

연구과제 결과보고서


과제번호	13
------	----

연구과제명	국문	GAN 기반의 수어 합성 모델		
	영문	GAN based Sign Language Synthesis Model		
책임연구자	공과대학	산업공학과	성명	황보진경
공동연구자	공과대학	전기·정보공학부	성명	김현준
공동연구자	공과대학	전기·정보공학부	성명	황보진경
연구 결과 요약	<p>본 연구는 발화자와 수어통역사의 동작을 합성하기 위한 딥러닝 모델을 구현하고 성능을 평가하는 것을 목표로 한다. 모델의 입력으로는 수어통역사의 동작에서 추출한 스켈레톤 이미지와 발화자의 이미지를 함께 사용하며, 합성된 이미지의 품질을 높이기 위해 GAN의 학습 기법과 함께 다양한 Loss 함수를 사용하였다. 본 연구를 통해 얻어진 모델은 효율적으로 수어 동작을 합성할 수 있으며, 생성된 이미지의 품질을 더욱 향상시켰다.</p>			
활용 방안	<p>본 연구의 결과물이 수어 통역사가 우측 하단에 작은 영역으로 표시되는 기존의 방법과 달리 화자가 직접 수어를 하므로 화자의 비언어적 표현을 유지하여 몰입도를 높이고 전달력이 손실되지 않으며 가시성이 좋을 것으로 기대하고 있으며, 나아가 청각장애인의 영상 콘텐츠 소비 접근성을 높이는데 도움을 주고자 한다. 본 연구의 결과는 음성이나 텍스트로부터 수어 동작 스켈레톤 이미지를 생성할 수 있는 연구가 이어진다면 그 활용성이 극대화될 것이라 기대한다.</p>			

※ 2페이지 이내로 작성 (최종 연구결과물은 15페이지 이상 별첨)

※ 글꼴 : 함초롱바탕, 글자크기 : 11, 줄 간격 : 160%

2021. 12 . 30.

책임연구자 : 황보진경  (인)

GAN based Sign Language Synthesis Model

황보진경 · 김현준 · 홍선우

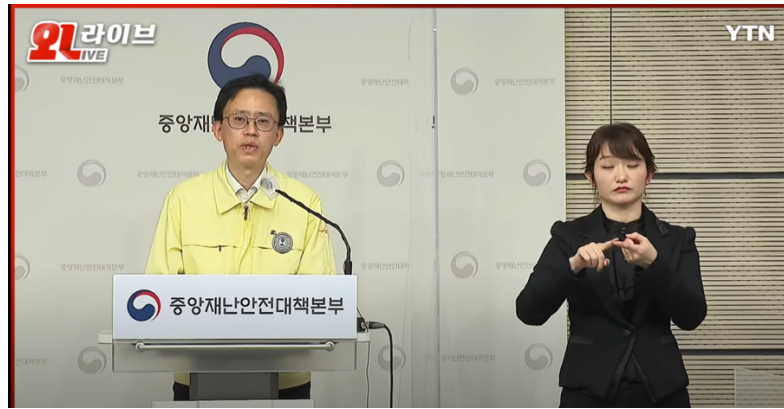
국문 초록

본 연구는 발화자와 수어 통역사의 동작을 합성하기 위한 딥러닝 모델을 구현하고 성능을 평가하는 것을 목표로 한다. 모델의 입력으로는 수어 통역사의 동작에서 추출한 스켈레톤 이미지와 발화자의 이미지를 함께 사용하며, 합성된 이미지의 품질을 높이기 위해 GAN의 학습 기법과 함께 다양한 Loss 함수를 사용하였다. 본 연구를 통해 얻어진 모델은 효율적으로 수어 동작을 합성할 수 있으며, 생성된 이미지의 품질을 더욱 향상시켰다. 본 연구의 결과물은 청각 장애인의 콘텐츠 접근성을 높이고 영상의 몰입도와 가시성을 높이는 데 활용할 수 있을 것이다.

Keywords: 수어 합성, 동작 합성, GAN, deep learning

1. Introduction

수어란 몸짓과 손짓에 의해 의사를 전달하는 언어로, 수어 단어를 손가락으로 표현하거나 표정이나 몸짓을 사용해 의미를 구성할 수 있다. 청각/언어 장애인들은 의사소통 시에 주로 수어를 비롯하여 구어, 필담 등의 방법을 사용하는데, 그 중에서도 수어의 사용 빈도가 가장 높다. 교육/의료-/직업/법률/방송 등 정확한 의사소통이 요구되는 상황에서 수어의 원활한 활용이 중요시되는데, 수어 사용자와 한국어 사용자의 의사소통 상황에서 메시지를 전달하기 위해서는 수어 통역사의 역할이 필수적이다. 그러나 비대면 확대 등의 영향으로 최근 동영상 콘텐츠 소비가 많아지는 흐름 속에서 청각장애인이 음성 언어 사용자와 동등한 콘텐츠 환경을 누리기는 쉽지 않다. 국립국어원의 ‘2017년 한국수어 사용 실태 조사’에 따르면 일상생활의 사용 비율과 달리 방송 및 인터넷을 접할 때 이용하는 서비스는 수어 통역이 36.9%로 자막(46.3%)보다 낮았으며, 15.1%는 서비스를 이용하지 못했다.



<그림 1 코로나19 재난방송의 Youtube 클립¹⁾>

<그림 1>은 코로나19 재난방송의 수어 통역 장면이다. 이와 같이 음성 언어를 사용하는 발화자(Speaker)와 수어 통역사가 병렬로 같은 크기로 배치되는 경우도 있으며, 방송에서의 수어통역은 대부분 수어 통역사 이미지가 오른쪽 하단에 작게 표시된다. 앞서 언급한 국립국어원의 연구에 따르면 수어 통역 서비스 이용 시에 내용이 이해되지 않는 원인은 ‘화면 크기가 작다’가 53%로 1위, 그리고 수어 속도가 빠르다가 14%로 2위였다. 우리는 이러한 조사 결과가 수어 통역사 이미지가 작게 병기될 수 밖에 없으며, 농인 이용자 입장에서 발화자와 통역사를 번갈아 보아야 하는 한계에서 기인한다고 판단하였다.

본 연구에서는 인공지능 기술을 이용하여 동영상의 원래 발화자가 직접 수어를 하는 것처럼 영상을 합성하는 방식으로 위 문제를 해결해보고자 하였다. 화자에 대한 집중도를 높이고, 표정과 같은 비언어적 표현을 손실하지 않기 위해서 수어 통역사의 이미지가 아닌 동작을 분석하여야 했으며, 손가락 하나하나의 모양 뿐 아니라 의미소로 작용하는 표정과 몸짓까지 놓치지 않아야 했다.

본 연구에서는 GAN(Generative Adversarial Network) 모델을 발전 시켜 pose content를 기존의 비디오에 합성하였다. 우리는 본 연구의 결과물이 수어 통역사가 우측 하단에 작은 영역으로 표시되는 기존의 방법과 달리 화자가 직접 수어를 하므로 화자의 비언어적 표현을 유지하여 몰입도를 높이고 전달력이 손실되지 않으며 가시성이 좋을 것으로 기대하고 있으며, 나아가 청각장애인의 영상 콘텐츠 소비 접근성을 높이는데 도움을 주고자 한다.

2. Related Work

이미지 생성 분야는 GAN의 등장 이후 급격하게 발전하였다. GAN은 데이터의 분포를 추론하고 생성하는 Generator와 생성된 데이터가 학습데이터일 확률을 예측하는 Discriminator로 이루어져 있다. Generator는 랜덤한 노이즈 z 로부터 생성된 데이터 y 가

¹ <https://www.youtube.com/watch?v=LhP-w9EFelY>

Discriminator를 속일 확률이 최대가 되도록, discriminator는 입력으로 들어온 데이터의 진위를 잘 구분할 수 있도록 두 행위자 최소 최대 게임(two-player minimax game) 전략을 사용하여 학습이 진행된다.[3]

노이즈 이외에 condition을 추가하여 원하는 이미지를 생성하기 위한 conditional GAN(cGAN)[5] 방법론은 이미지 편집, 비디오 생성, 질감 합성 등 다양한 분야에 활용되었다. 본 연구는 특정 이미지를 원하는 이미지로 바꾸기 위한 Image-to-Image Translation 분야에 초점을 맞추고 있다. 서로 다른 도메인 간의 이미지를 변환하기 위해 다양한 프레임워크가 등장하였다. pix2pix model은 입력과 출력 이미지 사이의 변환 함수를 cGAN을 통해 학습하는 방법론이다.[4] CycleGAN은 labeling이 되어 있지 않은 서로 다른 두 도메인 간의 이미지를 변환할 수 있는 방안을 소개하였다.[7] 이 외에도, 도메인의 개수만큼 generator가 필요하다는 CycleGAN의 단점을 보완하기 위해 StarGAN이 등장하였다.[2]

본 연구는 수어 동작을 하는 사람의 연속적인 이미지들(동영상)을 합성해야 하기 때문에 keypoint/skeleton guided image-to-image translation task에 해당한다. GestureGAN은 주어진 이미지에서 다른 정보는 그대로 유지한 채 손 모양만을 모양, 사이즈, 위치가 다른 또다른 손 모양으로 바꾸는 Hand Gesture-to-Gesture Translation 과제를 수행하였다. 이를 위해 GAN을 사용하여 object keypoint에 기반한 이미지 생성을 하고자 하였으며 channel pollution 문제를 해결하는 등 뛰어난 성능을 보여 주었다.[6]

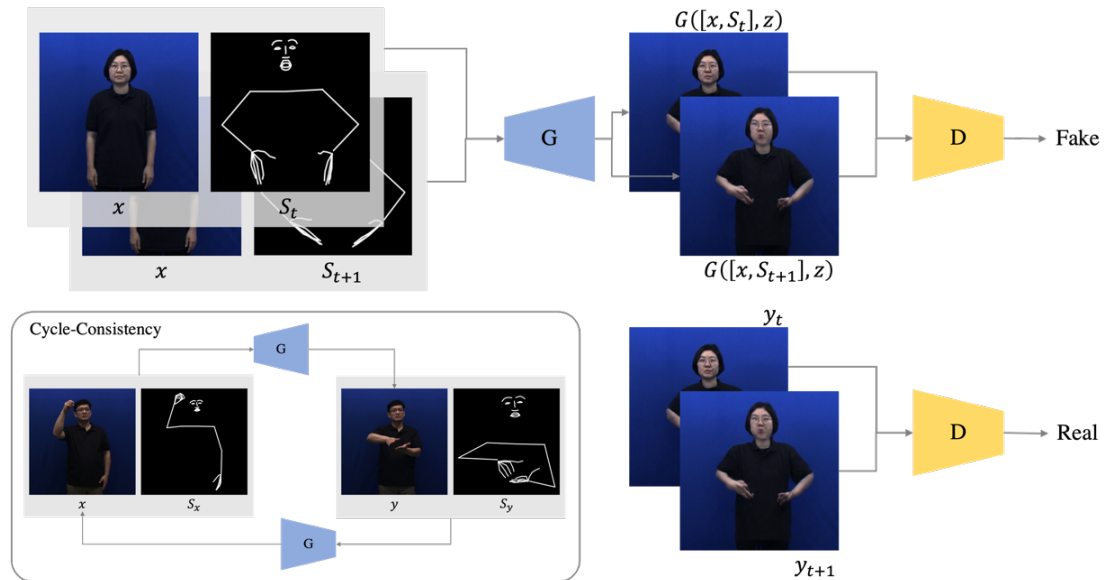
또한, 본 연구에서 다루는 문제는 하나의 이미지가 아니기 때문에 동영상에서의 pose transfer task를 다루었다. Everybody Dance Now는 source에서 target으로 동작을 자동으로 옮겨주는 video-to-video translation task를 해결하였다. motion을 표현하는 데이터로서 pose stick figure와 같은 keypoint-based pose를 사용하여 source 영상과 target 영상 사이의 mapping을 학습하였다. 특히 영상에서의 자연스러움을 보장하기 위하여 temporal smoothing을 제안하였다. Temporal smoothing은 frame-by-frame으로 두 영상 사이에서 motion을 옮길 때 서로 인접한 frame 간에 격차를 줄이는 방식으로 더욱 안정된 영상을 생성하는 방법이다. 예를 들어 t번째 frame 이미지를 생성할 때는, t번째 pose stick figure와 앞서 생성되었던 t-1번째 frame 이미지를 input으로 사용한다. 또한 따라서 discriminator가 생성된 이미지를 평가할 때도 실제와의 차이점뿐 아니라 앞서 생성된 가짜 이미지와의 연결성 또한 평가하게 된다. Temporal smoothing은 GAN의 모델에 인접한 frame간의 coherence를 강화하는 방향으로 변화를 주었다고 할 수 있다.[1]

3. Method

3.1. Model Architecture

본 연구의 목표는 수어 통역사와 발화자가 입력으로 들어 왔을 때, 발화자의 모습을

유지한 채로 수어 통역사의 동작을 따라하도록 하는 것이다. 이를 위해 크게 두 가지 단계를 거치게 된다. 첫번째 단계는 수어 통역사의 동작을 스켈레톤 모형으로 파악하는 것이다. 두번째 단계는 수어 통역사의 동작과 첫번째 단계에서 파악한 수어 통역사의 스켈레톤 모형을 합성하는 것이다. 첫번째 단계는 기존에 많이 연구된 pose estimation task로 Openpose 라이브러리를 이용하였다. 본 연구는 두 번째 단계에 초점을 맞추어 진행되었다. 본 연구에서 제안하는 수어 동작 합성 모델의 전체적인 구조는 <<그림 2>>와 같다.



<그림 2 수어 동작 합성 모델의 구조도>

Generator는 U-Net[4]의 구조를 차용하였다. U-Net은 모래시계 형태의 인코더와 디코더로 이루어져 있으며, 같은 계층의 레이어 사이에 skip connection을 추가하여 디테일한 정보를 유지하는 데에 장점을 가지고 있다. 발화자의 이미지와 수어 통역사의 동작을 의미하는 스켈레톤 이미지를 입력으로 받아서 발화자가 동작을 따라하는 이미지를 합성한다.

Discriminator는 PatchGAN[4]의 구조를 차용하였다. PatchGAN은 특정 크기의 patch 별로 진위 여부를 판단한 후, 결과들을 종합하여 최종 진위 여부를 판별한다. 연속적인 동작을 생성하기 위해서 discriminator는 연속적인 두 개의 이미지 프레임을 입력으로 사용한다.

3.2. Objective Function

3.2.1. Adversarial Loss

본 모델의 훈련 방식은 GAN을 기반으로 하고 있다. 일반적인 GAN의 목적함수는 (1)과 같다. [3]

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_y[\log D(y)] + \mathbb{E}_z[\log(1 - D(G(z)))] \quad (1)$$

cGAN[5]은 원하는 이미지를 생성하기 위해 추가적인 정보 x 를 입력으로 사용한다. 위와 유사하게 (2)의 목적함수를 사용하고, 훈련을 통해 얻은 해는 $G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D)$ 가 된다.

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))] \quad (2)$$

본 연구에서는 수어 통역사의 동작 정보를 담고 있는 스켈레톤 이미지를 condition으로 사용한다. OpenPose 라이브러리를 사용하여 수어 동작을 표현하기 위한 신체 주요 특징점의 좌표를 추출하였다. 얼굴에 대한 특징점 총 54개, 양손에 대한 특징점 각 25개, 몸통 및 팔에 대한 특징점 총 9개를 사용하였다. 수어는 손동작뿐만 아니라 표정도 의미 전달에 중요한 요소로 사용되기 때문에 손과 얼굴의 특징점들을 다량 사용하였다. 특징점의 위치 뿐만 아니라 각 특징점 사이의 연결 여부도 동작을 표현하는 데 중요한 요소이기 때문에 이를 스켈레톤 이미지 형태로 추출하였다. 즉, 검은색 배경 이미지 위에 각 특징점들을 흰 점으로 나타내고, 특징점 사이의 연결이 있는 경우 너비가 5pixel인 흰색 선으로 나타내었다.

스켈레톤 이미지 S 를 condition으로 사용하여 Generator는 이미지를 생성하고, Discriminator는 입력으로 받은 이미지가 condition에 대해 일치하는지를 판별한다. 따라서 목적함수를 (4)와 같이 재정의하였다.

$$\mathcal{L}_{S_y}(G, D, S_y) = \mathbb{E}_{[x, S_y], y}[\log D([x, S_y], y)] + \mathbb{E}_{[x, S_y], z_1}[\log(1 - D([x, S_y], G([x, S_y], z_1)))] \quad (3)$$

$$\mathcal{L}_S(G, D, S_x, S_y) = \mathcal{L}_{S_y}(G, D, S_y) + \mathcal{L}_{S_x}(G, D, S_x) \quad (4)$$

수어 동작의 연속성을 반영하기 위해 연속적인 이미지 프레임 사이의 관련성을 반영하기 위해 [1]과 같이 temporal smoothing을 추가하였다. Generator가 발화자의 이미지 x 와 t 번째 프레임의 스켈레톤 이미지 S_t 에 대해 $G([x, S_t], z_t)$ 와 $t+1$ 번째 프레임의 스켈레톤 이미지에 대해 $G([x, S_{t+1}], z_{t+1})$ 를 생성한다. Discriminator는 $(x, S_t, S_{t+1}, y_t, y_{t+1})$ 는 참으로, $(x, S_t, S_{t+1}, G([x, S_t], z), G([x, S_{t+1}], z))$ 는 거짓으로 판별한다. 따라서 (5)와 같은 목적함수를 사용한다.

$$\begin{aligned} \mathcal{L}_{smooth}(G, D, S_t, S_{t+1}) = & \mathbb{E}_{[x, S], y}[\log D(x, S_t, S_{t+1}, y_t, y_{t+1})] \\ & + \mathbb{E}_{[x, S], z}[\log(1 - D(x, S_t, S_{t+1}, G([x, S_t], z), G([x, S_{t+1}], z)))] \end{aligned} \quad (5)$$

3.2.2. Overall Loss

이 외에도 생성된 이미지의 품질을 높이기 위해 [6]에서 제안한 것과 같이 Color Loss와 Cycle-Consistency Loss, Identity Preserving Loss를 추가하였다. 기존에는 pixel 단위에서 생성된 이미지와 실제 이미지의 L1 혹은 L2 loss를 비교한 것과 달리 channel pollution 문제를 해결하기 위해 (7)과 같이 R, G, B 채널 각각에서 L1 혹은 L2 loss를 비교하는 color loss를 사용한다.

$$\mathcal{L}_{Color_{\{1,2\}}^c}(G, S_x, S_y) = \mathbb{E}_{[x^c, S_y], y^c, z} [\|y^c - G([x^c, S_y], z)\|_{\{1,2\}}] + \mathbb{E}_{[x^c, S_x], x^c, z} [\|x^c - G([y^c, S_x], z)\|_{\{1,2\}}] \quad (6)$$

$$\mathcal{L}_{Color_{\{1,2\}}}(G, S_x, S_y) = \mathcal{L}_{Color_{\{1,2\}}^R} + \mathcal{L}_{Color_{\{1,2\}}^G} + \mathcal{L}_{Color_{\{1,2\}}^B} \quad (7)$$

CycleGAN[7]에서 사용한 cycle-consistency loss는 target domain과 source domain 사이에서 정확히 매칭되는 훈련 데이터 쌍이 없더라도 훈련이 가능하다. 그러나 변환시키고자 하는 domain의 개수만큼 generator가 필요하다는 단점이 있다. StarGAN[2]에서는 이를 개선하여 같은 generator를 여러번 사용하여 입력으로 들어온 이미지를 다양한 domain으로 변환할 수 있다. 본 연구에서는 이와 유사하게 하나의 generator를 사용하여 다양한 수어 동작을 수행할 수 있도록 변환하기 위해 (8)과 같이 cycle-consistency loss를 사용한다.

$$\mathcal{L}_{cyc}(G, S_x, S_y) = \mathbb{E}_{x, y, S_x, S_y, z} [\|x - G(G([x, S_y], z), S_x, z)\|_1] + \mathbb{E}_{x, y, S_x, S_y, z} [\|y - G(G([y, S_x], z), S_y, z)\|_1] \quad (8)$$

생성된 이미지에서 원래 발화자의 정보를 유지하기 위해서 (9)와 같이 identity preserving loss를 사용한다. 원래 화자의 얼굴을 인식하기 위해 VGG network F를 사용하여 발화자의 외모와 관련된 feature를 추출하여 생성된 이미지에서의 추출한 feature와 비교한다.

$$\mathcal{L}_{identity}(G, S_x, S_y) = \mathbb{E}_{y, S_x, z} [\|F(x) - F(G([x, S_y], z))\|_1] + \mathbb{E}_{x, S_y, z} [\|F(y) - F(G([y, S_x], z))\|_1] \quad (9)$$

따라서 최종적인 목적함수는 (10)과 같다.

$$\mathcal{L} = \mathcal{L}_{smooth}(G, D, S_t, S_{t+1}) + \lambda_1 \mathcal{L}_{Color_{\{1,2\}}}(G, S_x, S_y) + \lambda_2 \mathcal{L}_{cyc}(G, S_x, S_y) + \lambda_3 \mathcal{L}_{identity}(G, S_x, S_y) \quad (10)$$

4. Experiments

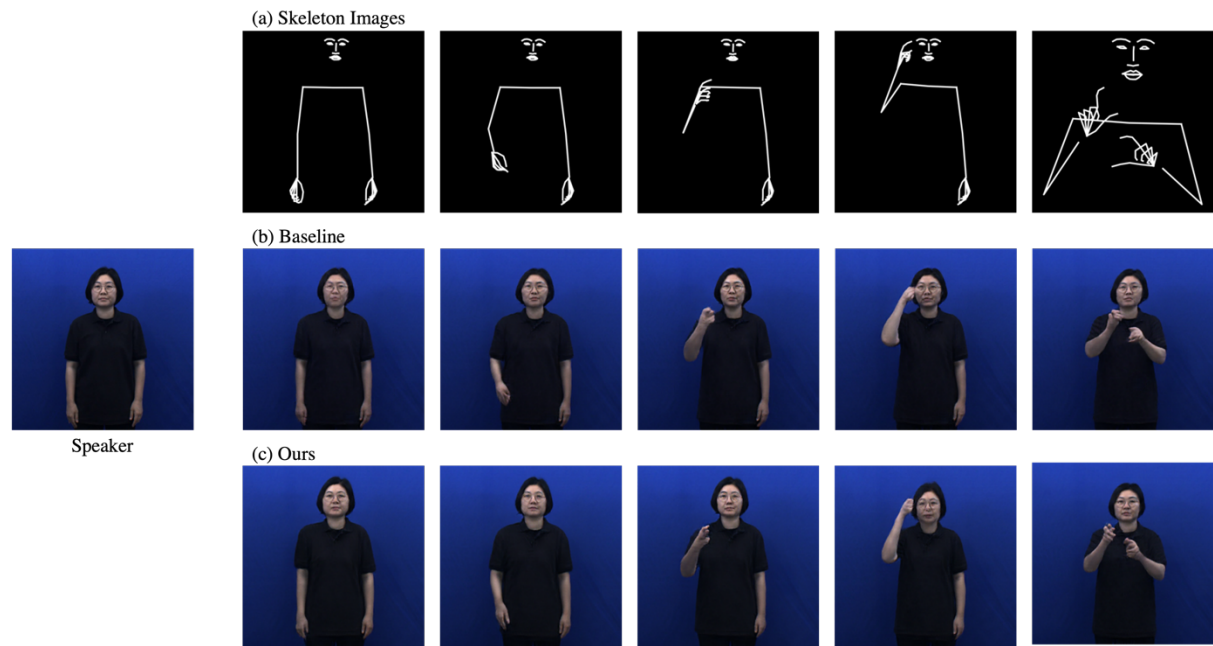
4.1. Experimental Setup

실험에 사용한 데이터셋은 AI Hub에서 제공하는 수어 영상 데이터셋이다.[8] 영상의 해상도는 $1920 \times 1080p$ 이고, 30 fps이다. 블루스크린을 배경으로 두고 검은색 옷을 입은 피사체로부터 2m 떨어진 위치에서 정면으로 촬영한 영상만을 사용하였다. 이 중 5인의 영상을 사용하였다. 정자세를 제외하기 위해 모든 영상에서 가운데 2초만을 추출하였다. 각 피사체별로 10,000개의 이미지 프레임을 랜덤하게 선택하여 총 50,000개의 이미지로 훈련 데이터셋을 구성하였다. 또한, 각 피사체별로 2,000개의 이미지 프레임을 훈련 데이터셋과 겹치지 않게 랜덤하게 선택하여 총 10,000개의 이미지로 테스트 데이터셋을 구성하였다.

hyperparameter는 실험적으로 설정하였다. optimizer는 Adam을 사용하였으며, 초기의 learning rate는 0.002, $\beta_1 = 0.5, \beta_2 = 0.99$ 이다. $\lambda_1 = 100, \lambda_2 = 10$ 으로 두었으며, λ_3 는 0.1에서 시작하여 0.5까지 증가하였다. Generator의 파라미터 수는 11.388 M개이고,

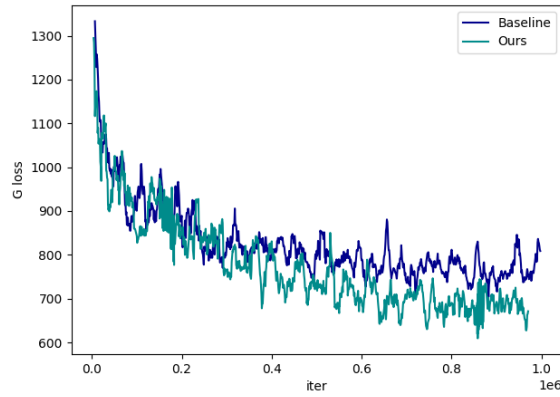
Discriminator는 스켈레톤 이미지의 입력 여부에 따라 두 개를 만들었고, 파라미터 수는 각각 2.768M개, 2.774M개이다. Pytorch를 사용하여 구현하였으며, 16G 메모리를 가진 NVIDIA V100 GPU 한 개로 20 epoch 훈련을 진행하였다.

4.2. Qualitative & Quantitative Results



<그림 3 "고민"을 의미하는 수어 동작의 합성 결과>

<그림 3>은 훈련 데이터셋에 존재하지 않는 발화자에 대해 랜덤하게 선택된 수어 동작 스켈레톤 이미지를 합성한 결과이다. 첫번째 행은 10프레임마다 추출한 타겟 스켈레톤 이미지이고, 두번째 행은 GestureGAN을 사용한 결과이고, 마지막 행이 본 연구의 결과이다. baseline과 비교하였을 때, 발화자의 얼굴 특징 보존, 손가락의 묘사 측면에서 디테일을 잘 표현하고 있음을 확인할 수 있다. 또한, <그림 4>는 훈련이 진행됨에 따라 Generator의 loss 값이 수렴하는 양상을 시각화한 그래프이다. baseline과 비교하였을 때, 본 연구에서 제안하는 모델이 추가적인 프레임 정보를 제공하기 때문에 수렴 속도가 소폭 빨라진 것을 확인할 수 있다.



<그림 4 Generator loss 의 수렴 그래프>

5. Conclusion

본 연구에서는 발화자와 수어 통역사의 동작을 합성하기 위한 딥러닝 모델을 구현하고 성능을 평가하였다. 이를 위해 Openpose 라이브러리를 이용하여 수어 통역사의 동작을 스켈레톤 이미지로 추출하고, 추출한 스켈레톤 이미지를 발화자의 이미지와 함께 모델의 입력으로 사용하였다. 합성된 이미지의 품질을 높이기 위해 다양한 loss 함수를 사용하고 GAN의 학습 기법을 사용하였다. 그 결과 기존의 연구들보다 효율적으로 수어 동작을 합성할 수 있었으며, 생성된 이미지의 품질도 향상되었다.

본 연구의 결과를 활용하게 되면 화자가 직접 수어 동작을 하게 됨으로써 영상의 몰입도와 가시성을 증가시킨다. 이 외에도 토론과 같이 다수의 화자가 동시에 대화하여 1인의 수어 통역사가 번역하기 힘든 상황을 해결하여 청각 장애인의 콘텐츠 접근성을 높일 수 있다는 의의가 있다. 그러나 컴퓨팅 자원의 한계로 사용한 훈련 데이터의 양이나 훈련 시간에 있어서 제약이 있었다. 훈련 데이터의 양을 늘이고 충분한 컴퓨팅 파워가 뒷받침된다면 뛰어난 일반화 성능을 자랑하는 모델을 얻을 수 있을 것으로 생각된다. 본 연구의 결과는 음성이나 텍스트로부터 수어 동작 스켈레톤 이미지를 생성할 수 있는 연구가 이어진다면 그 활용성이 극대화될 것이라 기대한다.

6. References

- [1] Chan, Caroline, et al., “Everybody dance now” , Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
- [2] Choi, Yunjey, et al., “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation” , Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.
- [3] Goodfellow, Ian, et al., “Generative adversarial nets” , Advances in neural

- information processing systems, pp. 2672 – 2680, 2014.
- [4] Isola, Phillip, et al., “Image-to-image translation with conditional adversarial networks” , Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.
 - [5] Mirza, Mehdi, and Simon Osindero, “Conditional generative adversarial nets” , arXiv preprint arXiv:1411.1784, 2014.
 - [6] Tang, Hao, et al., “Gesturegan for hand gesture-to-gesture translation in the wild” , Proceedings of the 26th ACM international conference on Multimedia, 2018.
 - [7] Zhu, Jun-Yan, et al., “Unpaired image-to-image translation using cycle-consistent adversarial networks” , Proceedings of the IEEE international conference on computer vision, 2017.
 - [8] <https://aihub.or.kr/aidata/7965>