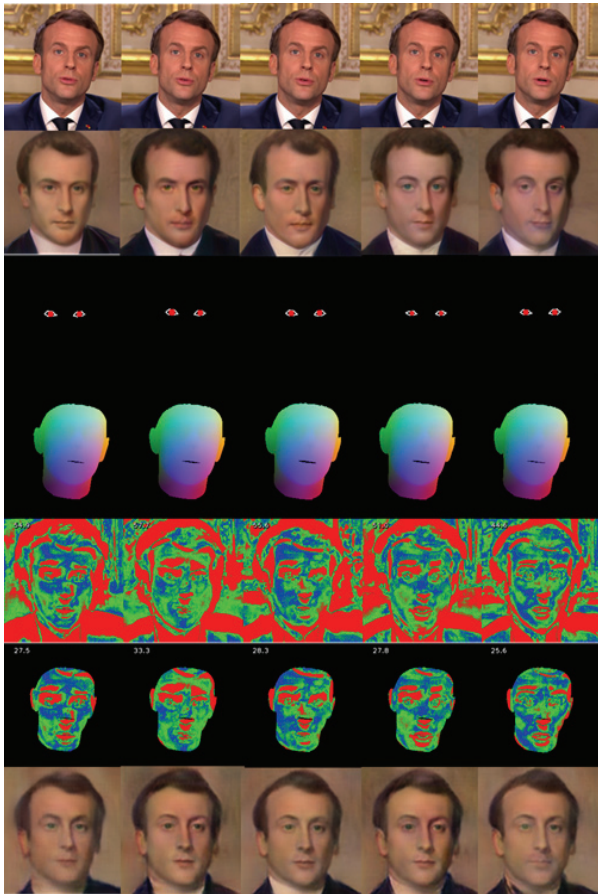


Fig. 7. Video-style Synthesized 3D Self-reconstruction Test Results for Macron

워크와 비교하였다. 스타일 변환 시, 두 데이터 셋 사이에 유사도를 계산하는 FID에 대한 성능 또한 낮게 측정되었다. 특징의 얽힘 정도를 평가하는 PPL에 대한 성능은 152.4로, 스타일갠보다는 좋은 스타일 영킴 분리(Disentanglement) 성능을 보였지만, 스타일갠2보다 낮은 성능을 냈다. 본 연구의 LPIPS는 0.344로 0.201인 DRIT++[34]의 LPIPS보다 좋은 성능을 하는 것을 확인했다. 자기 재구성 테스트에서 스타일갠2를 이용한 스타일 변환과 비디오 스타일 합성의 결과를 비교해봤을 때 정량적으로나 정성적으로나 스타일 변환의 성능이 좋은 것은 스타일갠2의 결과물이다. 하지만, 정규화된 3차원 좌표의 평균값과 카메라 파라미터를 통해 3차원의 전처리 작업을 거친 비디오 스타일 합성 네트워크의 결과 이미지는 포즈, 표정, 정체성에 대한 유사도가 높다. 또한, 동영상 스타일 합성에 대한 결과물이 기존의 스타일 합성 네트워크보다 안정적임을 확인했다.

5. 결 론

본 연구에서는 기존의 동영상 합성 네트워크인 헤드 두 헤드 네트워크를 기반으로 하여 스타일 변환 및 합성 네트워크



Fig. 8. Face Video Style Synthesis Result Image

의 동영상 합성의 한계점을 극복하고자 생성 네트워크를 확장한다. 본 논문의 네트워크에서는 동영상 합성을 위해 스타일갠을 통한 스타일 합성과 동영상 합성 네트워크를 통해 스타일 합성된 비디오를 생성하기 위해 네트워크를 학습시키고, 그 모델을 바탕으로 다양한 실험을 수행하였다. 다양한 관점에서 특징점 임베딩을 하기 위하여, 스타일갠2와 헤드투헤드 네트워크의 구조를 사용하였으며, 페이스 포렌식 데이터셋의 스피치 비디오에서 프레임 추출하여 이를 얼굴의 2차원적, 그리고 3차원적으로 전처리를 한다. 이 때, 스타일갠2의 학습 모델을 통해 프레임 별 스타일을 변환하고, 이를 통해 스타일 합성된 얼굴에 원본 얼굴의 정보를 이용해 재구성한다. 동영상 합성 시 자기 얼굴 재연 테스트를 통하여 비교적 자연스럽고 개연성있는 데이터를 생성하여 성능을 높일 수 있다. 또한, 원본 비디오로부터 자기 얼굴 재연을 통해 개연성과 일관성이 유지되는 안정적인 스타일 합성 비디오를 생성할 수 있음을 보였다. 앞으로의 연구에서는 스타일 변환 이전에 얼굴에 대한 3차원의 전처리를 통해서 얼굴의 3차원 랜드마크나 카메라 파라미터, 시선 추적에 대한 디테일을 얻고, 이를 동영상 합성에 충분히 이용한다면 머리와 얼굴 정보를 통한 비디오 스타일 합성에서도 안정적인 비디오 데이터를 생성할 수 있을 것으로 판단된다.

References

- [1] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June*, 2019. <https://doi.org/10.1109/CVPR.2019.00453>.
- [2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020. <https://doi.org/10.1109/CVPR42600.2020.00813m>.
- [3] M. J. Chong, "GANs N' roses: Stable, controllable, diverse image to image translation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021.
- [4] L. Tran and X. Liu, "On learning 3D face morphable model from in-the-wild images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.43, No.1, pp.157-171, 2021. <https://doi.org/10.1109/TPAMI.2019.2927975>.
- [5] T. C. Wang et al., "Video-to-video synthesis," *Advances in Neural Information Processing Systems, 2018-December*, 2018.
- [6] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," *Proceedings of the IEEE International Conference on Computer Vision, 2019-October*, 2019. <https://doi.org/10.1109/ICCV.2019.00009>.
- [7] M. R. Koujan, M. C. Doukas, A. Roussos, and S. Zafeiriou, "Head2Head: Video-based neural head synthesis," *Proceedings - 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020*, 2020. <https://doi.org/10.1109/FG47880.2020.00048>.
- [8] M. C. Doukas, M. R. Koujan, V. Sharmanska, A. Roussos, and S. Zafeiriou, "Head2Head++: Deep facial attributes re-targeting," *arXiv e-prints arXiv: 2006.10199*, 2020.
- [9] MetFace dataset [Internet], <https://github.com/NVlabs/metfaces-dataset>.
- [10] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. <https://doi.org/10.1109/CVPR.2018.00068>.
- [11] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoder," *arXiv preprint arXiv:2003.05991*. 2020.
- [12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.
- [13] I. J. Goodfellow et al., "Generative adversarial nets," *Advances in Neural Information Processing Systems*, 3(Jan.), 2014. https://doi.org/10.3156/jsoft.29.5_177_2.
- [14] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional GANs," *International Conference on Learning Representations*, 2016.
- [15] X. Huang and S. Belongie, "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization," *Proceedings of the IEEE International Conference on Computer Vision*, 2017. <https://doi.org/10.1109/ICCV.2017.167>.
- [16] X. Han, L. Zhang, K. Zhou, and X. Wang, "ProGAN: Protein solubility generative adversarial nets for data augmentation in DNN framework," *Computers and Chemical Engineering*, Vol.131, 2019. <https://doi.org/10.1016/j.compchemeng.2019.106533>.
- [17] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, Vol.14, 2013. <https://doi.org/10.1184/R1/6475463.V1>.
- [18] N.-A. Lahonce, Flickr-Faces-HQ Dataset (FFHQ), Nvidia, 2020.
- [19] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in Neural Information Processing Systems*, Vol.31, 2018.
- [20] D. A. Pisner and D. M. Schnyer, "Support vector machine," *In Machine Learning: Methods and Applications to Brain Disorders*, 2019. <https://doi.org/10.1016/B978-0-12-815739-8.00006-7>.
- [21] A. Mathiasen and F. Hvilshøj, "Fast fréchet inception distance," *arXiv preprint arXiv:2009.14075*, 2020.
- [22] O. Nizan and A. Tal, "Breaking the cycle-colleagues are all you need," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020. <https://doi.org/10.1109/CVPR42600.2020.00788>.
- [23] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Pix2Pix," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [24] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint arXiv:1703.10593*, 2017.
- [25] J. Han, J. Tao, and C. Wang, "FlowNet: A deep learning framework for clustering and selection of streamlines and stream surfaces," in *IEEE Transactions on Visualization and Computer Graphics*, Vol.26, No.4, pp.1732-1744, 2020, doi: 10.1109/TVCG.2018.2880207.

- [26] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. <https://doi.org/10.1109/CVPR.2017.179>.
- [27] L. Yuan, C. Ruan, H. Hu, and D. Chen, "Image inpainting based on Patch-GANs," in *IEEE Access*, Vol.7, pp.46411-46421, 2019, doi: 10.1109/ACCESS.2019.2909553.
- [28] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "LSGAN," *Proceedings of the IEEE International Conference on Computer Vision*, 2017.
- [29] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "MTCNN," *IEEE Signal Processing Letters*, Vol.23, No.10, 2016.
- [30] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6D object pose and size estimation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. <https://doi.org/10.1109/CVPR.2019.00275>.
- [31] B. J. B. Rani and L. M. E. Sumathi, "Survey on applying GAN for anomaly detection." *2020 International Conference on Computer Communication and Informatics (ICCCI)*, 2020. <https://doi.org/10.1109/ICCCI48352.2020.9104046>
- [32] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution." *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9906 LNCS, 2016. https://doi.org/10.1007/978-3-319-46475-6_43.
- [33] H. Y. Lee et al., "DRIT++: Diverse Image-to-Image Translation via Disentangled Representations," *arXiv preprint arXiv:1905.01270*, 2019.
- [34] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "GANSpace: Discovering interpretable GAN controls," *Advances in Neural Information Processing Systems*, Vol.33, pp.9841-9850, 2020.
- [35] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *arXiv preprint arXiv:1804.01005*, 2018.
- [36] J. H. Lee, M. J. Sung, J. W. Kang, and D. Chen, "Learning dense representations of phrases at scale," *arXiv preprint arXiv:2012.12624*, 2020.



최희조

<https://orcid.org/0000-0002-4879-6111>
 e-mail : heejo0624@seoultech.ac.kr
 2013년 동국대학교 영어영문학과(학사)
 2020년 ~ 현재 서울과학기술대학교
 IT미디어공학과 석사과정
 관심분야 : Computer Vision, GAN,
 Style Synthesis Network



박구만

<https://orcid.org/0000-0002-7055-5568>
 e-mail : gmpark@seoultech.ac.kr
 1980년 한국항공대학교 전자공학과(학사)
 1984년 연세대학교 전자공학과(석사)
 1986년 연세대학교 전자공학과(박사)
 2016년 ~ 현재 서울과학기술대학교
 IT전자미디어공학과 교수
 관심분야 : Computer Vision, Digital Broadcasting