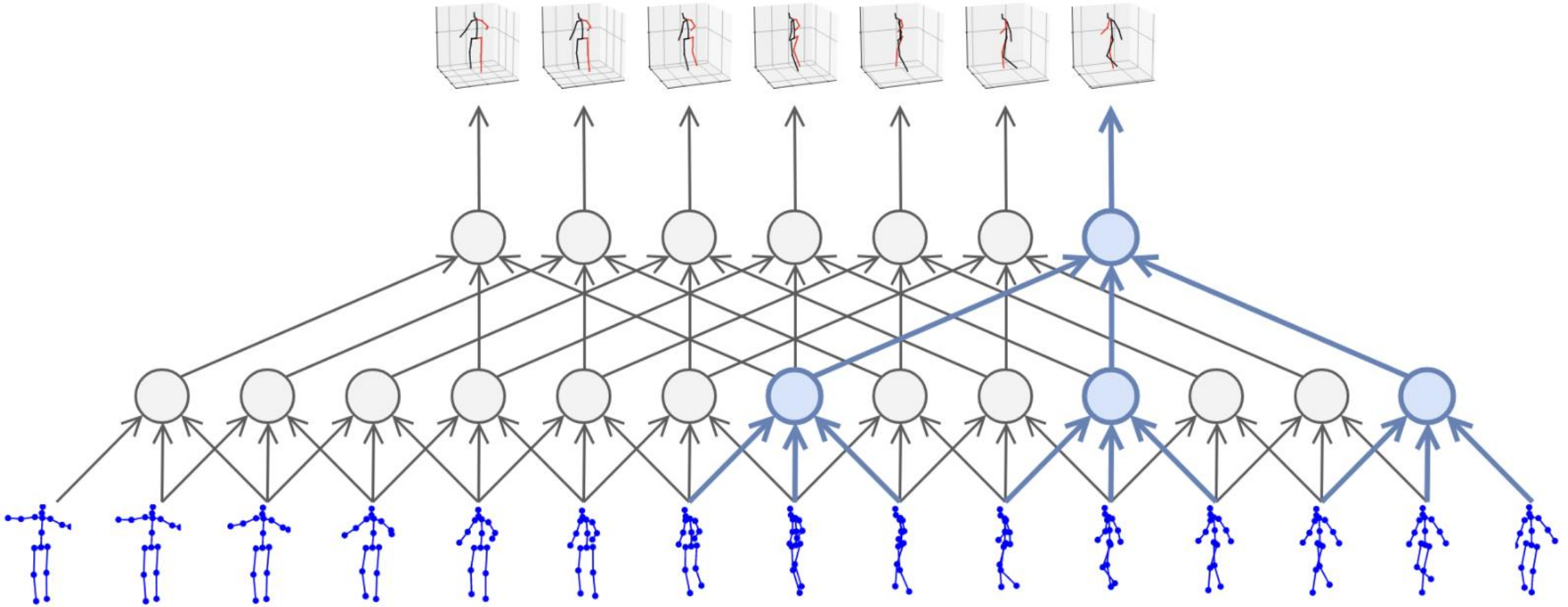


3D human pose estimation in video with temporal convolutions and semi-supervised training

In CVPR 2019

Facebook AI Research

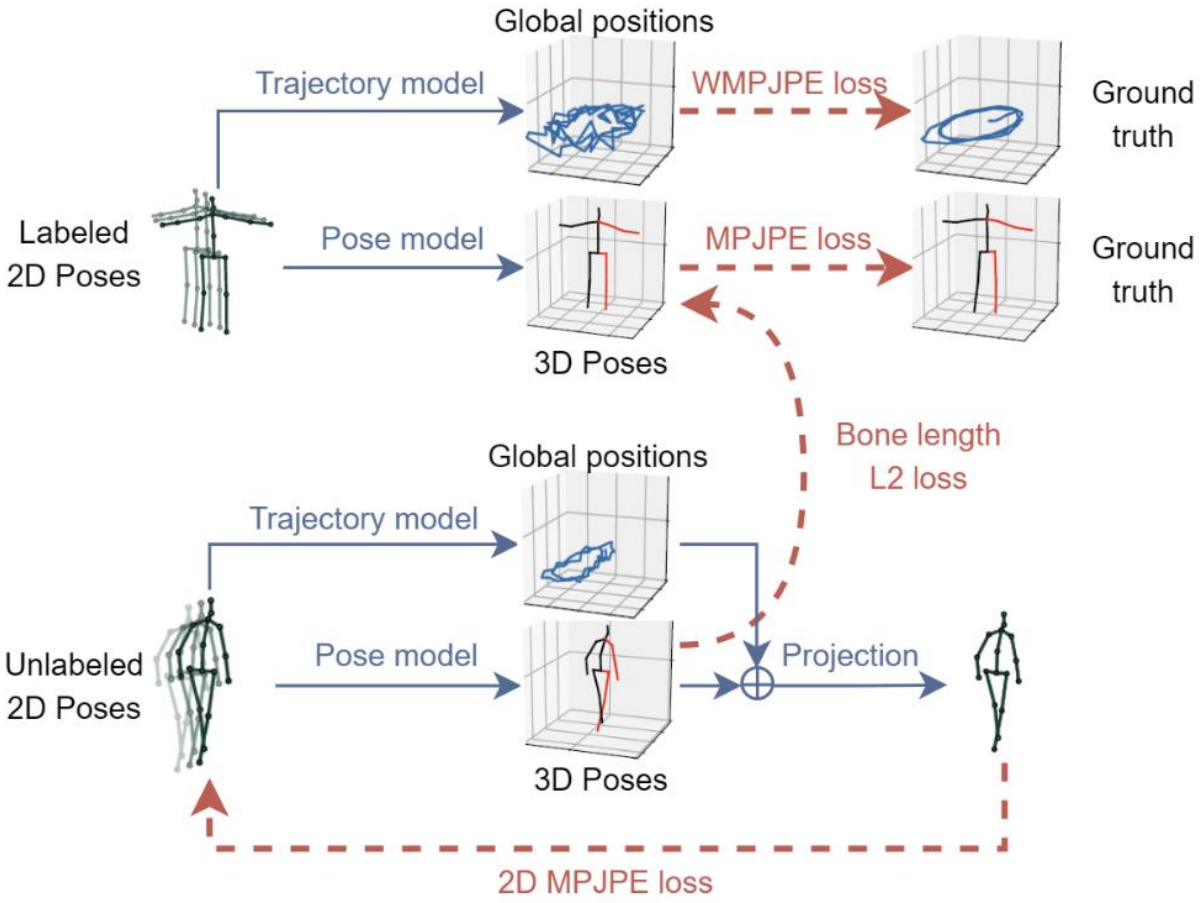
Contents

1. Purpose

2. background
   - 2-1. TCN
   - 2-2. casual network
   - 2-3. dilated TCN

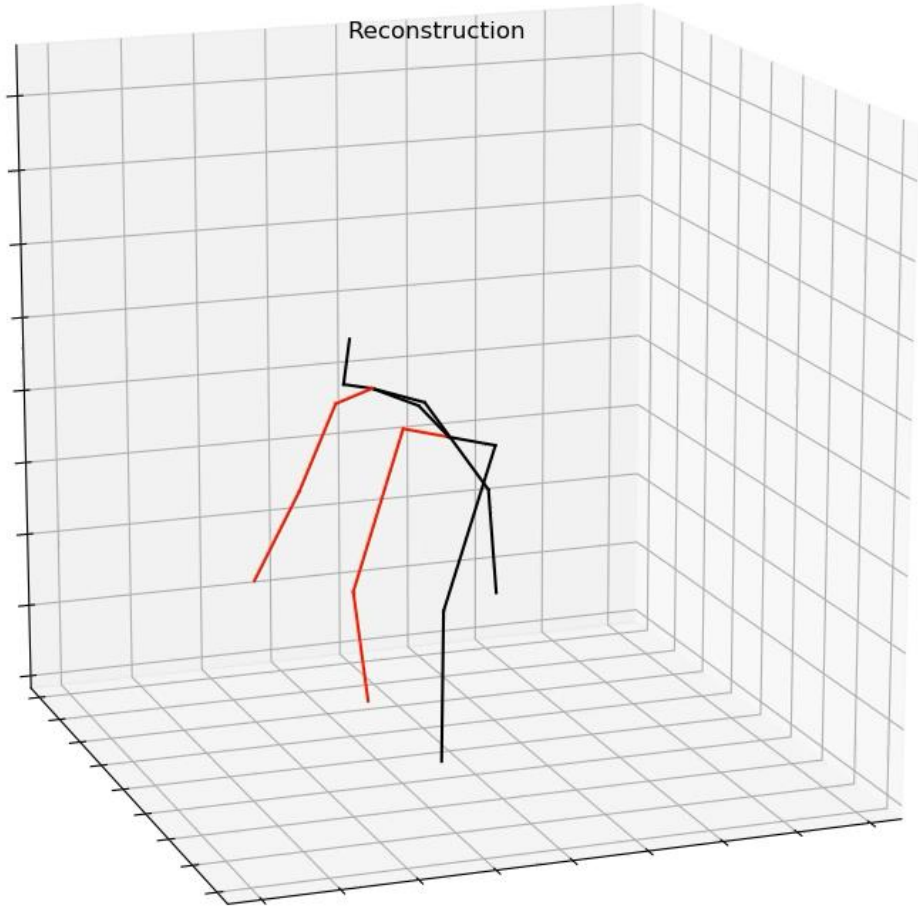
3. Total Flow
   - 3-1. Network
   - 3-2. semi-supervised

4. evaluation



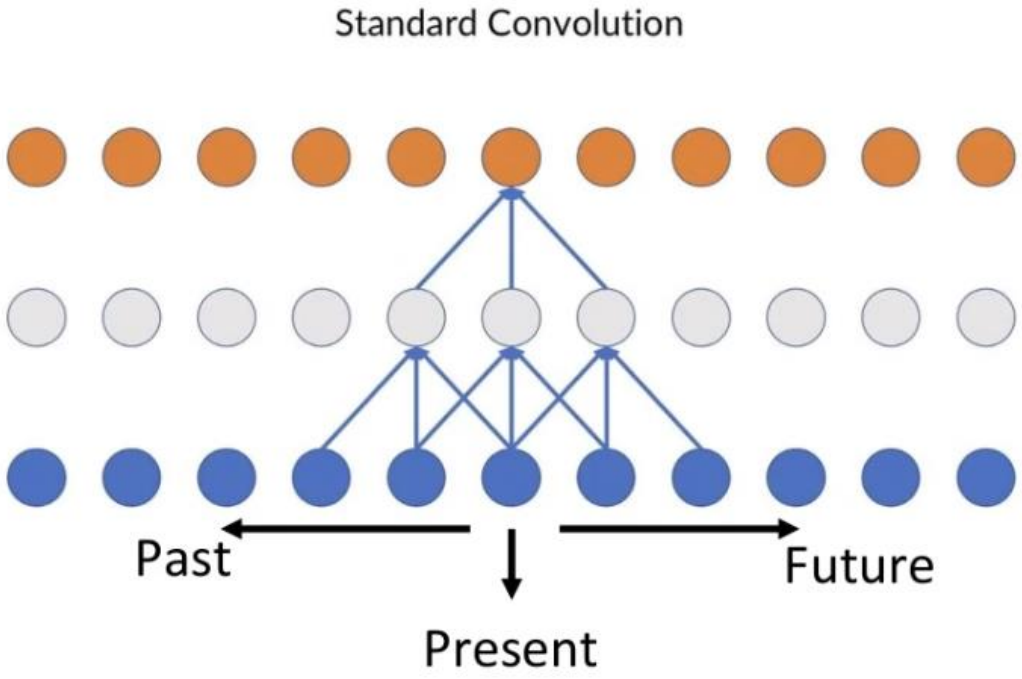
Purpose

3D human pose estimation in video

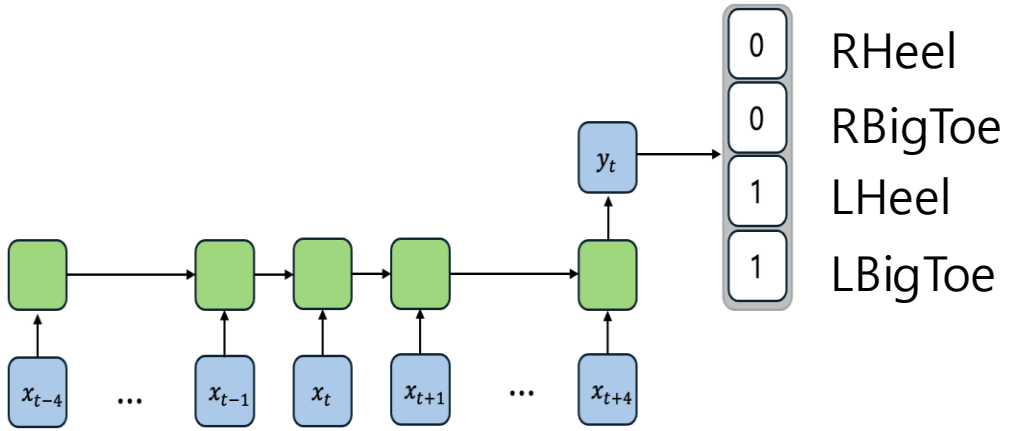


background – TCN

Temporal Convolution Network



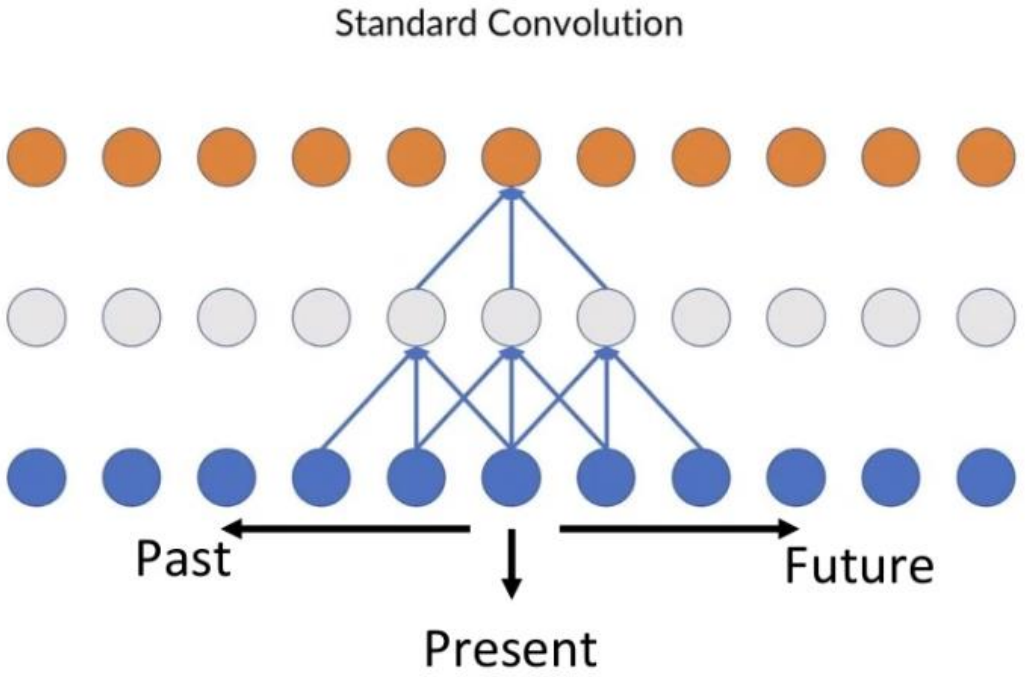
<TCN>



<RNN>

background – TCN

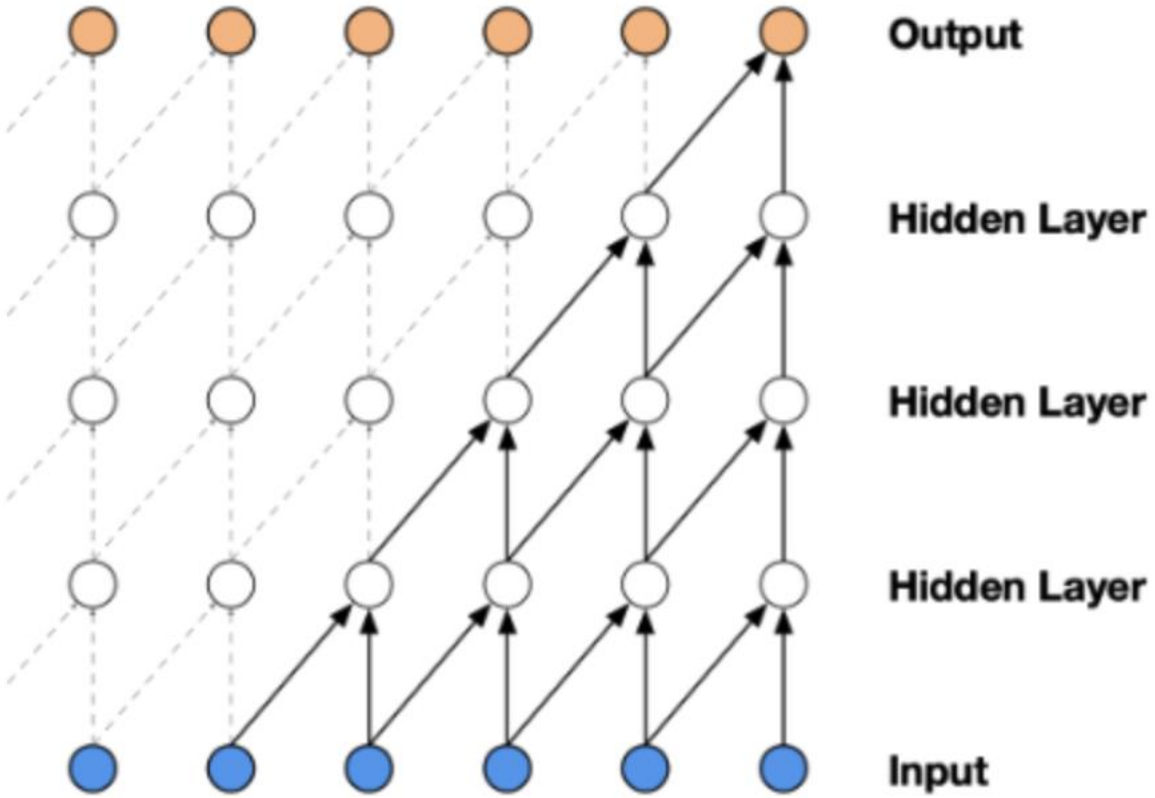
Temporal Convolution Network



<Advantages>

1. Parallelism
2. Flexible receptive field size
3. Stable gradients(exploding / vanishing gradient problem along the time axis)

background – Causal TCN



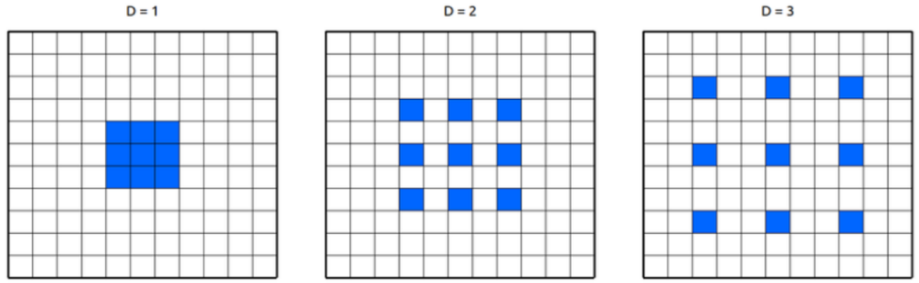
input sequence : $x_0, \dots, x_T$

output sequence : $y_0, \dots, y_T$

y_t <- input($x_i, \dots, x_t$)

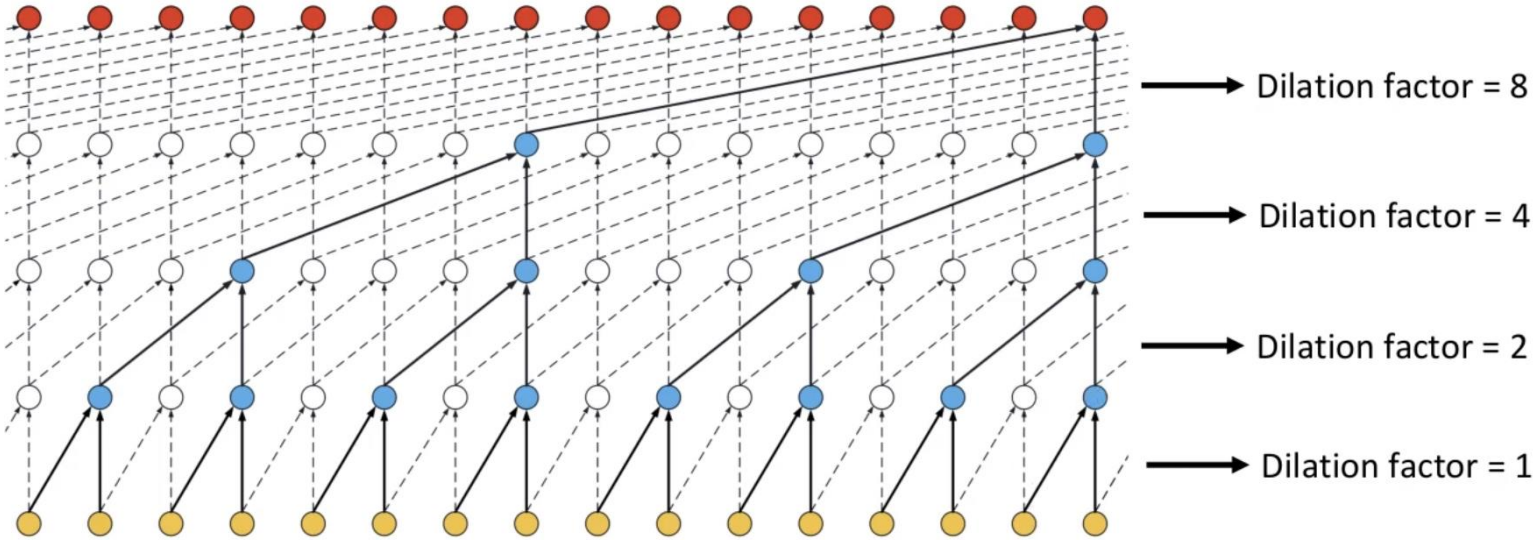
background – Dilated TCN

Dilated Convolution



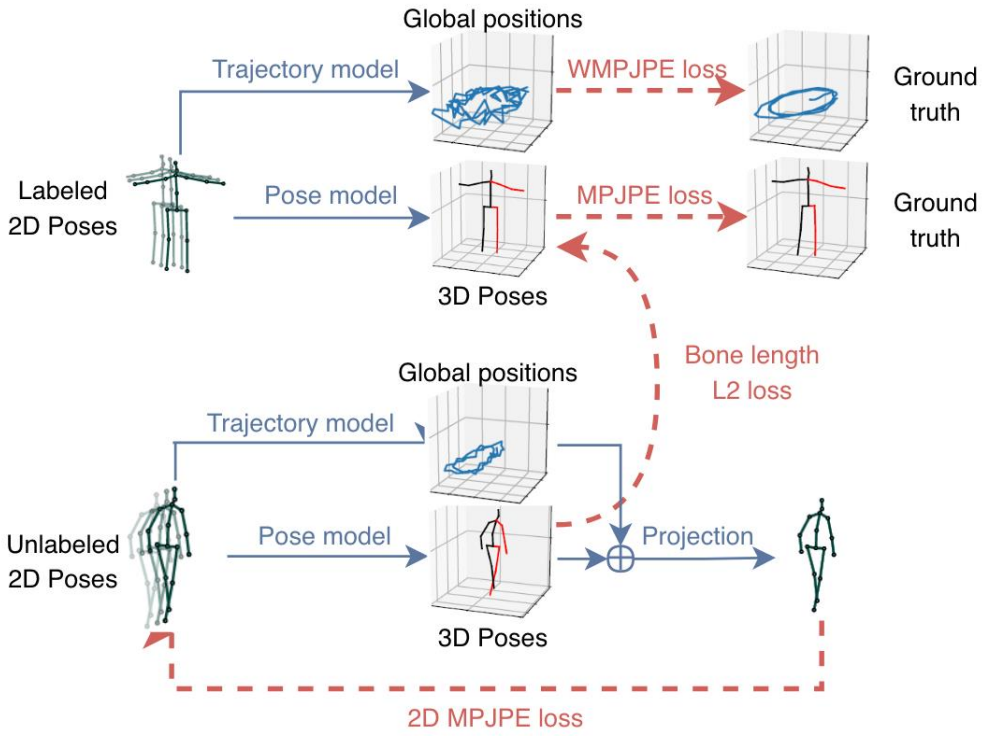
<Advantages>
An effective way to make the Receptive Field large -> increase in computation x

Dilated TCN



Total Flow

temporal convolutions and semi-supervised training



<data>

1. labeled 2D Poses <-> mocap data
2. Un-labeled 2D Poses

$$\text{Bone length L2 loss} = \sum_{j=1}^{J} \left(\|\mathbf{b}_j^{\text{pred}} - \mathbf{b}_j^{\text{gt}}\| \right)^2$$

$$\text{MPJPE} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{p}_i^{\text{pred}} - \mathbf{p}_i^{\text{gt}}\|$$

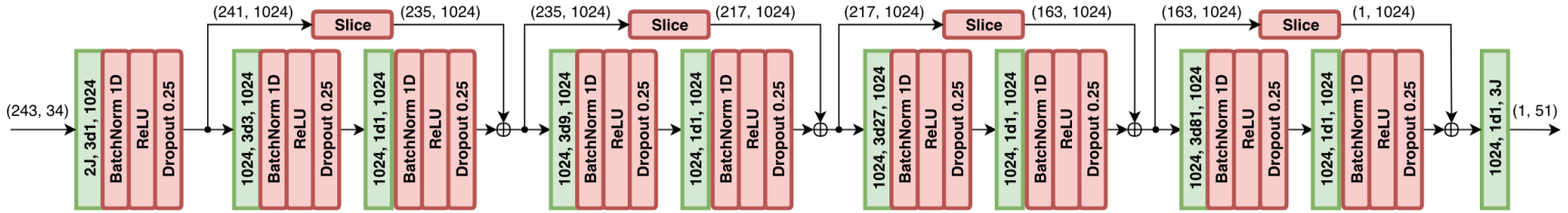
$$\text{WMPJPE} = \frac{1}{\mathbf{y}_z} \|f(\mathbf{x}) - \mathbf{y}\|$$

y_z: $gt\ depth\ in\ camera\ space$

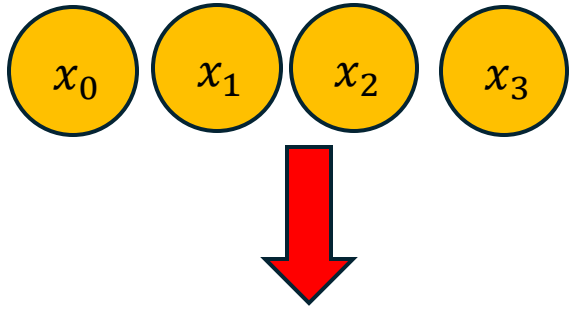
<each network>

The two networks have the same architecture but do not share any weights

Network – Temporal dilated convolutional model



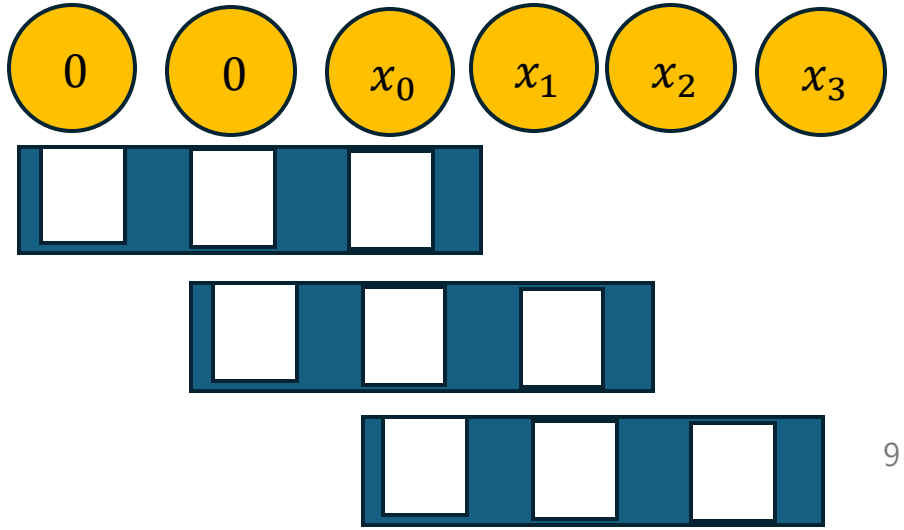
apply zero-padding, k = 3



input : (x, y) x joint x Timesteps

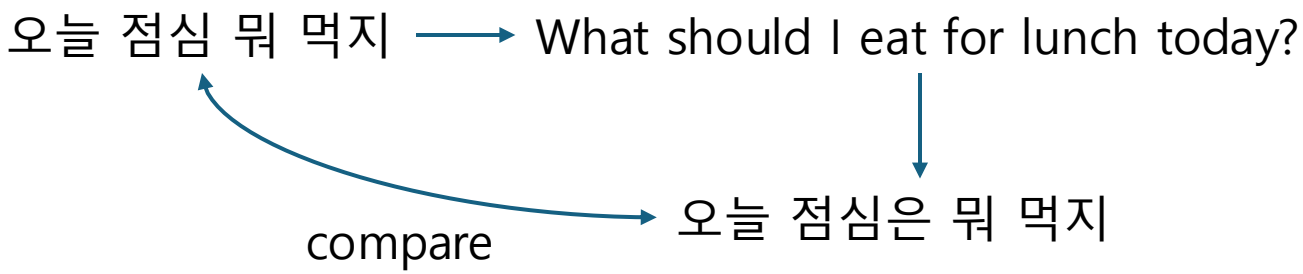
output : 3D poses for all frame

hyper-parameter)
kernel size : k
dilation factor : D



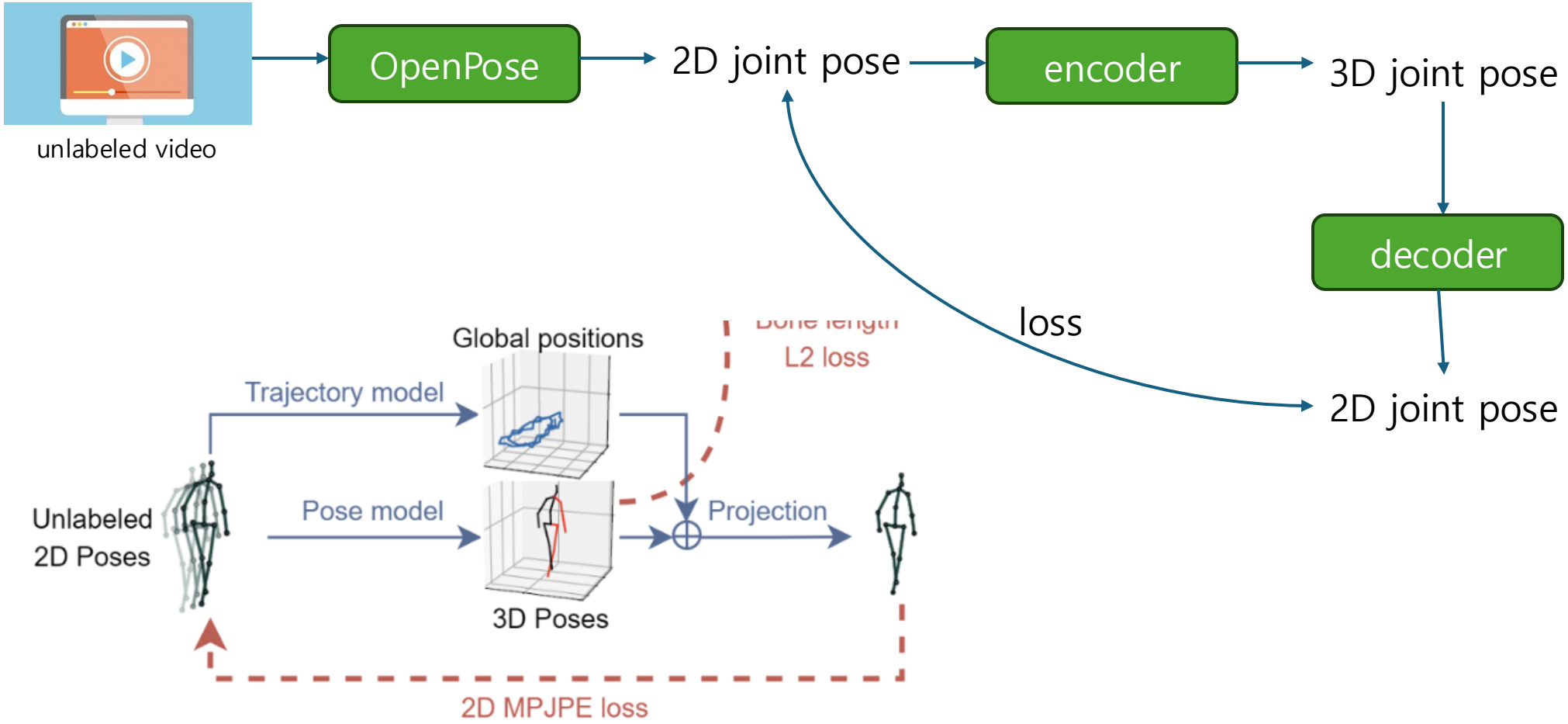
Network – Semi-supervised approach

main idea : cycle consistency



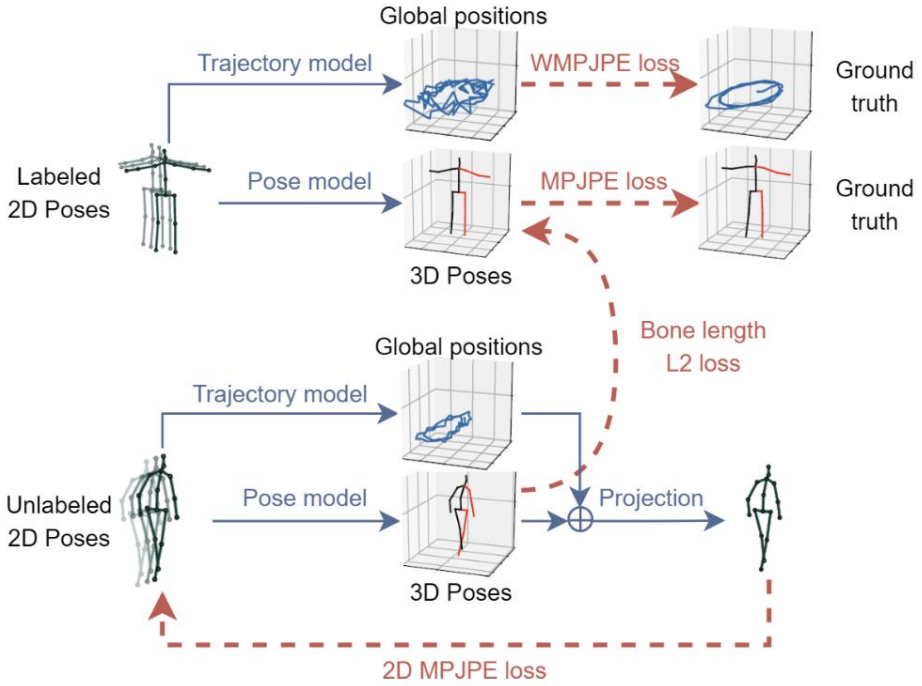
Cycle consistency)
maintaining its consistency when returning to its original state through a cyclical process

Network – Semi-supervised approach



Network – Trajectory

Trajectory model : For perspective projection



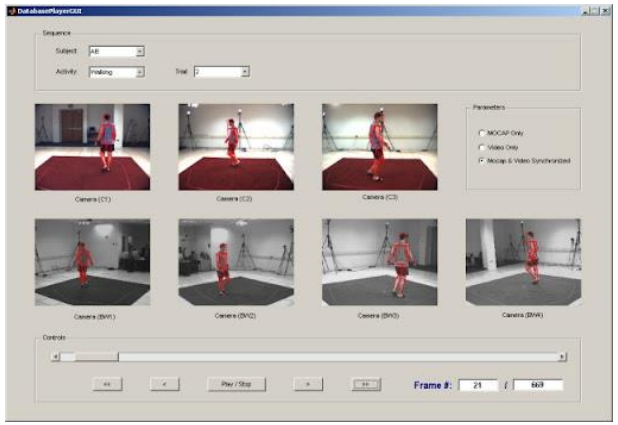
2D pose on the screen depends on…

1. the global position of the human root joint(trajectory)
2. 3D pose

<reason>

1. without global position, the subject would always be reprojected at the center of the screen with a fixed scale.

Evaluation – Datasets and Evaluation



Human3.6M

- 3.6 million video frames for 11 subjects,
of which seven are annotated with 3D poses

HumanEva-I

much smaller dataset, with three subjects recorded from three camera views

Evaluation – Datasets and Evaluation

	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).

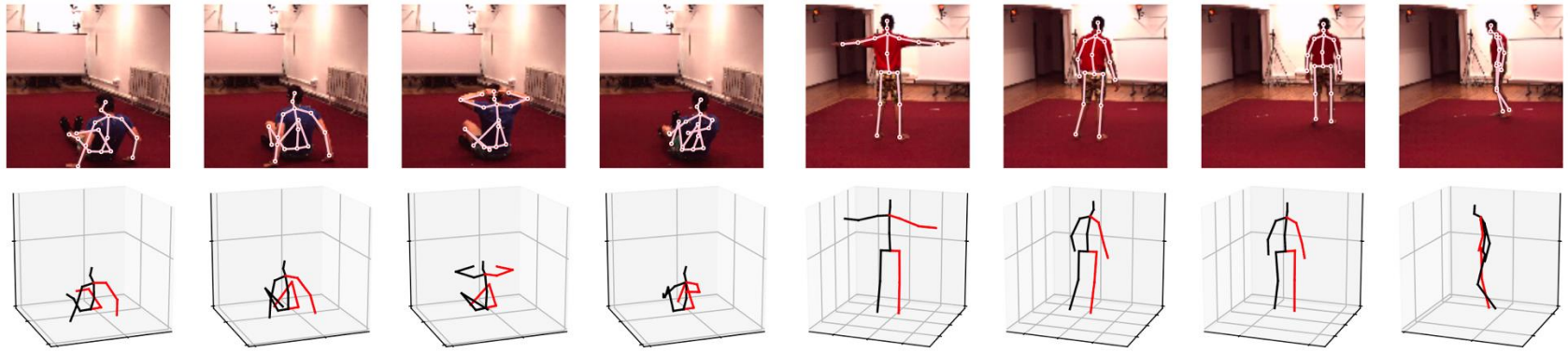


Figure 4: Qualitative results for two videos. **Top:** video frames with 2D pose overlay. **Bottom:** 3D reconstruction.