[3D human pose estimation in video with temporal convolutions and semi-supervised training]

- 2차원 키포인트 궤적(trajectory)에서 dilated temporal convolution을 기반으로 비디오에서 3차원 인간 자세 추정을 위한 간단하고 효율적인 접근 방식을 제시하였다는 것

- 이 모델은 계산 복잡성과 모델 매개 변수 수 측면에서 동일한 수준의 정확도에서 RNN 기반 모델보다 더 효율적임을 보여주었음.

- 레이블이 없는 비디오를 활용하고, 레이블이 있는 데이터가 부족할 때 효과적인 semi-supervised 방식을 도입한 것

- 이전의 semi-supervised 접근 방식과 비교할 때 Ground Truth 2D annotation 또는 Camera Intrinsic Prameter 가 있는 multi-view image 보다는 extrinsic camera parameter 만 필요함.

- SOTA 와 비교했을 때 본 논문의 접근 방식은 supervised 및 semi-supervised 방식 모두에서 가장 우수함

- 우리의 supervised model은 학습을 위해 추가로 레이블이 지정된 데이터를 사용하더라도 다른 모델보다 성능이 뛰어남

1. Temporal dilated convolution model

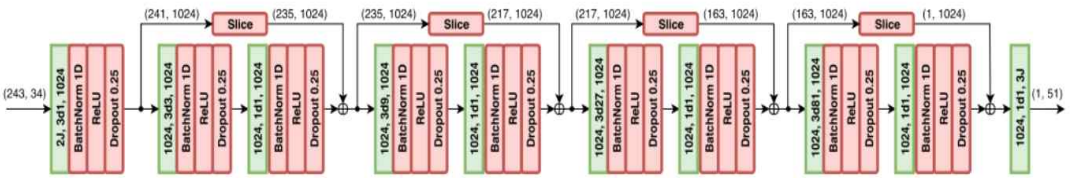


Figure 2: An instantiation of our fully-convolutional 3D pose estimation architecture. The input consists of 2D keypoints for a recpetive field of 243 frames ($B = 4$ blocks) with $J = 17$ joints. Convolutional layers are in green where `2J, 3d1, 1024` denotes $2 \cdot J$ input channels, kernels of size 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.

- input data : 243 (frame) * 34 (17 joints * 2 dim (x, y))
- 4 residual blocks, 0.25 dropout, 243 frames, filter size 3, output feature 1024

2. Semi-supervised approach

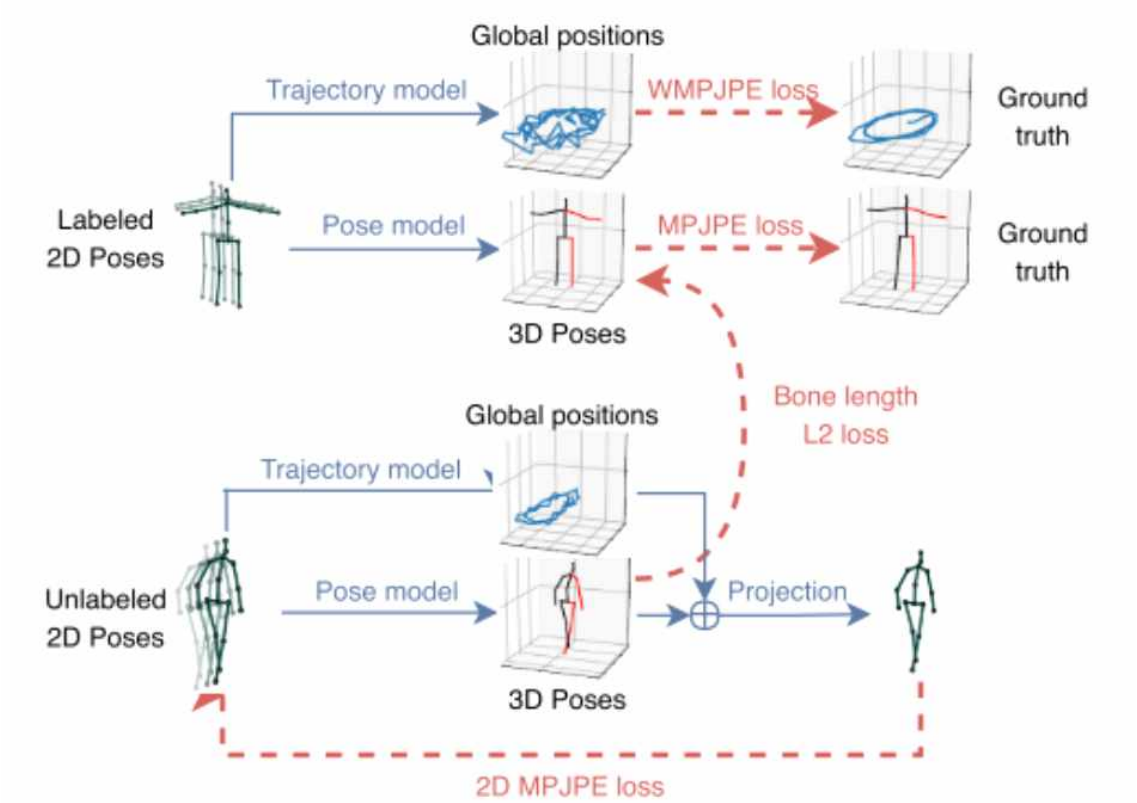


Figure 3: Semi-supervised training with a 3D pose model that takes a sequence of possibly predicted 2D poses as input. We regress the 3D trajectory of the person and add a soft-constraint to match the mean bone lengths of the unlabeled predictions to the labeled ones. Everything is trained jointly. WMPJPE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

- Supervised, Unsupervised loss를 모두 계산하고 동시에 최적화

- Supervised loss : Ground truth 3d joint 활용

- Unsupervised loss
  : Autoencoder 문제로 접근
  : encoder : 3d pose estimator
  ; 3d joint가 다시 projected back 되었을 때 reconstruction loss 사용
  : bone length를 L2 loss로 추가

- Trajectory model
  : 2d pose를 활용해 3d trajectory를 생성하는 네트워크
  : 2d -> 3d mapping을 위해 trajectory 추가로 활용
  : unlabeled data를 back projection할 때 3d trajectory까지 고려해서 reconstruct
  : back projection이 올바르게 작동 가능

- loss function (supervised loss)
  : 3d ground truth와 MPJPE 계산
  : global trajectory loss
    # camera에서 ground truth depth의 역수를 취한 값을 가중치로 사용
    # Weighted Mean Per-Joint Position Error (WMPJPE) 사용

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

3. Dataset and evaluation

- Dataset : Human 3.6M, HumanEva-I

- Evaluation (3 protocols)

: Protocol 1
밀리미터(mm) 단위의 평균 관절 위치 오차(MPJPE, mean per-joint position error)로 예상 관절 위치와 GT 관절 위치 사이의 평균 유클리디안 거리(Euclidean distance)를 사용한다.

: Protocol 2
변환(translation), 회전(rotation), 스케일(scale) 즉 P-MPJPE 에서 GT와 정렬한 후의 오류를 사용한다.

: Protocol 3
semi-supervised 실험을 위해 예측된 자세를 스케일 관점에서만 GT와 일치시킨 것(N-MPJPE)을 사용한다.

- 2d pose estimation : Mask R-CNN, Cascaded pyramid network

4. Result

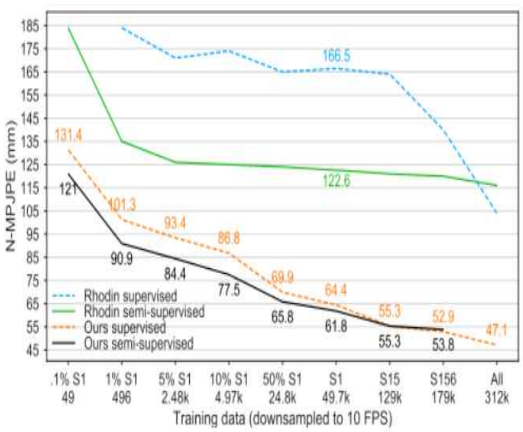
	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Pavlakos *et al.* [41] CVPR'17 (∗)	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Tekin *et al.* [52] ICCV'17	54.2	61.4	60.2	61.2	79.4	78.3	63.1	81.6	70.1	107.3	69.3	70.3	74.3	51.8	63.2	69.7
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	62.9
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	59.1
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	60.4
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	56.2
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	58.6
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	53.2
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	58.3
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	52.8
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	51.8
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	49.0
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**	**46.8**
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	47.1

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

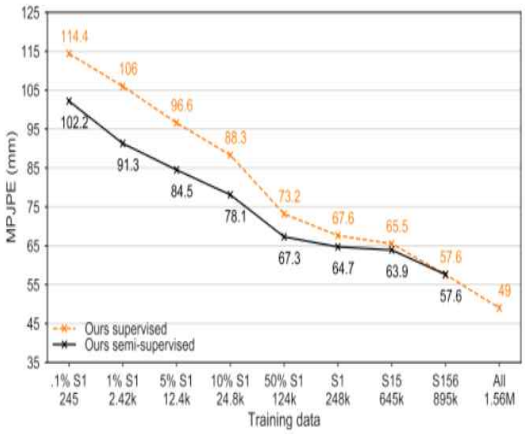
	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Martinez *et al.* [34] ICCV'17 (∗)	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Sun *et al.* [50] ICCV'17 (+)	42.1	44.3	45.0	45.4	51.5	53.0	43.2	41.3	59.3	73.3	51.0	44.0	48.0	38.3	44.8	48.3
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	45.7
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	41.8
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	37.7
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	44.1
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	40.0
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	38.1
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**	**36.5**
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	36.8

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

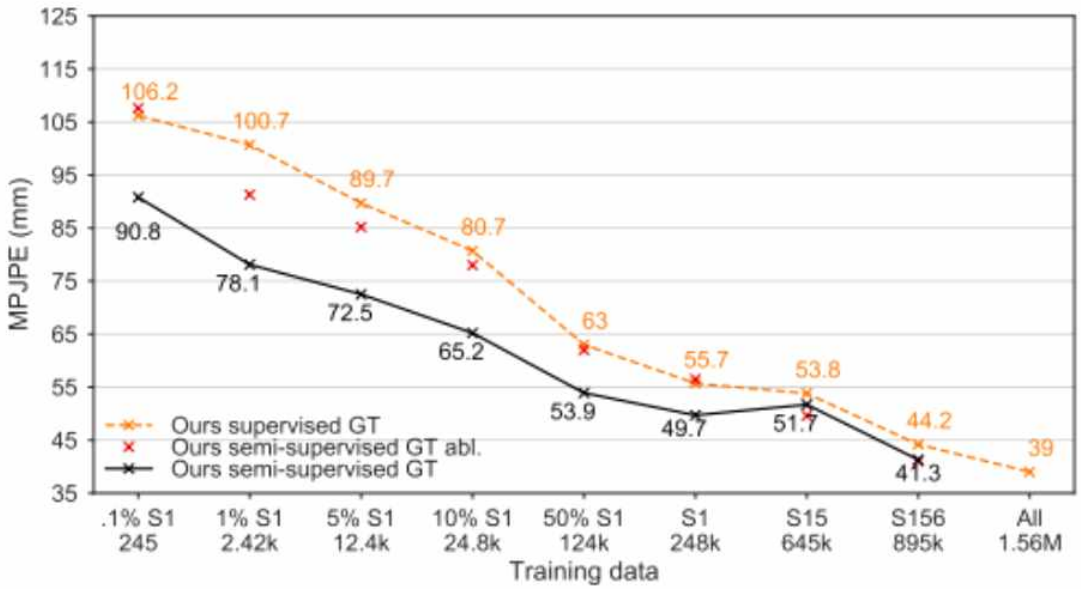
Table 1: Reconstruction error on Human3.6M. **Legend:** (†) uses temporal information. (∗) ground-truth bounding boxes. (+) extra data – [50, 40, 56, 33] use 2D annotations from the MPII dataset, [40] uses additional data from the Leeds Sports Pose (LSP) dataset as well as ordinal annotations. [50, 33]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



(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.