Hand to Hand

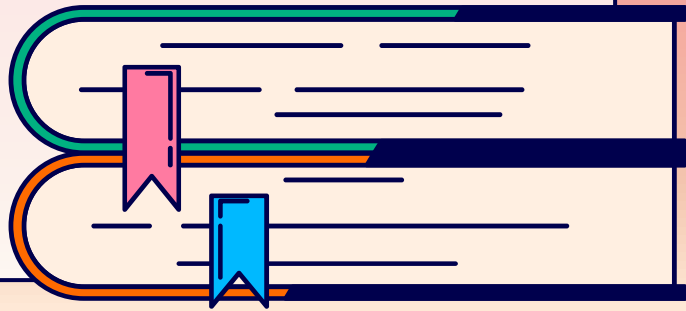
청각장애인의 즐거운 외출을 위한 수어 생성 모델

** members **

11기 이도연　13기 고유경　13기 김미성　13기 정민준
13기 조혜원　14기 강의정　14기 서아라　14기 이정은

Contents

주제 선정
배경



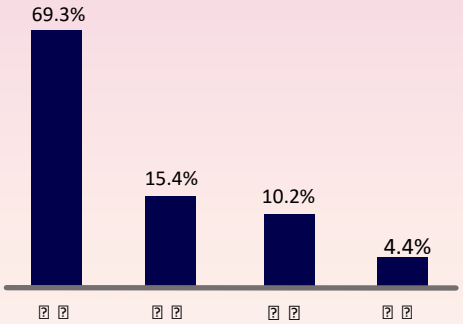
주제 선정

✦ 배경

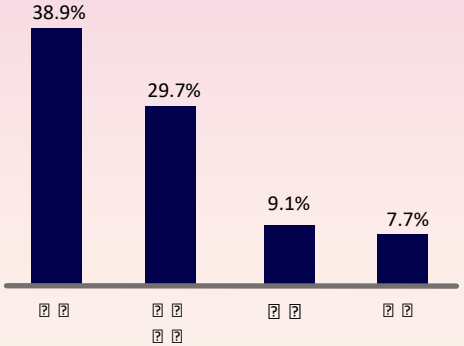
- 2016□ □□□□□□□□ □□□□ □□□ □□□ □□□□□□ □ □□ □□ □ □□ □□□ □□□□ □□□□ □□□ ,□□ □
□□□ □□□ □□□ □□□□□□ □□ □□ □□

- □□□□ □□□ □ □□ □□□□□□□ □□□□□□□ □□ □□□ □□ □□□□□□ □□□ □□□□ □□□□ □□□ □□□□□ □
□ □□□□ □□□□ 160□ □ □□

□□ □□□□ □□□□ □□



□□ □□□ □□□□ □□□□



□□□□□□ □□ □□□ □□□□□□ □□ ,□□ -□□□□ -□□ -□□ □□□ □□ □□□ □□□□□□ □□

주제 선정

✦ 배경

□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

□ □ □ □ □ □

소리가 들리지 않아 겪는
일상생활 속 위험에 대한 에피소드

□ □ <□ □ □ □ □ □ □ >

주제 선정

배경

[illegible]

[illegible]

...

[illegible]

*[illegible] : CBS [illegible], 2017

주제 선정

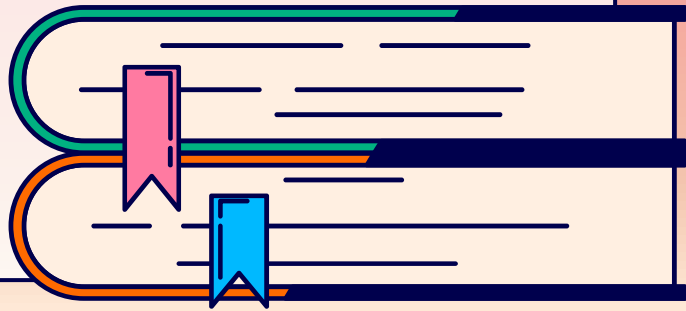
✦ 배경

‘ ’

…

* : CBS , 2017

데이터 수집 및 전처리



데이터 수집 및 전처리

□ □ □ □

: □
□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

□ □ □ □ □

: 문장 및 단어를 수어로 표현한 촬영 원본 데이터 제공
: 교통 및 지하철에서 자주 사용되는 대표 문장 26개 시나리오

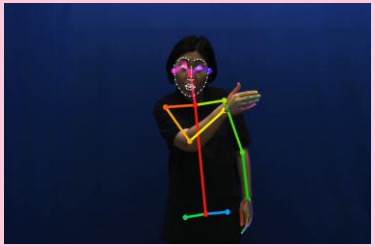
데이터

✦ 구조

REAL

video
word_keypoint
word_morpheme
sen_keypoint
sen_morpheme
data_info

- video → F,L,R,U,D 5 각도에 대한 촬영
- word_keypoint → Pose (25), Face (68), Left hand (21), Right hand (21)
- word_morpheme → Start , End , Attributes => meta data



데이터

✦ 구조

- word_morpheme

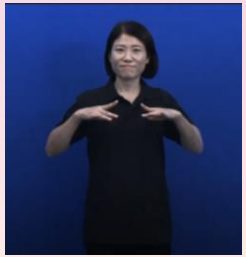

```
{
    "metaData": {
        "url": "https://blackolivevideo.blob.core.windows.net/sign-language/1123_gdchoi0992/NIA_SL_SEN0514_SYN05_D.mp4",
        "name": "NIA_SL_SEN0514_SYN05_D.mp4",
        "duration": 3.233,
        "exportedOn": "2020/11/26"
    },
    "data": [
        {
            "start": 1.113,
            "end": 1.708,
            "attributes": [
                {
                    "name": "에어컨"
                }
            ]
        },
```

데이터 전처리

과정

1. 영상 데이터 학습을 위한 이미지 추출

- 영상과 keypoint 정보를 mapping 시키기 위해 30fps로 영상을 잘라 이미지를 추출

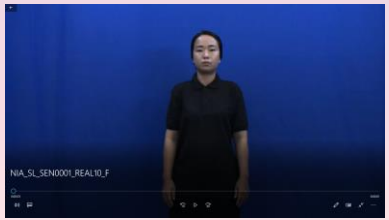


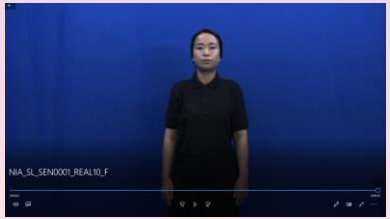
데이터 전처리

✦ 과정

2. 영상의 앞 뒤에 반복되어 나타나는 정자세 제거

- 영상으로부터 추출된 이미지에서 아래 예시와 같은 다량의 정자세 이미지 발견
- 따라서 파일의 start와 end 정보를 이용하여 영상 앞 뒤에 반복되어 나타나는 정자세 이미지 제거





3. Confidence 값 제거

- x, y, keypoint 의 정확도를 나타내는 confidence 값은 불필요 하다고 생각되어 제거

```
NIA_SL_SEN0330_REAL10_F_0046_keypoints, 0.5260547, 0.24576735, 0.52297366, 0.40648782, 0.45401025, 0.406592, 0.40645018,
0.57825255, 0.46017823, 0.44752344, 0.59503907, 0.40389758, 0.61956054, 0.62188363, 0.6302246, 0.8153117, 0.5230078, 0.8534462,
0.47245264, 0.8561875, 0.57510257, 0.8507162, 0.50916016, 0.22118576, 0.54291016, 0.21576823, 0.48618653, 0.2485243, 0.55975586,
0.24851823, 0.4875288, 0.22562501, 0.48869628, 0.24292448, 0.4902539, 0.26022398, 0.4914209, 0.27752343, 0.49453613, 0.2941302,
0.500376, 0.3058941, 0.5073779, 0.31696615, 0.516333, 0.32319358, 0.5256738, 0.3259618, 0.5342383, 0.32250175, 0.54280275,
```

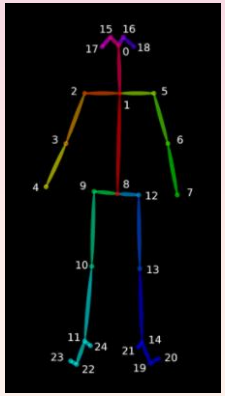
데이터 전처리

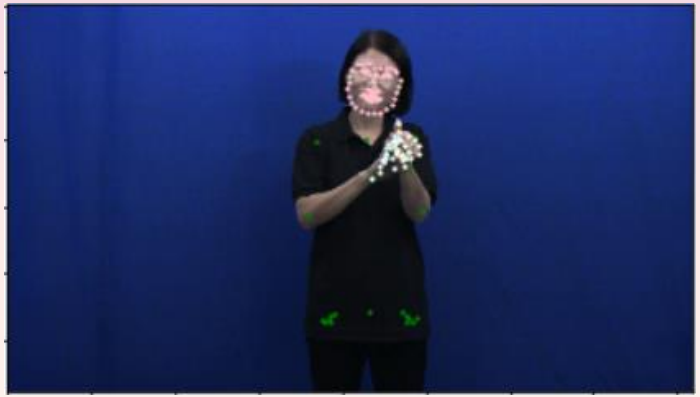
✦ 과정

4. 0으로 표기된 confidence 값의 의미와 처리

- Keypoint json 파일 정보 중 confidence 값이 0으로 된 값 발견
- 제공받은 pose_data는 25 개의 특징점을 나타내는데 실제 영상속에는 사람의 다리, 발 부분은 존재하지 않음
- 즉, 이 부분에 해당 하는 pose의 confidence 값이 0으로 표시되었다고 볼 수 있다.
- 따라서 하반신 부분에 해당하는 x,y pose keypoint 제거


```
959.834,
0.0,
1217.84,
949.834,
0.0,
1187.84,
959.834,
0.0,
943.797,
962.929,
0.0,
933.797,
952.929,
0.0,
963.797,
962.929,
0.0
```

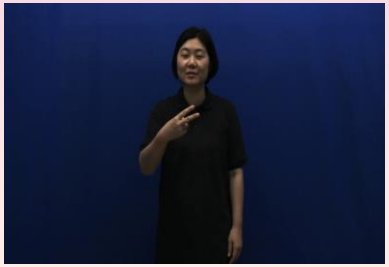


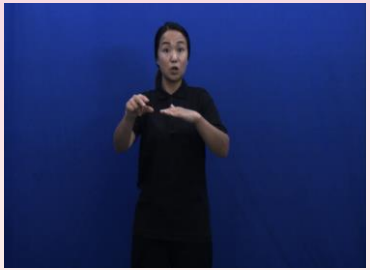
데이터 전처리 과정

5. Keypoint Normalization

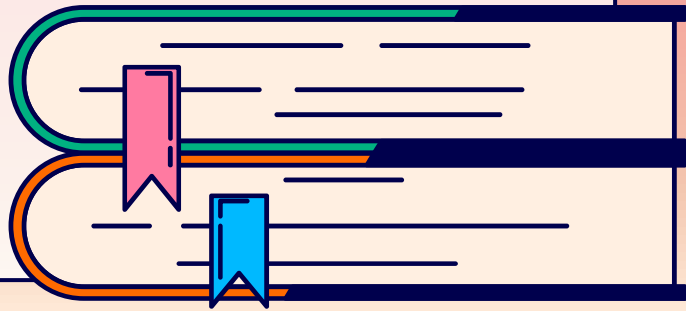
- Keypoint값을 0-1사이로 만드는 min max normalization을 함
- 또한 해당 데이터는 같은 단어/문장에 대해 여러 사람들로 구성된 dataset이므로 사람별로 영상의 특성이나,
- pose scale 정도가 다를 수 있어 사람에 따른 차이를 normalize 필요
- 여기선 사람을 구분하는 one-hot 추가

INDEX	형태소
NIA_SL_WORD0001_REAL02	고민
NIA_SL_WORD0001_REAL05	고민



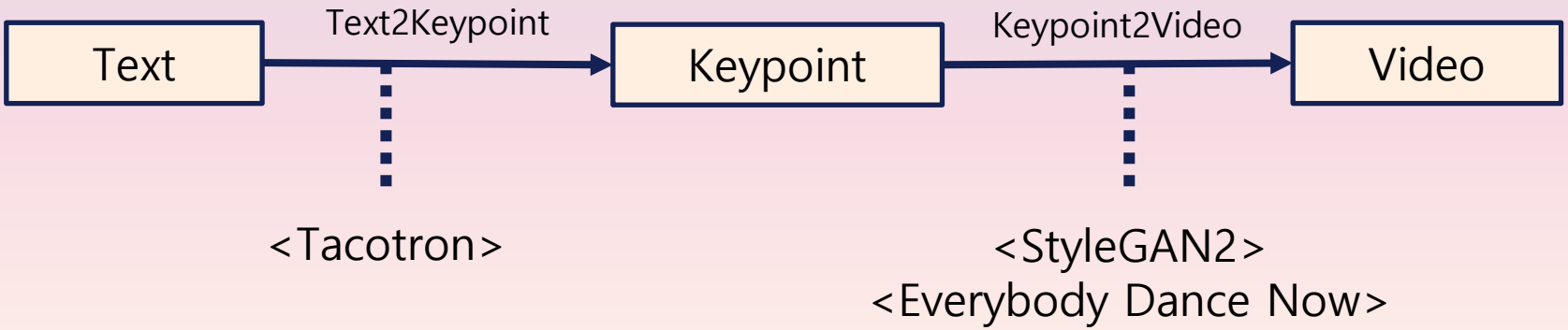


Model



✦ Model

주어진 텍스트를 수어 영상으로 번역하는 모델 구조를 두단계로 제안



✦ Model (Text2Keypoint – Tacotron)

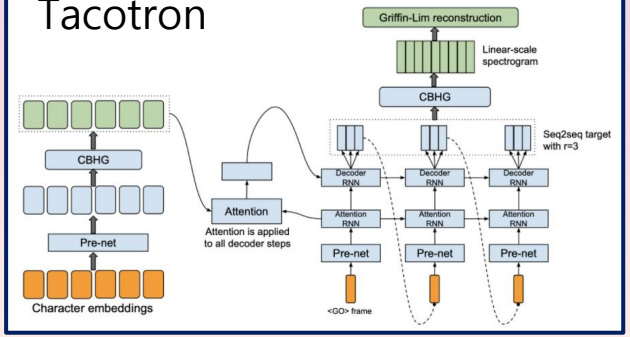
Input

Text/형태소

Ex. "지하철 환풍구에 가방이 빠졌어요!"
-> 지하철 환풍구 가방 빠지다

Model

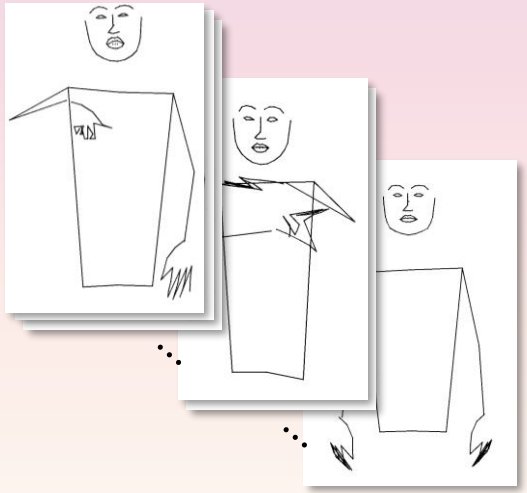
Tacotron



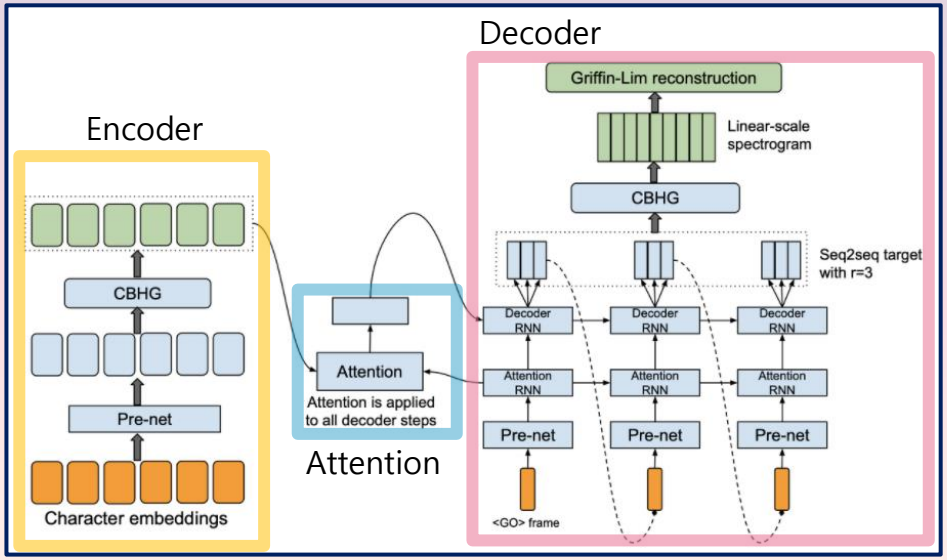
Output

Keypoint sequences

시퀀스 별 키포인트 값 254개
(오른손 21, 왼손 21, 포즈 25, 얼굴 68)

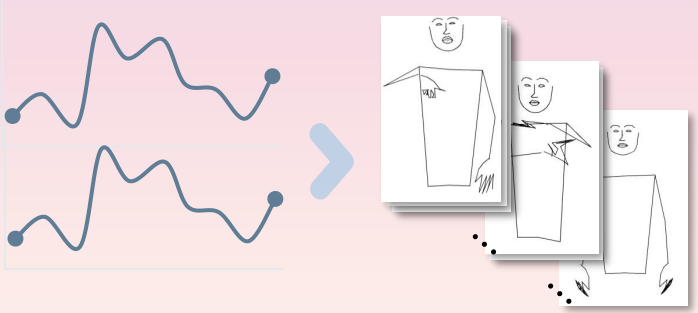


Model (Text2Keypoint – Tacotron)



Attention을 적용한
Encoder – Decoder 구조의 seq2seq 모델

< Q. 왜 음성 모델을 사용했나요?>



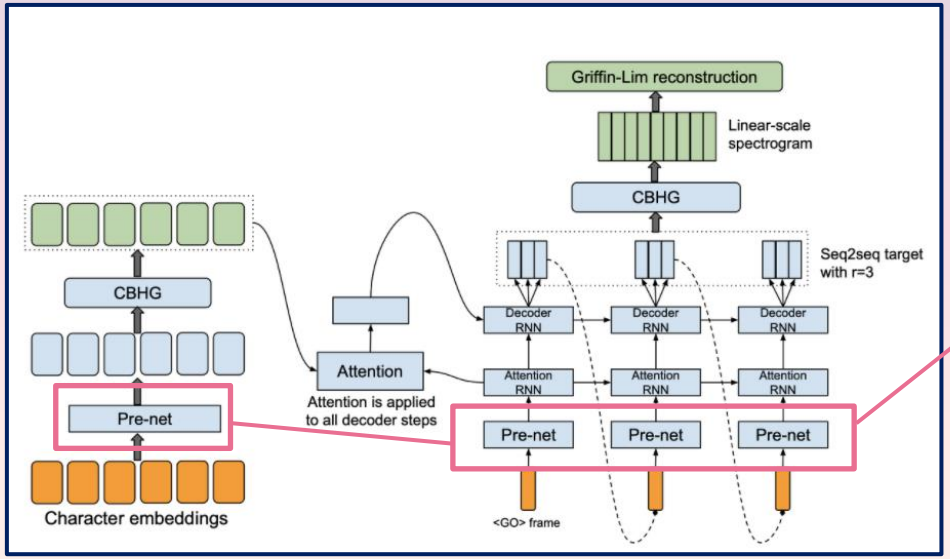
음성 신호 시퀀스

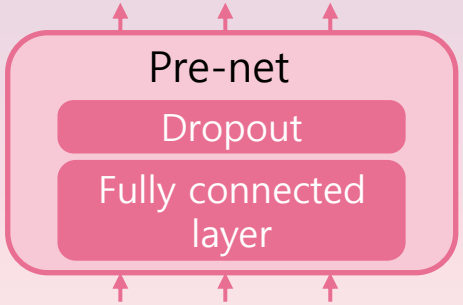
수어 keypoint 시퀀스

Text를 입력으로 받아 음성신호를 반환하는 TTS(Text to Speech) 모델

✦ Model (Text2Keypoint – Tacotron)

1. Pre-Net

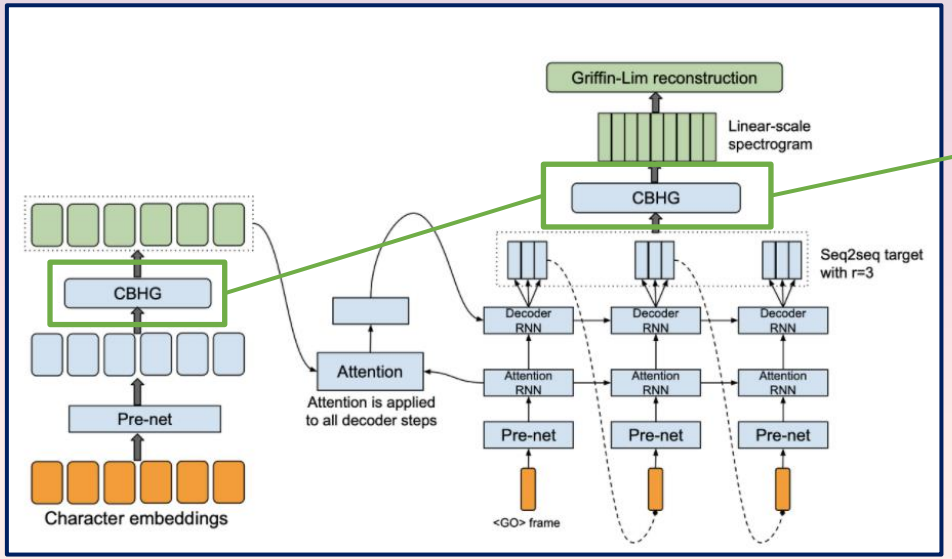


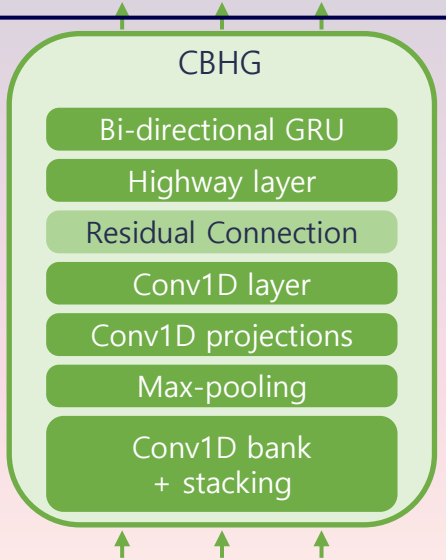


- encoder와 decoder 앞단에 위치
- fully connected layer + dropout으로 구성
- Overfitting을 방지하기 위한 구조

Model (Text2Keypoint – Tacotron)

2. CBHG



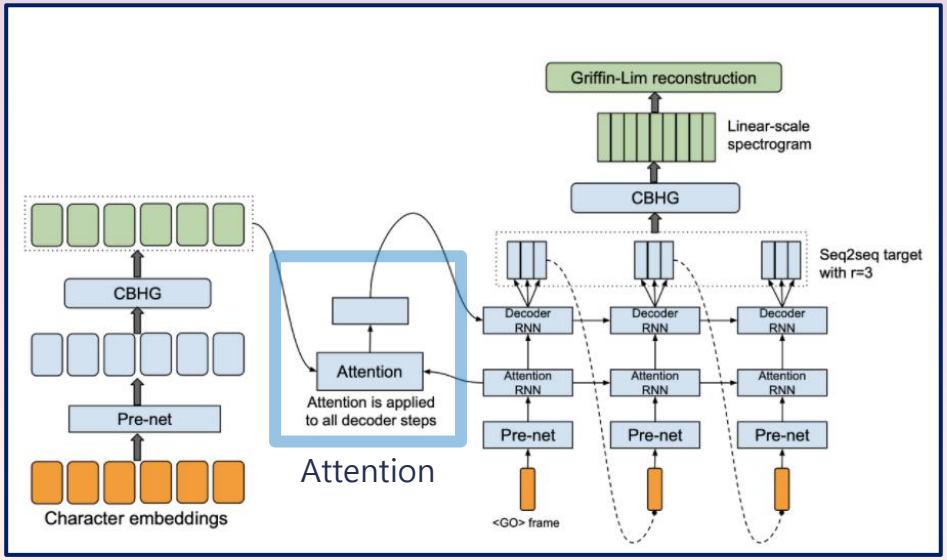


[Encoder] overfitting 방지, mispronunciations(오발음) 감소

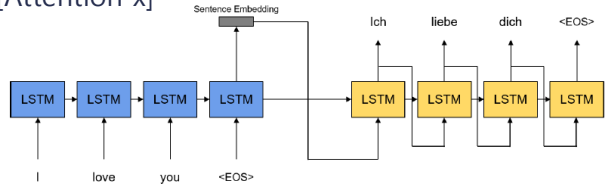
[Decoder] 개별 프레임의 양쪽 정보를 가지고 있어 prediction error를 수정하는데 용이

✦ Model (Text2Keypoint – Tacotron)

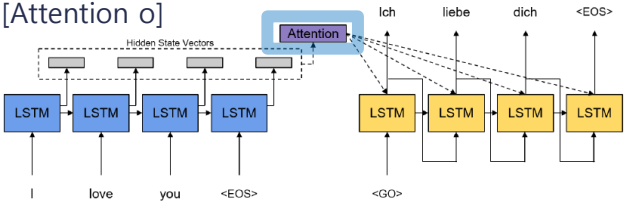
3. Attention



[Attention x]



[Attention o]



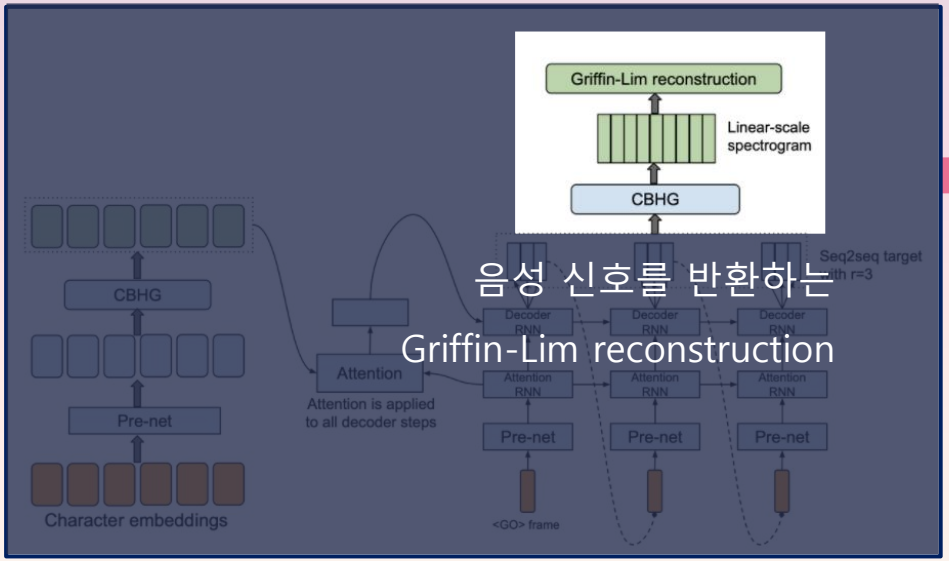
입력 시퀀스가 길어지면 기울기 소실, 정보 손실 문제 발생

Attention: 디코더에서 출력 시퀀스를 예측하는 매 시점마다 인코더에서 전체 입력 문장을 다시 한 번 참고하는 방식
(입력 문장을 전부 동일한 비율로 참고하지 않고 해당 시점 단어와 연관있는 부분에 attention을 주어 집중 참고)

Model (Text2Keypoint – Tacotron)

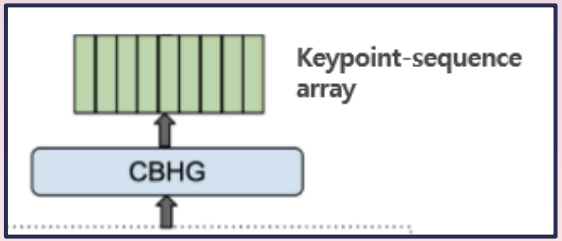
4. 수정사항

1) Input: Text, Output: Keypoint sequence



음성 신호를 반환하는

Griffin-Lim reconstruction



output의 형태가

linear-scale spectrogram에서

keypoint sequence array가 되도록

수정

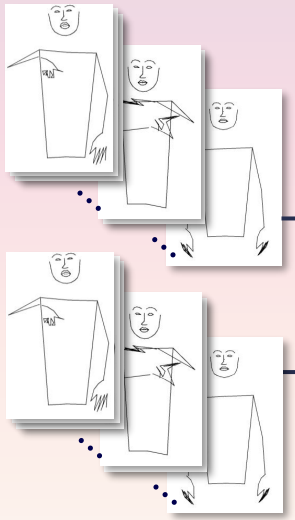
Model (Text2Keypoint – Tacotron)

4. 수정사항

2) Evaluation metric: DTW Score

Real

Pred



Dynamic Time Warping (DTW)

두 시계열 데이터 간

Minimum Distance Matrix(최소 거리 행렬)을 통해 유사도를 측정.

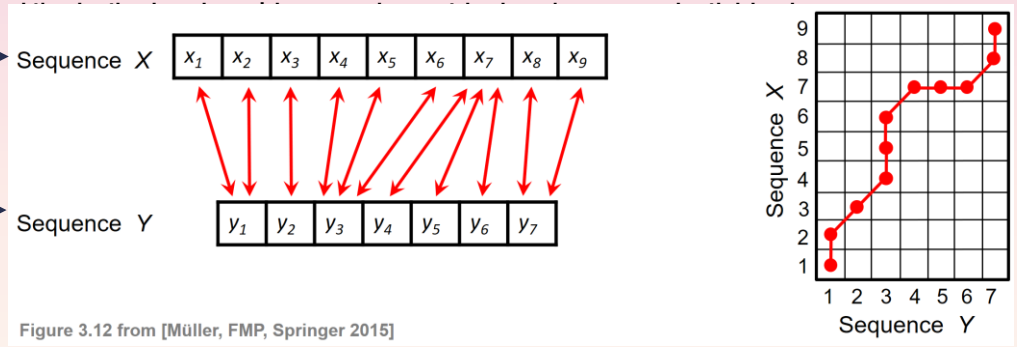
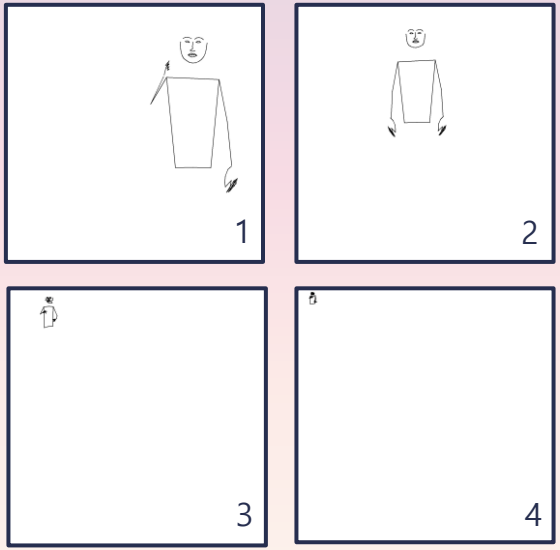


Figure 3.12 from [Müller, FMP, Springer 2015]

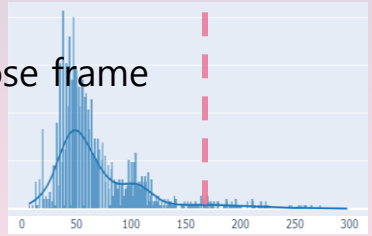
Model (Text2Keypoint – Tacotron)

4. 수정사항

3) CE/Masking or padding with last, static pose frame



후반 프레임으로 갈 수록 작아지는 형상
(키포인트 값이 0에 수렴)



문장 별 프레임 개수

[원인] Zero-Padding의 영향

170개 이상: 전체 학습 데이터의 2%

-> 문장이 짧을 수록(0으로 패딩된 시퀀스가 많을수록) 키포인트값 감소

Zero-padding의 영향을 최소화하기 위해 도입한 Masking Layer

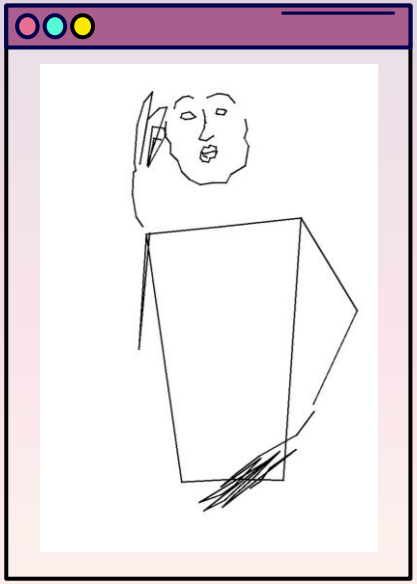
Counter Value(0~1)

Frame 1 (시작) [0.49 0.33 0.51 0.49 0.33 ... 0.23 0.78 0.0],
Frame 2 [0.49 0.33 0.51 0.49 0.33 ... 0.23 0.78 0.2],
Frame 3 [0.49 0.33 0.51 0.49 0.33 ... 0.23 0.78 0.4],
⋮ ⋮
Frame 28 (끝) [0.49 0.33 0.51 0.49 0.33 ... 0.23 0.78 1.0],
Frame 29 [0 0 0 0 0 0 ... 0 0 0],
⋮ ⋮
Frame 170 [0 0 0 0 0 0 ... 0 0 0]

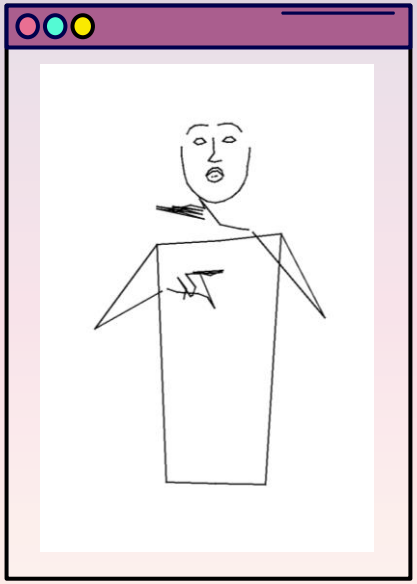
Zero-padding된 프레임은 loss 연산 등 학습 과정에서 제외하도록 Masking

Result (Text2Keypoint – Tacotron)

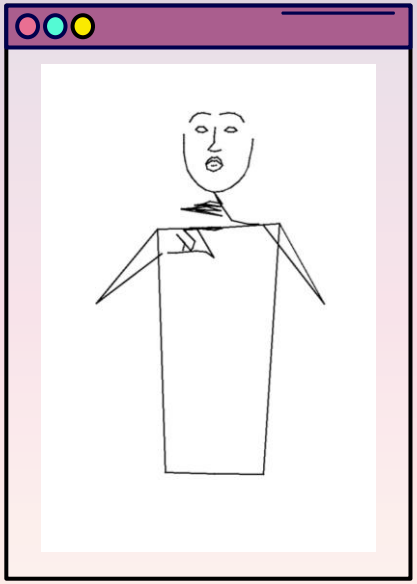
Epoch 1000



Epoch 2000

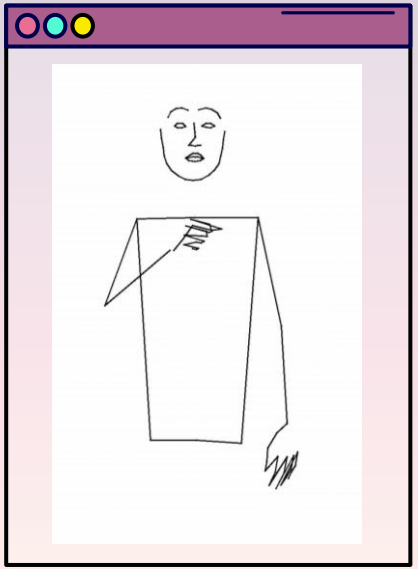


Epoch 5000

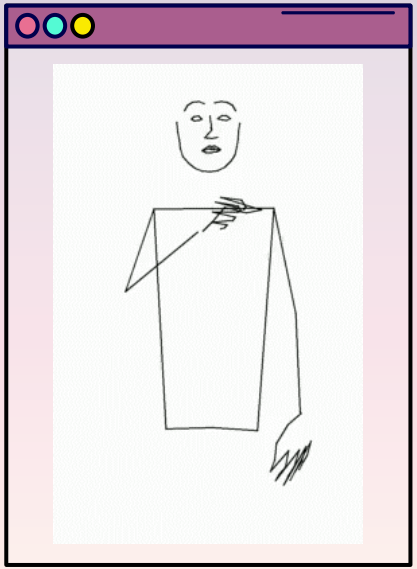


Epoch 10000

✦ Result (Text2Keypoint – Tacotron)



<Original>

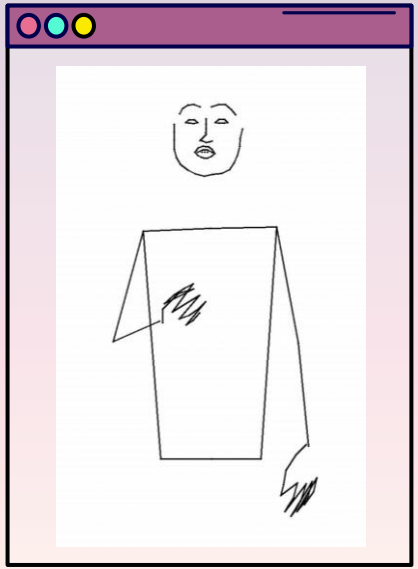


<Result>

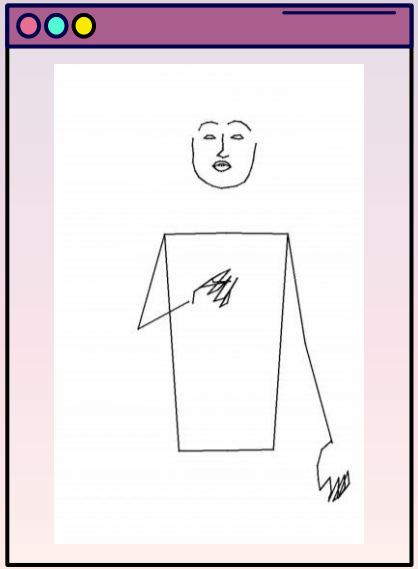
명동에서 내리다

(명동 내리다)

DTW Score: 3.84



<Original>



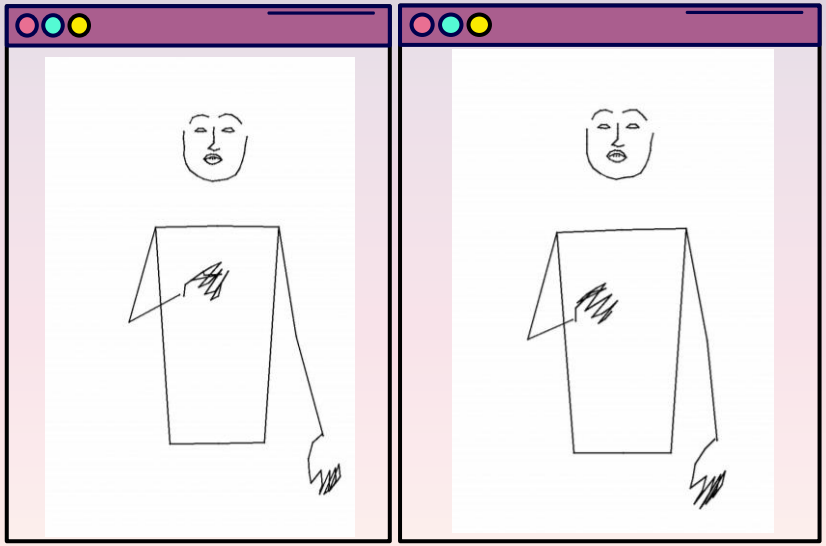
<Result>

서울역까지 가려면 지하철 몇호선으로 환승 해야하나요?

(서울 기차 가다 목적 지하철 바꾸다 몇호)

DTW Score: 10.91

✦ Result (Text2Keypoint – Tacotron)



<Tacotron>

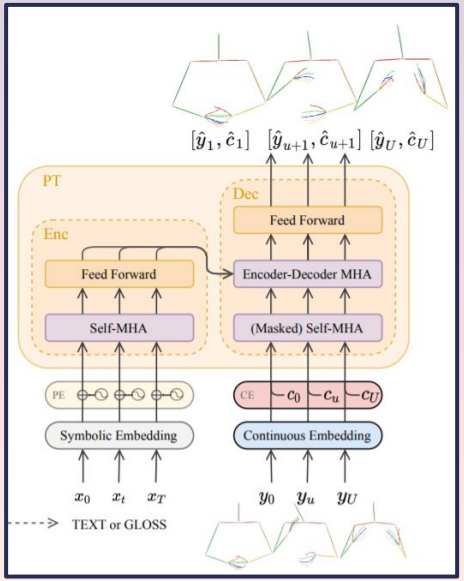
DTW Score: 10.91

<Transformer>

DTW Score: 1.20

서울역까지 가려면 지하철 몇호선으로 환승 해야하나요?

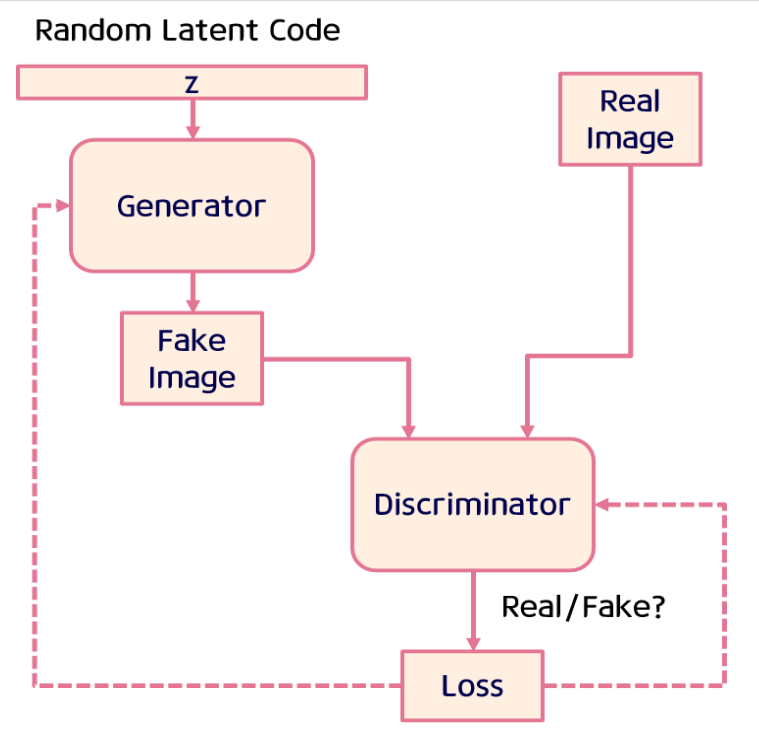
(서울 기차 가다 목적 지하철 바꾸다 몇호)



(2020, ECCV) Progressive Transformers for End-to-End Sign Language Production

https://github.com/BenSaunders27/ProgressiveTransformersSLP

Model (Keypoint2Video –StyleGAN2)



GAN

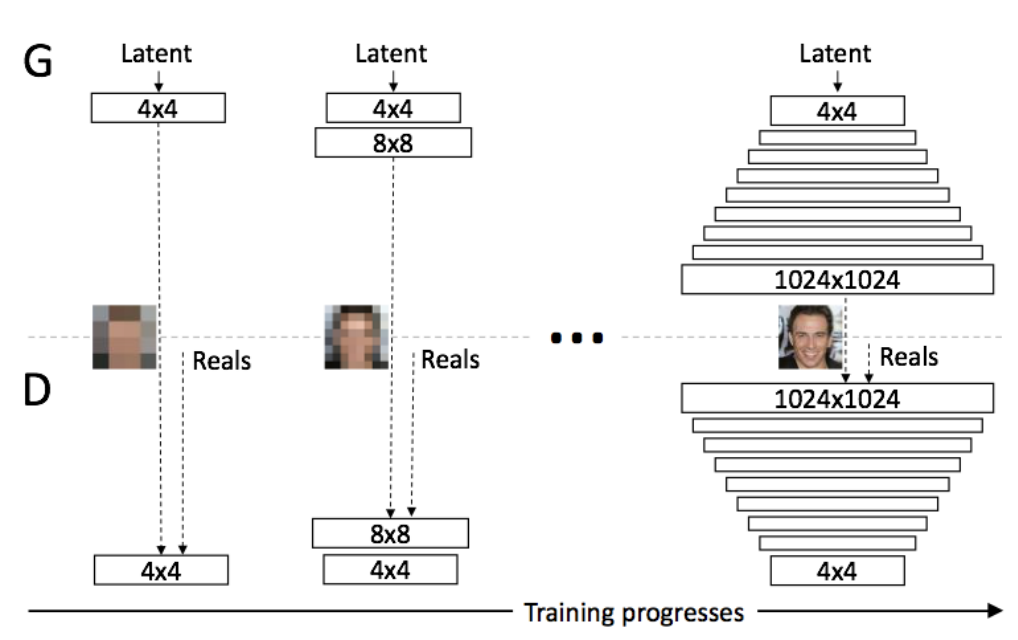
Generator와 Discriminator 사이의
경쟁적 학습을 이용한 생성모델

- Generator: 실제 이미지를 학습해 거짓 이미지 생성

- Discriminator: 실제 이미지와 Generator가 만들어낸 거짓 이미지를 판별

2014년 GAN이 처음 등장한 이후
안정적인 학습, 고화질의 사실적인 이미지 생성,
아웃풋을 컨트롤해 다양한 이미지 생성 등
다양한 과제들에 대한 연구가 진행되고 있다.

Model (Keypoint2Video –StyleGAN2)



PGGAN

- StyleGAN의 baseline이 되는 모델

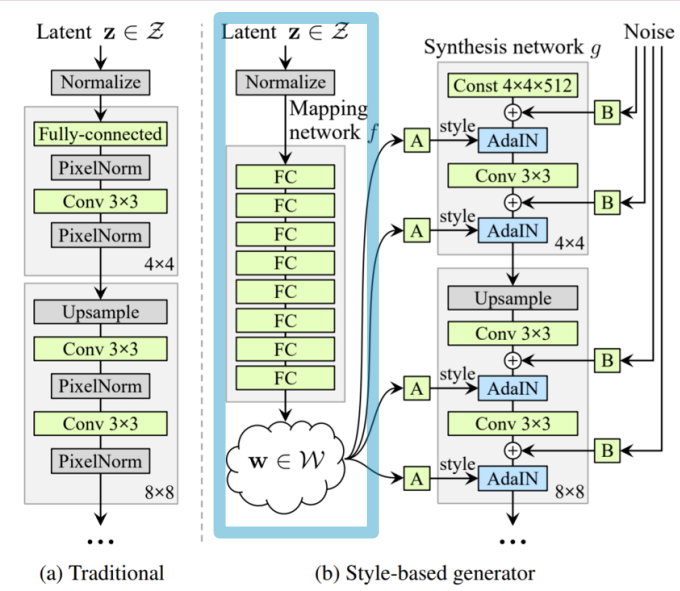
- Generator와 Discriminator의 점진적인 학습이 특징

고품질의 이미지를 생성하지만, 대부분의 모델과 같이 생성된 이미지의 구체적인 특징을 컨트롤하는 능력은 제한적이다. 특징들은 entangled 되어 있기에 input의 조정은 동시에 여러 특징에 영향을 미치게 된다.

Progressive Growing of GANs for Improved Quality, Stability, and Variation

✦ Model (Keypoint2Video –StyleGAN2)

StyleGAN



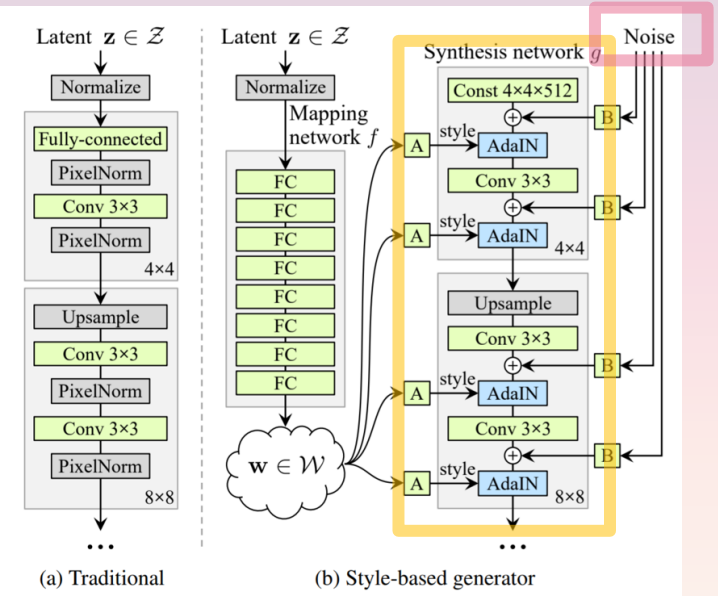
1. Mapping Network

- latent z를 intermediate vector w로 변환
- 고정된 distribution을 따를 필요 없어져
  w를 이용하여 visual attribute 조절이 쉬워짐
- Disentanglement

A Style-Based Generator Architecture for Generative Adversarial Networks

Model (Keypoint2Video –StyleGAN2)

StyleGAN



2. Synthesis Network

- w로 AdaIN을 통해 style을 입힘

- 각 layer의 style이 특정한 visual attribute를 담당하도록

3. Noise

- AdaIN을 통해 큼직한 Style을 학습한다면 Noise를 통해 이미지의 세세한 부분을 바꿈 (Stochastic Variation)

Model (Keypoint2Video –StyleGAN2)

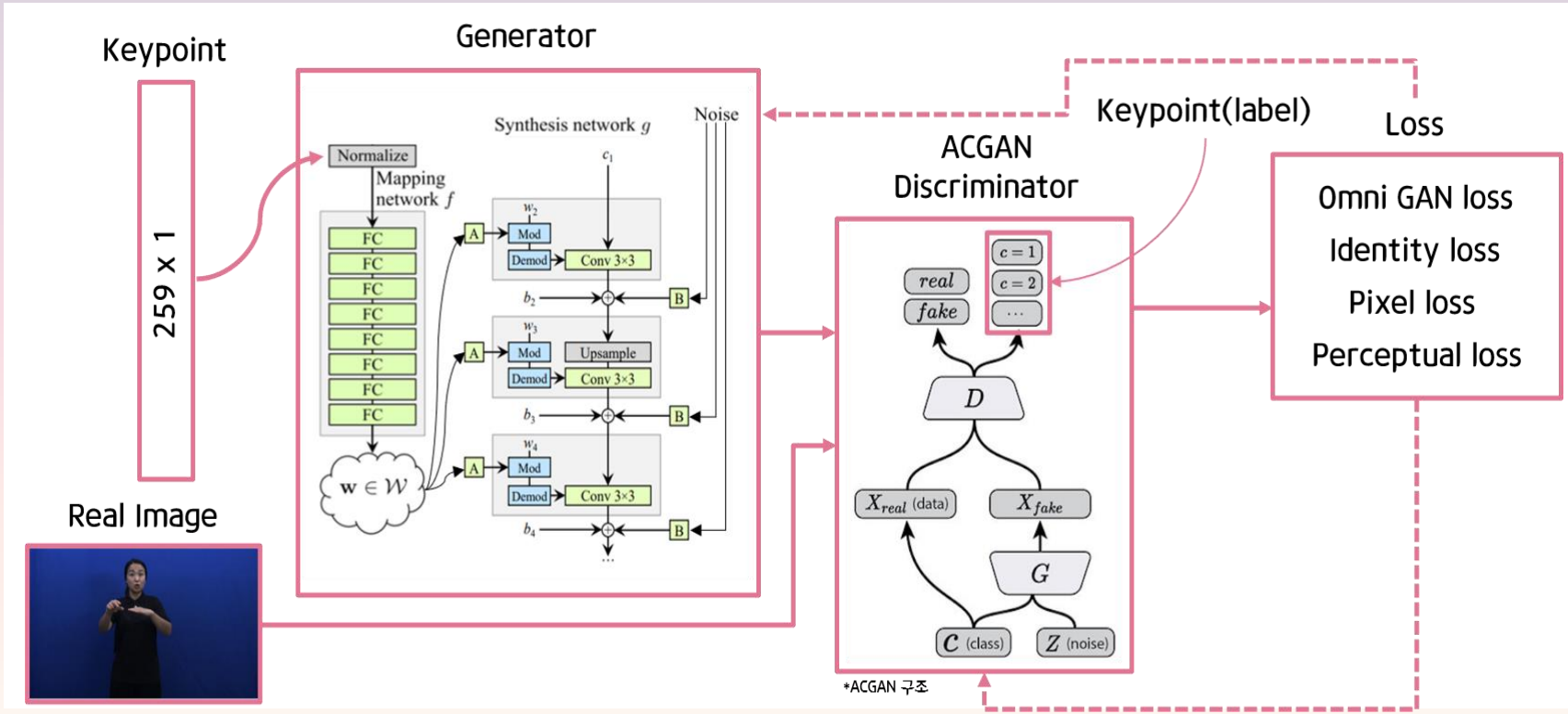
StyleGAN2

AdaIN으로 인해 생긴 물방울 모양의 노이즈

Progressive Growing으로 인해
일부 피쳐들이 얼굴의 움직임을 따르지 않아
부자연스러운 문제

이러한 문제들을 해결하기 위해 AdaIN 대신 가중치를 정규화하고, Progressive Growing을 제거하고 고품질의 이미지를 생성하기 위한 다른 방법을 통해 성능을 개선했다.

✦ Model (Keypoint2Video –StyleGAN2)



✦ Model (Keypoint2Video –StyleGAN2)

1. Mapping Network의 Input으로 keypoints (Conditional GAN)

Keypoints 254개

화자구분 one-hot

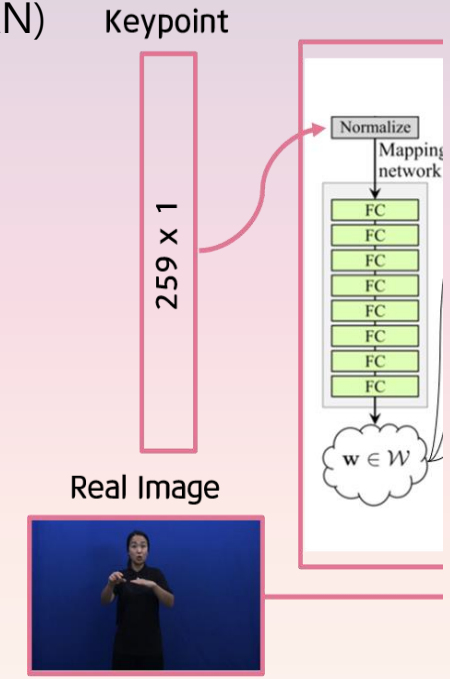
Y좌표

0.49686524, 0.24293488, 0.49387696, 0.43915623, …,

X좌표

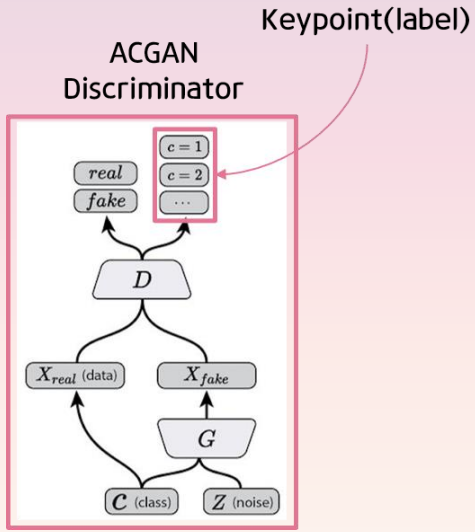
0.37626466, 1.0098177, 0, 0, 0, 1, 0

One-hot



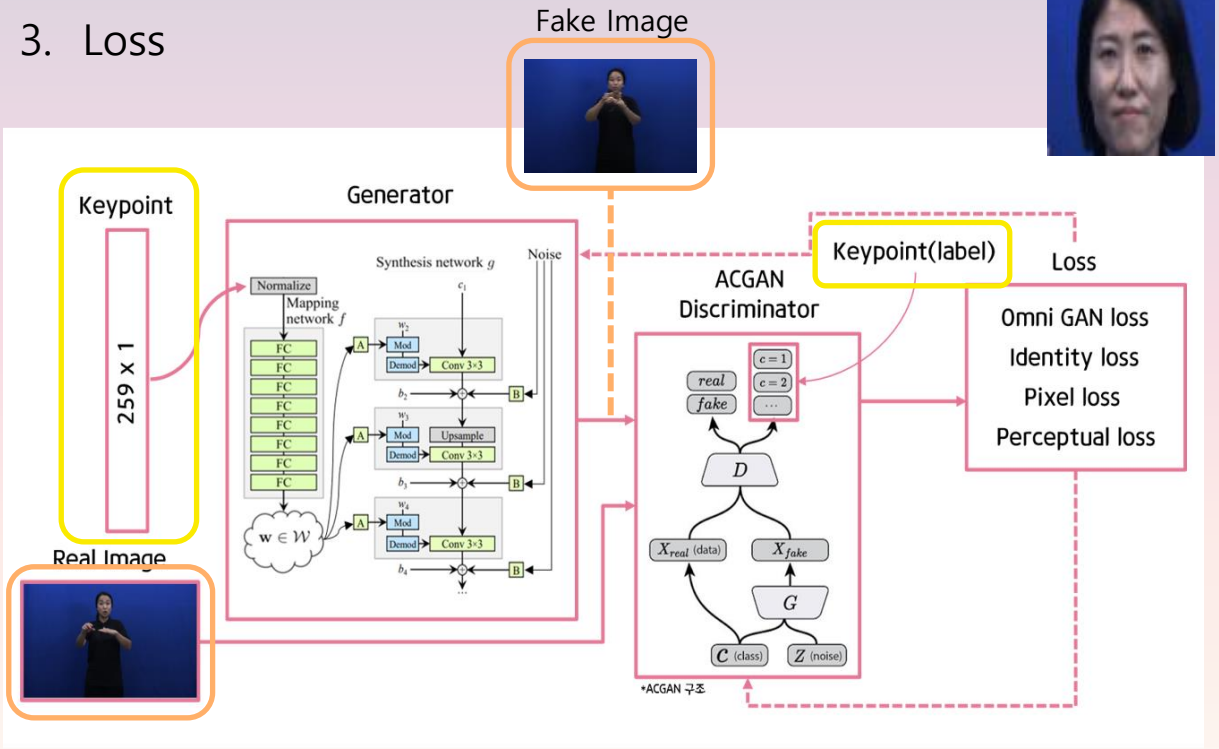
✦ Model (Keypoint2Video –StyleGAN2)

2. Discriminator의 마지막에 layer를 추가해
   Real/Fake 이외에도 Keypoint에 해당하는 output (ACGAN)



Model (Keypoint2Video –StyleGAN2)

3. Loss



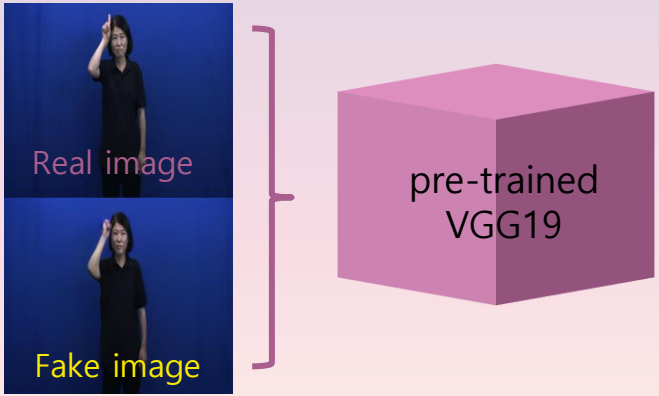

다양한 Loss에 대한 실험

- Omni GAN loss
- Label 사이의 identity loss
- Image 사이의 Pixel loss

- Perceptual loss
- Image 전체댕
- 손, 얼굴 crop

Model (Keypoint2Video –StyleGAN2)

3. Loss – perceptual loss



$$Loss = \sum_{i}^{batch} \underbrace{S(R_i) - S(F_i)}_{\text{Style loss}} + \underbrace{C(R_i) - C(F_i)}_{\text{Content loss}}$$

R_i : $Real\ image$
F_i : $Fake\ image$

사전 학습된 딥러닝 모델에 통과 시킨 후 얻은 feature map 사이의 손실을 최소화 하는 방향으로 파라미터를 학습

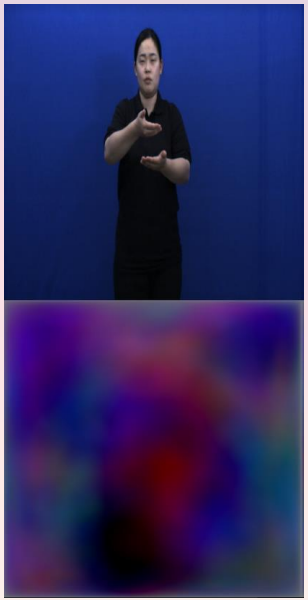
Loss = content loss + style loss

우리는 Real Image와 Generate(Fake) Image간의
Perceptual loss를 계산
- 전체 이미지
- 손, 얼굴 crop 이미지

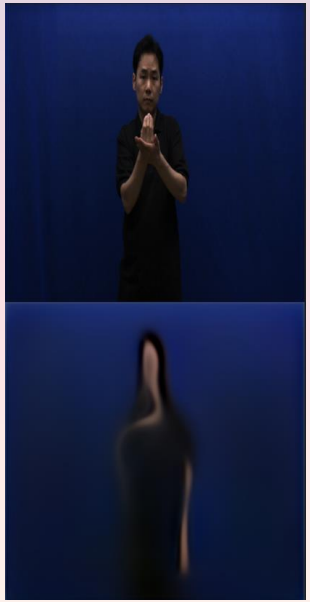
Perceptual Losses for Real-Time Style Transfer and Super-Resolution

Result (Keypoint2Video –StyleGAN2)

1. step 별 결과



100



1800

4900

11100

50700

149800

234600

237800

241000

전체 이미지에서 손이 차지하는 부분은 작으면서 손가락 하나하나까지 많은 키포인트 정보를 가지고 있다보니 손의 학습이 부족함, 손만 따로 생성할 필요성을 느낌

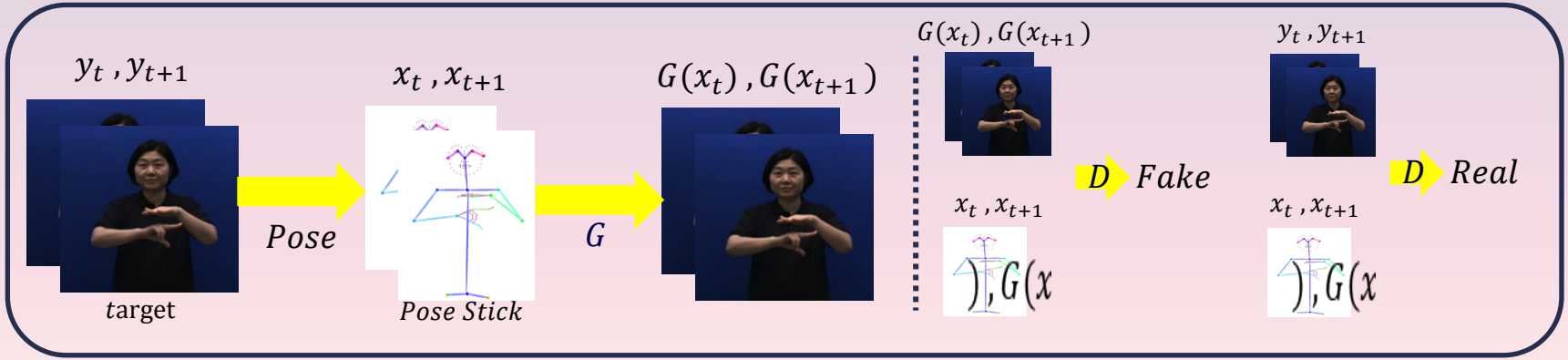
Result (Keypoint2Video –StyleGAN2)

2. 화자 구분에 따른 결과



같은 keypoint 동작에 one-hot 만 다르게 주었을 때 같은 동작을 하고 있는 다른 사람이 생성됨을 볼 수 있다.

Model (Keypoint2Video – Everybody Dance Now)

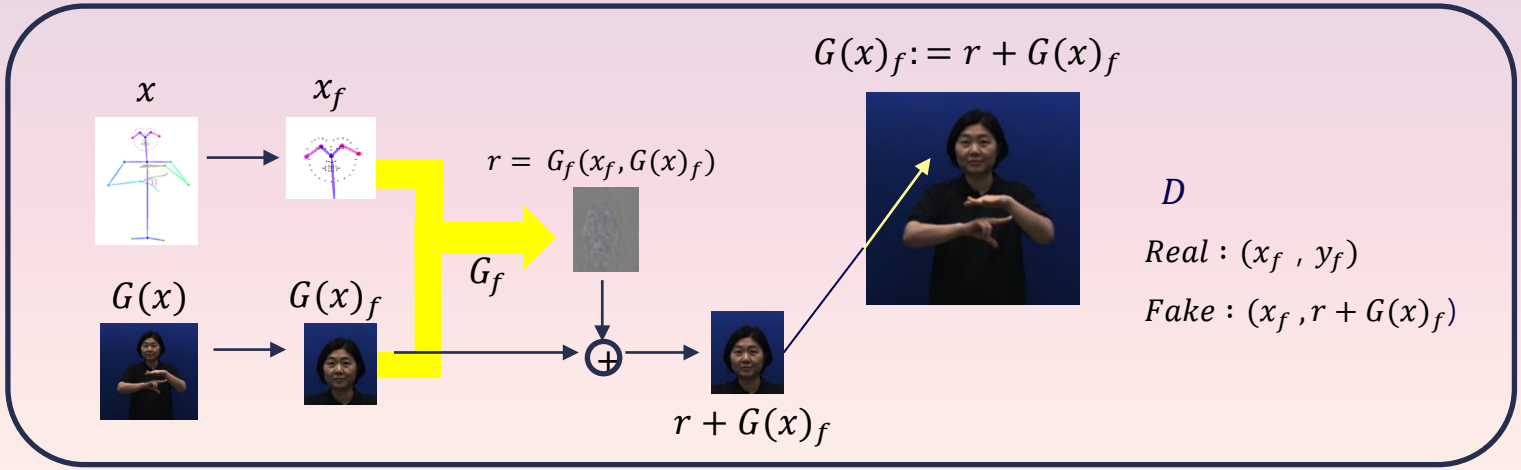


Generator : synthesizes images of a person given a pose stick figure.

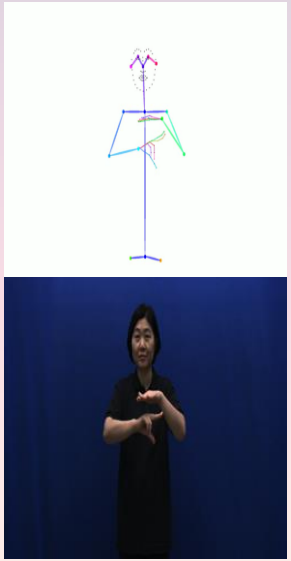
Discriminator : generate image, target image 구분

Model (Keypoint2Video – Everybody Dance Now)

Face GAN 👩 & Hand GAN ✋



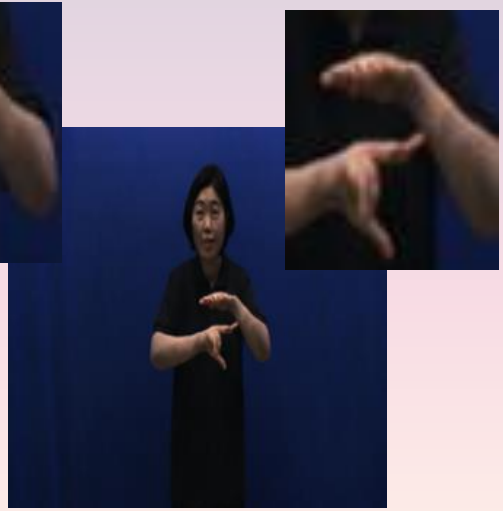
✦ Result (Keypoint2Video – Everybody Dance Now)



origina
l

초기 step

predict



With face & hand gan

Result (Keypoint2Video – Everybody Dance Now)

Original

mse: 243
psnr: 24.3
ssim: 0.889

Predict

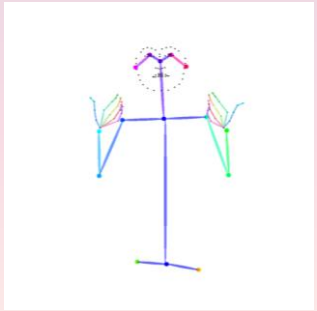
mse : 245,
psnr:24.4,
ssim:0.887

결과 및 결론

Everybody Dance Now 와 StyleGAN2 를 비교했을 때, 다음과 같은 개선이 필요해 보임

StyleGAN2

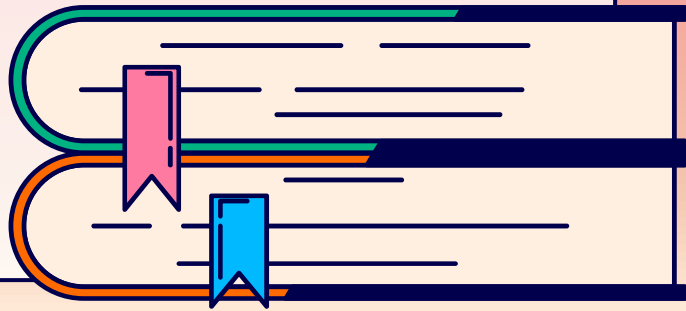
1. Input을 keypoint 대신 keypoint 이미지로 넣어보는 방식
2. crop한 손과 얼굴에 대해서 GAN 한번 더 적용



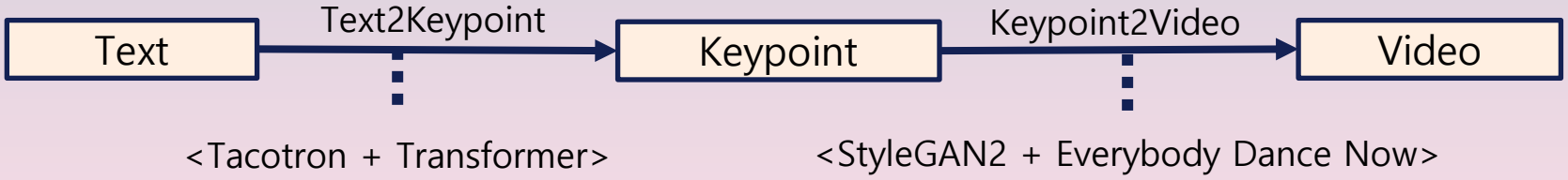
Everybody Dance Now

1. 손과 얼굴을 위한 새로운 GAN 구조 적용
   ex) hand keypoint -> real/fake

결과 및 결론

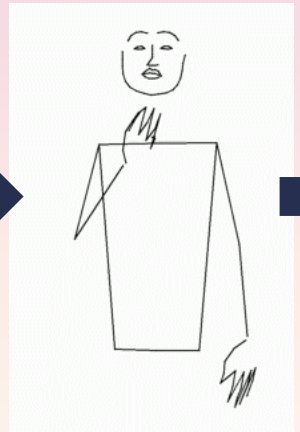


결과 및 결론

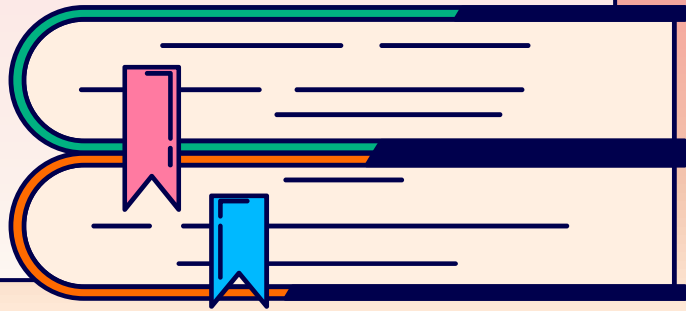


서울내 정기권 지하철표를
어디서 살수 있나요?

서울
지하철
1달
표
사다
어느 곳



활용 방안



활용 방안

화웨이(Huawei)가 만든 **스토리사인(StorySign)**이란 앱

- 인공지능 기술을 통해 사회적 약자의 생활을 개선할 수 있는 방향으로 연구 분야 및 서비스가 지속적으로 확대 중임.
- 많은 수의 농인들은 한국어를 읽거나 이해하지 못하고 수어를 통해서만 의사소통이 가능한 경우가 대부분임. 따라서 농인들이 사회에서 다양한 활동을 하기 위해서는 시간과 공간의 제약이 없는 한국어와 수어의 통역이 필요함.
- 수어 인식 인공지능모델 개발 등을 통해 상기 문제에 대해 일부는 도움이 될 것으로 기대됨.

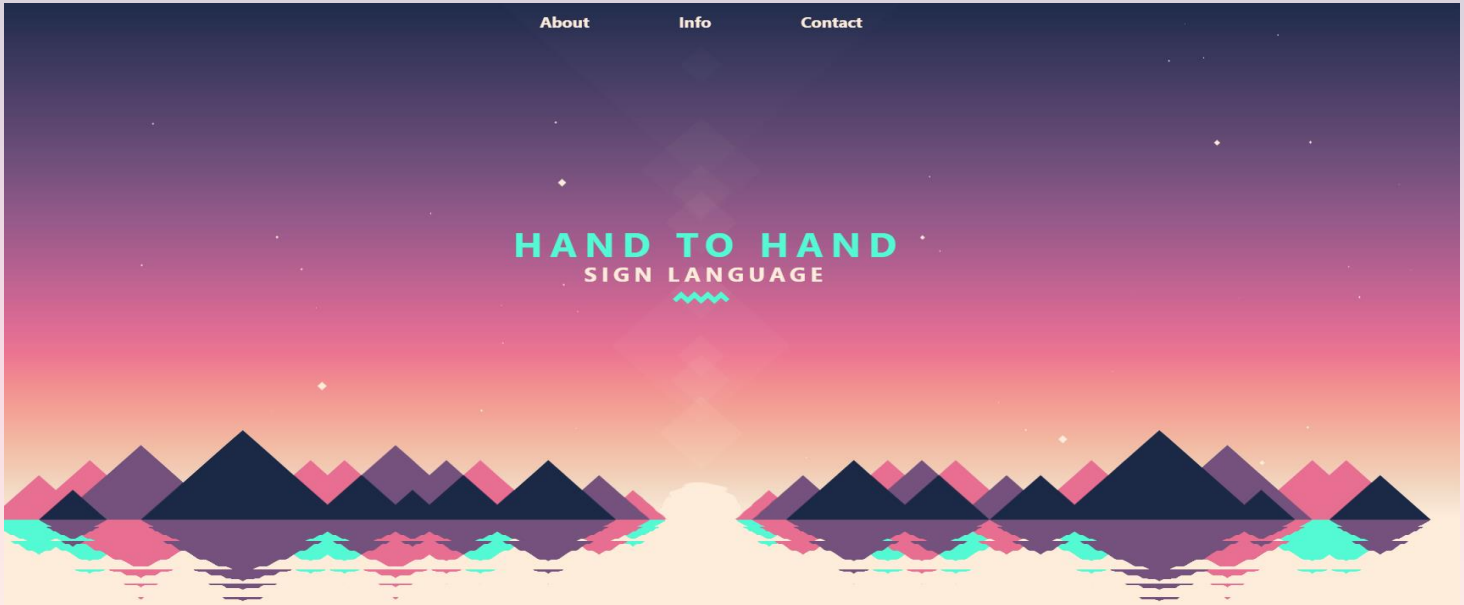
Hand In Hand Challenge

수어(手語/Sign language) 데이터를 활용한
베리어프리 서비스 디자인 챌린지
총상금 900만원(총 3팀)

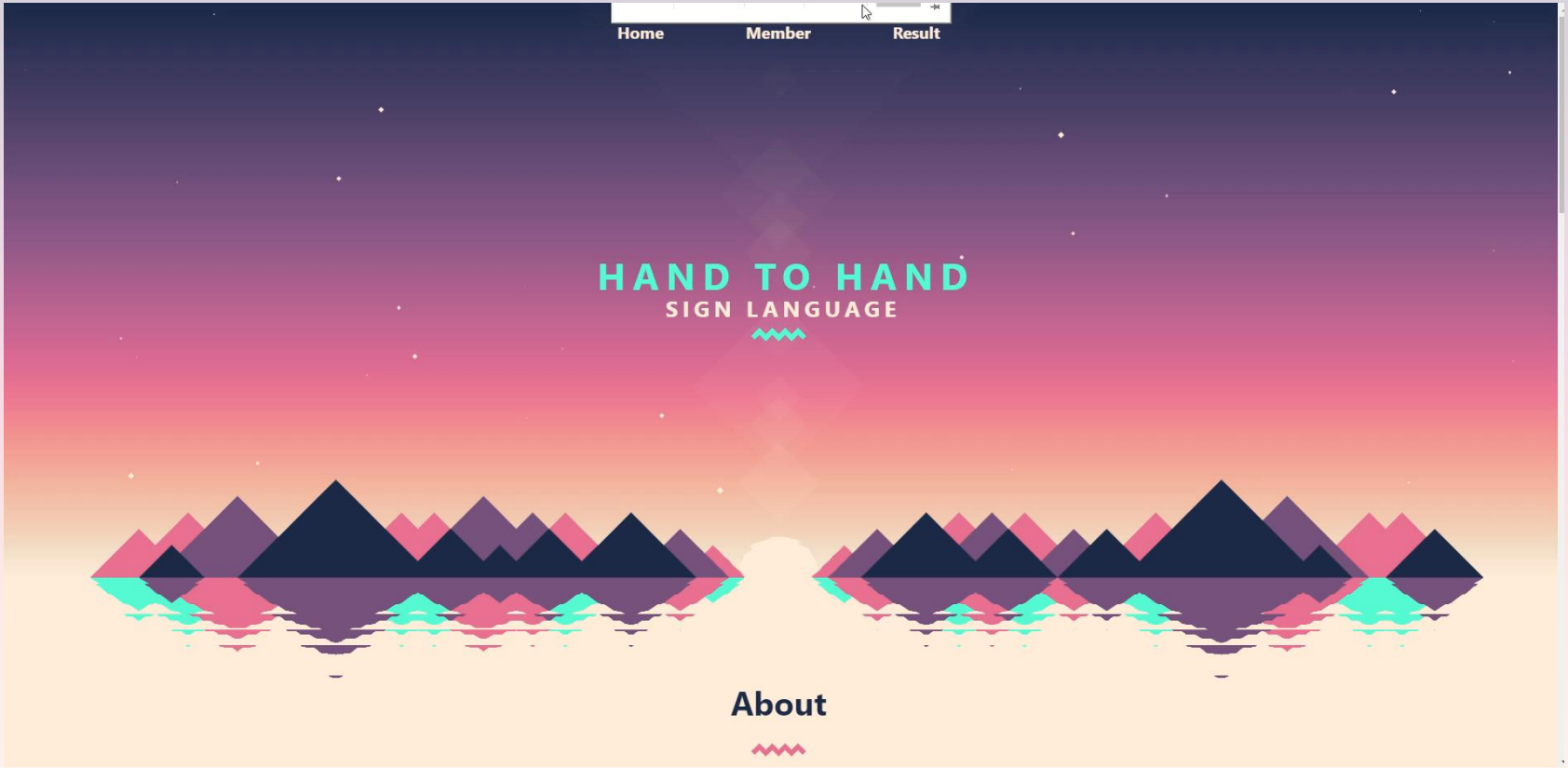
참가의향서 접수하기

대상	최우수상	우수상
500만원	300만원	100만원

- 결선 진출팀 전원 기계식 키보드 제공
- 결선 진출팀에 DGX A100 1core 1week 제공 커먼 컴퓨터 후원

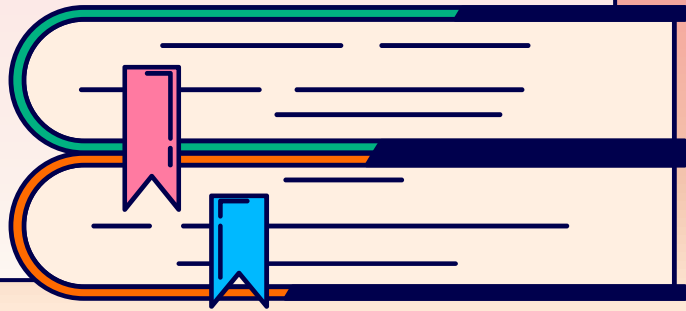


https://hand-to-hand.kro.kr/


Home
Member
Result
HAND TO HAND
SIGN LANGUAGE
About

감사합니다 :)

참여 멤버 소개

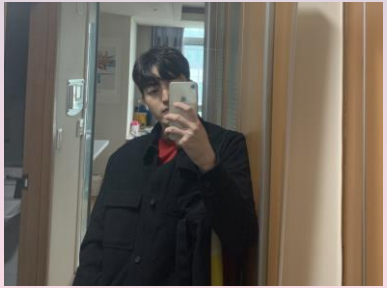


참여 멤버 소개

<11기
이도연>

<13기
고유경>

<13기
김미성>



<13기
정민준>

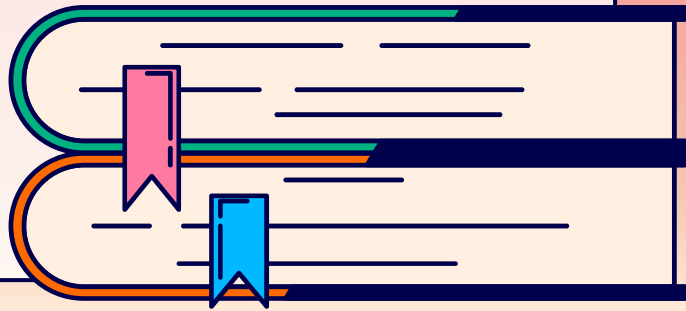
<13기
조혜원>

<14기
강의정>

<14기
서아라>

<14기
이정은>

참여 멤버 소개

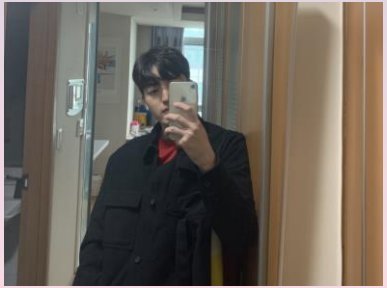


참여 멤버 소개

<11기
이도연>

<13기
고유경>

<13기
김미성>



<13기
정민준>

<13기
조혜원>

<14기
강의정>

<14기
서아라>

<14기
이정은>