

---

# **Skeletor: Skeletal Transformers for Robust Body-Pose Estimation**

---

Yongwon Jo

School of Industrial and Management Engineering, Korea University



KOREA  
UNIVERSITY



DMQA

# Contents

---

❖ Research Purpose

❖ Skeletor

❖ Experiments

❖ Conclusion

# Research Purpose

---

## ❖ Skeletor: Skeletal Transformers for Robust Body-Pose Estimation (arXiv, 2021)

- 영국 Surrey 대학에서 발표한 논문이며 2021년 9월 27일 기준 1회 인용
- 비지도 학습 방식으로 3D Pose estimation 모델을 사전 학습하는 방식 제안

## Skeletor: Skeletal Transformers for Robust Body-Pose Estimation

Tao Jiang, Necati Cihan Camgöz, Richard Bowden  
Centre for Vision, Speech and Signal Processing  
University of Surrey, Guildford, UK

{t.jiang, n.camgoz, r.bowden}@surrey.ac.uk

# Research Purpose

---

## ❖ Skeletor: Skeletal Transformers for Robust Body-Pose Estimation (arXiv, 2021)

- 단일 RGB 영상에서 3D Pose estimation 모델 학습은 여러 어려움 존재
  - 촬영하고자 하는 대상을 정확하게 촬영할 수 없음 (Jittering)
  - 색깔이 왜곡되거나 촬영 대상이 일부 가려짐이 발생할 수 있음 (Occlusion)
  - 영상 내 Frame 간 대상의 연속성을 가지기 어려움(Non-consistency)
- 3D Pose estimation 학습을 위한 Ground truth 수집이 어려움
- Depth 카메라나 여러 카메라를 사용해 공간 정보를 획득이 필요하지만 현실 적용이 어려움

# Research Purpose

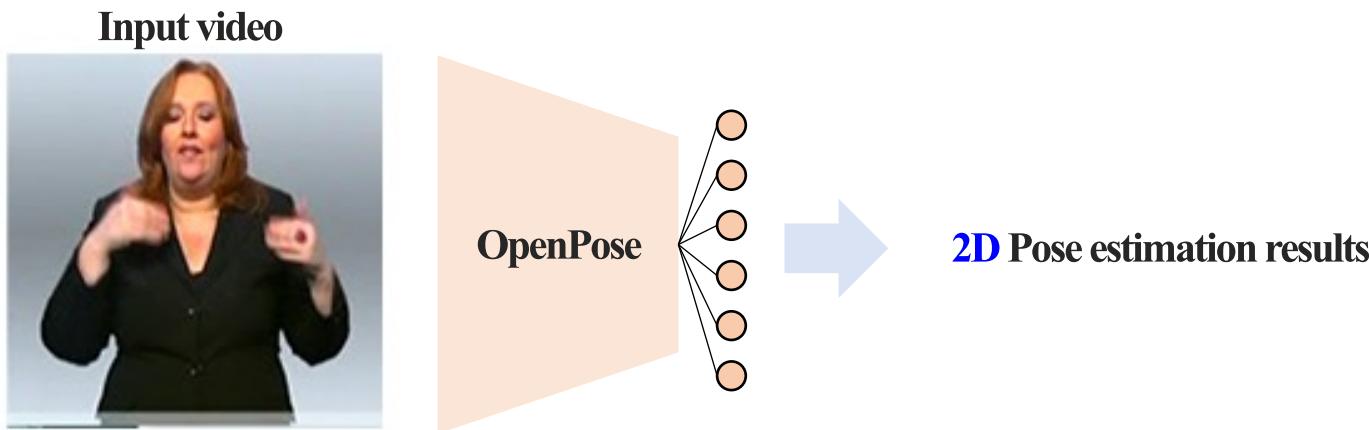
- ❖ **Skeletor: Skeletal Transformers for Robust Body-Pose Estimation (arXiv, 2021)**
  - 단일 RGB 영상에서 3D Pose estimation 모델 학습은 여러 어려움 존재
  - 3D Pose estimation 학습을 위한 Ground truth 수집이 어려움
  - Depth 카메라나 여러 카메라를 사용해 공간 정보를 획득이 필요하지만 현실 적용이 어려움
- ❖ **Skeletor: Skeletal Transformers for Robust Body-Pose Estimation (arXiv, 2021)**
  - 비지도 학습 기반 3D Pose estimation 모델을 학습하는 것이 본 논문의 목표
  - 자연어 처리 분야에서 흔히 사용되는 BERT와 유사한 방식으로 사전 학습을 진행
  - 2D 및 3D Pose estimation 모델의 예측 값을 적극적으로 활용
  - 사전 학습된 가중치를 전이시켜 Downstream task 진행하고 성능 평가 진행 (수화)

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

# Skeletor

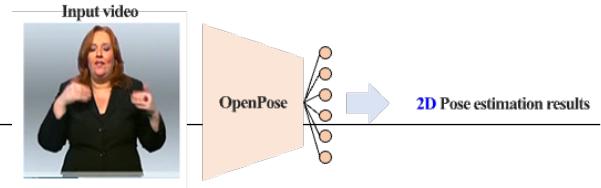
## ❖ Skeletor (Unsupervised 3D pose estimation model)

- 2D/3D 관절 좌표가 없는 상황을 가정하고 모델 학습을 시작
- OpenPose 모델을 사용해 2D Pose estimation 결과 값을 추출
  - OpenPose 모델은 2D Pose estimation이 가능하며 사전 학습된 모델 존재
  - 실제 영상에 Annotation하지 않고 예측 결과를 사용하기에 Unsupervised pose estimation
  - 2D Pose estimation 모델을 사용하던지 관계없으나 성능이 뛰어나다는 가정이 반드시 필요



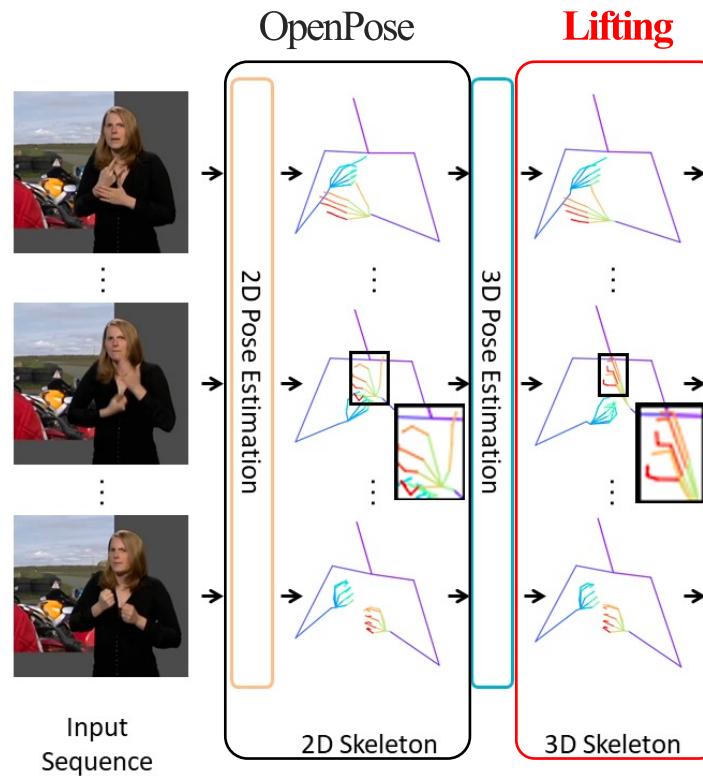
- Cao, Z., Hidalgo, G., Simon, T., Wei, S. F., & Sheikh, Y. (2019) OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1), 172-186.

# Skeletor



## ❖ Skeletor (Unsupervised 3D pose estimation model)

- OpenPose에서 획득한 2D pose estimation 결과가 Skleleton 입력 데이터로 사용
- 2D Pose estimation 결과를 Lifting 기법을 사용해 3D Pose estimation 결과 추출



# Skeletor

---

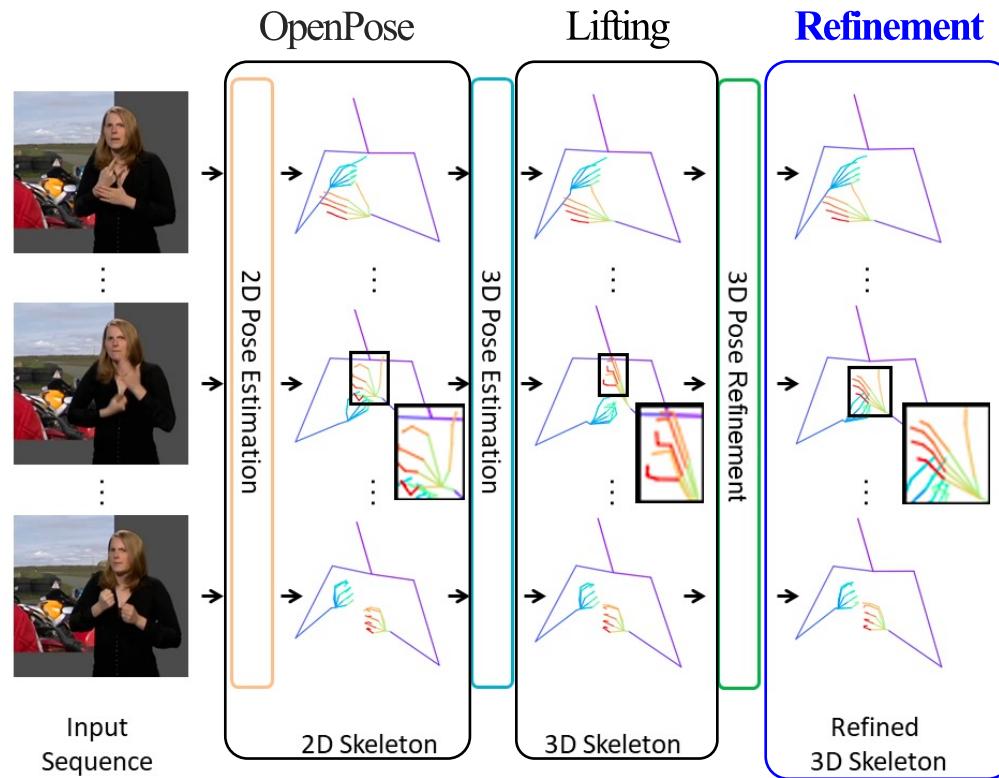
## ❖ Skeletor (Unsupervised 3D pose estimation model)

- OpenPose에서 획득한 2D pose estimation 결과가 Skleleton 입력 데이터로 사용
- 2D Pose estimation 결과를 Lifting 기법을 사용해 3D Pose estimation 결과 추출
  - i. 2D Pose estimation 결과를 추출
  - ii. 2D Pose estimation에 대해 3D Uplift 방법을 사용해 우선 인간 머리의 3D 좌표를 추출
  - iii. 영상 내 모든 프레임에 대한 2D Pose estimation 결과 내 뼈들의 길이 평균을 계산
  - iv. 머리부터 다른 관절까지 순차적으로 진행하며 아래의 요소들 간 Mean squared error (MSE)로 학습
    - a) 3D 관절 좌표를 다시 2D로 변환한 것과 2D 관절 좌표 사이 MSE
    - b) 최종 프레임 내 Trajectory 길이와 모든 프레임 내 뼈 길이 사이 MSE

# Skeletor

## ❖ Skeletor (Unsupervised 3D pose estimation model)

- Lifting 기법을 사용해 예측한 3D Pose estimation 결과에 변형 진행 (프레임 단위, 관절 단위)
- 변형된 3D Pose estimation 결과를 정확하게 복원하는 Refinement 과정 진행
- Refinement 과정에서 Transformer 기반 심층 신경망 모델 존재



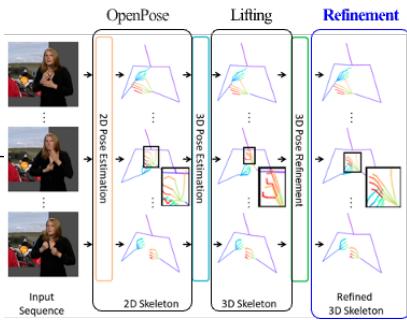
# Skeletor

---

## ❖ Skeletor (Unsupervised 3D pose estimation model)

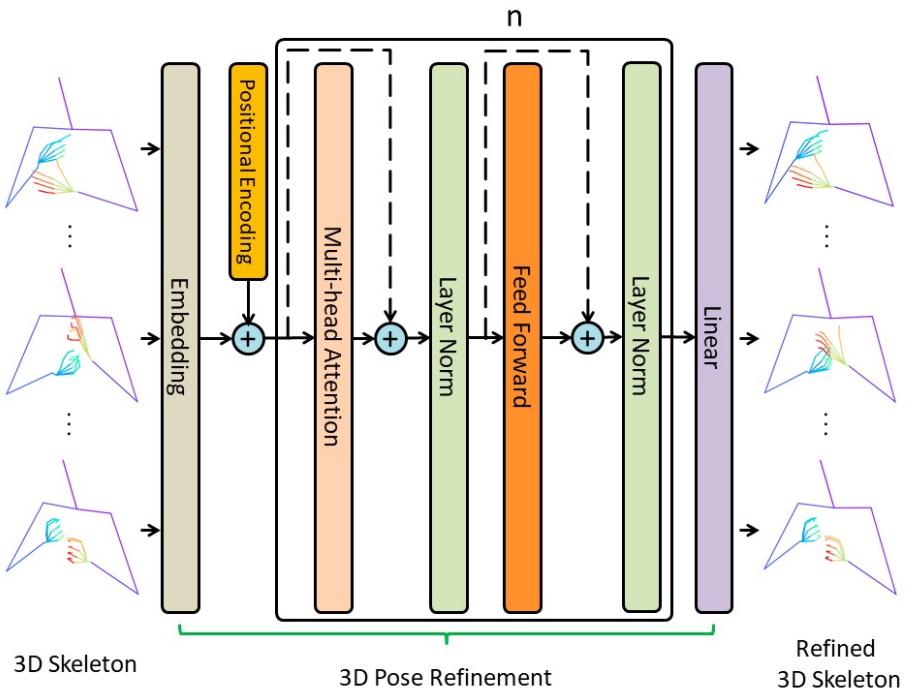
- Lifting 기법을 사용해 예측한 3D Pose estimation 결과에 변형 진행 (프레임 단위, 관절 단위)
  - 프레임 단위 Masking: 영상 내 특정 프레임을 제거하고 이를 복원하는 Refinement 과정 진행
  - 관절 단위 Masking: 여러 프레임 내 특정 관절을 제거하고 이를 복원하는 Refinement 과정 진행
  - 관절 위치에 Noise를 부여하고 Noise를 제거하는 Refinement 진행
- 변형된 3D Pose estimation 결과를 정확하게 복원하는 Refinement 과정 진행
- Refinement 과정에서 Transformer 기반 심층 신경망 모델 존재

# Skeletor



## ❖ Skeletor (Unsupervised 3D pose estimation model)

- 변형된 3D Pose estimation 결과를 정확하게 복원하는 Refinement 과정 진행
- Refinement 과정에서 Transformer 기반 심층 신경망 모델 존재
- 일반적인 Transformer layer를 사용하며 총 8개 layer로 구성
- 마지막에는 linear layer를 사용해 최종 Refinement 진행



# Experiments

---

## ❖ Unsupervised 3D pose refinement 결과 비교 (프레임 단위 제거)

- 전체 프레임 중 5%, 10%, 15%, 20%, 25% 단위 프레임을 제거하고 이를 Refinement 진행
- 10% 프레임 제거 시 실제 Test 단계에서 성능이 가장 좋음

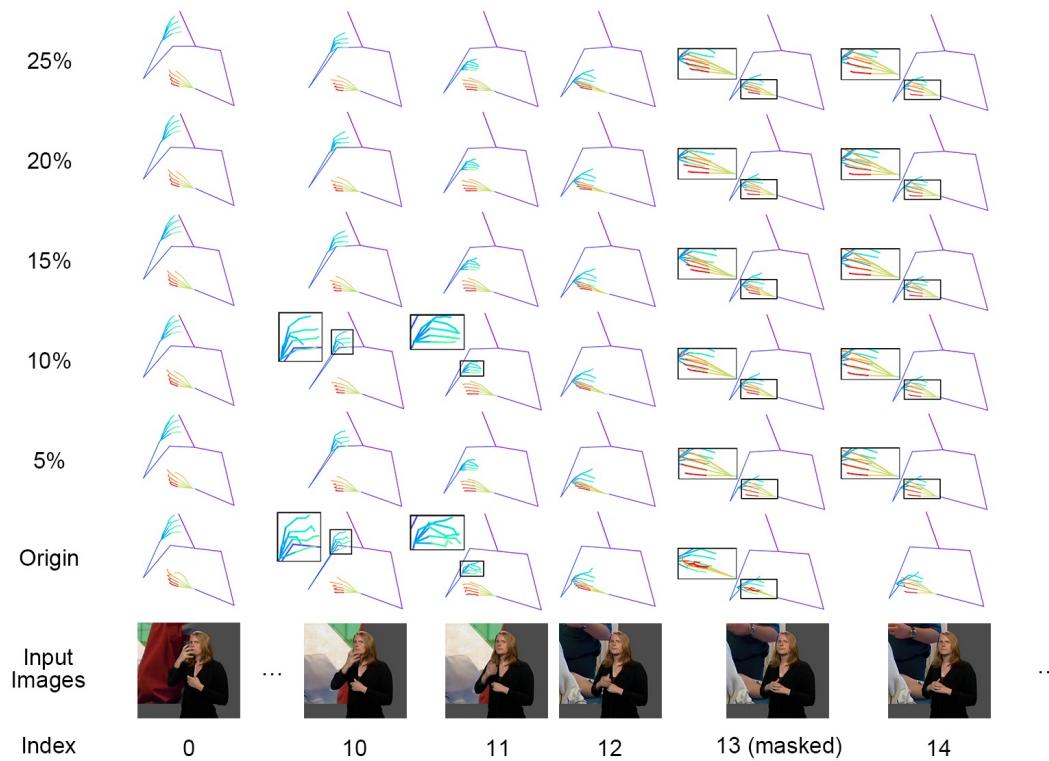
Table 1. The left column is the frame masking percentage the model is trained with. The middle column is the MSE evaluated on the development dataset at different training steps. The right column is the minimum (min), average (ave) and maximum (max) MSE evaluated on test data.

Frame mask	Development (# iterations)				Test		
	50,000	100,000	200,000	300,000	min	ave	max
5%	3.086	2.124	1.195	1.103	0.147	1.833	9.628
10%	<b>2.613</b>	<b>1.404</b>	<b>0.642</b>	<b>0.581</b>	<b>0.090</b>	<b>0.875</b>	<b>8.421</b>
15%	5.048	1.844	0.919	0.874	0.255	1.266	9.862
20%	5.518	2.196	1.034	0.856	0.264	1.250	9.459
25%	4.842	2.469	1.116	0.971	0.210	1.147	8.273

# Experiments

## ❖ Unsupervised 3D pose refinement 결과 비교 (프레임 단위 제거)

- 전체 프레임 중 5%, 10%, 15%, 20%, 25% 단위 프레임을 제거하고 이를 Refinement 진행
- 프레임 13은 제거되었지만 해당 모델이 정확히 예측하고 있음을 확인
- 프레임 10, 11에 존재하는 Noisy 역시 제거하고 있음을 확인 가능



# Experiments

---

## ❖ Unsupervised 3D pose refinement 결과 비교 (관절 단위 제거)

- 전체 관절 중 5%, 10%, 15%, 20%, 25% 단위 관절을 제거하고 이를 Refinement 진행
- 10% 관절 제거 시 실제 Test 단계에서 성능이 가장 좋음

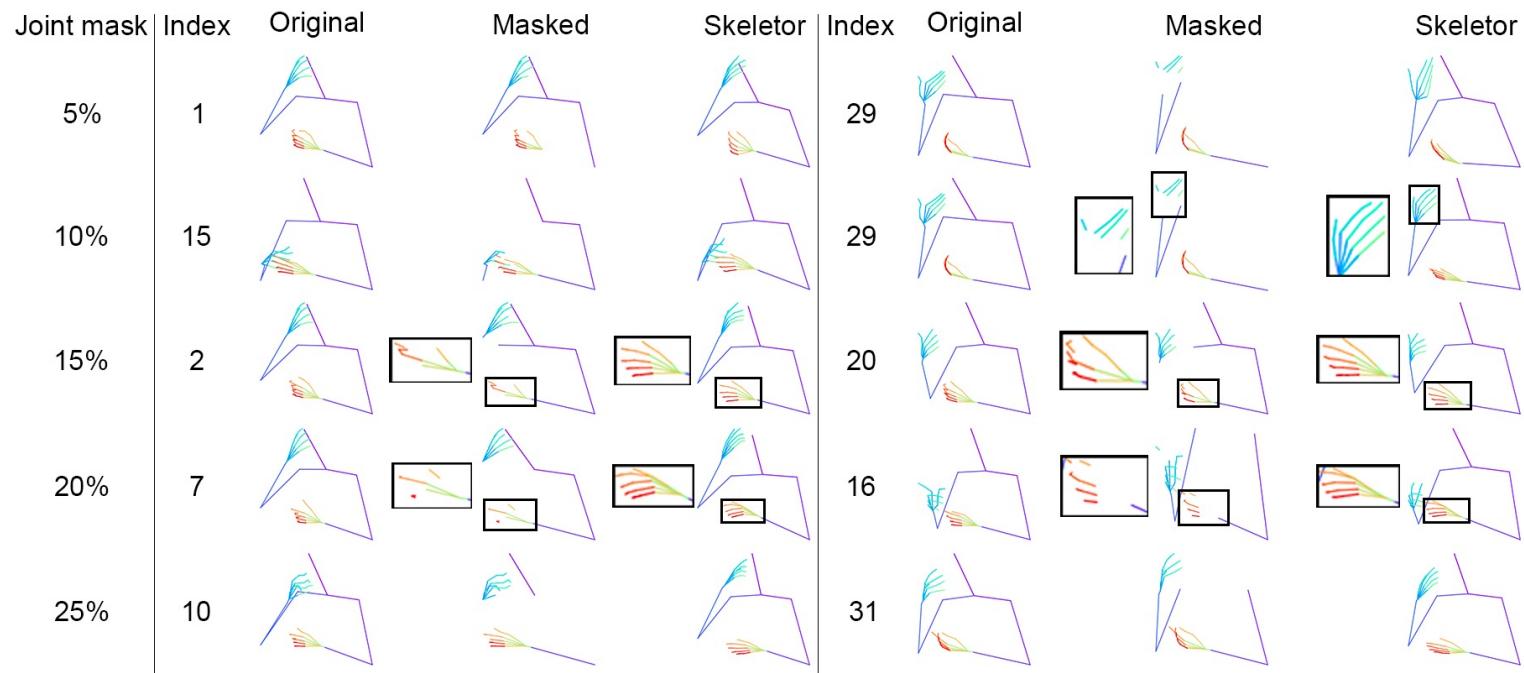
Table 2. The MSE results on the development and test datasets for models with different joint mask percentages.

Joint mask	Development (# iterations)				Test		
	50,000	100,000	200,000	300,000	min	ave	max
5%	<b>17.532</b>	15.146	13.655	13.645	2.019	4.543	12.484
10%	18.192	<b>14.493</b>	<b>13.585</b>	<b>13.521</b>	<b>1.869</b>	<b>3.555</b>	<b>9.590</b>
15%	20.759	15.594	14.585	14.640	1.979	4.181	15.478
20%	23.323	18.590	17.881	17.914	2.606	6.390	42.758
25%	23.453	17.816	16.452	16.418	2.314	5.031	22.171

# Experiments

## ❖ Unsupervised 3D pose refinement 결과 비교 (관절 단위 제거)

- 전체 관절 중 5%, 10%, 15%, 20%, 25% 단위 관절을 제거하고 이를 Refinement 진행
- 관절이 제거되었지만 대부분 정확히 Refinement 되는 것을 확인 가능
- 두 결과를 통해 행동 자체에 대한 인식과 관절 간 관계를 정확히 학습한 것으로 평가 가능



# Experiments

---

## ❖ Unsupervised 3D pose refinement 결과 비교 (프레임 단위 노이즈 부여)

- Noise 수준을 0.1부터 0.2씩 증가시키며 Refinement 성능 비교
- 전체 프레임 중 15% 프레임에만 관절 전체에 노이즈 부여
- 0.5 이상에서는 정확히 Refinement 되지 않는 것을 정량적으로 확인 가능

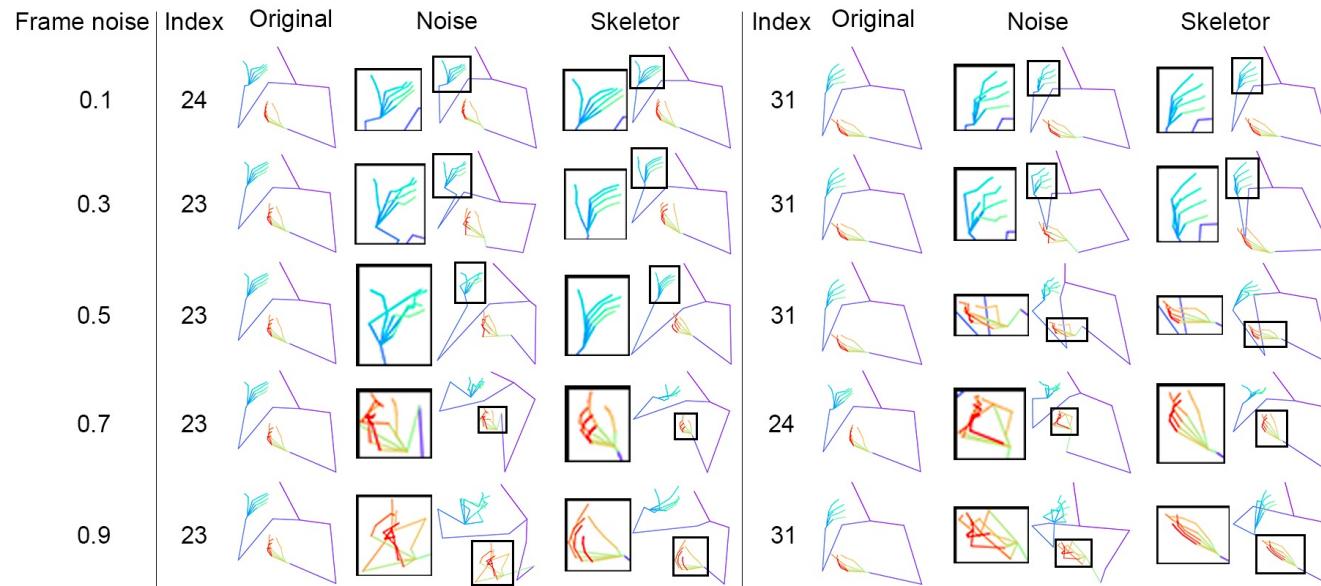
Table 3. *Skeletor* can be used to correct the skeleton with noise applied to the whole skeleton.

Noise Strength ( $s$ )	Frame-level		
	min	ave	max
0.1	0.059	0.117	1.408
0.3	0.154	0.215	1.488
0.5	0.345	0.409	1.691
0.7	0.585	0.689	1.983
0.9	0.841	1.100	2.331

# Experiments

## ❖ Unsupervised 3D pose refinement 결과 비교 (프레임 단위 노이즈 부여)

- Noise 수준을 0.1부터 0.2씩 증가시키며 Refinement 성능 비교
- 전체 프레임 중 15% 프레임에만 관절 전체에 노이즈 부여
- 정성적 평가로도 Noise 수준이 증가됨에 따라 Refinement 성능이 감소함을 확인 가능



# Experiments

---

## ❖ Unsupervised 3D pose refinement 결과 비교 (관절 단위 노이즈 부여)

- Noise 수준을 0.1부터 0.2씩 증가시키며 Refinement 성능 비교
- 전체 관절 중 15% 관절에만 노이즈 부여
- 관절 Refinement 성능이 준수하다는 것을 확인 가능

Table 3. *Skeletor* can be used to correct the skeleton with noise applied to the whole skeleton.

Noise Strength ( $s$ )	MSE on Test set		
	min	ave	max
0.1	0.059	0.117	1.408
0.3	0.154	0.215	1.488
0.5	0.345	0.409	1.691
0.7	0.585	0.689	1.983
0.9	0.841	1.100	2.331

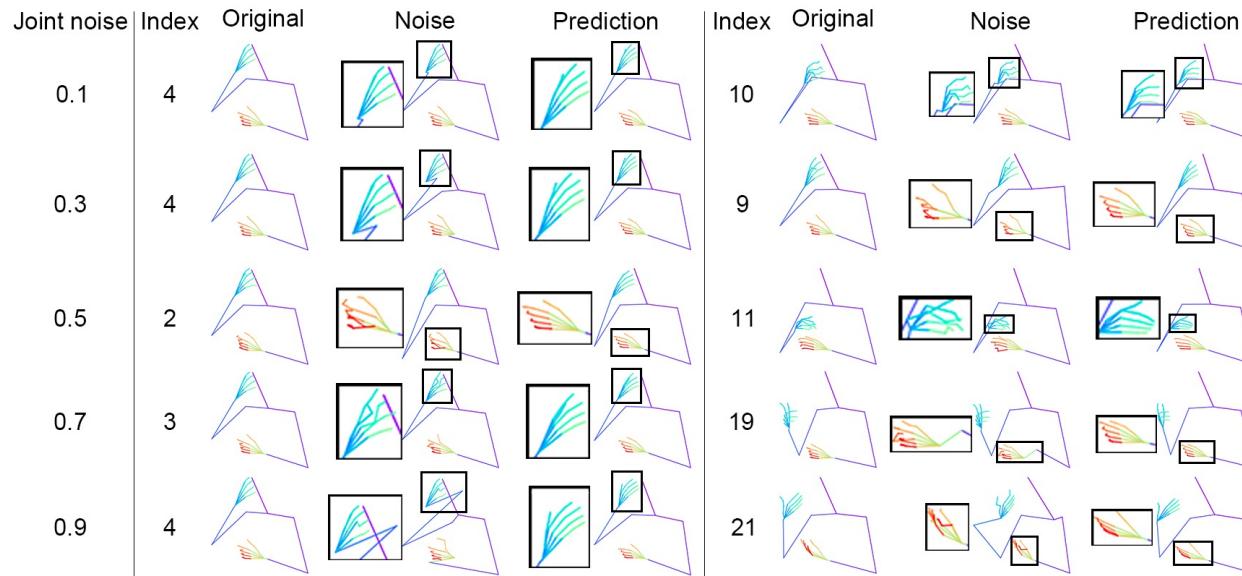
Table 4. *Skeletor* can be used to correct the skeleton with different levels of noise applied to the joints.

Noise Strength ( $s$ )	MSE on Test set		
	min	ave	max
0.1	0.225	0.452	4.573
0.3	0.272	0.508	4.786
0.5	0.364	0.616	5.126
0.7	0.496	0.770	5.209
0.9	0.656	0.965	5.483

# Experiments

## ❖ Unsupervised 3D pose refinement 결과 비교 (관절 단위 노이즈 부여)

- Noise 수준을 0.1부터 0.2씩 증가시키며 Refinement 성능 비교
- 전체 관절 중 15% 관절에만 노이즈 부여
- 관절 Refinement 성능이 준수하다는 것을 확인 가능



# Experiments

## ❖ Downstream task (Sign language translation)

- Sign language translation은 수(Hand)화 영상을 해석하는 문제
- 입력 데이터는 수화하고 있는 사람을 촬영한 영상이며 출력 데이터는 수화의 의미(Text)
- Phoenix 14T 데이터셋(수화)을 사용해 Downstream task 성능 평가 진행
- From scratch로 학습한 모델보다 성능이 뛰어난 것을 확인 가능
- 또한, Sign2Text와 같이 영상에서 수화 의미를 예측하는 방법보다 뛰어난 성능을 보임

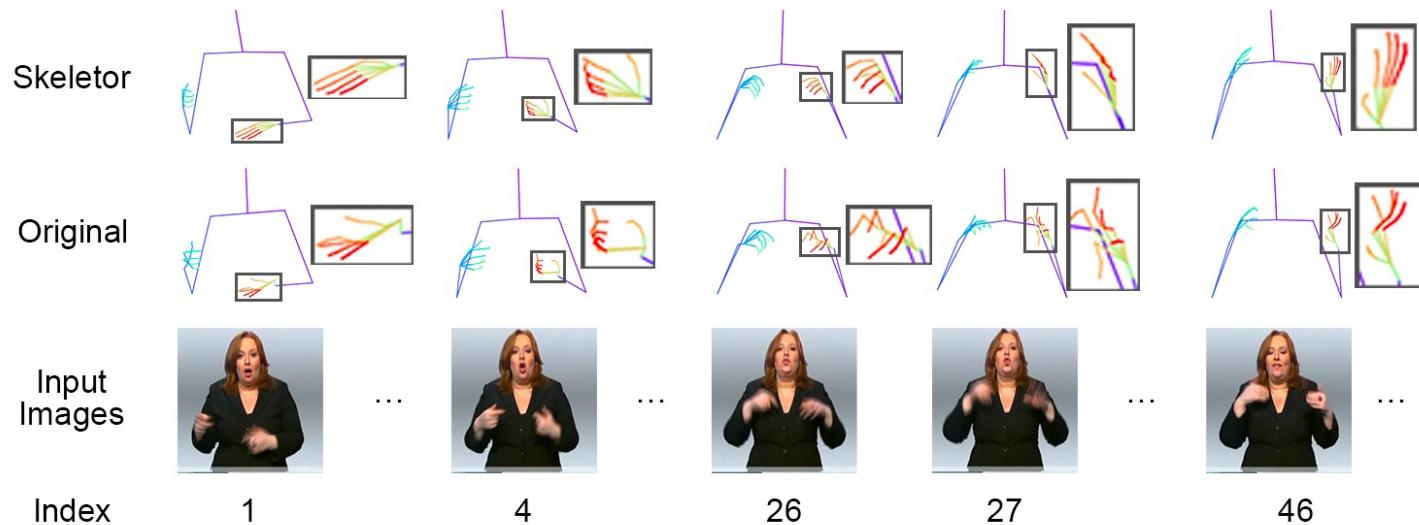
Table 5. Comparison with the Sign2Text baseline sign language translation model, using the back translation evaluation.

	DEV SET					TEST SET				
	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE
Sign2Text [3]	9.94	13.16	19.11	31.87	31.80	9.58	12.83	19.03	<b>32.24</b>	31.80
Raw 3D Skeleton Estimates	9.74	12.75	18.38	31.43	31.82	8.85	11.62	16.94	30.22	29.89
<i>Skeletor</i>	<b>10.91</b>	<b>14.01</b>	<b>19.53</b>	<b>31.97</b>	<b>32.66</b>	<b>10.35</b>	<b>13.49</b>	<b>19.11</b>	31.86	<b>31.80</b>

# Experiments

## ❖ Downstream task (Sign language translation)

- Sign language translation은 수(Hand)화 영상을 해석하는 문제
- 입력 데이터는 수화하고 있는 사람을 촬영한 영상이며 출력 데이터는 수화의 의미(Text)
- Phoenix 14T 데이터셋(수화)을 사용해 Downstream task 성능 평가 진행
- From scratch로 학습한 모델보다 성능이 뛰어난 것을 확인 가능
- 또한, Sign2Text와 같이 영상에서 수화 의미를 예측하는 방법보다 뛰어난 성능을 보임



# Conclusion

---

## ❖ Conclusion

- 사전 학습된 2D Pose estimation 모델 기반 3D Pose estimation 사전 학습 방식 제안
- 해당 방식은 손의 행동 인식에 대한 특징과 관절에 대한 특징을 모두 추출 가능
  - i. 손의 행동 인식 → Global representation 학습 가능
  - ii. 관절에 대한 특징 추출 → Local representation 학습 가능
- 영상에 대한 레이블이 아닌 예측 결과로 모델을 학습했다는 것에서 비지도 학습

## ❖ 본 논문에 대한 나의 생각

- Self-supervised learning의 Denoising autoencoder, inpainting 기반 사전 학습과 유사(Pretext task)
- Pose estimation을 위한 Contrastive learning 기반 사전 학습 방안에 대한 탐색 필요
- Pose estimation을 위한 Self-supervised learning, Data augmentation에 대한 연구도 가치가 있을 것

---

# Thank you