Journal of The Institute of Electronics and Information Engineers Vol.60, NO.5, May 2023 https://doi.org/10.5573/ieie.2023.60.5.60
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2023-60-5-7

계층별 트랜스포머를 활용한 비선형 스타일 결합 기반 생성적 스타일 변환 기법

(Generative Stylization using Non-linear Style Mixing via Layer-wise Transformers)

이 준 형*, 김 준 우*, 오 희 석**, 지 준**

(Junhyoung Lee, Junwoo Kim, Heeseok Oh, and Jun Ji©)

요 약

Stylization은 주어진 이미지를 원하는 특정 스타일로 변환하는 기술이다. 초기의 연구에서는 그래디언트를이용한 연구가 진행되었으나, 최근에는 고품질의 얼굴 이미지 생성이 가능한 생성 모델의 발전으로, 생성망을 활용한 stylization 기술이 발전되어왔다. 이 중 JoJoGAN은 GAN inversion을 통해 찾아낸 참조 스타일의 잠재 벡터와 임의의 노이즈의 가중합 통해 fine-tuning이 가능한 학습 과정을 소개했으며, 한 장의 참조 이미지만을 사용해 우수한 스타일 변환이 가능함을 보였다. 하지만 기존 가중합 기반 스타일 결합은 단순 선형 결합 형태로써 스타일 코드를 표현하는 매니폴드가 제한되는 문제점을 갖는다. 따라서 해당 잠재공간을 활용한 stylization 결과물은 미세한 표현을 효과적으로 반영하지 못함과 동시에 다양성에서 한계를 갖는다. 이러한 문제점을 해결하고자 본 논문에서는 두 가지 방법을 제안한다. 첫 번째로 다양성과 생성 품질을 높이기 위한 비선형적 스타일 결합으로써 스타일 잠재 공간을 폭 넓게 구성하는 방법이다. 이는 확장된 스타일 공간에서의 다양한 샘플링을 가능하게 함으로써 결과물의 품질 및 다양성을 향상시킨다. 두 번째로 스타일 결합에 있어서 공간적 장기 의존성을 파악해 의미 있는 스타일 특징의 조합을 만들어낼 수 있도록 계층별 Transformer 구조를 활용했다. 즉, coarse, middle, fine 계층에 대해 별도로 독립적인 신경망을 두어 스타일 코드를 인코딩한다. 제안하는 방법을 통해 전체적 외형은 원본의 외형을 따르도록 하고 미세영역에 대해 참조 스타일을 반영하도록 유도함으로써, 정량적/정성적/주관적 실험 결과를 통해 개선된 stylization 성능을 확인했다.

Abstract

Stylization is the method of transforming a given image into a specific desired style. In early studies, studies using gradients were conducted, but recently, stylization using generative networks have been developed along with the development of a generative model capable of generating high-quality facial images. Among them, JoJoGAN introduced a learning process that enables fine-tuning through the weighted sum of the reference style latent vector and random noise found through GAN inversion and showed that excellent style transformis possible using only one reference image. However, the existing weighted sum-based style combination has a problem in that the manifold expressing the style code is limited as a simple linear combination form. Therefore, the result of stylization using the potential space does not reflect the fine expression effectively and at the same time has a limit in diversity. To solve this problem, this paper proposes two methods. First, it is a method to broadly compose a style latent space by combining non-linear stylization to increase the diversity and quality of creation. This allows for different sampling in an expanded style space, improving the quality and variety of the output. Second, we used a hierarchical Transformer structure to identify spatial long-term dependencies in style combinations to generate meaningful combinations of style features. In other words, style codes are encoded by separate independent neural networks for coarse, medium and fine layers. In the proposed method, improved stylization performance was confirmed through quantitative/qualitative/subjective experiments by inducing the overall appearance to follow the original appearance and to reflect the reference style for fine details.

Keywords : Transformer, Style mixing, Stylization

*비회원, 한성대학교 IT융합공학과(Department of IT Convergence Engineering, Hansung University)
**정회원, 한성대학교 AI응용학과(Department of Applied Artificial Intelligence, Hansung University)
© Corresponding Author(E-mail : jun@hansung.ac.kr)
※ 본 연구는 한성대학교 교내학술연구비 지원과제임.
Received ; November 22, 2022 Revised ; February 5, 2023 Accepted ; March 16, 2023

Ⅰ. 서 론

최근 소셜미디어 혹은 모바일 어플리케이션을 통해 입력된 사용자 얼굴 사진을 특정 스타일로 변환하는 stylization 기술이 대중적으로 보급되었다. Stylization 응용은 현재 장기 실종자의 현재 외모 추정, 가상공간 내 아바타 생성 및 영화 등 영상 콘텐츠의 후반 작업과 같은 특정 전문 분야에도 적극적으로 활용되고 있다.

Stylization 연구 분야의 초기 기술로 그래디언트를 이용한 neural style transfer[1~3]가 널리 활용되었으나, 최근에는 GAN(generative adversarial network)을 위시한 생성망의 발전을 통해 고해상도의 얼굴 이미지 생성이 가능해짐으로써, 기 학습된 생성망을 스타일 결합(style mixing) 엔진으로 활용하는 stylization 알고리즘들이 개발되고 있다[4, 5].

하지만 스타일 결합을 통한 stylization 기술은 미세한 스타일들을 잘 구별하지 못해 학습된 데이터 분포로부터 생성된 샘플이 참조영상 내 타겟 캐릭터의 미세 특징을 효과적으로 반영하지 못한다는 문제점을 갖는다. 또한 제한된 형태만의 stylization을 수행함으로써 다양성(diversity) 측면에서 한계를 지니며, 이를 극복하기 위해서는 다양한 참조영상을 포함하는 대규모의 학습 데이터셋이 요구된다. 이를 해결하고자 Chong과 Forsyth는 하나의 참조영상만으로 기존 스타일 결합 엔진의 fine-tuning이 가능한 학습 scheme인 JoJoGAN[6]을 소개하였다. GAN inversion을 통해 찾아낸 참조영상의 잠재 벡터는 정규분포로 정의한 사전(prior)분포에서 샘플링한 임의의 노이즈와 가중치로 제어되는 스타일 결합을 통해 새로운 스타일 코드를 생성한다.

$$s' = \alpha s + (1-\alpha) z \tag{1}$$

즉, 수식 (1)에서 가중합을 통해 다양한 스타일 생성이 하나의 참조영상만으로도 가능하도록 하였고, 이는 데이터셋 증강 효과를 가짐으로써 미세 수준의 스타일을 반영함으로써 StyleGAN[7]의 생성망을 통한 높은 품질의 stylization 샘플 획득을 달성하였다.

하지만 이러한 가중합 형태의 스타일 결합은 스타일 코드가 임베딩된 스타일 공간을 구성함에 있어 선형 span에 의존함으로써, 두 벡터와 사이의 보간(interpolation)으로 이루어진 제한적 스타일 코드만을 생성할 수 있다는 단점이 있다. 그림 1(a)에서 보여주듯 랜덤 스타일 벡터와 참조 이미지의 스타일 벡터의 가중합을 통해 스타일 코드를 생성하며, 스타일 코드의 매니폴드가 제한되어 있다. 이는 최종적인 stylization 결과물의 미세 수준 표현과 다양성을 저해하는 원인으로 작용한다.

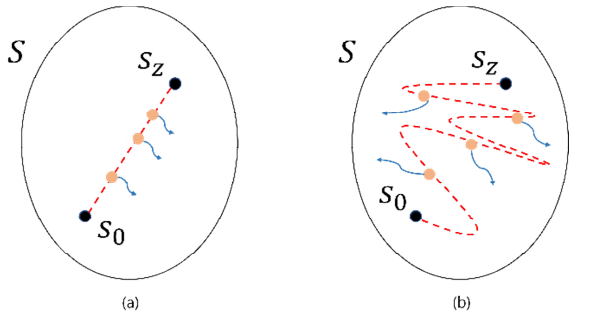

그림 1. 스타일 공간 내, 스타일 코드 획득을 위한 결합 방법 비교: (a) 선형 스타일 결합; (b) 비선형 스타일 결합

Fig. 1. Comparison of style mixing methods to obtain style codes: (a) Linear style mixing; (b) Non-linear style mixing.

본 논문에서는 선형 스타일 결합의 단점을 보완할 수 있도록 보다 넓은 스타일 공간을 표현할 수 있는 비선형 스타일 결합 알고리즘을 개발하였다. 우리는 특히 명시적인(explicit) 스타일 결합 방법을 이용하는 것 보다 최적의 내재적인(implicit) 비선형 스타일 결합을 신경망을 통해 학습시키고자 하였다. (b)에서는 비선형 스타일 결합을 통해 스타일 벡터를 생성함으로써 (a)보다 큰 스타일 공간에서의 스타일 코드 샘플링이 가능함을 개념적으로 가시화하였다.

이 때, 제안하는 방법은 스타일 결합에 있어 공간적 장기 의존성을 파악함으로써 의미 있는 스타일 특징 조합을 추정하고자 Transformer 구조[8, 9]를 활용하였으며 이에 캐릭터화의 결과, 즉 스타일 결합이 수행되어 생성된 캐릭터화 완성도가 향상됨을 확인할 수 있었다. 더불어 우리는 스타일에 따라 보다 원활한 조작이 가능함과 동시에 품질 높은 샘플 생성이 가능한 모델 구축을 위하여 스타일 코드를 계층별로 학습시켰다. 즉, 스타일 공간은 coarse(얼굴 윤곽, 외형 등), middle(머리 모양, 눈 모양 등), fine(머리 색, 피부 질감 등)계층을 가지고 별도로 독립적인 신경망을 두어 학습을 통해 계층별로 스타일 코드를 인코딩한다.

우리는 실험을 통해 제안하는 방법이 결과적으로 다양한 스타일 코드를 통해 생성망을 학습할 수 있어 풍부한 표현력을 지님과 동시에 우수한 품질의 stylization이 수행된 결과를 확인할 수 있었다.

본 논문의 구성은 Ⅱ장에서 stylization을 통한 캐릭터화 연구에 대한 기존 연구를 소개하며 문제점을 살펴

본다. Ⅲ장에서는 제안하는 비선형 스타일 결합을 통한 알고리즘 방법을 구체적으로 다룬다. Ⅳ장에서는 실험 결과를 토대로 기존 모델과의 정량적/정성적 성능 비교로써 개선된 stylization 성능을 확인하였다.

Ⅱ. 관련 연구

1. Image-to-Image translation 기반 stylization

GAN을 활용한 image-to-image translation 관점에서의 다양한 스타일 변환 모델들이 소개되었다. Pix2Pix[10]에서는 상이한 이미지 도메인 간 스타일 변환을 다루는 문제를 지도학습 방법론으로 정의하였고, 흑백영상을 컬러로 변환하거나 스케치를 실사로 변환하는 등의 다양한 응용이 가능함을 보였다. CycleGAN[11]에서는 cyclic consistency loss를 도입함으로써 스타일 변환에 있어 unpair 데이터셋으로도 스타일의 분포를 학습할 수 있는 방법을 고안하였다. CartoonGAN[12]에서는 cartoon 스타일화에 있어 부드러운 음영과 선명한 엣지를 강조하기위한 edge-preserving/smoothing이 반영된 목적 함수를 설계하였다. 하지만 기존 방법들의 경우, 스타일 텍스쳐가 강하게 반영될 시 원본 영상의 특징을 대부분 소실하게 되는 tradeoff가 존재하였다. AnimeGAN[13]에서는 이에 콘텐츠, 색, 음영에 대한 각각의 leveraging이 가능한 손실함수를 추가함으로써 향상된 품질의 스타일 변환이 가능함을 보였다.

이후 스타일 변환에 관한 연구는 사전분포와 GAN이 내재적으로 표현하는 스타일의 잠재공간(latent space)에 대한 논의가 주를 이루었다. UNIT[14]에서는 상이한 두 도메인이 서로 공유하는 잠재공간이 존재한다는 가정을 두어 두 영상을 하나의 잠재공간으로 매핑하고, 이를 영상으로 복원함으로써 더욱 자연스러운 스타일 변환이 가능함을 실험적으로 증명하였다. MUNIT[15]은 이러한 개념을 멀티 도메인으로 확장하여 다양성 측면에서 향상된 결과물을 생성할 수 있음을 보였으며, U-GAT-IT[16]에서는 attention 모듈을 도입해 원본 도메인과 타겟 도메인의 변환에 의미론적(semantic)으로 중요한 스타일 정보가 잠재공간으로 적절히 임베딩되도록 설계하였다.

2. 스타일 벡터의 swapping을 통한 스타일 변환

StyleGAN[7]은 데이터로부터 학습된 스타일 공간을 이용해 영상의 스타일을 계층별로 손쉽게 컨트롤할 수 있는 장점을 지니고 있으며, 다양한 스타일 변환 모델들이 전이 학습을 통해 이미지 조작에 활용하고 있다. 이러한 접근론에서는 주로 레이어 swapping 방법을 이용하며, Toonify[4]에서는 서로 다른 두 도메인의 데이터로부터 학습된 StyleGAN2[17] 생성망 내 특정 스타일 계층에 해당하는 가중치를 서로 바꾸어 캐릭터 스타일 반영을 달성했다. 또한 Cartoon-StyleGAN[5]에서는 StyleGAN2의 전이 학습 시, 초기 스타일 벡터와 생성망에 대한 가중치를 고정시킴으로써 타겟 영상이 원본 영상의 전역적 구조를 따를 수 있도록 설계하였다. 이러한 레이어 swapping 방법은 주로 낮은 해상도에서는 원본 이미지에 대해 학습된 생성망의 스타일을, 높은 해상도에서는 타겟 이미지에 대해 학습된 생성망의 스타일을 이용함으로써 콘텐츠와 텍스쳐의 밸런스를 유지하는 캐릭터화를 수행한다.

하지만 swapping 방법은 스타일의 단순한 레이어 임베딩 특성상 구조적으로 미세한 스타일을 섬세하게 표현하지 못하고, 몇몇 특정 부분에서의 색상 혹은 형태가 불완전한 스타일 결합에 의해 왜곡되어 나타난다는 문제점을 갖는다. 서론에서 언급한 JoJoGAN[6]에서는 한 장의 참조 이미지를 가지고 학습을 진행하며, 참조 이미지에 대한 GAN inversion을 수행하여 스타일 공간 내 참조 스타일 코드를 획득한다. 이후, FFHQ[7] 데이터셋에 기 학습된 StyleGAN2 생성망에서의 랜덤 스타일 벡터를 샘플링함으로써 타겟 스타일 벡터와 결합하여 추가적인 fine-tuning을 진행해 캐릭터로의 스타일 변환을 수행한다.

Ⅲ. 제안 방법

1. 비선형 스타일 결합을 통한 계층별 코드 생성

JoJoGAN에서의 선형 스타일 결합 방식은 그림 2에서 보여지는 바와 같이 참조 이미지의 앞머리와 눈썹과 같은 미세 스타일이 왜곡되어 도출되는 문제점을 가진다. [22]에서는 선형적 변환에서의 한계점을 변화의 요인이 없는 얼굴이 아닌 도메인에서 비선형 잠재 벡터 조작을 통해 높은 품질의 수정을 가능하게 했다. [23]에서는 선형 잠재 공간에 비선형 스타일을 투영하여 스타일을 결합함으로써 원본의 미세한 부분을 보존하면서 참조 스타일과 질감을 반영하게 했다. 따라서 비선형 스타일을 결합의 성공적이었던 결과를 보아 본 논문에서는 확장된 잠재 공간에서 원본과 참조 영상의 잠재 벡터를 계층적 Transformer를 활용한 비선형적 결합을 통해 다양성 및 품질 높은 캐릭터화를 수행하며 선

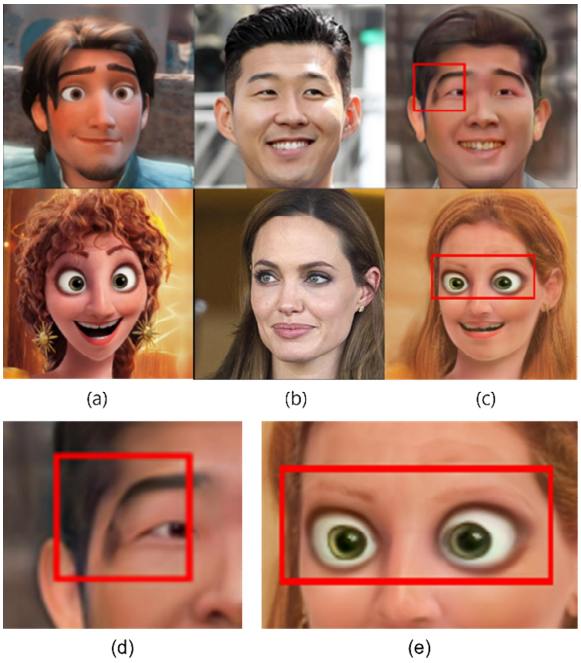

그림 2. 선형 스타일 결합 방법의 단점 및 이로 인한 스타일 변환 왜곡: (a) 참조 영상; (b) 원본 영상; (c) JoJoGAN 기반 stylization 결과 (붉은 박스 내부는 스타일 왜곡이 나타난 부분); (d)와 (e) (c)를 확대한 영상

Fig. 2. Drawbacks of linear style mixing, and the resulting distortions: (a) reference images; (b) source images; (c) outputs of JoJoGAN (the red boxes denote the distorted area); (d) and (e) extended (c) images.

형적 스타일 결합에서의 한계점 해결을 도모한다.

그림 3에서는 본 논문에서 제안하는 방법을 전체적으로 도식화하였다. 정규분포에서 샘플링된 랜덤 벡터는 mapping 네트워크를 통과해 랜덤한 스타일 벡터들 $Z \in R^{(18\times512)}$로 매핑된다. 이와 별개로 GAN Inversion을 통해 참조영상의 스타일 벡터들 $W \in R^{(18\times512)}$을 얻는다. 여기서 스타일 벡터 획득을 위한 GAN Inversion 과정은 Restyle[18]의 방법을 이용했다. 획득한 두 종류의 벡터들은 coarse, middle, fine 스타일 계층을 담당하는 각각 독립적으로 구성된 Transformer로 입력되어 스타일 코드 구성을 위해 결합된다. 이 때, coarse 스타일은 생성망의 앞쪽 레이어들에 전달되는 Z와 W 내 가장 앞부분의 4×512 크기 벡터들에 해당한다. middle 스타일 계층에서는 그 이후 4×512 크기의 벡터를 활용하고, fine 계층은 나머지 10×512 크기의 벡터가 담당한다. 최종적으로 각 계층별 벡터의 신경망을 통한 비선형 결합의 내재적 학습을 통해 스타일 벡터 $W' \in R^{(18\times512)}$의 인코딩이 가능하다.

이처럼 본 논문에서는 스타일 잠재 공간에서의 비선형적 결합을 유도함으로써 데이터 증강 효과를 도모하여 우수한 캐릭터화 성능을 달성 가능한 생성망의 학습이 가능하도록 접근하였으며 비선형 결합의 학습을 위한 구체적인 scheme을 남은 절에서 다룬다.

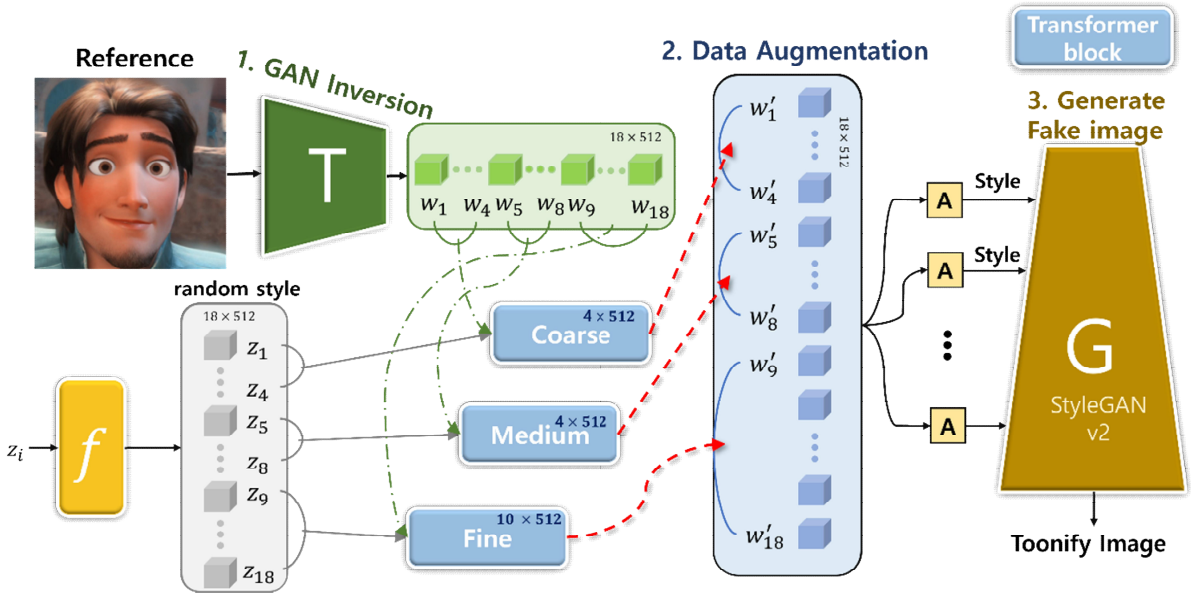

그림 3. 제안 방법의 전체 개요(TransMixing): 계층별 Transformer를 활용한 비선형 스타일 결합을 통한 스타일 코드 획득 및 생성망의 fine-tuning

Fig. 3. Overview of the proposed scheme(TransMixing): style code composition through non-linear style mixing using the layer-wise Transformers and fine-tuning the generative network.

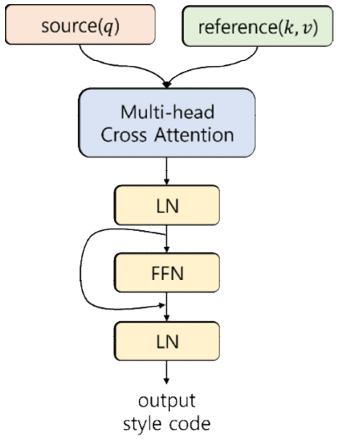

그림 4. 비선형 스타일 결합을 위한 수정된 Transformer 레이어
Fig. 4. The modified Transformer layer for non-linear style mixing.

2. 네트워크 구조

비선형 스타일 결합을 내재적인 함수로써 학습하기 위한 신경망은 ViT[8]의 디코더를 변형하여 활용하였으며, 이를 그림 4에 나타내었다. 제안하는 Transformer 기반 스타일 결합 연산의 핵심은 참조 이미지의 스타일과 랜덤 스타일의 상호 의존성 도출을 위한 cross-attention 과정에 있으며, multi-head cross-attention 모듈에 입력되는 query는 GAN inversion을 통해 찾아낸 스타일 공간에서의 참조 스타일 벡터이고, key와 value는 랜덤 스타일 벡터에 해당한다. 수식 (2)에서는 cross-attention을 구성한 식으로 query와 key을 통해 유사도를 구하고 value를 weight-sum을 통해 원본과 타겟의 비선형적인 결합이 진행된다. 즉, 스타일 계층마다 별도의 Transformer를 활용함으로써 계층별로 참조 이미지의 스타일과 랜덤 스타일 간 전역 연관성을 측정하여 스타일 공간을 구성한다.

$$\mathrm{softmax}\left(\frac{Q\times K^{T}}{\sqrt{d_k}}\right)\times V \tag{2}$$

3. 스타일 벡터를 활용한 생성망 one-shot 학습

이전 절까지 서술한 방법을 토대로, 비선형 결합으로써 증강된 스타일 벡터 기반의 one-shot 생성망 fine-tuning이 가능하며, 본 논문에서는 목적함수로써 향상된 화질의 결과물 생성을 위해 feature matching loss[19]를 도입하였다.

$$L(G(w';\theta),\ y)=\|D(G(w';\theta))-D(y)\|_1 \tag{3}$$

그림 5와 같이 생성자가 만들어낸 가짜 이미지와 참조 이미지에 관해 판별자의 특징을 스케일별로 추출하며 (resblock 2, 4, 5, 6 활성값 이용), 이들의 차이를 l_1 norm으로 계산하여 최소화하는 방향으로 생성망을 학습시켰다. 이는 이진 분류를 위해 판별자가 매핑하는 표현 공간의 다중 스케일 특징을 생성에 반영하기 위함이며, 생성 결과물의 디테일을 향상시킴으로써 인지적 화질의 상승을 유도한다.

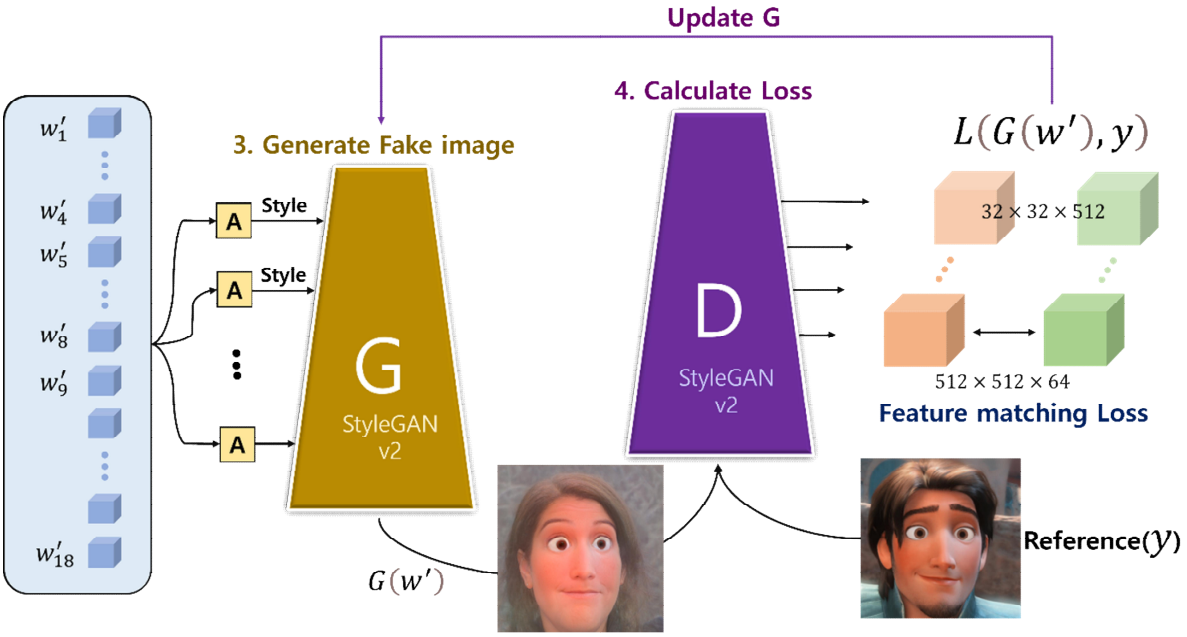

그림 5. Feature matching loss를 이용한 생성망 학습
Fig. 5. Training generative network uses Feature matching loss.

그림 6. 기존 기술과 제안 기술의 스타일 변환 결과: (a) 참조 영상; (b) 원본 영상; (c) JoJoGAN 결과; (d) TransMixing1: middle과 fine 계층에서 transformer를 적용한 모델; (e) TransMixing2: fine 계층에서만 transformer를 적용한 모델

Fig. 6. stylization results achieved by the proposed and previous methods: (a) reference image; (b) source image; (c) result of JoJoGAN; (d) result of TransMixing1: style-mixing Transformers were applied to both the middle and fine layers; (e) TransMixing2: style-mixing Transformers were applied to fine layers.

Ⅳ. 실험 결과

1. 정성적 결과

본 논문에서 제안하는 기술의 우수성을 확인하기 위해 두 가지 TransMixing 변형 모델에 대해 실험을 진행하였다.

√ TransMixing1: middle과 fine 계층에만 스타일 결합한 모델

√ TransMixing2: fine 계층에만 스타일 결합한 모델

원본 이미지의 외형을 최대한 유지할 수 있도록 하기 위해 coarse 계층을 제외한 middle과 fine 계층에서의 스타일 결합을 하는 TransMixing1, TransMixing2 방법의 생성 결과를 JoJoGAN과 비교하였으며, 그림 6에서 이를 정성적으로 확인할 수 있다.

그림 6에서 기존 기술의 결과물 (c)의 머리카락 색, 질감과 같은 미세한 부분의 참조 스타일과 치아, 눈동자 모양 같은 외형적인 부분의 원본 스타일이 소실되었다. 하지만 제안하는 스타일 결합의 학습을 통한 결과물인 (d), (e)은 좀 더 자연스럽고 세밀한 부분에서의 향상된 품질의 생성이 가능해짐을 확인 가능하다.

(d)와 (e)를 비교하면 상대적으로 (e)보다 (d)에서 눈썹, 눈동자, 입 모양 등 외형적인 부분이 참조 스타일의 형태를 반영해 원본 영상의 스타일을 소실하는 경향이 나타난다. 따라서 (e)가 더 stylization에 적합했다고 할 수 있고, 또 (d)와 (e)를 가지고 사용자의 선호도를 파악하기 위해 4의 유저스터디를 통한 결과도 확인할 수 있다.

그림 7은 제안 기술을 통해 각 성별에 맞는 캐릭터로 스타일 변환을 수행한 결과로 원본 영상에서의 인물에 원하는 스타일이 적절히 반영된 고품질의 캐릭터화가 수행된 결과를 확인할 수 있다.

2. 정량적 결과

정량적 수치로써 제안하는 모델의 우수성을 입증하고자 FID(Fréechet inception distance)[20], IS(inception score)[19], LPIPS(learned perceptual image patch similarity)[21] 지표를 활용하였으며, 표 1은 JoJoGAN과 제안하는 모델에 대한 측정 결과이다. FID는 참조 이미지의 분포와 생성된 영상의 분포에 대한 Wasserstein-2 거리로써 품질을 계산하는 지표이다. IS는 생성 결과의 분류 확률을 통해 다양성을 함께 반영하는 품질 지표이다. LPIPS는 분류 네트워크가 추출하는 특징 공간에서의 거리를 계산함으로써 인지적 화질을 표현하는 평가 지표이다.

제안하는 기술(TransMixing1)이 JoJoGAN을 상회하는 IS수치를 보여주며, 이는 더 다양하고 고품질의 결과물 생성이 가능함을 의미한다. 또한 제안 기술(TransMixing2)이 보다 높은 LPIPS를 가짐으로써, 시각적으로 보다 그럴듯한 결과물을 생성함을 객관적으로 확인할 수 있다.

3. 비교 실험 (Ablation study)

최적의 스타일 결합 네트워크를 찾기 위해 서로 다른 두 종류의 스타일 벡터 결합을 위한 Transformer 레이어를 각 계층별로(coarse, middle, fine) 적용할 때, 어떤 계층에서의 결합이 가장 효율적일지 측정하는 실험이다.

그림 8에서 (c)는 참조 이미지의 눈동자 스타일이 과도하게 반영되어 원본 이미지의 특징을 잃은 결과를 보

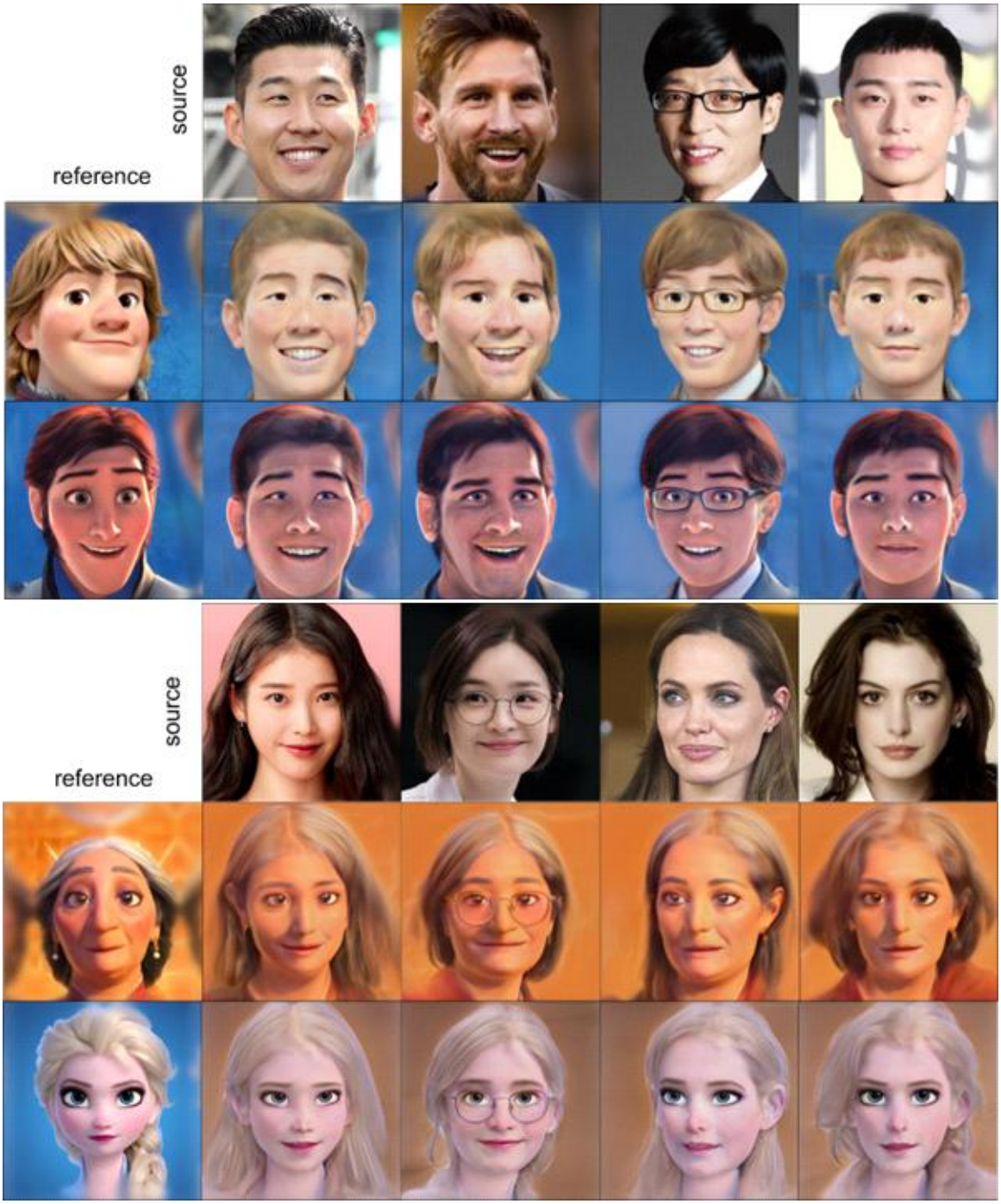

그림 7. 남자, 여자 도메인에 대한 스타일 변환 결과. 첫 번째 열과 첫 번째 행은 각각 참조영상과 원본영상에 해당
Fig. 7. Stylization outputs for male and female domains. The first column and the first row correspond to the reference and the source images, respectively.

였다. (d)는 얼굴 형태가 참조 이미지를 과도하게 반영하여 원본 이미지의 얼굴 형태를 유지하지 못했다. (e)는 (d)에 비해 향상된 캐릭터화를 수행하나 눈썹과 머리카락과 같은 fine 스타일의 컬러가 회색조로 저해되었다. (f)는 JoJoGAN의 한계점(그림 2)을 원본 이미지의 눈 스타일을 확장된 스타일 공간에서 반영함으로써 해결했다. 따라서 coarse 계층은 큰 외형 변화로 인해 참조 이미지와 유사하지만 원본 이미지와 많이 달라지기 때문에 스타일 결합에서는 적합하지 않았고, middle과 fine 계층에서의 스타일 결합이 효과적이고 우수한 스타일 변환 결과를 도출함을 확인하였다.

표 1. 스타일 변환 결과의 FID, IS, LPIPS 측정
Table 1. FID, IS, and LPIPS measurements for style mixing outputs.

	FID↓	IS↑	LPIPS↑
JoJoGAN	**104**	3.53	0.38
TransMixing1	110	**3.64**	0.38
TransMixing2	125	3.28	**0.40**

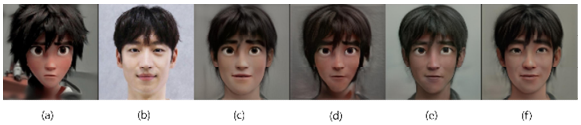

그림 8. 계층별 Transformer 적용 결과: (a) 참조 이미지; (b) 원본 이미지; (c) JoJoGAN 결과; (d) coarse, middle, fine 레이어에 Transformer 적용; (e) middle, fine 레이어에 Transformer 적용; (f) fine 레이어에 Transformer 적용
Fig. 8. Effects of applying layer-wise Transformer at each layer: (a) Reference image; (b) Source image; (c) result of JoJoGAN; (d) result when applying Transformer to coarse, middle, and fine layers; (e) result when applying Transformer to middle, and fine layers; (f) result when applying Transformer to fine layers.

4. 유저스터디 (User study)

제안하는 방법의 우수성을 증명하기 위해 우리는 해당분야 비전문가 일반인 61명(남자 38명, 여자 23명)을 대상으로 설문조사를 진행했다. 실험 디자인 설계는 다음과 같다. 남성, 여성 캐릭터로 설문을 진행하였으며, 실험자가 생각하기에 우수한 캐릭터화 결과물을 선택하도록 하였다. 이 과정에서 가설에 대한 단서는 일체 제공하지 않았으며, 실험자 스스로의 기준에 따라 캐릭터화의 품질에 대한 평가를 진행했다. 그리고 원본 이미지의 얼굴과 어떤 결과가 더욱 유사한지 선택하는 질문 부여를 통해 원본 이미지가 캐릭터화가 수행된 결과물에서도 잘 보존되는지에 대한 유사성 평가를 함께 진행하였다. 이를 통해, JoJoGAN과 우리의 모델(TransMixing1, TransMixing2)을 비교해 어떤 방법이 캐릭터화가 잘 되었는지를 평가했다.

그림 9에서 (d)을 보면 (a)의 참조 이미지로 모델을 학습해 (b)의 원본 이미지로 캐릭터화를 수행한 결과 이미지 (c), (d), (e)에서 확인할 수 있다. 표 2는 각 질문당 응답했던 결과를 나타낸 것이다. 남자 도메인에 해당하는 결과를 봤을때 캐릭터화 결과물의 품질이 좋은지 평가하는 질문에 우리의 모델인 TransMixing1(36.1%),

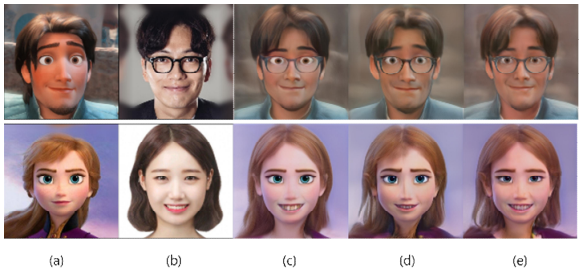

그림 9. 설문조사에서 모델별로 비교한 스타일 변환 결과의 이미지: (a) 참조 이미지; (b) 원본 이미지; (c) JoJoGAN 결과; (d) TransMixing1 결과; (e) TransMixing2 결과
Fig. 9. Utilized stimuli for the user study: (a) Reference image; (b) source image; (c) result of JoJoGAN; (d) result of TransMixing1; (e) result of TransMixing2.

표 2. 품질과 유사성에 관한 설문조사 결과 (단위: %)
Table 2. Result of the user study regarding quality and plausibility. (value: %)

	남자		여자	
	품질	유사성	품질	유사성
JoJoGAN	27.9	34.4	11.5	8.2
TransMixing1	36.1	23	63.9	14.8
TransMixing2	36.1	42.6	24.6	77

TransMixing2(36.1%)가 JoJoGAN(27.9%) 보다 월등히 높은 응답수를 나타냈다. 따라서 품질 측면에서 주관적 화질 선호도가 높음을 확인할 수 있다. 유사성에 해당하는 질문에는 제안하는 방법인 TransMixing2(42.6%)가 JoJoGAN(34.4%) 보다 높은 응답수를 기록하였다. 따라서 캐릭터화를 했을 때의 원본 이미지를 제안하는 방법이 JoJoGAN보다 잘 보존하는 결과를 얻었다. 특히 middle 계층에서 스타일 결합된 TransMixing1에서는 fine 계층보다는 덜 미세한 부분의 스타일인 입 모양, 눈썹의 모양 반영되었기에 참조 영상의 특징이 반영되어 원본 영상의 유사성이 떨어졌다. 그에 비해 TransMixing2 모델은 fine 계층에서만의 스타일 결합으로 타겟 영상의 미세한 부분만 반영하여 TransMixing1보다 좀 더 원본 영상의 특성을 유지할 수 있어 유사성 결과도 높게 나오는 결과를 확인할 수 있다. 따라서 캐릭터화가 JoJoGAN에 비해 품질과 유사성 질문에서제안 방법이 높은 선호도 결과를 얻어 우수한 캐릭터화를 수행했다고 결론지을 수 있다.

V. 결 론

우리는 선형적 스타일 결합의 한계점을 극복하고자 캐릭터화 수행과 관련한 연구를 수행하였다. 기존 미세 영역에서의 캐릭터 특징이 보존되지 못하던 단점을 문제로 정의하고 Transformer로 내재적으로 구성된 비선형적 스타일 결합 방법을 통해 보다 넓은 잠재 공간을 구성하였으며, 확장된 잠재공간에서의 스타일코드 샘플링을 통해 학습을 진행하는 방법으로 문제를 해결하였다. 이로써 스타일 변환 결과가 기존 기술 대비 정량적으로 우수한 품질을 도출함을 확인할 수 있었으며, 참조 이미지의 외형적인 모습 또한 잘 유지함으로써 설문 조사를 통해 정성적으로 높은 선호도를 확인하였다. 우리는 추후 본 기술을 발전시켜 가상의 메타버스 공간 내 캐릭터를 생성함으로써 3차원의 스타일 변환을 달성함을 목표로 연구 중이다.

REFERENCES

[1] L. A. Gatys, A. S. Ecker and M. Bethge, "Image style transfer using convolutional neural networks," *in Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2414-2423, 2016.

[2] B. Jeon and J. Jeong, "Blocking artifacts reduction in image compression with block boundary discontinuity criterion," *IEEE transactions on circuits and systems for video technology*, vol. 8, no. 3, pp. 345-357, Jun. 1998.

[3] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin and J. Shlens, "Exploring the structure of a real-time, arbitrary neural artistic stylization network," *ArXiv preprint arXiv:1705.06830*, May. 2017.

[4] J. N. M. Pinkney and D. Adler, "Resolution dependent gan interpolation for controllable image synthesis between domains," *arXiv preprint arXiv:2010.05334*, Oct. 2020.

[5] J. Back, "Fine-Tuning StyleGAN2 For Cartoon Face Generation," *arXiv preprint arXiv:2106.12445*, Jun. 2021.

[6] M. J. Chong and D. Forsyth, "Jojogan: One shot face stylization," *in European Conference on Computer Vision*, pp. 128-152, Oct. 2022.

[7] T. Karras, S. Laine and T. Aila, "A style-based generator architecture for generative adversarial networks," *in Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 4401-4410, 2019.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D.Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, Oct. 2020.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[10] P. Isola, J. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1125-1134, 2017.

[11] J. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *in Proceedings of the IEEE international conference on Computer Vision*, pp. 2223-2232, 2017.

[12] Y. Chen, Y. Lai and Y. Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 9465-9474, 2018.

[13] J. Chen, G. Liu and X. Chen, "AnimeGAN: a novel lightweight GAN for photo animation," *in International symposium on intelligence computation and applications*, pp. 242-256, May. 2020.

[14] M. Liu, T. Breuel and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in neural information processing systems*, 2017.

[15] X. Huang, M. Liu, S. Belongie and J. Kautz "Multimodal unsupervised image- to-image translation," *in Proceedings of the European conference on Computer Vision*, pp. 172-189, 2018.

[16] J. Kim, M. Kim, H. Kang and K. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to- image translation," *arXiv preprint arXiv:1907.10830*, Apr. 2020.

[17] T. Karras, S. Laine, M. Aittala, J. Hellsten,

J. Lehtinen and T. Aila, "Analyzing and improving the image quality of stylegan," *in Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 8110-8119, 2020.
[18] Y. Alaluf, O. Patashnik and D. Cohen-Or, "Restyle: A residual-based stylegan encoder via iterative refinement," *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6711-6720, 2021.
[19] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, 2016.
[20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, 2017.
[21] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 586-595, 2018.
[22] V. Khrulkov, L. Mirvakhabova, I. Oseledets and A. Babenko, "Latent Transformations via NeuralODEs for GAN-based Image Editing", *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14428-14437, 2021.
[23] Z. S. Liu, V. Kalogeiton and M. P. Cani, "Multiple Style Transfer via Variational AutoEncoder", *in IEEE International Conference on Image Processing*, pp. 2413-2417, 2021.

저 자 소 개

이 준 형(비회원)
2023년 한성대학교 IT응용시스템 공학과 학사 졸업.
2004년 모대학교 전자공학과 박사 졸업.

<주관심분야 : 이미지생성. 컴퓨터비전>

김 준 우(비회원)
2023년 한성대학교 IT응용시스템 공학과 학사 졸업.

<주관심분야 : 인공지능. 컴퓨터비전, 딥러닝>

오 희 석(정회원)
2017년 연세대학교 전기전자공학과 박사 졸업
2017년～2017년 삼성전자 DMC 연구소 책임연구원
2017년～2020년 한국전자통신 연구원 선임연구원
2020년～현재 한성대학교 IT융합공학부 조교수
<주관심분야 : 영상처리, 컴퓨터비전, 혼합현실, 심층생성모델 등>

지 준(정회원)
1989년 서울대학교 자원공학과 석사 졸업
1995년 Stanford University, Geophysics 박사 졸업
1996년 3DGeo Inc. 선임연구원
1997년~현재 한성대학교 AI응용학과 교수
<주관심분야 : 인공지능, 딥러닝, 머신러닝>