Journal of Korea Multimedia Society Vol. 27, No. 5, May 2024(pp. 563-576)
https://doi.org/10.9717/kmms.2024.27.5.563

# 애니메이션 얼굴 변환을 위한 적응형 특징 융합 연동 다중 스케일 어텐션 활용 비지도 GAN 모델 개발

한아영[†], 김경태[††], 최재영[†††]

# Multi-scale Attention GAN with Adaptive Feature Fusion for Unsupervised Animation Face Generation

A Yeong Han[†], Kyeong Tae Kim[††], Jae Young Choi[†††]

## ABSTRACT

In this paper, we propose a novel generator architecture to translate a selfie photo into an animation face image in an unsupervised way. Recent face translation works often fail to preserve the characteristics of facial parts of selfie photos while transferring to animation images. To cope with this, the proposed method develops new Generative Adversarial Network (GAN) architecture composed of (1) simple cycle contents loss, (2) multi-scale assisted self-attention, and (3) adaptive feature fusion. The goal of using cycle contents loss is to make the GAN preserve a wide range of selfie image contents, which includes hair shape, facial expression and shape of face. In addition, multi-scale assisted self-attention complements the existing attention by using its various scales of self-attentions. Adaptive feature fusion helps model to understand the importance of each self-attention map among the multiple self-attention maps generated by the multi-scale self-attention module. Our extensive and comparative experiments on Selfie2Anime and Photo2Anime datasets have been performed to demonstrate the effectiveness of our method over other state-of-the-art methods.

**Key words:** Image Translation, Cross-Domain Image Generation, Unsupervised Learning

## 1. 서 론

이미지 변환이란 한 범주에 속하는 입력 이미지를 또 다른 범주에 속하는 출력 이미지로 매핑하는 컴퓨터 비전의 한 분야다. 예를 들면, 사람의 얼굴 사진을 바탕으로 애니메이션 얼굴 그림으로 변환하는 일을 말한다. 최근 연구에 따르면 이미지 변환은 이미지 손상 복원[1], 슈퍼 샘플링[2], 흑백사진의 채색[3]과 같은 넓은 범용성 가능성으로 인해 이미지 간 변환 분야가 머신러닝(Machine Learning)과 컴퓨터 비전(Computer vision)분야에서 많은 연구자들의 관심을 받고 있다. 그중에도 제로섬(Zero-sum)게임에서 비롯된 적대적 생성적 신경망[4]을 기반으로 한 이미지 변환 분야가 활발하게 연구되고 있다.

※ Corresponding Author : Jae Young Choi, Address: (17035) 81 Oedae-ro, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do, Korea, TEL : +82-31-330-4906, FAX : +82-31-330-4906, E-mail : jychoi@hufs.ac.kr
Receipt date : Apr. 12, 2024, Revision date : May. 7, 2024
Approval date : May. 13, 2024
[†] Department of Computer Engineering, Hankuk University of Foreign Studies
(A Yeong Han, E-mail : lucky@hufs.ac.kr)
(Kyeong Tae Kim, E-mail : rudxo9251@gmail.ac.kr)

※ This research was supported by Hankuk University of Foreign Studies Research Fund.
※ This research was supported by the “Development of technology using analysis of ocean satellite images (20210046)” funded by the Korea Institute of Marine Science & Technology Promotion (KIMST).
※ This research was supported by the “Development of Wave Overtopping quantitative observation technology” funded by the Korea Institute of Marine Science & Technology Promotion (KIMST) (No. 20220180).

비지도 학습 기반 이미지 변환이 상당히 발전했음에도 불구하고 사람의 얼굴 사진을 애니메이션 변환과 같이 형상적 변화가 크게 요구되는 대상 간의 이미지 변환은 다음과 같은 두 가지 문제를 지니고 있다. 첫째, 실제 사람의 얼굴과 애니메이션의 얼굴은 외형적으로 아주 큰 분포의 차이를 지니고 있다. 사람의 얼굴을 이루는 눈, 코 그리고 입과 같은 부분들은 사람마다 다양한 모양과 크기를 가지고 있지만 애니메이션에서 사람의 얼굴은 상대적으로 유사한 모양과 대체적으로 작은 코와 입을 가지며 눈의 경우, 사람의 눈보다 아주 크게 과장된 경우가 많다. 이러한 특성으로 인해 기존 비지도 학습 기반으로 하는 생성적 적대 신경망 연구들은 사람 얼굴 사진의 지역적 특징들을 애니메이션으로 변환할 때 무시하거나 큰 형상적 변환을 이루지 못하게 된다. 이는 사람 얼굴 사진의 고유한 개성들을 반영하지 못하는 애니메이션 얼굴 사진을 생성하게 하거나 전혀 애니메이션 얼굴 같지 않은 이미지를 생성하는 결과를 가져온다. 두 번째 문제점은 이미지 변환 문제에 집중하여 원본 이미지의 특성은 무시한 채 다른 분류의 이미지로 변환하는 것에 있다. 예를 들면 사람의 얼굴에서 나타나는 표정과 머리 모양의 개성을 무시하고 애니메이션 얼굴에 해당하는 임의의 이미지를 생성한다.

위 문제점들로 인해 이미지 변환에서 원본으로 사용되는 이미지의 특성을 반영하면서 동시에 목표하는 분야의 이미지로 변환하기에 어려움이 있다. 기존의 연구들은 일반적으로 이미지 변환 자체에 집중한 나머지 원본으로 사용되는 이미지 고유의 특성, 예를 들면 얼굴의 경우 표정, 머리 모양 또는 바라보는 방향 등을 보존하는 이미지 변환은 미흡한 실정이다. 이로 인해 위에서 언급된 원본 이미지의 고유 특성을 제한적으로 보존하면서 이미지 변환을 수행하게 된다.

이러한 문제를 해결하기 위해, 이 논문에서는 주기 특성 손실 함수(Cycle Contents Loss)와 적응형 특징맵 융합 기법을 통한 다중 스케일의 셀프-어텐션 모듈을 사용하는 생성적 적대 신경망[4]을 제안한다. 본 논문에서 제안한 기술의 차별성은 아래와 같이 요약된다.

• 새로운 주기 특성 손실 함수를 제안한다. 이미지 변환 단계에서 원본 이미지의 지역적 특성이 지나치게 요약 또는 손상되지 않고 변환될 수 있도록 GAN의 손실함수에 주기 특성 손실 함수를 추가한다. 주기 특성 손실 함수는 변환 단계에서 특징 맵이 지나치게 요약되어 지역적 특성이 손실되는 것을 억제하며 사진-애니메이션 변환에서 사람의 특성 있는 입 또는 코가 애니메이션의 단순한 작은 입 또는 코로 표현되는 현상을 막는다.

• 다중 스케일 기반의 셀프-어텐션 모듈을 제안한다. 기존의 GAN 기반 모델의 중간 단계에 다중 스케일 기반 셀프-어텐션 모듈을 적용한다. 해당 모듈은 정사각형 모양의 어텐션[5]뿐만 아니라 가로 및 세로가 각각 1/2인 비율의 어텐션 정보를 전달한다. 이는 단일 스케일로 얻을 수 없는 지역적 특성들의 상관관계를 GAN모델에 전달하는 역할을 하고 이미지 변환 단계에서 원본 이미지의 지역적 특성과 상관관계를 변환된 이미지에 적용할 수 있도록 GAN 모델을 유도하는 역할을 한다.

• 적응형 특징 맵 융합 기법을 상기 다중 스케일 기반 셀프-어텐션 모듈에 적용한다. 다양한 스케일의 어텐션 정보를 GAN에 전달하면 한 번에 많은 정보가 전달되게 된다. 이때 지나치게 많은 정보로 인한 GAN 모델의 학습 속도 저하 및 손실 함수의 발산을 막고 안정적인 학습을 위해 모든 스케일의 어텐션 맵에 적응형 특징 맵 융합 기법을 적용한다. 이는 손실 함수에 연결되어 특징 맵 별로 중요도를 학습 단계에서 스스로 판별하도록 하여 너무 많은 정보로 방해받지 않고 안정적으로 학습하도록 유도하는 역할을 한다.

## 2. 연구 배경

### 2.1 생성적 적대 신경망

생성적 적대 신경망(Generative Adversarial Network)[4]은 다양한 이미지 생성 및 변환에서 우수한 성능을 보여주고 있다. 생성적 적대 신경망의 성공의 열쇠는 생성자와 식별자 사이의 적대적 학습으로 볼 수 있다. 두 $X, Y$의 분류에 속하는 이미지를 $x$와 $y$라고 할 때, 생성자는 $X$에서 임의로 추출한 이미지 $x$를 바탕으로 $Y$에 속하도록 $y'$를 생성한다. 판별자는 $Y$에 속하는 임의의 이미지 $y$와 생성자가 생성한 $y'$가 $X$ 또는 $Y$중 어떤 분류에 속하는지 판별한다. 이때 생성자는 자신이 생성한 $y'$를 판별자가 Y에 속하는

y라고 속는 것을 목표로 학습하고, 판별자는 생성자가 생성한 $y'$와 실제 $Y$에 속하는 $y$가 서로 다름을 감별하는 것을 목표로 학습한다. 서로의 목적이 상반되는 결과를 요구함으로써 학습이 진행됨에 따라 생성자와 판별자는 서서히 정교한 결과를 얻도록 유도되는 점을 이용한다. 일반적으로 생성적 적대 신경망 학습에 사용되는 손실 함수는 아래와 같다.

$$\min_G \max_D F(D,G) = \mathbf{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbf{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

$D$는 판별자에 해당하며 입력으로 받은 $x$가 생성자에 의해 변환된 이미지가 아니라면 1의 값을, 생성자에 의해 변환되어 목표 분류에 속하게 된 이미지라면 0의 값을 가진다. 따라서 $D(G(z))$는 노이즈 $z$를 입력으로 생성자에 만들어진 이미지를 받고 있으므로 생성자의 변환이 $D$를 충분히 속일 수 있다면 1의 값을 가지고, 속일 수 없다면 0의 값을 가지게 된다. 식(1)에서 $D$가 함수 $F$를 최대화하기 위해선 우변의 모든 항을 최댓값으로 가져야 한다. 즉, $D(x)$가 1이 되어야 하며 이는 판별자가 생성자에 의해 변환된 이미지가 아닌 이미지를 1로 판별 하도록 학습하게 된다. 생성자인 $G$는 식(1)을 최소화하기 위해선 우변의 $1 - D(G(z))$를 최솟값으로 만들어야 한다. 따라서 생성자 $G$는 $D(G(z))$가 1의 값을 가지도록 학습한다.

## 2.2 CycleGAN

Zhu et al.(2017)은 이미지 간 변환 쌍이 없는 데이터에 대해 이미지 간 변환을 수행하는 CycleGAN[6]을 제안했다. 즉, 기존의 이미지 간 변환에선 GAN[4]의 학습을 위해 이미지 간 변환 쌍이 필수적으로 요구되었다. 그러나 CycleGAN[6]은 이미지 간 변환 쌍이 존재하지 않고 원본과 변환 대상에 해당하는 데이터를 변환하고 다시 복원하는 방법을 통해 이미지 간 변환을 수행한다. 이에 따라 기존의 GAN[4]에서 생성자와 식별자가 하나씩 존재했던 반면에, CycleGAN[6]은 이미지 간 변환을 위한 생성자와 식별자가 한 쌍, 그리고 변환된 이미지를 복원하기 위한 생성자와 식별자가 한 쌍으로, 총 두 쌍으로 이루어져 있다.

Fig. 1은 CycleGAN[6]의 간략한 구조와 전체 신경망과 데이터 이동 과정을 간략히 나타낸 것이다. $x$와 $y$는 각각 생성자 $G_1$, $G_2$에 입력으로 들어가는 원본 이미지와 변환을 목표로 하는 데이터를 의미하며, $x'$, $y'$는 생성자를 통해 변환된 출력 이미지이다. Cycle-consistency 손실 함수는 이미지 변환 단계에서 다시 복원될 수 있도록 제약 조건을 주며 원본 이미지의 주요 특성들이 이미지 변환 단계에서 복원될 수 있는 수준으로 크게 요약 또는 손실되지 않으면서 이미지 변환을 수행하도록 생성적 적대 신경망 모델을 유도한다.

## 2.3 Self-Attention의 이미지 처리

자연어 처리 분야에서 큰 영향력을 미쳤던 Self-Attention[5] 기법은 자연어뿐만 아니라 다른 여러 분야에도 영향을 미쳤다. 그중에도 이미지 처리에서 Self-Attention[5]은 쿼리-키 관계의 분석을 그대로 사용할 수 있다.

Fig. 2는 그 과정을 시각화하며, 먼저 특징 맵을 3개 복제한 뒤 각각 채널 간 합성곱 연산을 진행한다. 이후 쿼리 역할을 할 특징 맵을 전치시킨 후, 복사했던 다른 특징맵과 요소별 곱셈을 실행한다. 이 결과물을 어텐션 맵이라고 할 때, 이는 쿼리에 존재하는 모든 픽셀에 대해 자기 자신에 대한 연관성을 학습할 수 있도록 하며 이러한 연관성은 행렬 곱이 실행된 뒤, 소프트맥스 함수를 통과하면서 구현된다. 소프트맥스 함수를 통과한 어텐션 맵은 가장 처음 복사했던 특징 맵 3개 중에 사용되지 않았던 특징 맵과 그대로 행렬 곱을 진행하며 이는 어텐션 맵에서의 중요도가 특징 맵에 마스크를 씌우는 것과 동일하게 동작하게 되어 어텐션이 낮은 부분의 특징 맵은 값이 작아지고 어텐션이 높은 부분의 특징 맵은 값이 유지되거나

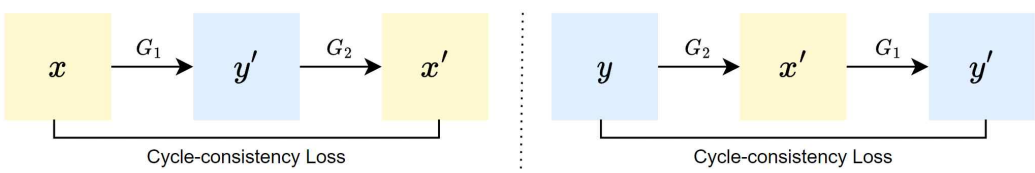


Fig. 1. Training with cycle-consistency loss function.

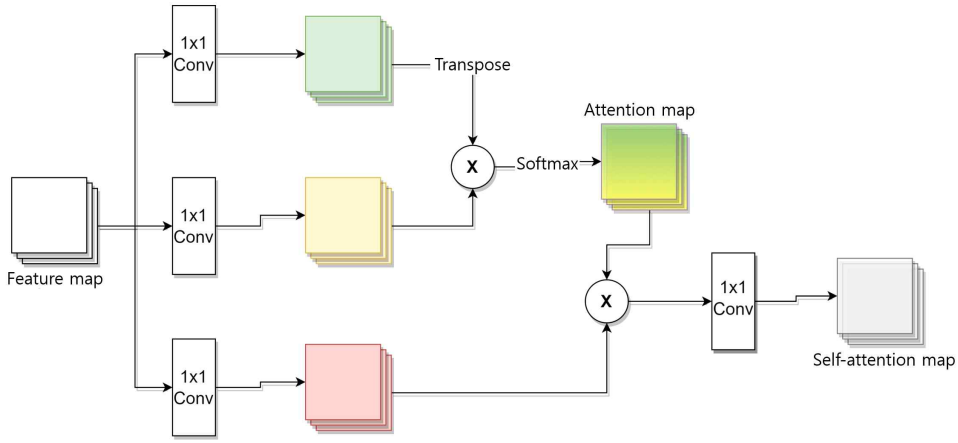


Fig. 2. Self-attention mechanism for image processing.

아주 적게 영향을 받는다.

## 3. 제안한 방법

Fig. 3은 제안한 다중 스케일 셀프-어텐션 기반 이미지 간 변환 GAN 모델의 구조이며 이미지 변환은 두 가지 단계로 나눌 수 있다. Fig. 3의 윗부분은 원본 이미지를 목표로 하는 분류에 해당하는 이미지로 변환하는 첫 번째 단계를 보여준다. 소스 도메인으로부터의 기존 이미지 $x$가 생성자 $G_{s-t}$를 통과해 타겟 도메인의 이미지로 변환된 $y'$를 얻는다. 이 이미지를 다시 생성자 $G_{t-s}$에 통과시켜 다시 소스 도메인의 이미지로 복구한 $x''$을 얻는다. 이때 $y'$가 $x$의 특성

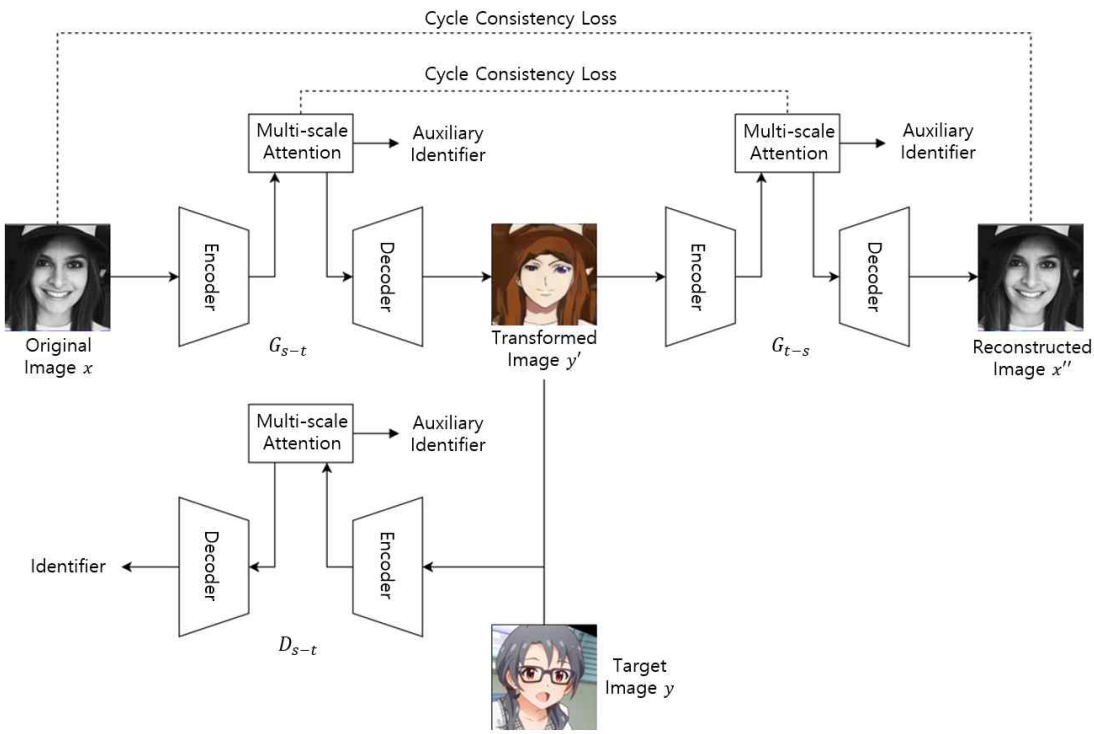


Fig. 3. Overview of the proposed model for image-to-image translation.

을 제대로 반영하여 이를 기반으로 복구하였을 때 $x''$가 기존의 이미지 $x$와 유사해지도록 유도한다. 본 논문에서 제안하는 다중 스케일 어텐션 모듈과 적응형 특징 맵 합성 기법은 첫 번째 단계의 중간단계에 적용되어 있다. 생성자는 인코더-디코더 구조로 구성되었으며 $G_{s-t}$는 소스 도메인의 이미지를 타겟 이미지로, $G_{t-s}$는 타겟 도메인의 이미지를 소스 도메인의 이미지로 변환한다. 아래의 두 번째 단계에서는 판별자 $D_{s-t}$를 통해 첫 번째 단계에서 변환된 이미지 $y'$와 목표로 하는 타겟 도메인의 실제 이미지 $y$인지 분류하여 $y'$가 타겟 도메인의 특징을 제대로 반영하도록 유도한다. 이 때 $y$는 $x$를 애니메이션으로 변환한 정답 영상이 아닌 타겟 도메인 내의 랜덤한 샘플 영상으로 변환 대상 데이터의 도메인 정보를 제공하는 용도로만 사용된다. 이는 신경망이 소스 이미지 $x$를 타겟 도메인의 스타일로 변환하기 위해 해당 도메인의 특성을 이해할 수 있는 정보를 필요로 하기 때문이다. 주기 특성 손실 함수는 기존의 생성적 적대 신경망의 손실 함수에 적용되어 있으며 두 제안 방법 모두 아래에서 자세히 설명한다.

### 3.1 다중 스케일 어텐션 모듈

기존의 생성적 적대 신경망 기반 이미지 간 변환 기법들은 이미지의 변환 과정에서 원본으로 사용되는 이미지의 중요 특성들을 보존하는 별다른 과정 없이 생성적 적대 손실 함수에 의존하여 학습되었다. 이와 달리 본 논문에서는 생성자 부분에 추가적인 셀프-어텐션 모듈을 적용하여 이미지 간 변환 단계에서 원본으로 사용되는 이미지의 주요 특성(얼굴 모양, 머리 형태, 표정 등)을 보존하고자 했다. 그러나 종래의 셀프-어텐션 모듈[5]은 필터의 정사각형 모양으로 인해 높이나 폭이 긴 형태와 같이 비율과 관련된 특징을 수용하기 어려운 문제가 있다. 이에 따라 본 논문에서는 Fig. 4와 같이 정사각형 모양의 어텐션뿐만 아니라 높이 또는 폭이 긴 형태의 어텐션 모듈을 추가하고 높이와 폭을 동시에 줄인 저해상도 어텐션 모듈을 추가했다. 이러한 추가적인 스케일의 어텐션 모듈은 기존의 정사각형 어텐션 모듈이 인지하기 힘든 폭이 긴 형태의 상관관계(눈과 눈) 또는 높이가 긴 형태의 상관관계(코와 입)를 효과적으로 인지할 수 있도록 한다. 다중 스케일 어텐션의 합성은 기존의 셀프-어텐션 기법을 포함하여 가로(Width) 풀링과 세로(Height) 풀링으로 폭과 높이가 절반 줄어든 특징 맵에 셀프-어텐션을 취하는 부분과 폭과 높이 모두 절반 줄어든 특징 맵에 셀프-어텐션을 취하는 부분을 포함한다. 어텐션 맵을 취한 이후엔 가장 크기가 큰 어텐션 맵을 기준으로 모든 어텐션 맵의 폭과 높이를 변경하여 출력한다.

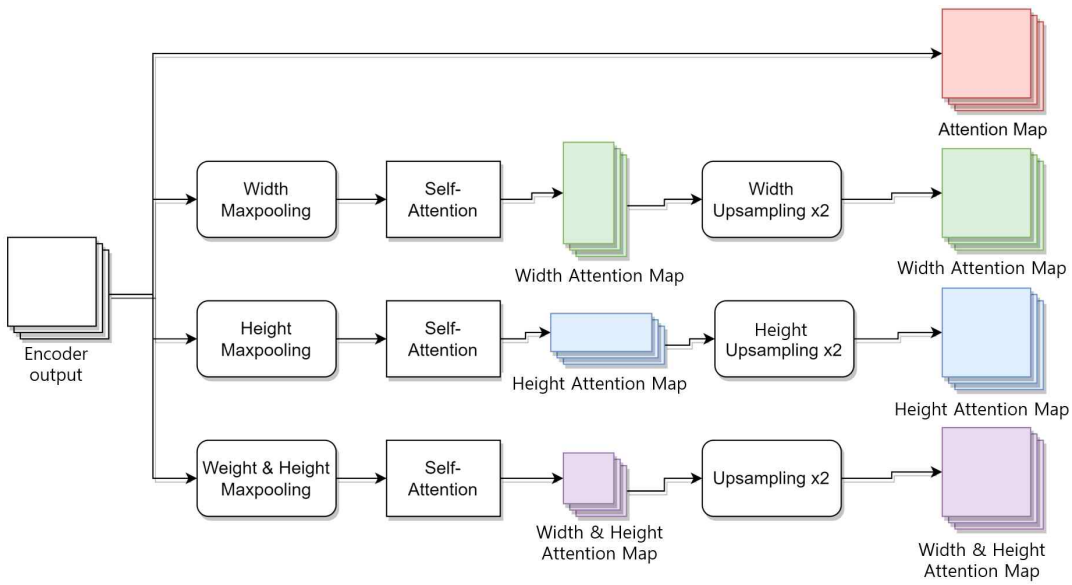


Fig. 4. Flowchart for multi-scale attention module.

### 3.2 적응형 특징 맵 합성

생성적 적대 신경망에 많은 정보를 제공하여도 이미지 간 변환에 유효한 의미를 지니는 정보는 상대적으로 적다. 이는 이미지에서 특히 두드러지게 나타나는 현상으로 변환의 대상이 되는 객체와 배경 간의 이미지 점유 넓이와 관련 있다. 생성적 적대 신경망이 학습 단계일 때, 이러한 불필요한 정보가 유효한 정보를 압도하지 않도록 다중 스케일 어텐션 모듈에서 생성되는 다양한 스케일의 어텐션 맵의 중요도를 학습 단계에서 스스로 파악하고 이를 가중치로 반영하는 기법을 적용한다. Fig. 5는 이 과정을 시각화해서 보여주며 중요도를 판단하는 매개변수는 생성적 적대 신경망의 손실 함수와 함께 학습된다. $w_1, w_2, w_3$은 각 채널의 중요도를 나타내는 학습 가능한 가중치이다.

### 3.3 손실 함수

본 논문이 제안하는 생성적 적대 신경망을 학습시키는 손실 함수는 아래와 같으며 학습 단계에서는 아래의 모든 손실 함수를 모두 더한 값을 기준으로 역전파 기법을 활용하여 신경망의 모든 가중치와 학습 가능한 매개변수를 갱신한다.

#### 3.3.1 적대적 손실 함수

기본적인 적대적 학습을 위한 손실 함수를 사용하고 보다 안정적인 학습을 위해 일반적인 생성적 적대 신경망 학습에 사용되는 손실 함수[4]가 아닌 최소 제곱법 GAN[7]을 적대적 학습을 위한 손실 함수에 사용했다. 주어진 사람의 사진 $X_s$의 이미지를 $x$라 하고 애니메이션 얼굴 이미지 $X_t$에 속하는 이미지를 $y$라고 할 때, 생성자는 $x$를 입력으로 하여 $X_t$에 속하도록 변환한 이미지 $G_{s-t}(x)$를 생성한다. 식별자는 $X_t$에 속하는 사람의 이미지 $y$와 $G_{s-t}(x)$로 변환된 이미지를 구별하는 역할을 하도록 한다.

$$L_{adv}^{s-t} = E_{y \sim X_t}[D(y)^2] + E_{x \sim X_s}[(1 - D(G_{s-t}(x)))^2] \quad (2)$$

#### 3.3.2 Cycle Consistency 손실 함수

본 논문에서는 정답이 없는 비지도 학습 기반으로 이미지 간 변환 문제를 해결하기 위해 Cycle Consistency Loss[6]를 채용했다. Fig. 6는 원본 이미지와 각 생성자에 의해 생성된 이미지 샘플을 통해 Cycle Consistency Loss의 학습 원리를 시각화하여 보여준다. 이 손실은 생성자에 의해 변환된 이미지 $G_{s-t}(x)$를 반대로 역변환 했을 때 이미지 $X_s$에 속하는 $x$와

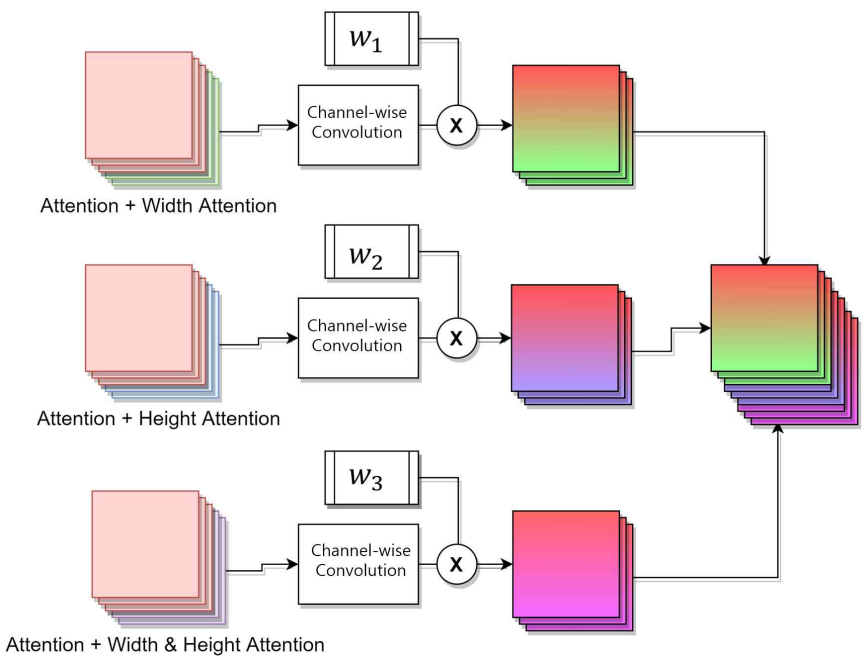


Fig. 5. Adaptive feature map fusion process.




Fig. 6. Visualization of cycle consistency loss training.

$G_{t-s}(G_{s-t}(x))$간에 차이를 비교한다. 이 손실 함수는 생성자가 이미지 간 변환을 수행할 때 역변환이 불가능할 정도로 주요 특성을 생략 및 훼손하지 않도록 제약을 거는 역할을 한다. 따라서 $X_s$에 속하는 이미지 $x$가 생성자 $G_{s-t}$에 의해 변환되고 다시 $G_{t-s}$에 의해 복원되었을 때 이미지가 성공적으로 복원 될 수 있도록 한다.

$$L_{cycle}^{s-t} = E_{x\sim X_s}[\,|\,x - G_{t-s}(G_{s-t}(x))\,|\,_1] \quad (3)$$

### 3.3.3 주기 특성 손실 함수

변환 단계에서 $X_s$에 속한 이미지 $x$의 특성을 보존하기 위해 $x$에서 얻은 특징 맵과 $G_{t-s}(x)$에서 얻은 특징 맵이 서로 유사한 특성을 가져야 한다고 가정했다. Cycle Consistency 손실 함수[6] 또한 $X_s$의 이미지 $x$와 복원된 이미지 $G_{t-s}(G_{s-t}(x))$간의 차이가 적게 제약을 걸었다면, 주기 특성 손실 함수는 $X_s$의 이미지 $x$로부터 얻은 특징 맵 $\mathrm{Enc}(x)$와 $G_{s-t}(x)$의 복원 단계의 $\mathrm{Enc}(G_{s-t}(x))$간의 차이가 적도록 제약을 건다. 이 손실 함수는 높은 수준의 특징 맵 단계에서 제약을 걸기 때문에 $X_s$의 이미지 $x$의 지역적 특성이 상대적으로 요약 및 손실되지 않고 전달하도록 하는 기능을 한다. 단, $s-t$ 변환 단계와 $t-s$ 변환 단계가 유사하지만 동일하지는 않아야 형상적 변환에 대한 유연성을 확보할 수 있다. 따라서 복원 과정의 이미지가 원본 이미지 $x$에 지나치게 가까워지지 않도록 Smoothed $L_1$ 손실을 사용한다.

$$L_{content}^{s-t} = E_{x\sim X_s}[\,|\,\mathrm{Enc}_{s-t}(x) - \mathrm{Enc}_{t-s}(G_{s-t}(x))\,|\,_1] \quad (4)$$

$\mathrm{Enc}_{s-t}$는 $X_s$의 이미지 $x$를 $X_t$에 속하는 이미지 $y$로 변환하는 과정에서 사용되는 생성자 $G_{s-t}$의 인코더를 뜻하며, $\mathrm{Enc}_{t-s}$는 복원 과정 $G_{t-s}$의 생성자에 속한 인코더를 뜻한다.

### 3.3.4 보조 식별자 손실 함수

생성자의 출력을 생성하는 디코더 부분의 입력의 절반은 다중 스케일 어텐션 모듈에서 전달받는다. 이때 무작위로 초기화된 가중치 값을 갖는 셀프-어텐션 모듈은 학습 초기에 유의미한 어텐션 맵을 얻지 못하고 생성자의 출력에 악영향을 미치고 식별자의 성능이 압도하는 일이 발생할 수 있다. 이에 따라 본 논문에서 제안하는 다중 스케일 어텐션 모듈에 추가적인 손실 함수를 부여한다. 식 5는 보조 분류기 손실 함수로 어텐션 모듈의 출력을 가져와 간단한 이진 분류를 수행하도록 한다. 이때 $\eta_G, \eta_D$는 각각 생성자, 판별자의 보조 분류기이며 생성자의 보조 분류기의 손실 함수에는 로그 함수를 사용하여 학습 초기의 손실 함수가 다른 손실 함수보다 큰 값을 가지도록 하여 신경망의 다른 부분 보다 보조 분류기에 연결된 다중 스케일 어텐션 모듈을 먼저 학습시켜 유효한 결과를 얻도록 한다. 손실 함수 계산에는 생성자의 보조 분류기와 식별자의 보조 분류기는 입력으로 들어온 어텐션 맵이 $X_s$ 또는 $X_t$에서 유래했는지 분류하며 이는 본 논문의 식별자가 학습하는 방법과 동일하다. $L_{G:aux}^{t-s}$ 도 이와 유사하게 설계되었다.

$$\begin{aligned} L_{G:aux}^{s-t} &= E_{x\sim X_s}[-\log(\eta_G(x))] + E_{y\sim X_t}[-\log(1-\eta_G(y))] \\ L_{D:aux} &= E_{y\sim X_t}[\eta_D(y)^2] + E_{x\sim X_s}[(1-\eta_D(G(x))^2] \end{aligned} \quad (5)$$

### 3.3.5 최종 손실 함수

$$L \quad (6)$$

최종적으로 상기 언급된 모든 손실 함수를 더하여 생성자, 보조 식별자, 식별자를 최적화하는 방향으로 학습하게 된다. 본 논문에서 $L_{adv} = L_{adv}^{s-t} + L_{adv}^{t-s}$ 의미하며 다른 손실 함수도 동일한 방법으로 정의된다. 위 식에서 매개변수의 값은 $\lambda_1$=1, $\lambda_2$=10, $\lambda_3$=10, $\lambda_4$=1000이다.

## 4. 실험 결과 및 고찰

### 4.1 실험 환경

본 논문에서 생성적 적대 신경망을 손실 함수로부터 최적화할 때 ADAM 최적화 기법[8]을 사용한다. 학습률은 $1e^{-4}$에서 0.5와 0.999의 $\beta_1$, $\beta_2$ 값을 사용했다. 또한 가중치 업데이트의 감쇄율은 $1e^{-5}$로 설정했다. 학습을 위한 데이터 배치는 1로 설정했으며 본 논문에서 사용된 모든 모델은 100만 회 반복 학습되었다. 데이터 증강을 위해서는 거울 대칭을 적용하고 256 * 256 크기로 무작위 잘라내기를 수행했다.

본 논문에서 제안하는 생성적 적대 신경망은 3세대 Intel 제온 Gold 6330N, RTX 3090 2개와 128GB RAM이 장착된 PC에서 진행했다. Ubuntu 20.04에서 Python 3.9와 Pytorch 1.11 버전을 CUDA 11.3과 함께 딥러닝 프레임워크로 사용했다. 기타 의존성 소프트웨어 요구사항은 Pytorch의 의존성에 따라 사용되었다.

### 4.2 데이터셋

학습에 사용되는 데이터셋은 CycleGAN[6], UGATIT[9]와 같은 최신 연구들은 사람-애니메이션 얼굴 이미지 간 변환 성능을 검증하기 위해 Selfie2Anime[9]와 Photo2Anime[10] 데이터셋을 사용했다. 두 데이터셋은 여성의 얼굴 사진과 애니메이션 얼굴 이미지를 포함하며 짝이 없는 형태로 구성되어 있으므로 비지도 기반 생성적 적대 신경망을 학습하고 평가하는 데 널리 사용되기 때문에 본 논문에서도 활용하였다. Selfie2Anime 데이터셋은 MUNIT[11], DRIT[12], AGGAN[13], UNIT[14], CycleGAN[6], CartoonGAN[15], UGATIT[9]의 모델에서 활용되었고, 상대적으로 최근 발표된 Photo2Anime 데이터셋은 CycleGAN[6], CartoonGAN[15], UGATIT[9]에서 사용되었다. 두 데이터셋 모두 사람의 얼굴 사진은 대부분 백인 여성으로 구성되어 있고, 애니메이션 얼굴 사진은 대부분이 앞머리를 기르고 있는 형태로 구성 되어있는 편향성이 다소 존재한다.

### 4.3 모델 구조

본 모델의 인코더와 디코더는 UGATIT[9] 모델에서 추출하였다. 인코더는 초기 입력 이미지로부터 고수준의 특징을 추출하기 위해 스트라이드 크기가 2인 컨볼루션 층과 4개의 잔차 블록(Residual Block)으로 이뤄져 있다. 본 연구에서는 인코더의 중간 단계에 자체적으로 개발한 다중 어텐션 모듈을 추가하여 특징을 향상시켰다. 디코더는 인코더와 다중 어텐션 모듈을 거친 특징 맵을 처리하기 위해 4개의 잔차 블록과 2개의 업샘플링 컨볼루션 레이어로 구성되었으며 AdaLIN 정규화[9]층을 추가하여 인코딩 단계에서 얻은 다양한 수준의 특징들이 적절히 정규화하였다. 다중 어텐션 모듈 내부에는 각각의 특징 맵 차원 (너비, 높이, 모든 방향)에 대해 셀프-어텐션이 구현되어 있으며, 각 방향 별로 총 3개의 컨볼루션 층과 업샘플링 층으로 이루어져 있다.

### 4.4 평가 지표

평가 지표로는 IS[16], FID[17], KID[18]를 선택했다. IS는 Inception Score의 약자로 Sharpness와 Diversity의 곱으로 계산된다. Sharpness는 Classifier가 이미지를 입력으로 받았을 때 얼마나 높은 Confidence Score를 가지고 분류를 하는지에 대한 지표이다. Diversity는 분류기가 특정 분류를 진행할 때 얼마나 여러 종류로 분류하는지에 대한 지표이다. IS는 이미지가 확신을 가지고 분류되며 여러 가지로 분류될수록 더 높은 점수를 얻으며 높을수록 이미지의 품질이 좋다고 평가할 수 있다.

### 4.5 실험 결과

Selfie2Anime와 Photo2Anime 데이터셋으로 비교 실험한 결과는 Fig. 7-9, Table 1, 2에 나타내었다. Fig. 8과 Fig. 9에서는 가장 유사한 성능 지표를 가지는 UGATIT[9] 모델의 결과와 비교하여 제안 모델은 좀 더 원본 이미지의 표정, 머리모양, 얼굴 형태와 전체적인 색상을 반영하는 결과를 얻을 수 있다는 것을 알 수 있다. 특히 Selfie2Anime 데이터셋에서는 표정에 대한 특성 보존이 두드러지게 나타났고, Photo2Anime 데이터셋에서는 머리 모양에 대한 특성 보존이 눈에 띄게 향상된 것을 볼 수 있다. 어텐션 맵 분석을 보여주는 Fig. 7에 따르면 어텐션 기법이 적용되지 않은 UGATIT, 기존 단일 어텐션 모듈 또는 제안 다중 어텐션 모듈이 적용되었을 때 어텐션 맵 결과와 그 변환 결과를 볼 수 있다. 어텐션 맵에서 적색에 가까울수록 해당 영역이 점점 강조되는 것을 의미하며 파란색일수록 상대적으로 강조하는 정도를 줄이

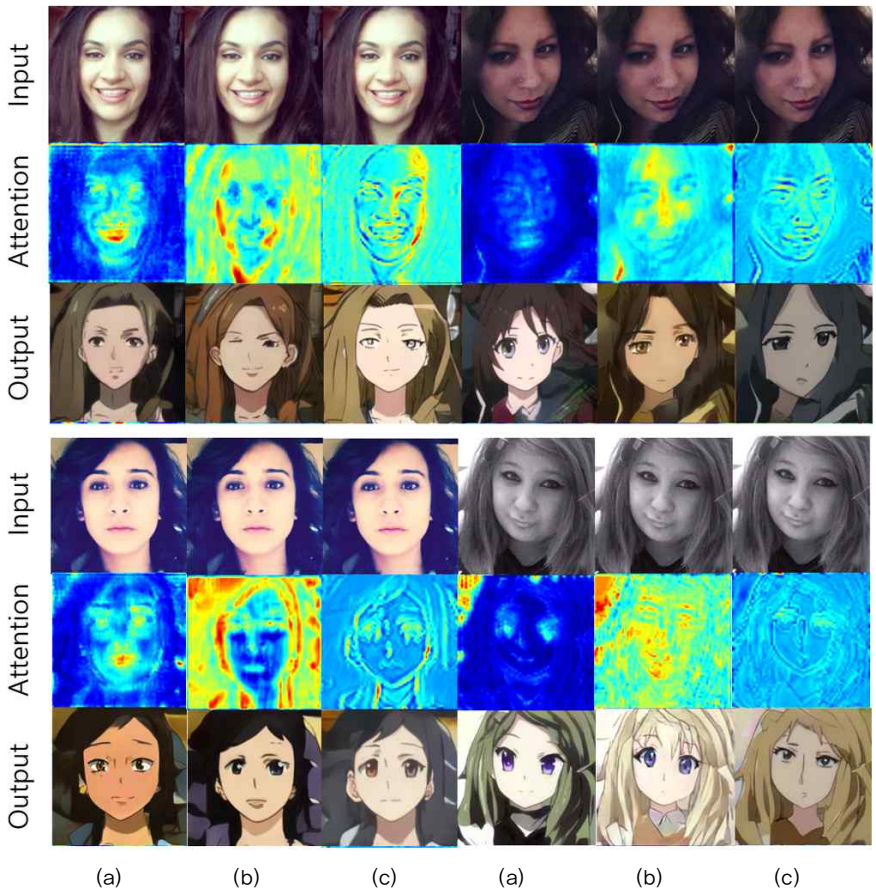


Fig. 7. Comparison of attention maps and image transfer results. (a) UGATIT without attention module, (b) with single attention, and (c) with proposed multi-scale attention.

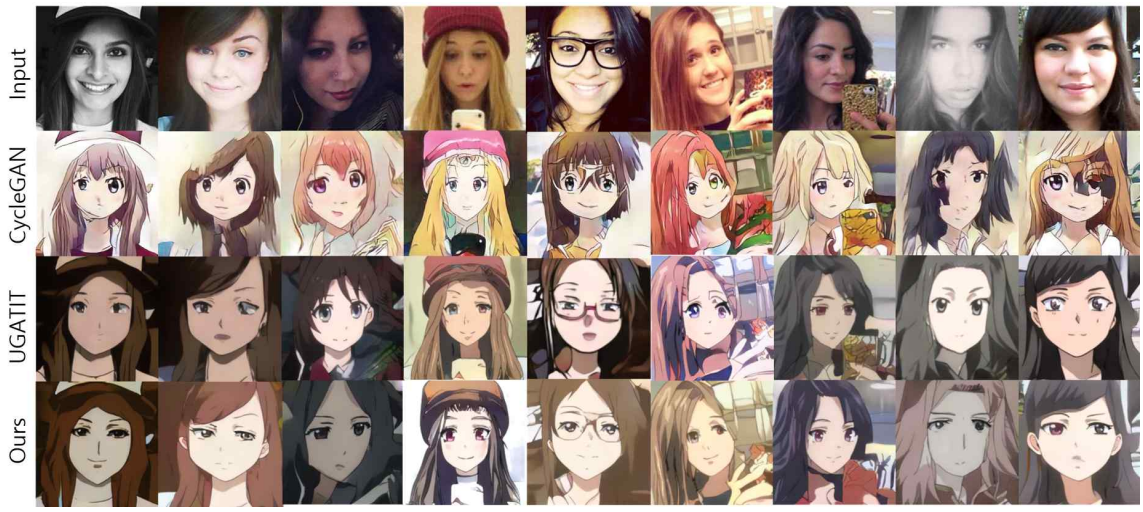


Fig. 8. Transfer results by CycleGAN, UGATIT and our method on Selfie2Anime dataset.




Fig. 9. Transfer results by CycleGAN, UGATIT and our method on Photo2Anime dataset.

Table 1. Evaluation metrics for the quality assessment of image translation on 100 test images from the Selfie2Anime dataset. We multiply KID by 100 and divide FID by 100 for better comparison.

| Method | IS ↑ | FID ↓ | KID ↓ |
|---|---|---|---|
| MUNIT [11] | - | - | 13.85 |
| DRIT [12] | - | - | 15.08 |
| AGGAN [13] | - | - | 14.63 |
| UNIT [14] | - | 1.59 | 14.71 |
| CycleGAN [6] | - | 1.46 | 15.85 |
| CartoonGAN [15] | - | 1.21 | 12.8 |
| UGATIT [9] | 1.50 | 1.37 | 11.61 |
| Ours | 1.65 | 1.17 | 9.70 |

Table 2. Evaluation metrics for the quality assessment of image translation on 898 test images from the Photo2Anime dataset. We multiply KID by 100 and divide FID by 100 for better comparison.

| Method | IS ↑ | FID ↓ | KID ↓ |
|---|---|---|---|
| CycleGAN[6] | - | 1.75 | 16.52 |
| CartoonGAN[15] | - | 1.53 | 16.08 |
| UGATIT[9] | 1.73 | 1.59 | 14.21 |
| Ours | 1.81 | 1.38 | 13.33 |

는 것을 의미한다. Fig. 7은 제안하는 다중 어텐션 모듈을 적용했을 때 얻는 어텐션 맵이 어텐션 모듈을 제외했을 경우와 기존 단일 어텐션 방법을 사용했을 때와 비교하여 얼굴의 지역 정보(얼굴 윤곽선, 눈, 코, 입등)를 더 뚜렷하게 강조함을 볼 수 있다. 이러한 결과는 생성자가 이미지 간 변환을 진행할 때 원본 이미지의 얼굴형, 표정, 얼굴 지역 정보 같은 고유 특성을 최대한 유지하며 애니메이션 스타일로 변환을 가능케 한다. Fig. 7에서 관찰된 제안 다중 어텐션 기법의 효과는 Fig. 9에서도 관찰할 수 있으며, 다중 어텐션 맵을 사용했을 때, 사용하지 않은 다른 두 모델에 비해 얼굴 윤곽, 형태 및 이목구비를 더 강조하는 모습을 보여준다. 이는 다중 어텐션 맵이 얻은 뚜렷한 얼굴의 지역 정보가 변환 결과에 반영되어 코와 입이 일부 변환되지 않고 얼굴형이 찌그러지는 다른 두 모델에 비해 다중 어텐션 모듈을 활용한 변환에서는 코와 입이 모두 표현되어 있고 얼굴 형태 또한 원본의 형태를 그대로 따른 변환이 되었음을 볼 수 있다.

제안 모델은 다른 최신 연구와 비교하여 모든 지표에서 성능이 최소 0.15에서 최대 1.91까지 성능이 향상되었으며 이는 10~15% 정도의 성능 향상을 보여주고 있다. 특히 KID 지표에서 성능향상이 두드러지게 나타나며 그 다음이 FID, IS 순으로 성능이 개선되는 것을 볼 수 있다. KID가 낮은 점수를 기록하는 이유는 제안 모델이 타 모델에 비해 변환하는 이미지의 특징이 실제 애니메이션 이미지의 특징과 유사한 것을 의미하여 전체적인 이미지 간 변환 품질이 다른 모델에 비해 우수하다는 것을 의미한다.

Table 3에서는 제안 기법을 하나씩 추가할 때마다 KID의 변화를 알 수 있다. 모두 미적용은 UGATIT

Table 3. Performance improvement by adding the proposed module on 100 test images from the Selfie2Anime dataset. We multiply KID by 100 for better comparison.

| Modules | KID ↓ |
|---|---|
| Baseline | 12.85 |
| +Content Loss | 13.08 |
| +Single-scale Attn. | 12.14 |
| +Multi-scale Attn. | 10.63 |
| +Adaptive Fusion | 9.70 |

[9]의 모델에서 제안한 어텐션 부분을 제외했을 때의 성능이다. 이후 어텐션 부분이 제외된 UGATIT[9] 모델에 제안 방법을 하나씩 추가한다. Content Loss를 추가했을 때 지표가 증가하여 변환 성능이 오히려 감소했다. 이는 Content Loss가 원본 이미지의 특성이 변환된 이미지에 남도록 제약을 건 부분이 KID와 같은 성능 지표에 나타난 것으로 볼 수 있다. 즉, 변환된 이미지 오히려 성능이 떨어진 것으로 볼 수 있다. 이러한 결과는 Fig. 10에서 이미지 변환 품질이 떨어진 것으로 성능이 약간 하락했음을 알 수 있다. 이후 본 논문에서 제안한 다중 스케일 어텐션 모듈을 단일 스케일로 변경하여 적용했을 때 KID가 약 0.9 감소한다. 또한 단일 스케일에서 다중 스케일로 어텐션 모듈을 변경했을 때 KID가 약 1.5 감소하는 것을 볼 수 있다. 이는 다중 어텐션 모듈이 본 논문에서 목표한 대로 원본의 특성과 형상적 상관관계를 충분히 생성자에 전달하여 변형된 이미지의 품질이 높아졌음을 알 수 있다. 최종적으로 다중 스케일 어텐션 모듈의 출력을 가중치 기반으로 융합하는 적응형 특징 맵 합성기법을 추가했을 때 KID가 약 0.8 하락하는 것을 볼 수 있다. 이는 적응형 특징 맵 합성 기법이 다중 스케일의 어텐션으로부터 얻은 여러 어텐션 맵의 중요도를 파악하고 이미지 간 변환에 효과적으로 이를 적용하여 성능이 향상한 것으로 볼 수 있다. 특히 Selfie2Anime 데이터셋의 샘플 결과에 의하면 원본 이미지의 표정을 더 많이 반영하는 결과를 얻을 수 있음을 알 수 있으며, Photo2Anime 데이터셋의 경우 머리 모양과 색, 얼굴 모양에서 더 원본 이미지

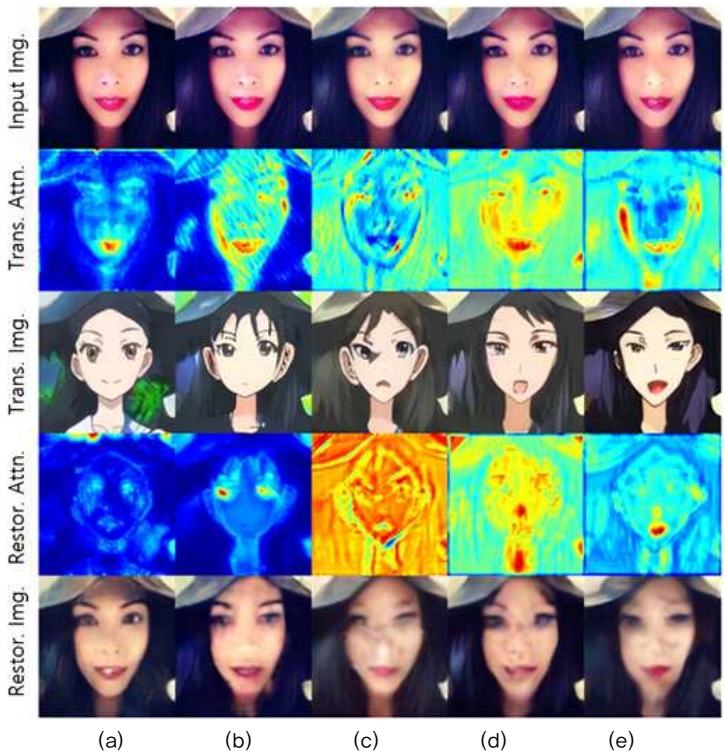


Fig. 10. Visualization of image-to-image translation by adding the proposed module. (a) UGATIT without attention, (b) +Content Loss, (c) +Single-scale Attn., (d) +Multi-scale Attn., and (e) +Adaptive Fusion.

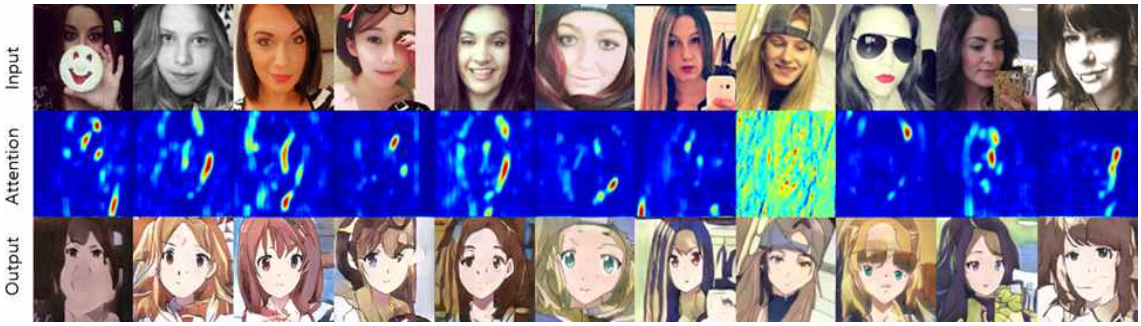


Fig. 11. Visual representation of incorrectly generated samples on Photo2Anime dataset.

의 특성을 반영하는 이미지 변환 결과를 얻을 수 있음을 알 수 있다.

제시한 실험 결과들을 바탕으로 다중 스케일 어텐션 모듈과 적응형 특징 맵 합성 기법, 그리고 주기 특성 손실 함수를 생성적 적대 신경망과 함께 활용하면 원본 얼굴 이미지의 표정, 머리모양, 색감 등의 주요 특성을 보존하면서 성공적으로 이미지 간 변환을 할 수 있고 최신 연구와 비교하여 IS, FID, KID와 같은 성능 지표를 기준으로도 성능의 개선이 가능함을 보였다.

특이 사항으로, 원본 영상에 앞머리가 없는 경우에도 대부분의 애니메이션 변환 결과에서 앞머리가 생성되는 경향을 관찰하였다. 또한 서양인이 아닌 사람 얼굴을 변환할 때 품질이 현저히 떨어지는 현상이 발생한다. 이는 Selfie2Anime 및 Photo2Anime 두 데이터셋에 포함된 애니메이션 영상의 대부분이 앞머리를 포함하고 있고 인물 영상의 대부분이 백인 여성으로 구성되어 있어서 발생하는 편향 때문인 것으로 보인다. 이러한 문제들은 앞머리가 없는 애니메이션 영상과 백인 여성이 아닌 다양한 인물의 영상을 추가로 확보하게 되면 해결될 수 있다. Fig. 11은 Photo2Anime 데이터셋에서 잘못 변환된 결과 샘플들을 보여준다. 예시에서와 같이 원본 영상에서 얼굴을 가리는 요소가 있는 경우 눈, 코, 입과 같은 중요한 특징들이 인식되지 않거나 잘못 해석 될 수 있다. 또한, 앞서 언급한 문제와 같이 원본 영상에 앞머리가 없는 경우 앞머리를 만드려는 경향과 원본의 패턴을 유지하려는 의도가 충돌하여 부자연스러운 결과를 도출하기도 했다.

## 5. 결 론

본 논문에서는 다중 스케일 어텐션 모듈과 적응형 특징 맵 합성 기법과 주기 특성 손실 함수를 활용하는 비지도 학습 기반 생성적 적대 신경망을 활용하는 이미지 간 변환 모델을 제안했다. 여러 스케일의 어텐션을 특정하는 모듈을 사용하고 이로부터 출력된 여러 어텐션 맵을 중요도에 따라 모델 스스로 가중치를 부여하는 방법과 원본 이미지의 주요 특성이 과도하게 생략되지 않도록 생성적 적대 신경망에 제약을 가하는 주기 특성 손실 함수를 사용한다. 실험 결과를 통해 본 논문에서 제안하는 기법들은 기존 기법들과 비교하여 IS, FID 및 KID의 성능 지표에서 약 10%에서 19%까지 더 우수한 결과를 원본으로 사용된 이미지의 특성 중에서도 Selfie2Anime 데이터셋에서는 원본의 표정이 반영되는 이미지 변환 결과를 얻을 수 있었고 Photo2Anime 데이터셋에서는 주로 원본 이미지의 얼굴 모양이나 머리 모양을 반영하는 이미지 변환 결과를 얻을 수 있었다. 제안 방법을 하나씩 추가하는 실험에서는 각각의 제안 방법이 생성적 적대 신경망에 어떠한 영향을 미치는지 성능 지표와 이미지 변환을 통해 알 수 있었다. 각각의 제안 방법은 단독으로 사용될 때 성능 저하를 유발하기도 했으나 제안 방법을 조합하여 활용하였을 때 최종적으로 성능 지표상에서 더 높은 성능을 보였고 이미지 변환 결과에서도 다른 모델과의 비교를 통해 원본 이미지의 특성을 더 많이 반영하는 이미지 변환 결과를 얻을 수 있음을 보였다. 상기 실험 결과를 통해 제안 방법을 생성적 적대 신경망에 조합하여 활용하면 비지도 학습 기반 이미지 간 변환 성능을 개선하고 원본 이미지의 특성을 반영하는 이미지 변환이 가능함을 입증하였다.

## REFERENCE


[1] J. Qiu and K. Xie, "A GAN-Based Motion Blurred Image Restoration Algorithm," *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, pp. 211-215, 2019.

[2] H. Ha and B.Y. Hwang, "Enhancement Method of CCTV Video Quality Based on SRGAN," *Journal of Korea Multimedia Society*, Vol. 21, No. 9, pp. 1027-1034, 2018.

[3] X. Liu, Z. Gao, and B. M. Chen, "MLFcGAN: Multilevel Feature Fusion-Based Conditional GAN for Underwater Image Color Correction," *IEEE Geoscience and Remote Sensing Letters*, Vol. 17, No. 9, pp. 1488-1492, 2019.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, Vol. 27, pp. 139-144, 2014.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, et al., "Attention is All You Need," *Advances in Neural Information Processing Systems*, Vol. 30, pp. 6000-6010, 2017.

[6] 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, 2017.

[7] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, and S. Paul Smolley, "Least Squares Generative Adversarial Networks," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2794-2802, 2017.

[8] D.P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv Preprint,* arXiv:1412.6980, 2014.

[9] J.H. Kim, M.J. Kim, H.W. Kang, and K.H. Lee, "U-GAT-IT: Unsupervised Generative Attentional Networks With Adaptive Layer-Instance Normalization for Image-to-Image Translation," *arXiv Preprint,* arXiv:1907.10830, 2019.

[10] B. Li, Y. Zhu, Y. Wang, C.W. Lin, B. Ghanem, and L. Shen, "AniGAN: Style-Guided Generative Adversarial Networks for Unsupervised Anime Face Generation," *IEEE Transactions on Multimedia*, Vol. 24, pp. 4077-4091, 2021.

[11] X. Huang, M.Y. Liu, S. Belongie, and J. Kautz, "Multimodal Unsupervised Image-to-Image Translation," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172-189, 2018.

[12] H.Y. Lee, H.Y. Tseng, J.B. Huang, M. Singh, and M.H. Yang, "Diverse Image-to-Image Translation Via Disentangled Representations," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 35-51, 2018.

[13] H. Tang, D. Xu, N. Sebe, and Y. Yan, "Attention-Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation," *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, 2019.

[14] M.Y. Liu, T. Breuel, and J. Kautz, "Unsupervised Image-to-Image Translation Networks," *Advances in Neural Information Processing Systems*, Vol. 30, pp. 700-708, 2017.

[15] Y. Chen, Y.K. Lai, and Y.J. Liu, "Cartoongan: Generative Adversarial Networks for Photo Cartoonization," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9465-9474, 2018.

[16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," *Advances in Neural Information Processing Systems*, Vol. 29, pp. 2234-2242, 2016.

[17] 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*, Vol. 30, pp. 6626-6637, 2017.

[18] M. Bińkowski, D.J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," *arXiv Preprint,* arXiv:1801.01401, 2018.



한 아 영

2021년 한국외국어대학교 컴퓨터·전자시스템공학부 학사
2022년~현재 한국외국어대학교 컴퓨터공학과 석사
관심분야 : 인공지능, 딥러닝, 이미지 분류



김 경 태

2016년 중원대학교 의료공학과 학사
2017년~2019년 한국외국어대학교 컴퓨터공학과 석사
2019년~현재 한국외국어대학교 컴퓨터공학과 박사
관심분야 : 머신러닝, 패턴인식, 영상처리, 인공지능



최 재 영

2011년 KAIST 전기및전자공학과 박사
2008년~2012년 토론토대학 연구원
2012년~2013년 펜실베니아대학 연구원
2013년~2014년 삼성전자 책임연구원
2014년~2016년 중원대학교 의료공학과 조교수
2023년~2024년 브리티시컬럼비아대학 컴퓨터공학과 방문교수
2016년~현재 한국외국어대학교 컴퓨터공학과 부교수
관심분야 : 딥러닝, 머신러닝, 패턴인식, 영상처리