Week17: 논문스터디

예습	
복습	
복습과제 날짜	
예습과제 날짜	
내용	

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE : Vi-T(Vision Transformer)

Paper:

https://arxiv.org/abs/2010.11929

Code:

https://github.com/google-research/vision_transformer

Abstract

CNN을 사용하지 않는 pure transformer이 이미지 분류 태스크에서 좋은 성능을 발위할 수 있다. Vision Transformer를 많은 양의 데이터에 사전 학습 한 다음, 이를 mid-size image, small-size image 인식 테스크 등의 다양한 테스크에 전이학습하여 evalution 진행, ViT는 기존 sota 성능을 보이는 cnn 기반 네트워크에 필적하는 좋은 결과가 보였다.

Self-attention기반 아키텍처는 NLP분야의 태스크에서 보편적으로 쓰이는 모델입니다(대표적으로 트랜스포머).일반적으로 **Large text corpus**에 사전학습시킨 다음, **smaller task-specific dataset**에 전이학습을 시키는 접근법이 널리 사용됩니다.**Transformer** 기반 모델들은 아래와 같은 장점을 가지기에 더욱 널리 사용되고 있으며, 모델과 데이터셋이 증가함에 따라 성능 또한 포화상태에 이를 기미가 아직은 보이지 않습니다.

Transformer-based 모델의 장점1. 연산이 효율적이다(computational efficiency).2. 확장성이 좋다(scalability). 특히 input sequence의 길이에 구애받지 않는다.

NLP 분야에서 트랜스포머의 **scaling successes**에 특히 영감을 받아 image에 트랜스포머를 **직접적으로 적용할 수 있는** 연구를 진행했습니다.**ViT의 과정**은 대략적으로 아래와 같이 기술할 수 있습니다.

1. 이미지를 image pathces로 쪼갠다.
2. 이 patches들을 linear하게 임베딩해 시퀀스를 생성한다.
3. 이렇게 생성한 시퀀스를 트랜스포머의 input으로 투입한다.

> 즉, ViT에서 하나의 이미지 패치는 하나의 token으로 작동합니다.

이런 방식으로 **이미지 분류**를 지도학습합니다.

ViT

original transformer의 구조를 대부분 따르다.

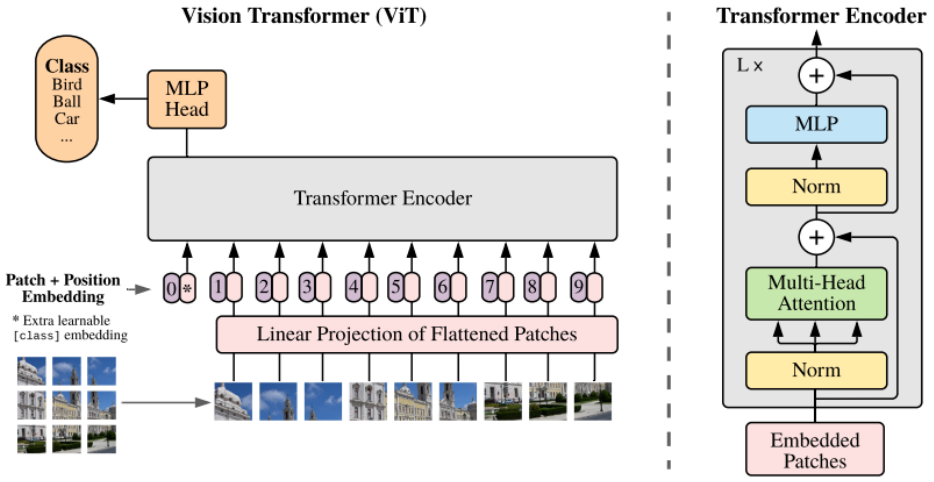
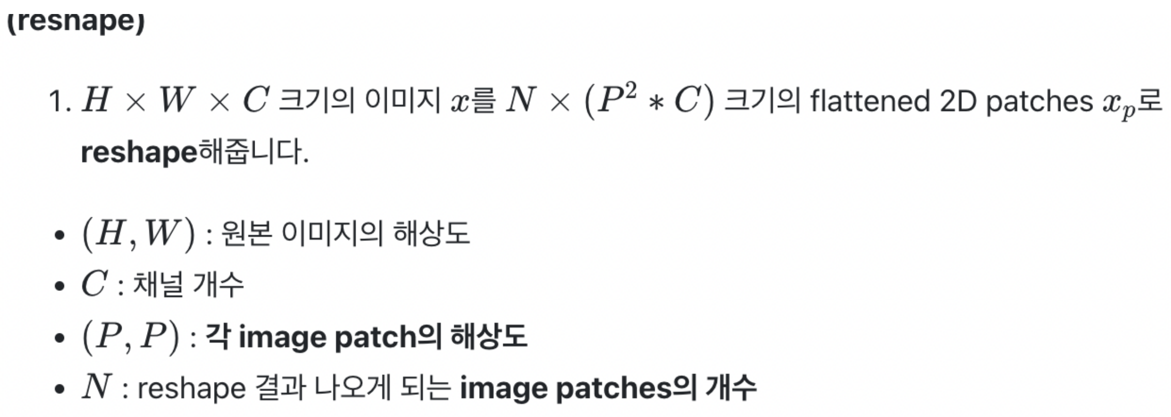


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable “classification token” to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).



(reshape)

1. $H \times W \times C$ 크기의 이미지 x를 $N \times (P^2 * C)$ 크기의 flattened 2D patches x_p로 **reshape**해줍니다.

- (H, W) : 원본 이미지의 해상도
- C : 채널 개수
- (P, P) : **각 image patch의 해상도**
- N : reshape 결과 나오게 되는 **image patches의 개수**

Reshape → projection→ class embeddings→ positional embeddings

Fine-Tuning and Higher Resolution

NLP에서 했던 것처럼 **ViT를 large dataset에 pre-trained**한 다음, **down stream tasks**에 **fine-tuning**을 진행합니다.이런 down stream task에 적용하기 위해서 pre-train 할 때에 **prediction head**를 없애고, *D*×*K*의 feed forward layer로 변경을 합니다.이 때 *K*는 downstream task의 class 개수입니다.

다만, 이는 pre-trained할 때의 이미지 해상도보다, 고해상도로 **down-stream task에 fine-tuning**할 때 효과적입니다.

Experiment

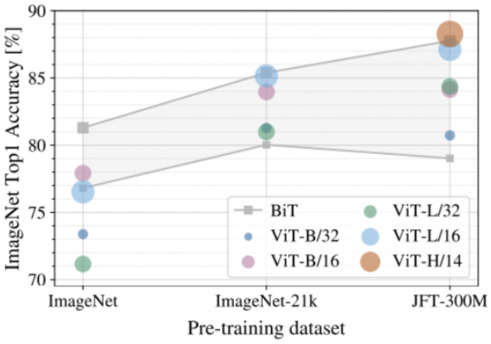


Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

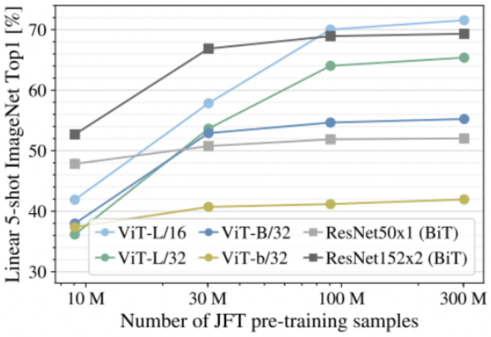


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

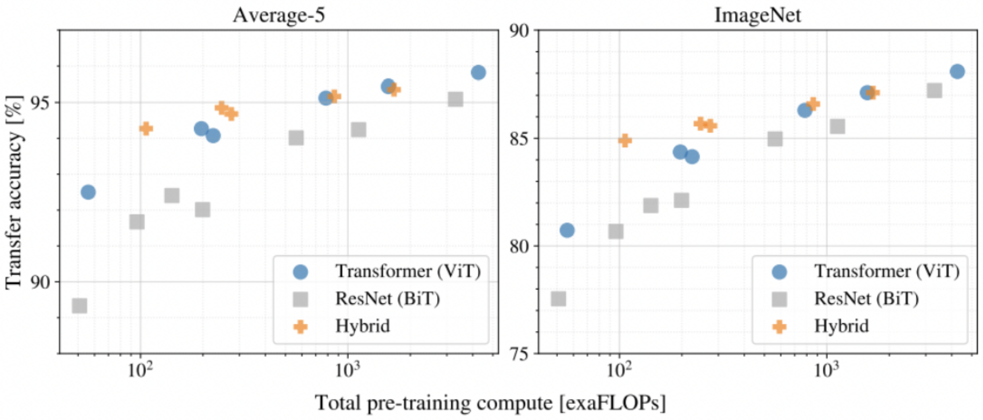


Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger models.

3D human pose estimation in video with temporal convolutions andsemi-supervised training

Abstract

Unlabeled video data를 이용한 semi-supervised learning을 이용했습니다. unlabeled video에서 예측한 2d keypoints를 이용하여 3D pose를 예측하고 이를 다시 2d keypoint로 back-project합니다.

목적 : input 영상 → 2d포즈를 3d 포즈로 변환

2d key points를 이용하여 dilated temporal convolutional model을 이용해서 3d 포즈로 변환. CNN을 사용하여 병렬 작용

contiribution

1. 2차원 trajectory에서 dilated temporal convolution에 기반하여 3d 포즈 예측
2. 레이블이 없을 때도 예측 가능/레이블이 부족해도 가능. semi-supervised learning 사용

cnn 모델

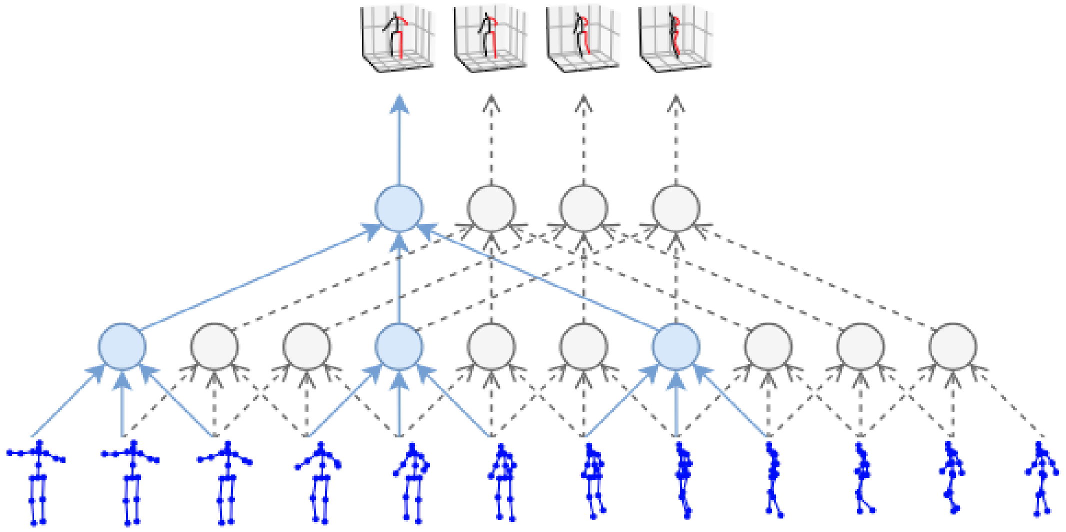


Figure 1: Our temporal convolutional model takes 2D keypoint sequences (bottom) as input and generates 3D pose estimates as output (top). We employ dilated temporal convolutions to capture long-term information.

Temporal Dilated Convolutional Model

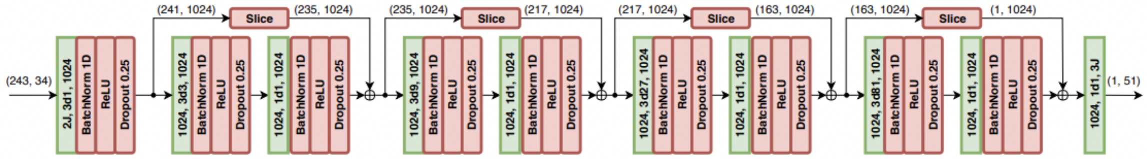


Figure 2: An instantiation of our fully-convolutional 3D pose estimation architecture. The input consists of 2D keypoints for a recpetive field of 243 frames ($B = 4$ blocks) with $J = 17$ joints. Convolutional layers are in green where `2J, 3d1, 1024` denotes $2 \cdot J$ input channels, kernels of size 3 with dilation 1, and 1024 output channels. We also show tensor sizes in parentheses for a sample 1-frame prediction, where `(243, 34)` denotes 243 frames and 34 channels. Due to valid convolutions, we slice the residuals (left and right, symmetrically) to match the shape of subsequent tensors.

초록: cnn 설명

빨강: batch norm, rectified linear units, dropouts

first input: 234 frame, 34 channel (17 joints *(x,y) coord(2d))

final output: 3D joint Coord (17*3(3d))

시나리오

이전시점, 미래시점을 모두 사용해서 학습 - symmetric convolutions

test에는 causal convolution사용 이전시점만 사용

Semi-supervised Approach

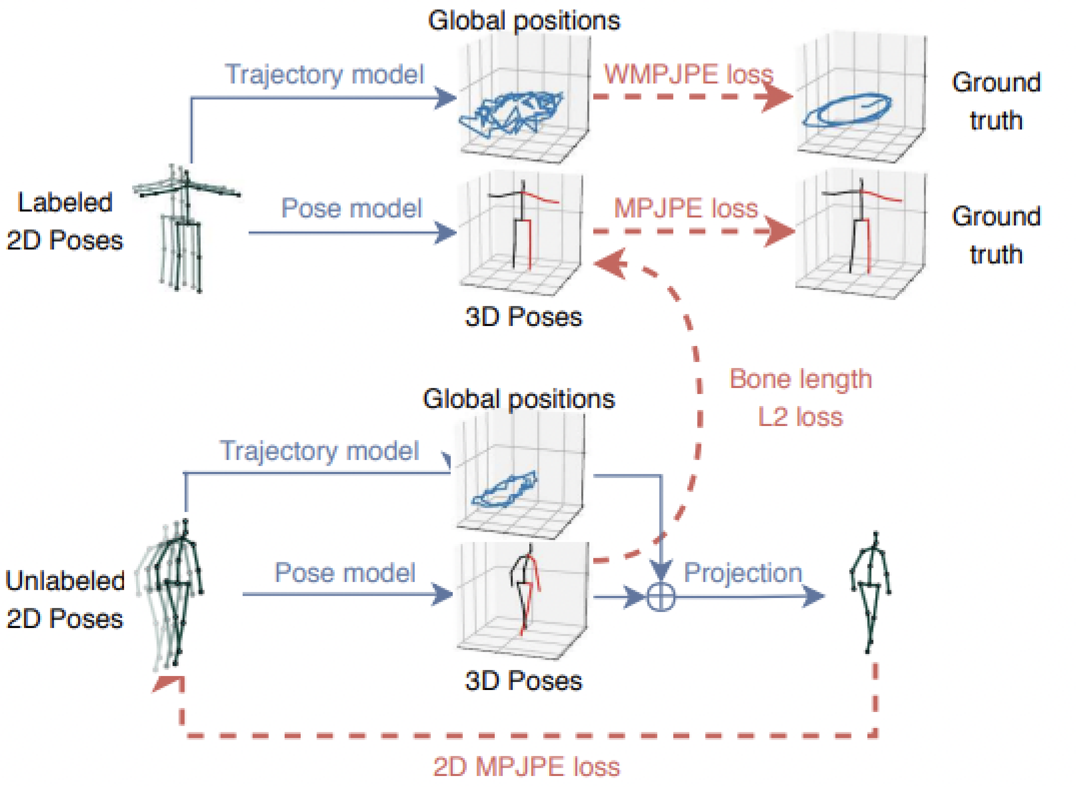


Figure 3: Semi-supervised training with a 3D pose model that takes a sequence of possibly predicted 2D poses as input. We regress the 3D trajectory of the person and add a soft-constraint to match the mean bone lengths of the unlabeled predictions to the labeled ones. Everything is trained jointly. WMPJPE stands for "Weighted MPJPE".

supervised:

input에 라벨된 데이터.

ground truth 에는 MPJPE loss 모델 사용 - 매칭되는 조인트의 평균값을 이용해서 로스 추출

trajectory model을 이용해서 3d 포즈모델을 예측할 때는 2디 포즈 시퀀스를 이용해서 사람의 움직임을 나타내고 그것에 대한 ground truth를 비교해서 WMPJPE loss계산

$$E = \frac{1}{y_z} ||f(x) - y||$$

unsupervised: 오토인코더

포즈 모델을 이용해서 3d pose를 생성하고 projection을 거쳐서 2d pose를 다시 생성. 중간 3d 포즈가 latency가 된다. 이걸 3d pose latency vector로 만들기 위해 bone length L2 loss를 사용(사람 관절 길이 조절)

trajectory모델에서 2d 포즈를 이용해서 3d trajectory를 만드는 네트워크.. unsupervised를 좀 더 학습잘하기 위해 pose model만 학습이 아니라 trajectory를 통해 한 번 더 한다.

Experiment

dataset: human3.6M, humanEva-I

2d pose estimation: mast R-CNN with ResNet-100, CPN with ResNet-50

3d pose estimation

- 프로토콜

Protocol 1 is the mean per-joint position error (MPJPE) in millimeters which is the mean Euclidean distance between predicted joint positions and ground-truth joint positions and follows

Protocol2 reports the error after alignment with the ground truth in translation, rotation, and scale (P-MPJPE)

Protocol 3 aligns predicted poses with the ground-truth only in scale (N-MPJPE) following for semi-supervised experiments.

- 결과

	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	A
Pavlakos *et al.* [41] CVPR'17 (∗)	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	7
Tekin *et al.* [52] ICCV'17	54.2	61.4	60.2	61.2	79.4	78.3	63.1	81.6	70.1	107.3	69.3	70.3	74.3	51.8	63.2	6
Martinez *et al.* [34] ICCV'17 (∗)	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	6
Sun *et al.* [50] ICCV'17 (+)	52.8	54.8	54.2	54.3	61.8	67.2	53.1	53.6	71.7	86.7	61.5	53.4	61.6	47.1	53.4	5
Fang *et al.* [10] AAAI'18	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	6
Pavlakos *et al.* [40] CVPR'18 (+)	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	5
Yang *et al.* [56] CVPR'18 (+)	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	43.6	60.1	47.7	5
Luvizon *et al.* [33] CVPR'18 (∗)(+)	49.2	51.6	47.6	50.5	51.8	60.3	48.5	51.7	61.5	70.9	53.7	48.9	57.9	44.4	48.9	5
Hossain & Little [16] ECCV'18 (†)(∗)	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	5
Lee *et al.* [27] ECCV'18 (†)(∗)	**40.2**	49.2	47.8	52.6	50.1	75.0	50.2	**43.0**	**55.8**	73.9	54.1	55.6	58.2	43.3	43.3	5
Ours, single-frame	47.1	50.6	49.0	51.8	53.6	61.4	49.4	47.4	59.3	67.4	52.4	49.5	55.3	39.5	42.7	5
Ours, 243 frames, causal conv. (†)	45.9	48.5	44.3	47.8	51.9	57.8	46.2	45.6	59.9	68.5	50.6	46.4	51.0	34.5	35.4	4
Ours, 243 frames, full conv. (†)	45.2	**46.7**	43.3	**45.6**	**48.1**	**55.1**	44.6	44.3	57.3	**65.8**	**47.1**	**44.0**	**49.0**	32.8	**33.9**	**4**
Ours, 243 frames, full conv. (†)(∗)	45.1	47.4	**42.0**	46.0	49.1	56.7	**44.5**	44.4	57.2	66.1	47.5	44.8	49.2	**32.6**	34.0	4

(a) Protocol 1: reconstruction error (MPJPE).

	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	A
Martinez *et al.* [34] ICCV'17 (∗)	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	4
Sun *et al.* [50] ICCV'17 (+)	42.1	44.3	45.0	45.4	51.5	53.0	43.2	41.3	59.3	73.3	51.0	44.0	48.0	38.3	44.8	4
Fang *et al.* [10] AAAI'18	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	4
Pavlakos *et al.* [40] CVPR'18 (+)	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	4
Yang *et al.* [56] CVPR'18 (+)	**26.9**	**30.9**	36.3	39.9	43.9	47.4	**28.8**	**29.4**	**36.9**	58.4	41.5	**30.5**	**29.5**	42.5	32.2	3
Hossain & Little [16] ECCV'18 (†)(∗)	35.7	39.3	44.6	43.0	47.2	54.0	38.3	37.5	51.6	61.3	46.5	41.4	47.3	34.2	39.4	4
Ours, single-frame	36.0	38.7	38.0	41.7	40.1	45.9	37.1	35.4	46.8	53.4	41.4	36.9	43.1	30.3	34.8	4
Ours, 243 frames, causal conv. (†)	35.1	37.7	36.1	38.8	38.5	44.7	35.4	34.7	46.7	53.9	39.6	35.4	39.4	27.3	28.6	3
Ours, 243 frames, full conv. (†)	34.1	36.1	34.4	**37.2**	**36.4**	**42.2**	34.4	33.6	45.0	**52.5**	**37.4**	33.8	37.8	**25.6**	**27.3**	**3**
Ours, 243 frames, full conv. (†)(∗)	34.2	36.8	**33.9**	37.5	37.1	43.2	34.4	33.5	45.3	52.7	37.7	34.1	38.0	25.8	27.7	3

(b) Protocol 2: reconstruction error after rigid alignment with the ground truth (P-MPJPE), where available.

Table 1: Reconstruction error on Human3.6M. **Legend:** (†) uses temporal information. (∗) ground-truth bounding box (+) extra data – [50, 40, 56, 33] use 2D annotations from the MPII dataset, [40] uses additional data from the Leeds Spo Pose (LSP) dataset as well as ordinal annotations. [50, 33] evaluate every 64th frame. [16] provided us with corrected resu over the originally published results [3]. Lower is better, best in bold, second best underlined.

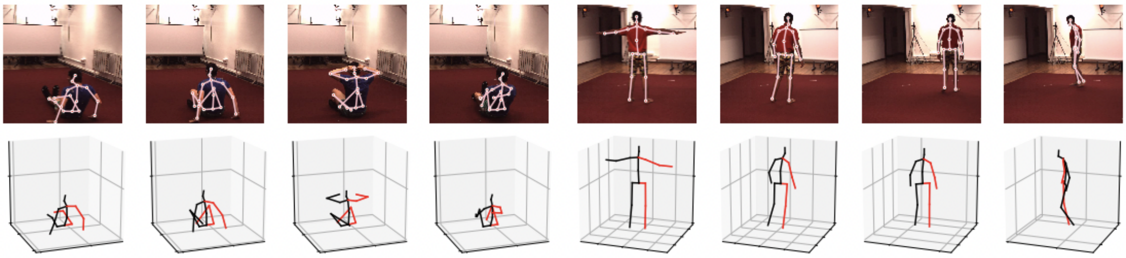
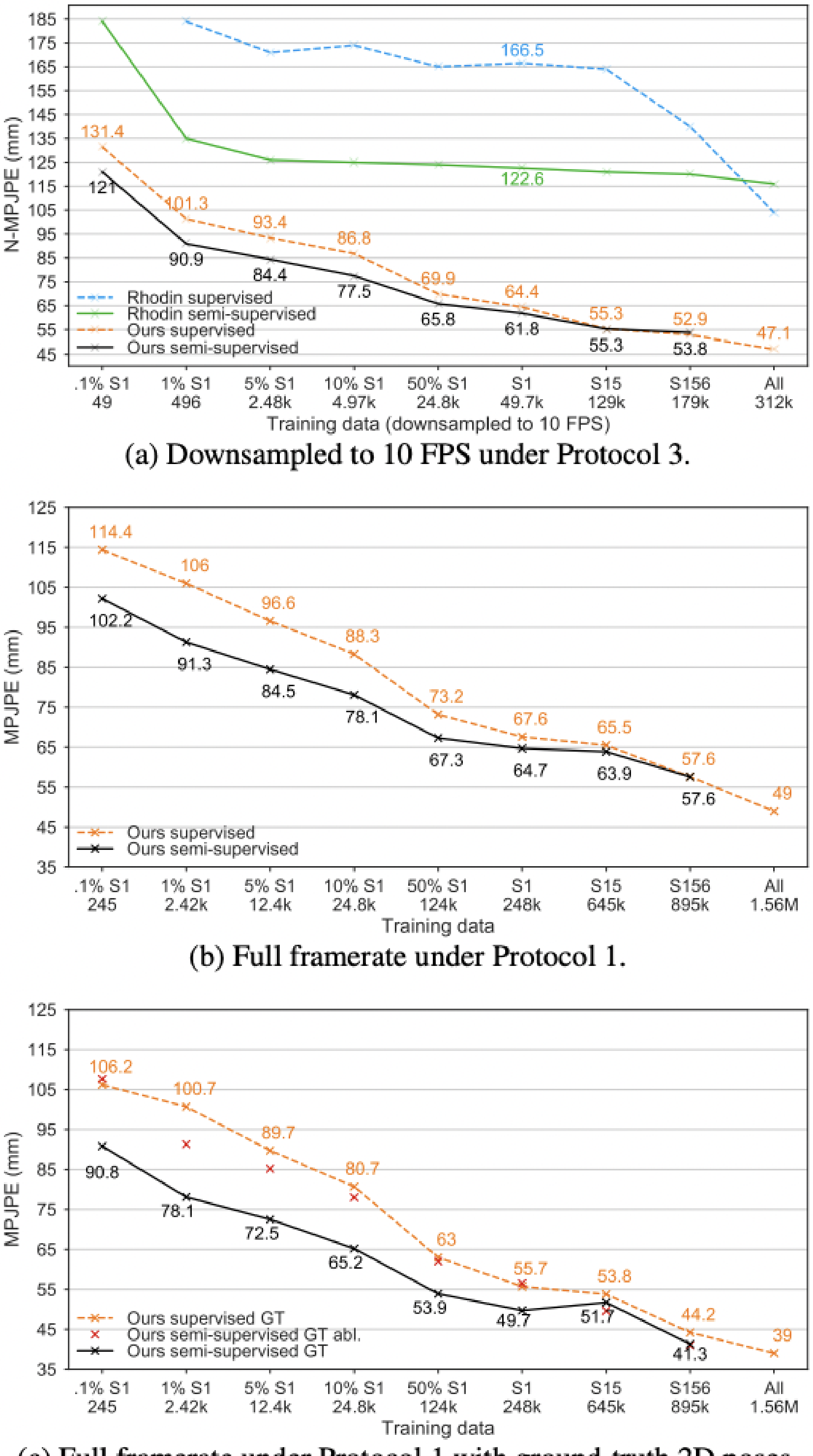


Figure 4: Qualitative results for two videos. **Top:** video frames with 2D pose overlay. **Bottom:** 3D reconstruction.

	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Single-frame	12.8	12.6	10.3	14.2	10.2	11.3	11.8	11.3	8.2	10.2	10.3	11.3	13.1	13.4	12.9	11.6
Temporal	3.0	3.1	2.2	3.4	2.3	2.7	2.7	3.1	2.1	2.9	2.3	2.4	3.7	3.1	2.8	2.8

Table 2: Velocity error over the 3D poses generated by a convolutional model that considers time and a single-frame baseline.



(a) Downsampled to 10 FPS under Protocol 3.

(b) Full framerate under Protocol 1.

(c) Full framerate under Protocol 1 with ground-truth 2D poses.

Figure 5: **Top:** comparison with [45] on *Protocol 3*, using a downsampled version of the dataset for consistency. **Middle:** our method under *Protocol 1* (full frame rate). **Bottom:** our method under *Protocol 1* when trained on ground-truth 2D poses (full frame rate). The small crosses ("abl." series) denote the ablation of the bone length term.

1 - 프로토콜3을 기반으로 다운샘플링. 다른 모델에 비해 로스값이 낮게 나옴을 확인