JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 5, pp. 37-46, May 2020
https://doi.org/10.9708/jksci.2020.25.05.037

MSaGAN: Improved SaGAN using Guide Mask and Multitask Learning Approach for Facial Attribute Editing

Hyeon Seok Yang*, Jeong Hoon Han*, Young Shik Moon*

*Student, Dept. of Computer Science and Engineering, Hanyang University, Ansan, Korea
*Student, Dept. of Computer Science and Engineering, Hanyang University, Ansan, Korea
*Professor, Dept. of Computer Science and Engineering, Hanyang University, Ansan, Korea

[Abstract]

Recently, studies of facial attribute editing have obtained realistic results using generative adversarial net (GAN) and encoder-decoder structure. Spatial attention GAN (SaGAN), one of the latest researches, is the method that can change only desired attribute in a face image by spatial attention mechanism. However, sometimes unnatural results are obtained due to insufficient information on face areas. In this paper, we propose an improved SaGAN (MSaGAN) using a guide mask for learning and applying multitask learning approach to improve the limitations of the existing methods. Through extensive experiments, we evaluated the results of the facial attribute editing in therms of the mask loss function and the neural network structure. It has been shown that the proposed method can efficiently produce more natural results compared to the previous methods.

▸ **Key words**: Facial attribute editing, Generative adversarial networks, SaGAN, Deep learning, Spatial attention mechanism

[요 약]

최근 얼굴 속성 편집(facial attribute editing)의 연구는 GAN(Generative Adversarial Net)과 인코더-디코더(encoder-decoder) 구조를 활용하여 사실적인 결과를 얻고 있다. 최신 연구 중 하나인 SaGAN(Spatial attention GAN)은 공간적 주의 기제(spatial attention mechanism)를 활용하여 얼굴 영상에서 원하는 속성만을 변경할 방법을 제안하였다. 그러나 불충분한 얼굴 영역 정보로 인하여 때로 부자연스러운 결과를 얻는 경우가 발생한다. 본 논문에서는 기존 연구의 한계점을 개선하기 위하여 유도 마스크(guide mask)를 학습에 활용하고, 다중작업 학습(multitask learning) 접근을 적용한 개선된 SaGAN(MSaGAN)을 제안한다. 폭넓은 실험을 통해 마스크 손실 함수와 신경망 구조에 따른 얼굴 속성 편집의 결과를 비교하여 제안하는 방법이 기존보다 더 자연스러운 결과를 효율적으로 얻을 수 있음을 보인다.

▸ **주제어**: 얼굴 속성 편집, GAN, SaGAN, 깊은 신경망, 공간적 주의 기제

• First Author: Hyeon Seok Yang, Corresponding Author: Young Shik Moon
*Hyeon Seok Yang (hsyang@visionlab.or.kr), Dept. of Computer Science and Engineering, Hanyang University
*Jeong Hoon Han (bghan@visionlab.or.kr), Dept. of Computer Science and Engineering, Hanyang University
*Young Shik Moon (ysmoon@hanyang.ac.kr), Dept. of Computer Science and Engineering, Hanyang University
• Received: 2020. 03. 26, Revised: 2020. 05. 04, Accepted: 2020. 05. 05.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

사람의 얼굴은 개인의 정체성을 나타낼 수 있기 때문에 다양한 연구에서 다루어지고 있다[1-3]. 그 중에서도 얼굴 속성 편집(facial attribute editing)은 얼굴 영상의 속성을 의도대로 변경하는 방법 및 모델을 연구하는 분야이다. 이 연구는 사진 편집, 미용, 디자인, 오락 등에 활용될 수 있다. 최근에는 생성적 모델인 GAN(Generative Adversarial Net)[4]과 인코더-디코더(encoder-decoder) 구조를 활용하여 자연스러운 결과를 생성해낼 수 있는 연구들이 이루어져 왔다[5-9].

하지만 많은 얼굴 속성 편집 연구들에서는 특정한 속성을 변경할 경우, 의도와 달리 연관성이 높은 다른 속성까지 변경되는 문제가 있다. 예를 들어, 수염 유무를 변경하면 성별이 함께 변경될 수 있다[5]. 이러한 결과는 얼굴 영상의 원하는 속성만을 변경하는 것을 방해하기 때문에 부적절한 결과다. 이 문제를 해결하기 위해 얼굴 속성 편집 시에 변경을 원하는 속성만 바꾸는 선별적 얼굴 속성 편집 연구들이 진행되었다[5-9]. 하지만 기존의 선별적 얼굴 속성 편집 연구에서도 변경하고자 하는 속성이 충분히 반영되지 않거나, 부자연스러운 결과가 나타나는 경향이 있다. 이러한 결과는 학습 시에 변경하고자 하는 속성 영역에 대한 정보가 충분히 제공되지 않은 것을 원인으로 볼 수 있다.

우리는 지난 연구[6]에서 최근의 선별적 얼굴 속성 편집 방법의 하나인 SaGAN(Spatial attention GAN)[7]을 기반으로 학습 시에 마스크 정보를 추가로 반영하고 SaGAN의 모델의 인코더 부분을 통합하는 모델을 제안하였다. 본 논문에서는 지난번 연구를 확장하여 마스크 손실(mask loss)과 다중작업 학습(multitask learning) 접근으로 기존 생성자의 네트워크를 통합한 편집 신경망을 제안한다. 또한 마스크 손실들과 신경망 구조들을 비교 평가하여 얼굴 속성 편집 결과의 품질과 학습 속도에 어떠한 영향을 주는지를 분석한다.

본 논문의 구성은 다음과 같다. II 장에서는 관련 연구를 설명하고, 제안하는 방법과 기존 방법의 차이점을 기술한다. III 장에서는 제안하는 방법의 구조와 세부 사항들에 관해서 설명한다. IV 장에서는 실험 세팅과 세팅별 실험 결과의 샘플을 살펴보고 분석하며, V 장에서는 실험 결과에 대한 종합적 분석을 하고 결론을 맺는다.

II. Related Works

1. Generative adversarial network

최근 얼굴 속성 편집의 대표적인 접근은 GAN[4]을 기반으로 한 접근이다. GAN은 생성 모델의 일종으로 기존 방법보다 사실적인 영상을 생성하는 데에 뛰어난 성능을 보이며, 얼굴 속성 편집, 영상 개선 등의 분야에 활용되고 있다[1,10].

기본적인 GAN[4]은 무작위 벡터(random vector)를 입력받아 사실적인 영상을 생성해낼 수 있다. GAN은 사실적인 영상의 학습을 위하여 실제 영상들과 생성자(generator)가 생성한 가짜 영상을 식별자(discriminator)가 식별해 내도록 하며, 생성자와 식별자를 경쟁적으로 학습시킴으로써 점차 더 사실적인 영상을 생성해낼 수 있다.

DCGAN[11]은 GAN의 계층(layer)을 컨볼루셔널 계층(convolutional layer)으로 변경하고 몇 가지 신경망 설계 요령을 적용하여 기존 GAN보다 안정적인 학습 결과를 생성하는 방법으로 제안되었다. 하지만 이러한 GAN 모델은 사용자 조작 없이 영상을 생성하므로 의도하는 결과를 만들어내는 데는 적합하지 않았다. 사용자가 원하는 결과를 생성하기 위해서는 사용자의 의도를 반영할 수 있는 모델이 필요하다.

cGAN[12]은 GAN의 조건적인 버전이다. GAN의 학습 시에 추가로 클래스 등의 속성 정보를 생성자와 식별자에 함께 제공하여 학습한다. 이를 통해 예측 시에 생성자에게 원하는 조건을 추가로 입력하여 사용자의 의도를 반영한 결과 영상을 생성할 수 있다.

2. Image-to-image translation

얼굴 속성 편집은 얼굴 영상에 원하는 속성을 반영해 변경한다. 이러한 작업은 이미지 변환(image-to-image translation)의 특수한 경우로 볼 수 있다. IcGAN(Invertible cGAN)[13]은 이미지 변환에 적용할 수 있는 모델 중 하나로써 인코더와 cGAN을 결합한 모델이다. 인코더는 cGAN의 역함수에 해당하며 실제 영상을 잠재 공간(latent space)과 조건적 표현으로 맵핑(mapping)할 수 있다. 이러한 IcGAN을 통해 영상을 원하는 조건을 갖도록 변경하는 것이 가능하다.

pix2pix[14]는 이미지 변환을 위한 일반적 방법으로 제안되었다. 이 신경망은 DCGAN의 설계 요령을 따랐으며, 식별자로 PatchGAN을 사용하였다. pix2pix는 입력 영상을 원하는 속성을 가진 영상으로 변경할 수 있다, 하지만 학습할 때 데이터 세트에 대응되는 정답 영상이 있어야 한다. 많은 경우에 서로 다른 도메인(domain)의 영상은 대응되는 영상을 갖지 않는다.

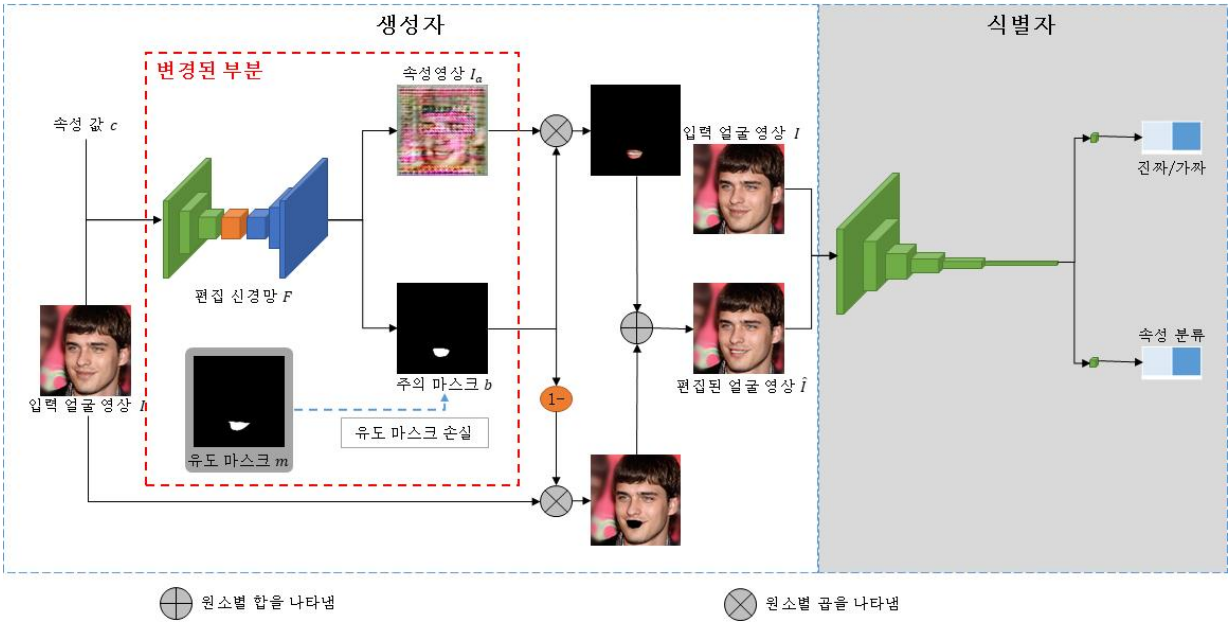

Fig. 1. Network structure of the proposed method

CycleGAN[15]은 pix2pix의 한계점을 해결하기 위하여 기계 번역에서 사용되는 순환 일관성(cycle-consistent)을 반영하여 대응되는 영상이 없는 경우에도 서로 다른 도메인 간의 변환을 할 수 있는 모델이다. 그러나 CycleGAN은 각 도메인 간 변환마다 하나씩의 모델의 학습이 요구된다. 만약 서로 변환해야 하는 도메인의 수가 많아질 경우에는 필요한 모델의 수가 기하급수적으로 증가하게 된다.

StarGAN[16]은 CycleGAN이 도메인간 변환 시에 다량의 모델을 요구하는 한계점을 해결하기 위하여 하나의 생성자를 학습할 때에 도메인 정보를 함께 제공함으로써 다양한 도메인을 하나의 모델로 학습하였고, 서로 유사한 데이터 세트를 활용해 동시에 학습하는 방법을 제안하였다.

앞서 소개한 몇몇 모델들[13,15,16]은 얼굴 도메인에 대해서 적용된 바 있으며, 얼굴 속성의 변환에 성과를 보였다. 그러나 이러한 모델들은 관심 속성을 변화시키는 것에 초점을 맞췄고, 그로 인해 의도하지 않은 부분이 함께 변화하는 한계가 있다.

3. Selective facial attribute editing

얼굴 속성 편집 분야에서는 의도한 것과 다른 속성이 변화하는 문제점을 개선하기 위하여 몇 가지 연구가 진행되었다[1,5-9].

ResGAN[5]은 전체 영상의 변경을 학습하는 대신에 잔차 영상(residual image)만을 학습하고, 잔차 영상이 희소하도록 규제함으로써 영상을 원하는 부분에 집중하여 변경 방법을 제안하였다. SaGAN[7]도 ResGAN과 동일한 문제의식을 공유하고 있다. 다만 SaGAN은 ResGAN과 달리 공간적 주의 기제(spatial attention mechanism)를 고려한 모델을 사용하였다. AttGAN[8] 또한 얼굴 속성 편집에서 원하는 것만 변경하는 것을 목표로 제안된 모델이다. 이 모델은 원하는 속성만 바꾸기 위해서 속성 분류 제약(attribute classification constraint)과 복원 학습(reconstruction learning)을 적용하였으며 하나의 모델로 다수의 속성 편집을 학습하고, 동시에 변경하는 것이 가능하다. 하지만 제약이 충분하지 않아서 여전히 의도하지 않은 부분이 변화하는 한계점이 있다.

제안하는 방법은 원하는 부분만을 한정적으로 수정하는 데에 강점을 보이는 SaGAN 모델을 기반으로 SaGAN의 성능과 효율성을 높이는 데 초점을 맞춘다. 기존의 SaGAN은 두 가지 측면에서 개선의 여지가 있다. 첫째로는 공간적 주의 기제를 사용하여 변경 영역을 효과적으로 한정하였고 변경 영역에 대한 추가적인 정보를 요구하지 않았으나, 변경할 영역에 대한 정보가 부족하여 변경 영역을 소극적으로 형성하거나 심지어 원본 영상이 그대로 반환되도록 학습되는 경우가 발생한다.

두 번째로 신경망 구조의 비효율성이 있다. 기존 SaGAN에서는 두 개의 생성자를 사용하여 편집된 얼굴 영상과 주의 마스크를 병렬로 학습한다. 그러나 두 신경망은 얼굴 영상의 동일한 부분에 대한 서로 다른 부분을 학습하는 것이므로 두 신경망이 학습할 특징은 유사한 성질을 공유할 것이다. 따라서 하나의 신경망으로 다중작업 학습을 수행하는 것이 더 효율적인 접근으로 보인다.

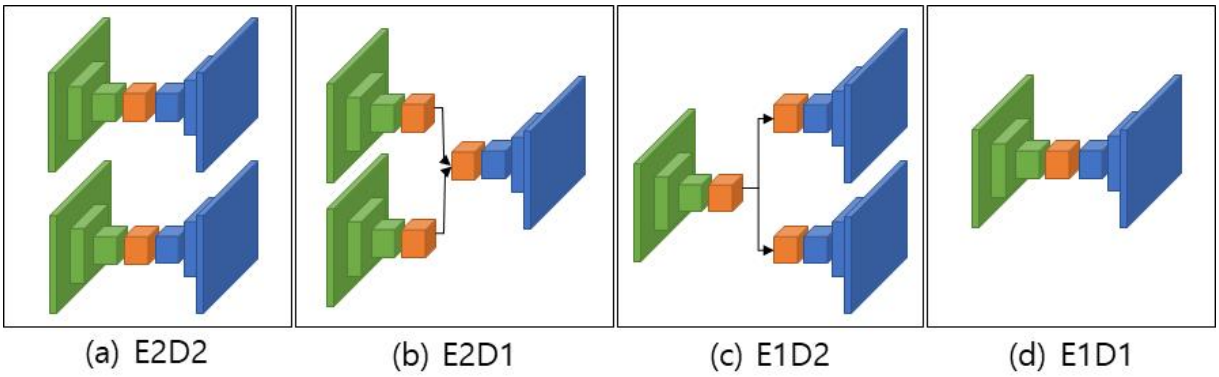

Fig. 2. The candidate structures of the editing network

제안하는 방법에서는 마스크를 학습에 반영함으로써 부족한 정보를 보완하고 신경망 구조를 변경하여 더 효과적이고 효율적으로 얼굴 속성 편집을 학습하는 방법을 제안한다.

얼굴 속성 편집의 기존 연구 중 관련성이 높은 연구로는 공간적 주의 기제를 적용하는 SaGAN[7]과 의미적 영상 분할을 적용한 연구[9]가 있다. SaGAN은 비지도 학습(unsupervised learning)을 사용하였기 때문에 마스크를 요구하지 않았으며, 대신 성능을 극대화하는 데에 한계가 있었다. 의미적 영상 분할을 적용하는 연구는 얼굴 속성 편집의 특성 상, 변경되는 영역이 일정하지 않은 경우에 적용에 어려움이 있다. 우리의 지난 연구[6]는 SaGAN과 비교할 때 마스크 정보를 추가적으로 활용하는 점이 차별점이며, 의미적 영상 분할 방식과 비교할 때는 공간적 주의 기제의 유연성을 갖는 점이 차별점으로, SaGAN의 유연성과 함께 의미적 영상 분할의 정답을 활용하는 복합적(hybrid) 방법으로 볼 수 있다. 이번 연구에서는 우리의 지난 연구[6]를 확장하여 같은 목적을 손실 함수와 신경망 구조에 대해 확장하고 정성적, 정량적 평가를 수행한다. 이를 통해 기존 SaGAN 모델을 성능과 효율 측면에서 실험적으로 개선한다.

4. Dataset

얼굴 속성 편집을 수행하기 위해서는 얼굴 영상과 속성 주석(attribute annotation)이 포함된 데이터 세트가 필요하다. 최근 얼굴 속성 관련 연구에서는 CelebA 데이터 세트와 LFW 데이터 세트가 주로 활용된다[1,17]. CelebA 데이터 세트는 10,177명의 유명인의 사진 총 202,559장으로 구성된다. 추가로 각 영상마다 5개의 표지점(landmark)과 40가지의 이진 속성(binary attribute)이 포함된다. LFW는 5,749명의 총 13,233장의 다양한 환경의 얼굴 영상으로 구성되어 있으며 적게는 40개, 많게는 73개의 이진 속성이 포함되어 있다. 하지만 위의 데이터 세트는 각 속성이 어떤 화소에 반영되어있는지를 알 수 있는 의미적 분할 마스크(semantic segmentation mask) 정보는 포함되어있지 않다.

의미적 분할 마스크가 포함된 최신 데이터 세트로는 CelebAMask-HQ가 있다[18]. 이 데이터 세트는 CelebA 데이터 세트의 고해상도 영상 30,000장으로 구성되어 있고 수동으로 작성한 19종류의 얼굴 영역 마스크(피부, 눈, 머리카락, 안경 등)를 포함한다. 본 연구에서는 의미적 분할 마스크를 포함하는 CelebAMask-HQ 데이터 세트를 활용한다.

III. The Proposed Scheme

1. System overview

제안하는 연구의 신경망 구조는 Fig. 1과 같다. 기본적인 구조는 SaGAN을 기반으로 하고, 두 신경망을 하나로 통합하며, 마스크 손실 함수를 추가한다.

신경망은 크게 생성자 G와 식별자 D로 구성된다. 생성자의 입력은 얼굴 영상 I와 변경할 속성 값 c이고, 출력은 편집된 얼굴 영상 $\hat{I}$이다. 수식은 아래와 같다.

$$\hat{I} = G(I, c). \tag{1}$$

먼저 입력은 생성자 내의 편집 신경망(editing network) F에 입력된다. 편집 신경망은 I의 변경된 속성 영상 I_a와 변경할 부분의 주의 마스크(attention mask) b를 생성한다. 수식은 아래와 같다.

$$I_a, b = F(I, c). \tag{2}$$

이후에는 주의 마스크 b를 기준으로 속성 영상 I_a의 주의 영역과 입력 영상 I의 변경할 속성과 무관한 영역을 얻고, 이 두 영역을 합쳐 편집된 얼굴 영상 $\hat{I}$을 생성한다. 식별자 D는 입력 영상 I와 생성자로 편집된 영상 $\hat{I}$을 식별하고, 속성을 분류한다. 수식은 다음과 같다.

$$\hat{I}= G(I,c) = I_b \bullet b + I \bullet (1-b). \quad (3)$$

2. Editing network

편집 신경망은 기존 SaGAN의 속성 조작 신경망(attribute manipulation network)과 공간적 주의 신경망(spatial attention network)을 통합한 신경망이다. SaGAN에서는 하나의 신경망을 사용해 학습하는 대신에 두 개의 신경망이 역할을 나누어 학습을 수행한다. 하지만 두 신경망이 하는 작업이 같은 얼굴 영상의 같은 속성을 처리한다는 점을 고려하면 비효율적인 구조일 수 있다. 본 논문에서는 기존 SaGAN의 신경망을 통합한 편집 신경망을 제안한다.

3. Objective function

제안하는 방법의 목적 함수(objective function)는 기존 SaGAN에 추가로 정확한 주의 마스크를 생성하기 위한 마스크 손실 함수를 추가한다. 마스크 손실 함수는 데이터 세트의 관련 마스크들로 구성한 유도 마스크(guide mask) m과 주의 마스크 b간의 차이를 손실 함수로 정의한다. 차이의 척도로는 차집합 오차(Set Difference Error, SDE)[5]를 사용한다. 적용된 마스크 손실 L_{mask}^G은 다음과 같다.

$$L_{m,id}^G = E_{I_{id}}[\max(m - b_{id}, 0)],$$
$$L_{m,dual}^G = E_{I_{dual}}[\max(m - b_{dual}, 0)], \quad (4)$$

$$L_{mask}^G = \alpha L_{m,id}^G + \beta L_{m,dual}^G. \quad (5)$$

m은 유도 마스크를 의미하고, id는 동일 속성으로 복원하는 경우이며, $dual$은 속성을 변환 후, 다시 원래 속성으로 복원한 경우이다. 차집합 오차는 생성된 주의 마스크 b가 유도 마스크 m보다 작은 경우에 더 넓은 형태로 생성하도록 강제하기 위한 것이다. 이는 기존 SaGAN이 필요한 것보다 작은 주의 마스크 b를 생성하는 경향을 보정하기 위한 의도로 고안되었다.

유도 마스크는 주의 속성 영역에 대응하는 마스크의 합집합을 사용한다. 본 논문에서는 입의 개폐 여부의 속성(Mouth_Slightly_Open)을 실험에 활용하였으며, 이를 위해 CelebAMask-HQ의 입술과 입에 대응하는 마스크들(윗입술, 입, 아랫입술)의 합집합을 유도 마스크로 사용하였다.

기존 SaGAN의 목적 함수(objective function)[7]에 마스크 손실 L_{mask}^G을 추가하여 변경된 생성자 G의 목적 함수는 다음과 같다.

$$\min_F L_G = L_{adv}^G + L_{ds}^G + L_{rec}^G + L_{mask}^G. \quad (6)$$

위 식에서 L_{adv}^G는 경쟁 손실(adversarial loss), L_{ds}^G은 속성 분류 손실(attribute classification loss), L_{rec}^G는 복원 손실(reconstruction loss)을 뜻한다.

식별자의 목적 함수는 기존의 SaGAN과 같으며 다음과 같다.

$$\min_{D_{src}, D_{cls}} L_D = L_{src}^D + L_{ds}^D. \quad (7)$$

L_{src}^D는 식별자의 경쟁 손실이고, L_{ds}^D는 속성 분류 손실이다.

IV. Experimental Results

1. Experimental settings

본 논문에서는 변경된 마스크 손실과 신경망 구조가 결과 영상의 품질에 어떠한 영향을 주는지를 실험하고 학습 시간이 어떻게 변화하는지를 측정한다. 평가는 얼굴 속성 편집 결과의 정성적 평가(qualitative evaluation), 정량적 평가(quantitative evaluation)를 수행하고, 신경망 구조별 학습 시간을 비교한다.

구체적 실험 구성은 마스크 손실은 마스크 손실을 쓰지 않는 경우(No mask), MAE(Mean Absolute Error), MSE(Mean Square Error), 그리고 SDE[6]까지 4가지를 실험하였다. 실험에서 고려한 편집 신경망의 후보 구조는 Fig. 2와 같이 4가지다. 편의상 인코더와 디코더의 수에 따라 ExDy(인코더 x개, 디코더 y개)의 형태로 표현한다. 기존 방법인 SaGAN는 마스크 손실을 사용하지 않는 경우(No mask)이면서 인코더와 디코더를 2개씩 사용한 경우(E2D2)와 같다. 최종적으로 기존 방법인 SaGAN을 포함하여 마스크 손실 4가지와 신경망 구조 4가지를 조합한 16가지 모델을 각각 8 에폭(epoch)씩 학습하였다. 그 과정에서 학습에 실패하여 원본 영상을 그대로 반환한 세 모델(No mask+E2D2, No mask+E2D1, MAE+E2D1)의 경우

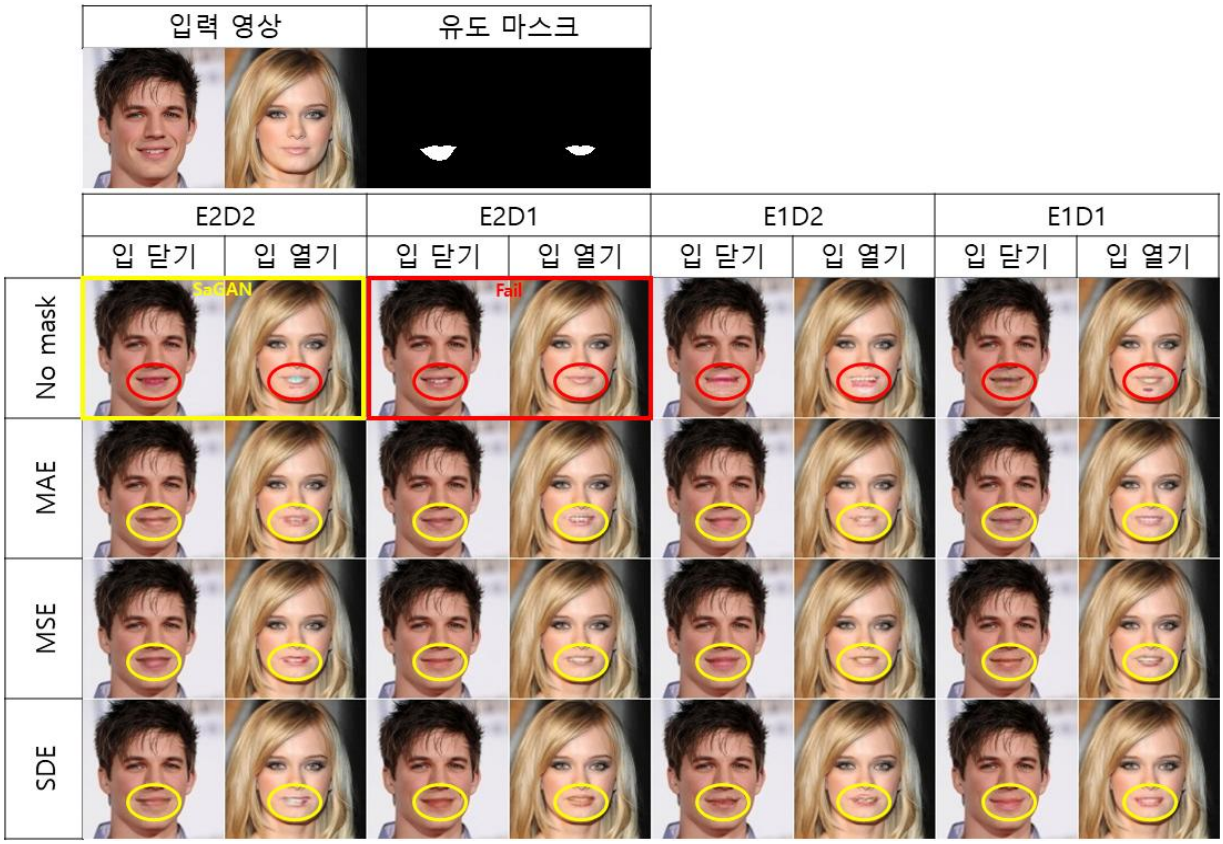

Fig. 3. Samples of edited image $\hat{I}$

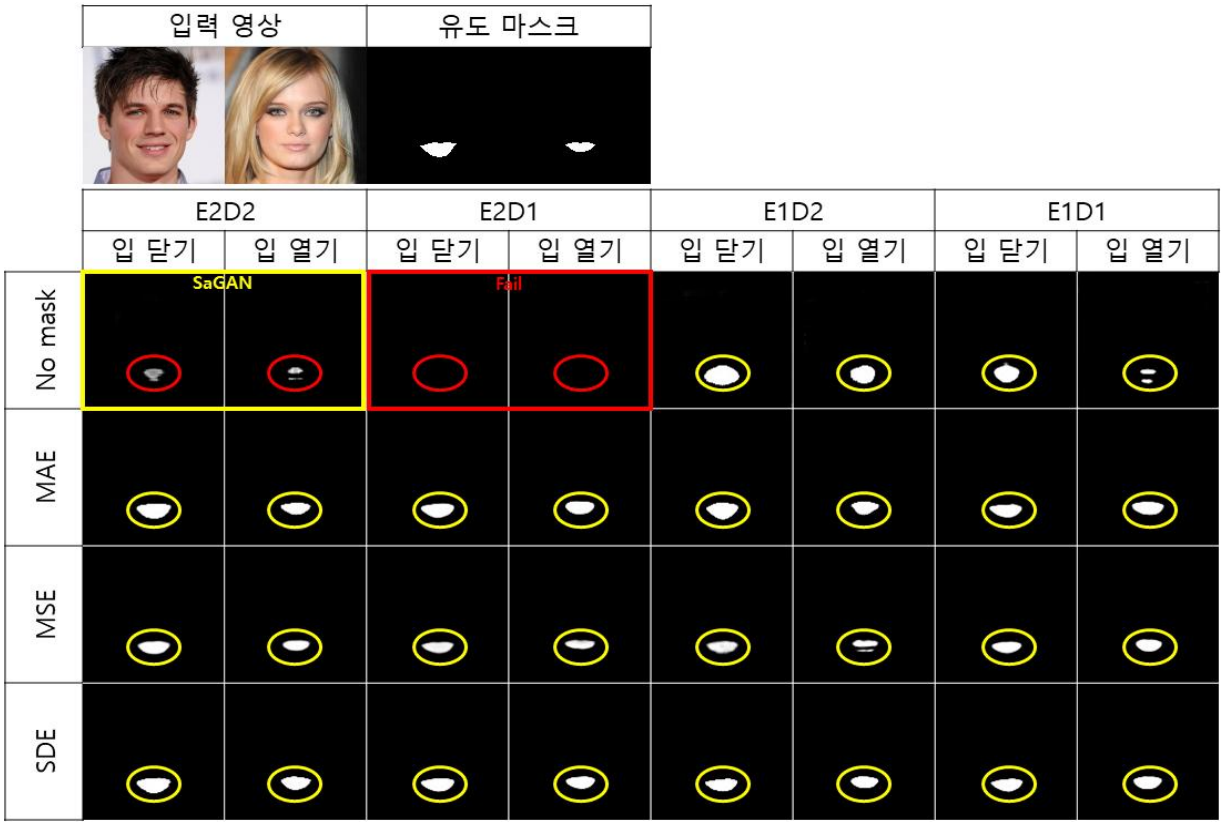

Fig. 4. Samples of attention mask b

는 추가로 학습을 시도하여 성공한 경우는 대체하였고, 총 세 번의 시도에도 실패한 한 가지 경우(No mask+E2D1)는 그대로 반영하였다. 마스크 손실과 신경망 구조 및 위에 언급된 예외적 사례에 대한 조치를 제외한 요인들은 일정하게 통제하였다.

사용한 데이터 세트는 CelebAMask-HQ이며, 데이터 세트에서 제공되는 학습 세트 구분을 그대로 활용하여 학습 세트 20,000장, 검증 세트 5,000장, 테스트 세트 5,000장으로 학습을 수행하였다. 테스트한 속성은 입 개폐(Mouth_Slightly_Open)이다.

제안하는 모델의 기본적인 파라미터 세팅은 SaGAN과 동일하다. 학습과 테스트 시의 입력 영상은 128×128로 정규화(normalization)하였다. 마스크 손실은 $\alpha = 100$와 $\beta = 20$으로 설정하였고, 배치 크기(batch size)는 16이다. 실험에 사용한 PC는 운영 체제는 Windows 10, CPU는 Intel core i7 6700 3.40GHz, 그래픽 카드는 NVIDIA GeForce RTX 2080 Ti, 메모리는 DDR 4 16Gb, 데이터 세트의 저장 장치는 SSD를 사용하였다. 또한 아나콘다(Anaconda) 가상 환경의 Python 3.7.5, TensorFlow 2.0.0에서 테스트하였다.

2. Qualitative evaluation

정성적 평가는 각 마스크 손실과 신경망 구조에 따라서 얼굴 속성 편집결과 및 중간 결과들이 어떤 형태와 특성을 갖는지를 살펴본다. Fig. 3-5는 마스크 손실과 편집 신경망의 구성을 달리하여 학습한 경우의 얼굴 속성 편집시의 편집된 영상 $\hat{I}$, 주의 마스크 b, 속성 영상 I_a을 나타낸다. 각 행은 서로 다른 마스크 손실로 학습한 경우이고, 각 열은 서로 다른 편집 신경망 구조로 학습한 경우를 속성 값별 샘플(sample)로 나타냈다. 기존 방법인 SaGAN은 첫 번째 행의 첫 번째 열(No mask & E2D2, 황색 사각형)과 동일하다.

Fig. 3-5의 첫 행(No mask)은 마스크 손실을 사용하지 않았을 경우다. 때문에 상대적으로 관련 속성의 위치 정보가 부족하여 Fig. 4의 No mask의 일부 경우(적색 원)에는 주의 마스크가 부정확하게 활성화되었고 이로 인해 부자연스러운 결과가 생성된 것을 볼 수 있다. 특히 No mask이면서 E2D1인 경우(적색 사각형)에는 주의 마스크가 전혀 활성화되지 않아서 원본 영상이 그대로 출력되었다. 반면 마스크 손실을 적용한 경우는 Fig. 4에서 주의 마스크가 활성화되어있음을 확인할 수 있다.

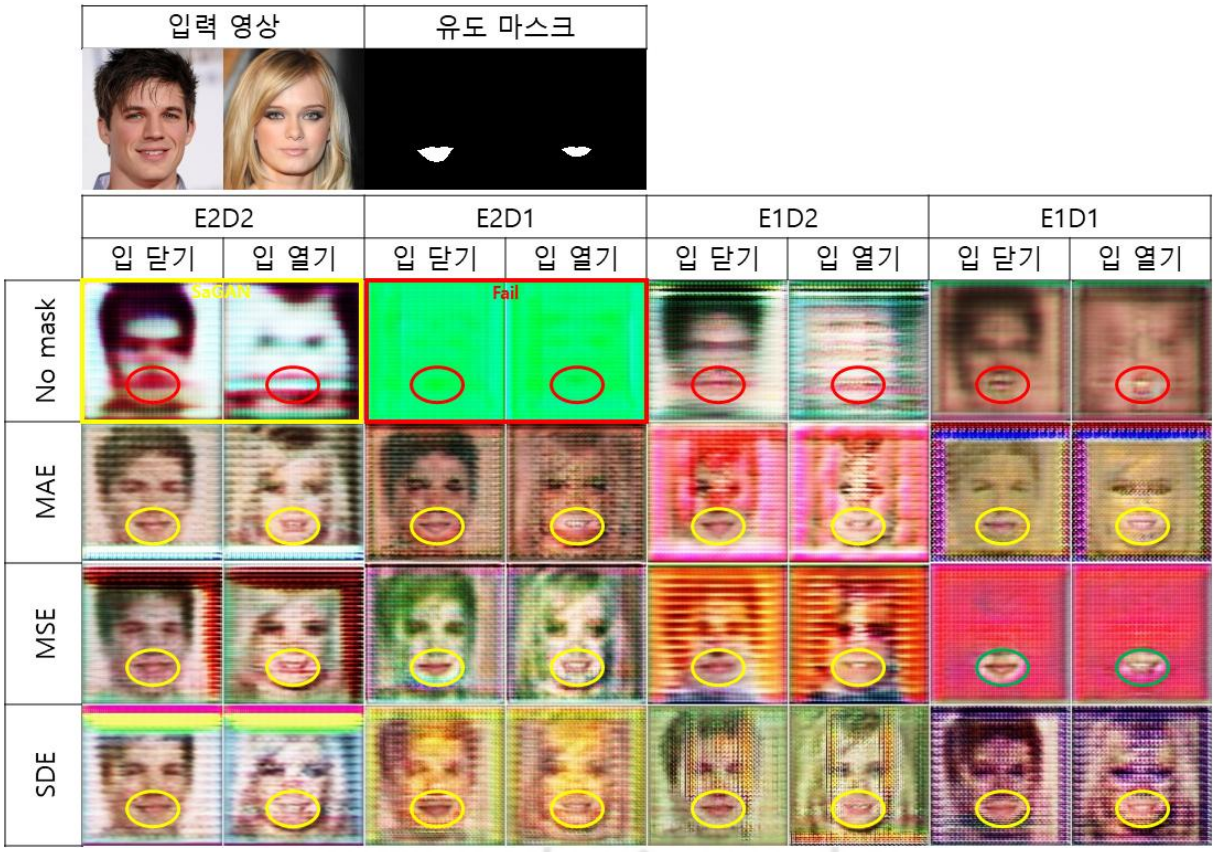

Fig. 5. Samples of attribute image I_a

마스크 손실을 적용한 경우들(MAE, MSE, SDE)은 적용하지 않은 경우에 비해 확연히 더 자연스럽고 원하는 속성이 반영된 결과를 얻었다. 결과는 샘플마다 품질의 편차가 있어서 특정한 방식이 일관되게 뛰어나거나 부족하다고 말하기는 어렵다.

Fig. 4의 주의 마스크를 유도 마스크와 비교해볼 때, No mask의 경우에는 모델마다 편차가 컸다. E2D2(SaGAN, 황색 사각형)에서는 국지적이고 약하게 마스크가 활성화되어 기존 속성이 충분히 변경되지 않았다. E2D1(적색 사각형)은 마스크가 전혀 활성화되지 않았다. E1D2는 유도 마스크보다 더 둥근 형태로 마스크가 얻어졌고 무관한 영역에서도 일부 활성화되는 경향이 있었다. E1D1은 E1D2와 E2D2의 중간 정도의 특성을 보였다. 마스크 손실을 적용한 경우는 주의 마스크와 유도 마스크의 유사성이 더 컸다. MAE와 SDE의 주의 마스크는 서로 유사한 경향을 띠었다. MSE는 활성화되는 영역이 상대적으로 작았고, 마스크의 엣지에서의 값의 기울기가 완만하였다. 세 마스크 손실 모두 E2D2에서의 마스크 활성 정도가 다른 경우보다 더 큰 경향이 있었다.

Fig. 5는 생성된 속성 영상이다. 주로 입 닫기의 경우에는 입술을 재현하는 붉은 색이 생성되고, 입 열기의 경우에는 치아를 재현하는 흰색이 생성되다가 점차 입의 형태를 생성해나간다. No mask의 경우(적색 원)에는 대체로 형태가 흐릿하고 색상 위주로만 표현되었다. 또한 색이 과장되어있어서 자연스러움이 부족하였다. 마스크 손실을 쓰는 경우는 대체로 입력 영상의 모습이 왜곡된 형태로 나타나면서 변경하고자 하는 영역인 입 주변만이 더 자연스럽게 생성되는 경향이 있다. MSE의 E1D1의 경우(녹색 원)에는 입 주변만 집중적으로 생성되는 경향이 있었다. 이는 MSE와 다중작업 학습으로 인해 속성과 관련된 부분의 생성에 집중하도록 학습되는 것으로 보인다.

3. Quantitative evaluation

얼굴 속성 편집은 그 특성으로 인해 정답을 확보하는 것이 불가능하다. 때문에 본 연구에서는 품질을 두 가지 측면으로 나누어 기존에 얼굴 속성 편집에 사용된 방식[1]을 차용하여 각기 측정한 결과를 종합적으로 분석한다. 얼굴 속성 편집의 정량적 평가는 원본 영상과의 유사성이 얼마나 보존되는지와 원하는 속성이 정확히 편집되었는지를 각기 평가한다. 원본 영상이 얼마나 보존되었는지는 구조적 유사성의 측정 방법으로 널리 사용되는 SSIM(Structural SIMilarity)로 측정하였다. 속성 편집은 변경된 속성을 직접 측정하는 것은 불가능하므로 별도로 학습된 속성 분류기(attribute classifier)를 활용해 간접적으로 속성 편집 정확도(attribute editing accuracy)를 측정하였다. 이러한 접근은 기존 얼굴 속성 편집 연구에서 사용된 바 있다[8]. 속성 분류기는 ResNet50을 기반으로

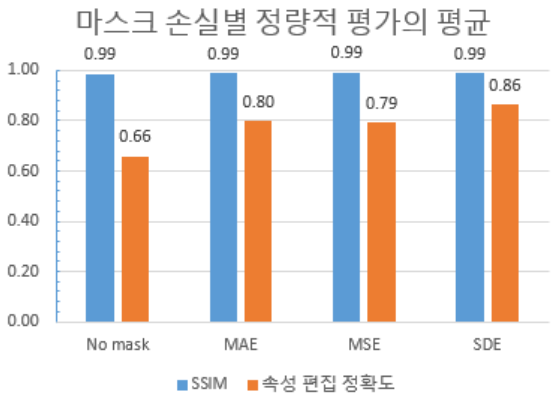

Fig. 6. Average of quantitative evaluation by mask loss

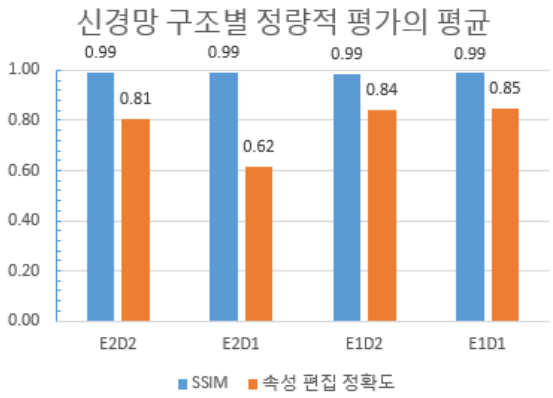

Fig. 7. Average of quantitative evaluation by neural network structure

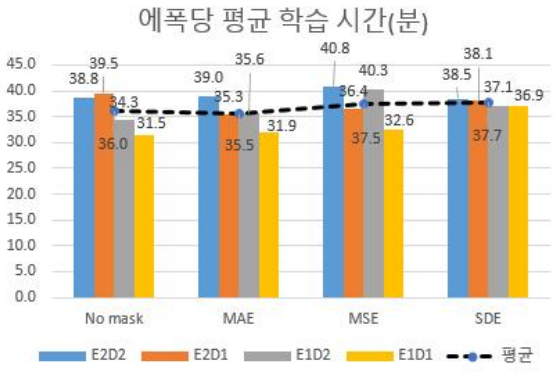

Fig. 8 Average learning time per epoch according to neural network composition

Mouth_ Slightly_Open 속성의 여부를 학습하였다. 데이터 세트는 CelebAMask-HQ의 학습 데이터 세트를 사용하였고, 50 에폭 학습하였다. 이 분류기는 테스트 세트에 대해 93.42%의 정확도를 얻었다. 분류기가 분류한 속성을 기준으로 변경하고자 의도한 속성이 일치하는 경우를 정확한 편집으로 간주하고, 불일치하는 경우는 부정확한 편집으로 간주하여 속성 편집 정확도를 계산하였다.

Fig. 6은 측정한 SSIM과 속성 편집 정확도를 마스크 손실별로 평균한 그래프이다. 공간적 주의 기제가 활용되었으므로 변경되는 영역 이외의 값은 유지되기 때문에 SSIM의 값은 대부분의 경우에 0.99 내외의 값을 얻었다. No mask의 경우에 속성 편집 정확도는 약 66%로 마스크 손실을 사용한 경우보다 크게 낮은 값이 측정되었다. MAE와 MSE는 각각 약 80%와 79%로 유사한 값을 획득하였다. SDE는 86%로 다른 방법보다 좀 더 높은 값을 얻었다. 이는 SDE가 다른 방법보다 놓친 부분을 더 학습하도록 촉진한 결과로 추정된다.

Fig. 7은 신경망 구조별로 계산한 각 정량적 평가의 평균이다. SSIM은 모두 0.99 내외의 값을 얻었다. 속성 편집 정확도는 E2D2는 약 81%를 얻었다. E2D1은 62%로 다른 구조보다 약 20%가량 낮은 값을 얻었다. E1D2와 E1D1은 약 84%, 85%로 비슷한 값을 얻었다. 기존 SaGAN의 구조인 E2D2와 비교할 때 E1D2이나 E1D1은 약 3~4%의 향상을 보였다. 이는 인코더를 공유함으로써 학습의 효율이 높아진 것으로 보인다.

4. Experiment of training time

Fig. 8은 모델별 학습시간을 그래프로 나타낸 것이다. 기존 방법인 SaGAN은 가장 좌측의 세팅(No mask, E2D2)이다. 손실 함수별 결과마다 E1D1이 가장 학습시간이 적게 걸리는 것을 확인할 수 있다. 이는 파라미터 수가 가장 작다는 측면에서 예상과 일치한다. 또한 대부분의 결과(No mask 제외)에서 E2D2가 가장 학습이 오래 걸렸다. 학습시간을 고려할 때는 E1D1 구조를 사용하는 것이 더 빠른 학습이 가능할 것으로 보인다. 손실 함수별 평균으로는 MAE가 에폭당 평균 35.5분으로 가장 빨랐으나 다른 손실 함수와의 차이가 크지는 않았다.

V. Conclusions

본 논문에서는 기존의 SaGAN의 주의 마스크가 충분히 활성화되지 않는 것을 개선하기 위하여 추가 정보로 유도 마스크를 반영하는 마스크 손실을 추가하고, 신경망 구조를 개선하여 더 자연스러운 얼굴 속성 변경 결과를 얻을 수 있음을 확인하였다. 또한 신경망 구조를 개선함으로써 신경망 학습 시간을 감소시키면서도 결과 영상의 품질은 저하되지 않음을 확인하였다. 결론적으로 SaGAN에 마스크 손실을 사용하고, E1D1 구조로 학습을 하는 것이 효율적인 것으로 판단된다.

REFERENCES

[1] X. Zheng, Y. Guo, H. Huang, Y. Li, and R. He, "A Survey to Deep Facial Attribute Analysis," International Journal of Computer Vision, pp. 1-33, Mar. 2020. DOI: 10.1007/s11263-020-01308-z

[2] T. J. Choi and H. M. Lee "An Algorithm for Converting 2D Face Image into 3D Model," Journal of The Korea Society of Computer and Information, Vol. 20, No. 4, pp. 41-48, Apr. 2015. DOI: 10.9708/jksci.2015.20.4.041

[3] S. C. Bae, Y. S. Lee, and S. W. Choi "Vision-based Authentication and Registration of Facial Identity in Hospital Information System," Journal of The Korea Society of Computer and Information, Vol. 24, No. 12, pp. 59-65, Dec. 2019. DOI: 10.9708/jksci.2019.24.12.059

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," Advances in neural information processing systems, pp. 2672-2680, Dec. 2014.

[5] W. Shen and R. Liu, "Learning Residual Images for Face Attribute Manipulation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4030-4038, Jul. 2017. DOI: 10.1109/CVPR.2017.135

[6] H. S. Yang, J. H. Han, Y. C. Cho, H. G. Lee, Y. Park, and Y. S. Moon, "Study on Performance Improvement of SAGAN using Mask," Proceeding of 2019 Korea Signal Processing Conference, pp. 2557-2560, Sep. 2019.

[7] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative Adversarial Network with Spatial Attention for Face Attribute Editing," Proceedings of the European Conference on Computer Vision (ECCV), pp. 417-432, Sep. 2018. DOI: 10.1007/978-3-030-01231-1_26

[8] Z. He, W. Zuo, and S. Shan, "AttGAN: Facial Attribute Editing by Only Changing What You Want." IEEE Transactions on Image Processing, Vol. 28, No. 11, pp. 5464-5478, May. 2019. DOI: 10.1109/TIP.2019.2916751

[9] H. S. Yang and Y. S. Moon, "Face Attribute Editing using AttGAN and Guide Mask," 2019 International Conference on Electronics, Information, and Communication (ICEIC), pp. 1-3, Jan. 2019. DOI:

10.23919/ELINFOCOM.2019. 8706471

[10] S. K. Woo, "Generation of Contrast Enhanced Computed Tomography Image using Deep Learning Network," Journal of The Korea Society of Computer and Information, Vol. 24, No. 3, pp. 41-47, Mar. 2019. DOI: 10.9708/jksci.2019.24.03.041

[11] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv preprint arXiv:1511.06434v2, pp. 1-16, Jan. 2016.

[12] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv preprint arXiv:1411.1784, pp. 1-7, Nov. 2014.

[13] G. Perarnau, J. V. D. Weijer, B. Raduanu, and J. M. Álvarez, "Invertible Conditional GANs for Image Editing," NIPS 2016 Workshop on Adversarial Training, pp. 1-9, Dec. 2016.

[14] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efos, "Image-to-Image Translation with Conditional Adversarial Networks," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125-1134, Jul. 2017. DOI: 10.1109/CVPR.2017.632

[15] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks," Proceedings of the IEEE international conference on computer vision, pp. 2223-2232, Oct. 2017. DOI: 10.1109/ICCV.2017.244

[16] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8789-8797, Jun. 2018. DOI: 10.1109/CVPR.2018.00916

[17] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," Proceedings of the IEEE International Conference on Computer Vision, pp. 3730-3738, Dec. 2015. DOI: 10.1109/ICCV.2015.425

[18] C. H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards Diverse and Interactive Facial Image Manipulation," arXiv preprint arXiv:1907.11922v2, pp. 1-20, Apr. 2020.

Authors

Hyeon Seok Yang received his B.S. degree in the Department of Electronics and Information Engineering from Yeungnam University, Korea, in 2010. He received the M.S. degrees in the Department of Computer Science & Engineering from Hanyang University, Korea, in 2012. He is studying for his PhD. degree in the Department of Computer Science & Engineering from Hanyang University, Korea. His research interests include computer vision, pattern recognition, and deep learning.

Jeong Hoon Han received his B.S. degree in the Department of Computer Science and Engineering from Hallym University, Korea, in 2016. He is currently working towards PhD. Degree at the Department of Computer Science and Engineering from Hanyang University, Korea, From 2016. His research interests include computer vision and machine learning.

Young Shik Moon received the B.S. and M.S. degrees in Electronics Engineering from Seoul National University and Korea Advanced Institute of Science and Technology, Korea, in 1980 and 1982, respectively, and PhD. degree in Electrical and Computer Engineering from the University of California at Irvine, CA, in 1990. From 1982 to 1985, he had been a researcher at the Electronics and Telecommunication Research Institute, Daejon, Korea. In 1992, he joined the Department of Computer Science and Engineering at Hanyang University, Korea, as an Assistant Professor, and is currently a Professor. Dr. Moon served as General Chair of 2014 IEEE International Symposium on Consumer Electronics, and worked as the President of the Institute of Electronics and Information Engineer, Korea.