Through Wall Human Pose Estimation Using Radio Signals

Migmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, Dina Katabi,

CVPR ‘18

OBJECTIVE

Estimate a 2D skeletal representation of the joints on the arms and legs, and keypoints on the torso and head while with occlusions (like wall)

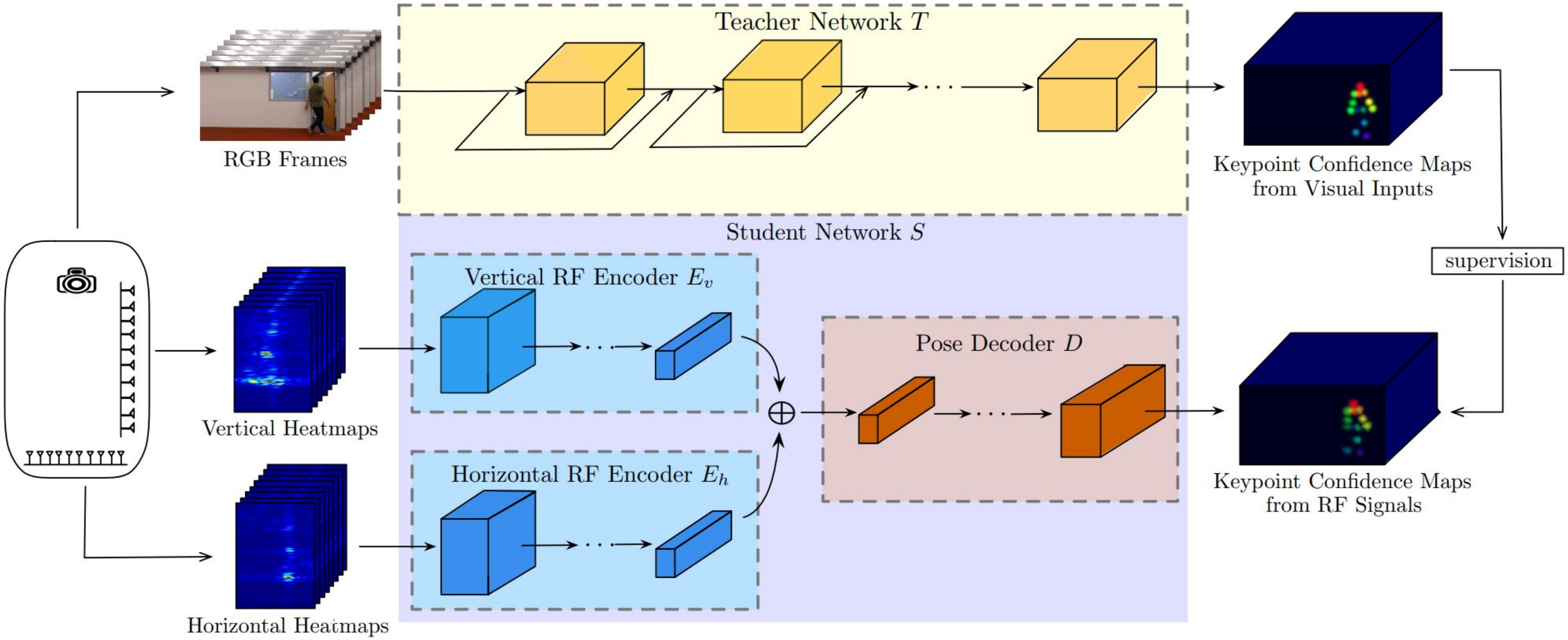
Motivation

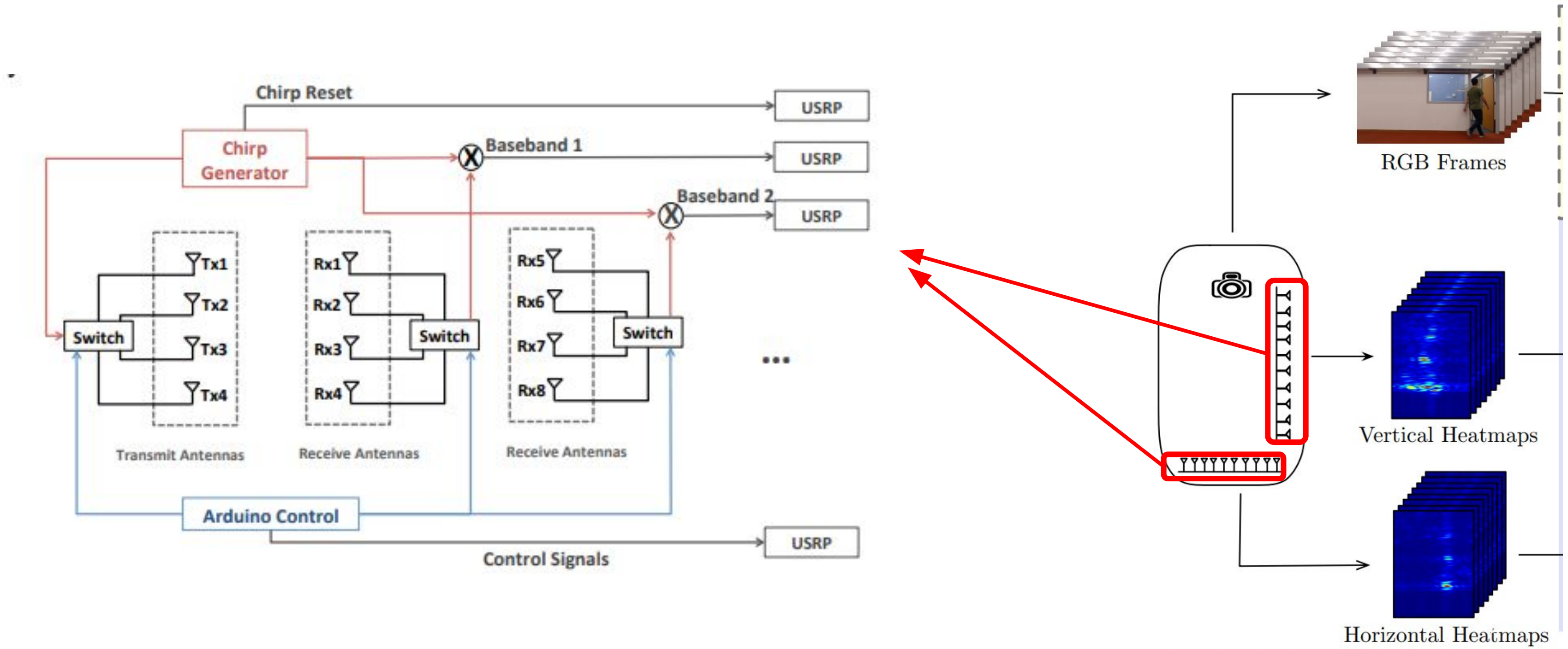
- Human Pose Estimation is an important task in Computer Vision
  - Surveillance
  - Activity Recognition
  - Gaming etc
- With camera occlusions are a big hindrance
- While RF signals can see through wall
  - 3D tracking via body radio reflections, Fadel Adib, 2014
  - Capturing the Human figure through a wall, Fadel Adib, 2015
  - Wfid: Passive device-free human identification using WiFi signal, F Hong, 2016

Related Work:

- Computer Vision:
  - Top-down: First detect people and then apply pose to each individual person
  - Bottom-up: First identify key-points and then group and associate them to form a person
- Wireless System:
  - High frequency based localization and people tracking : Uses mmWave, but fail to penetrate walls
  - Lower Frequency based: Uses GHz signals like WiFi to track and it can penetrate through walls
  - Device free tracking uses reflections to localize and track people

METHOD


Teacher Network T
RGB Frames
Keypoint Confidence Maps
from Visual Inputs
supervision
Student Network S
Vertical RF Encoder E_v
Vertical Heatmaps
Horizontal RF Encoder E_h
Horizontal Heatmaps
Pose Decoder D
Keypoint Confidence Maps
from RF Signals



Chirp Reset
USRP
Chirp Generator
Baseband 1
USRP
Baseband 2
USRP
Tx1
Tx2
Tx3
Tx4
Switch
Rx1
Rx2
Rx3
Rx4
Switch
Rx5
Rx6
Rx7
Rx8
Switch
Transmit Antennas
Receive Antennas
Receive Antennas
Arduino Control
USRP
Control Signals
RGB Frames
Vertical Heatmaps
Horizontal Heatmaps

Resolutions for the RF

10cm resolution in distance ⇒ 3GHz of Bandwidth

(They use 5.46 − 7.24 GHz ⇒ 2GHz)

15^{o} resolution in angle ⇒ 8 antenna in both horizontal and vertical axes

100 frames.. So inputs are 100xMxN for images

For horizontal and vertical heatmaps these will be 200xMxK and 200xNxK

(70μWatts of Transmit Power)

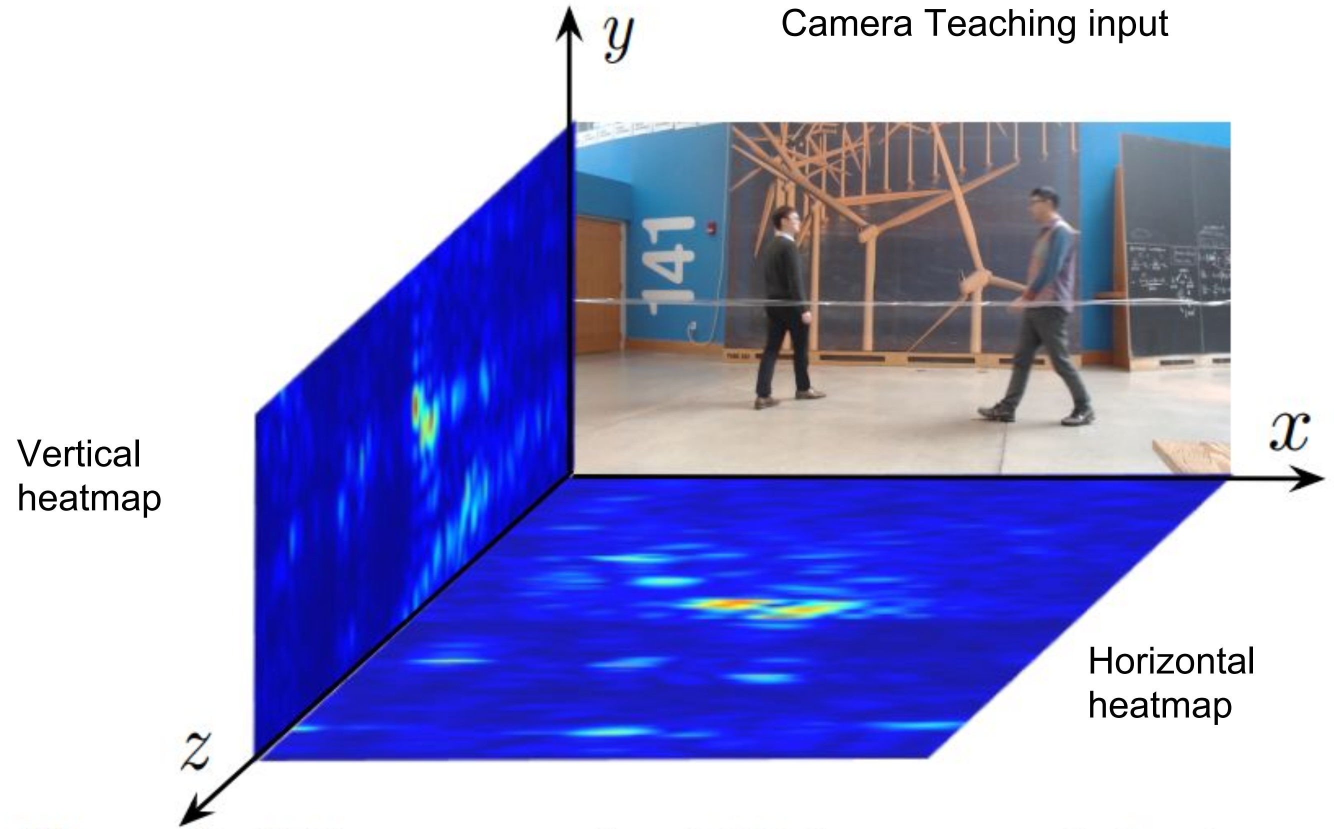
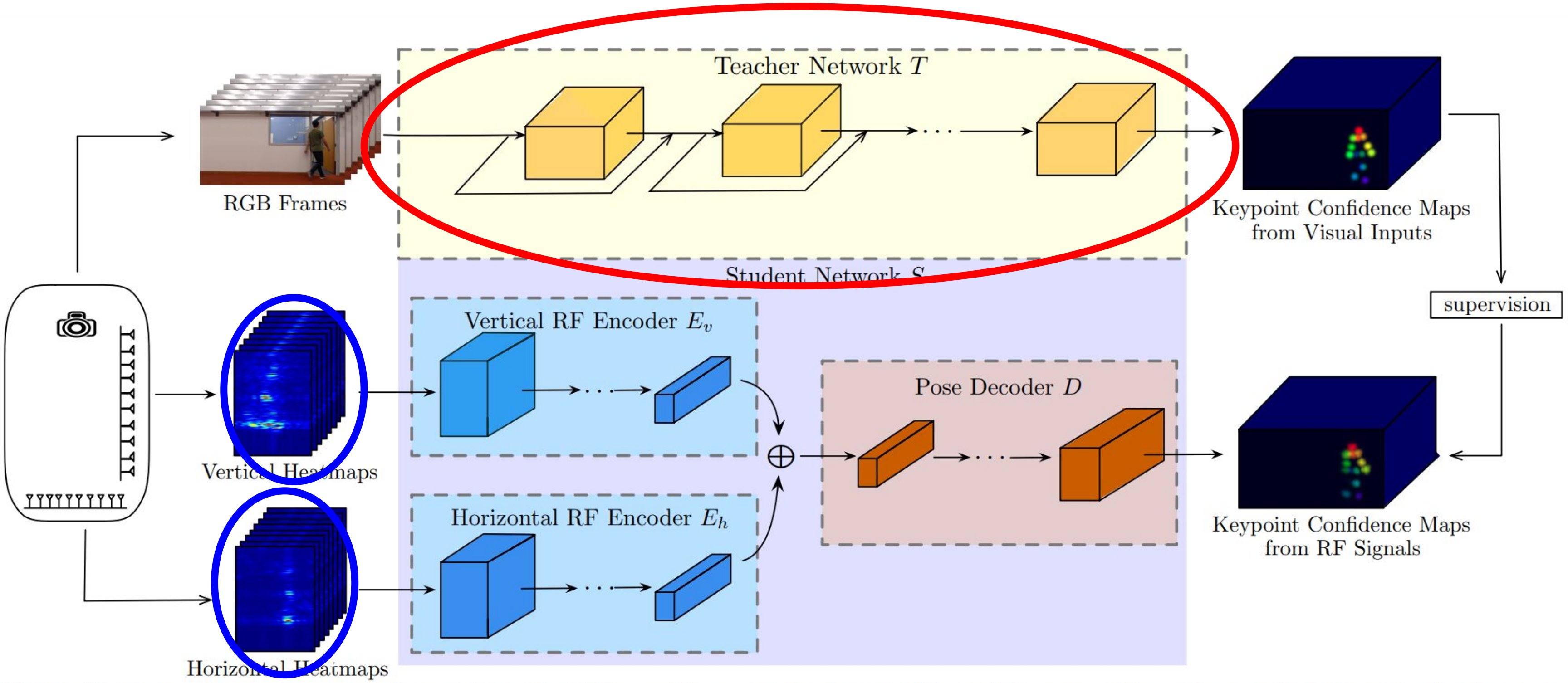


Figure 2: RF heatmaps and an RGB image recorded at the same time.

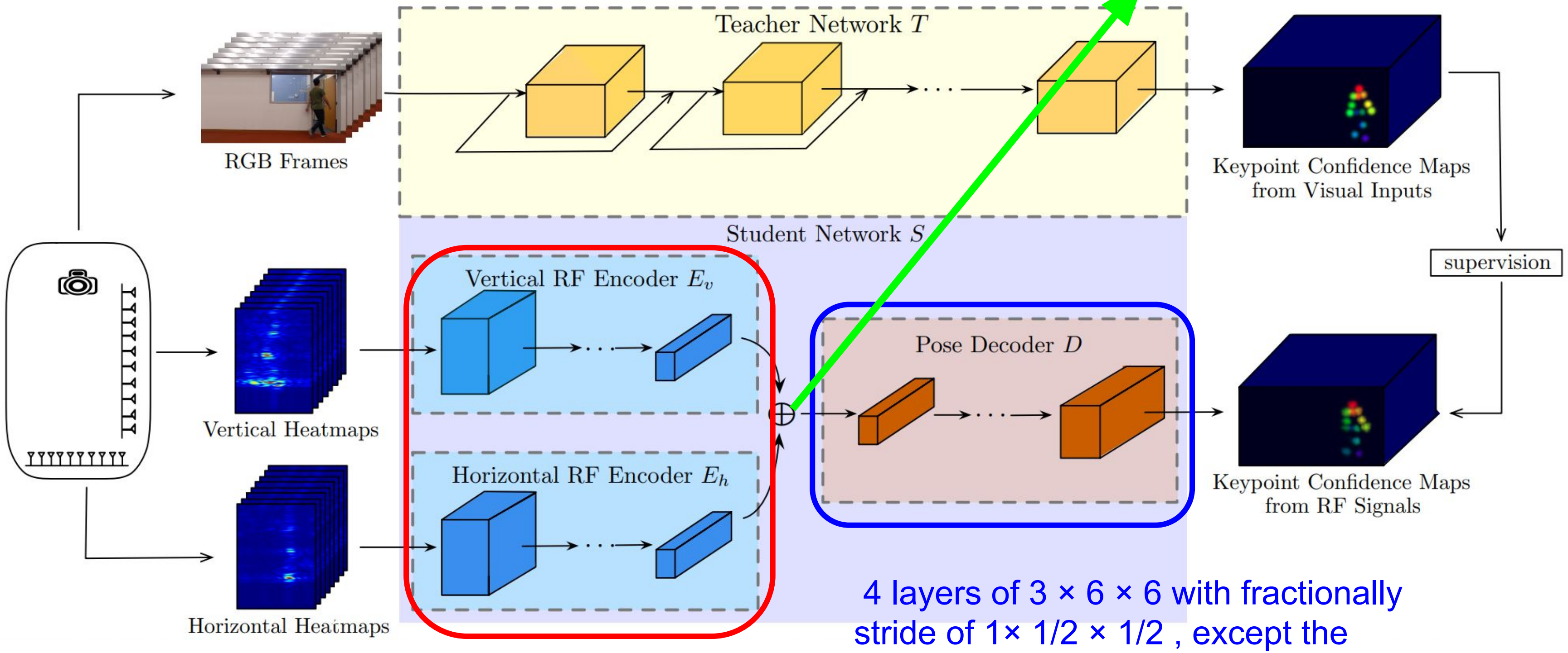
PRE-TRAINED (OpenPose)



Are complex channels with two real valued channels one for each real and imaginary parts

- Horizontal and vertical heatmaps are Complex heatmaps
    - different networks for both real and imaginary parts
- These are represented as two different channels (so.. 2*100xMxK (2*100xNxK) for horizontal (vertical) streams)

channel-wise concatenation of
horizontal and vertical RF encodings



4 layers of 3 × 6 × 6 with fractionally
stride of 1× 1/2 × 1/2 , except the
last layer has one of 1 × 1/4 × 1/4

10 layers of 9 × 5 × 5
spatio-temporal convolutions with
1 × 2 × 2 strides

Loss Function

$$\min_{\mathbf{S}} \sum_{(\mathbf{I},\mathbf{R})} L(\mathbf{T}(\mathbf{I}), \mathbf{S}(\mathbf{R})) \quad (1)$$

We define the loss as the summation of binary cross entropy loss for each pixel in the confidence maps:

$$L(\mathbf{T}, \mathbf{S}) = -\sum_{c} \sum_{i,j} \mathbf{S}_{ij}^{c} \log \mathbf{T}_{ij}^{c} + (1 - \mathbf{S}_{ij}^{c}) \log (1 - \mathbf{T}_{ij}^{c}),$$

Dataset

50 hrs of data collection at 50 different locations

Offices, coffee houses lecure and seminar halls across MIT

RESULTS

Methods	Visible scenes			Through-walls		
	AP	AP^{50}	AP^{75}	**AP**	AP^{50}	AP^{75}
RF-Pose	62.4	**93.3**	70.7	**58.1**	**85.0**	**66.1**
OpenPose[10]	**68.8**	77.8	**72.6**	-	-	-

Table 1: Average precision in visible and through-wall scenarios.

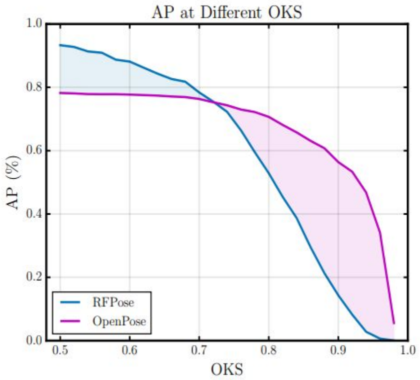
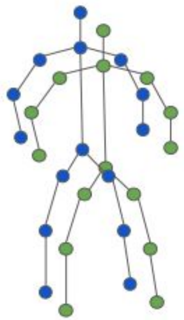


Figure 5: Average precision at different OKS values.

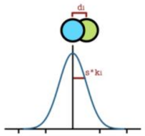
Methods	Hea	Nec	Sho	Elb	Wri	Hip	Kne	Ank
RF-Pose	**75.5**	**68.2**	62.2	56.1	51.9	**74.2**	63.4	54.7
OpenPose[10]	73.0	67.1	**70.8**	**64.5**	**61.5**	71.4	**68.4**	**68.3**

Table 2: Average precision of different keypoints in visible scenes.

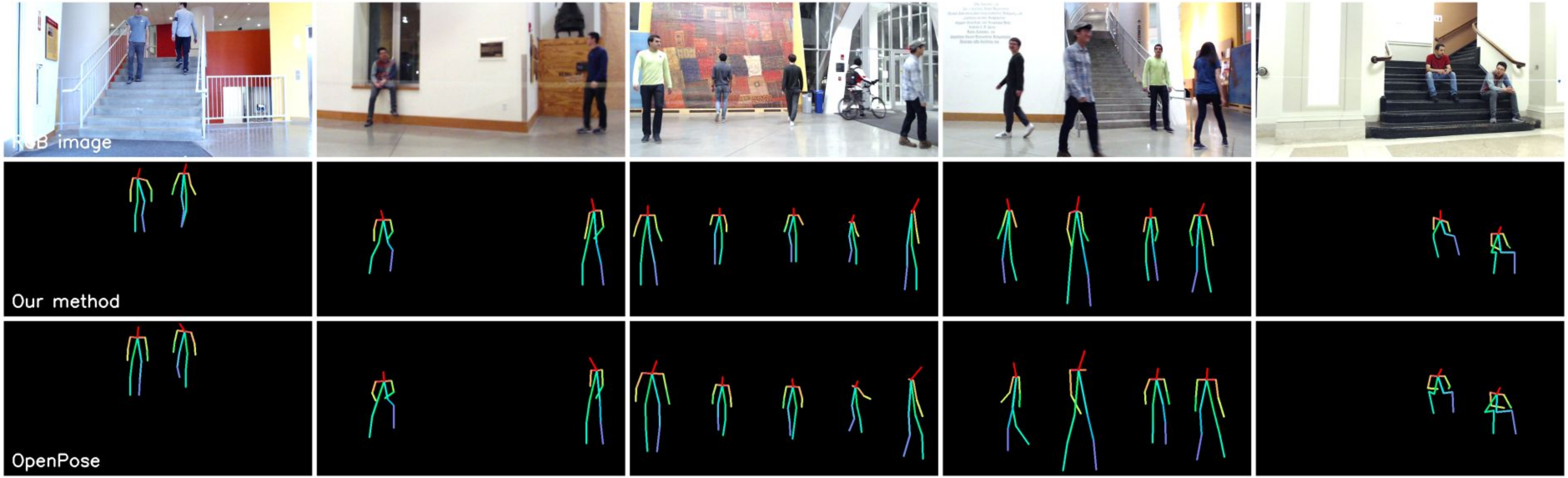


Object Keypoint Similarity

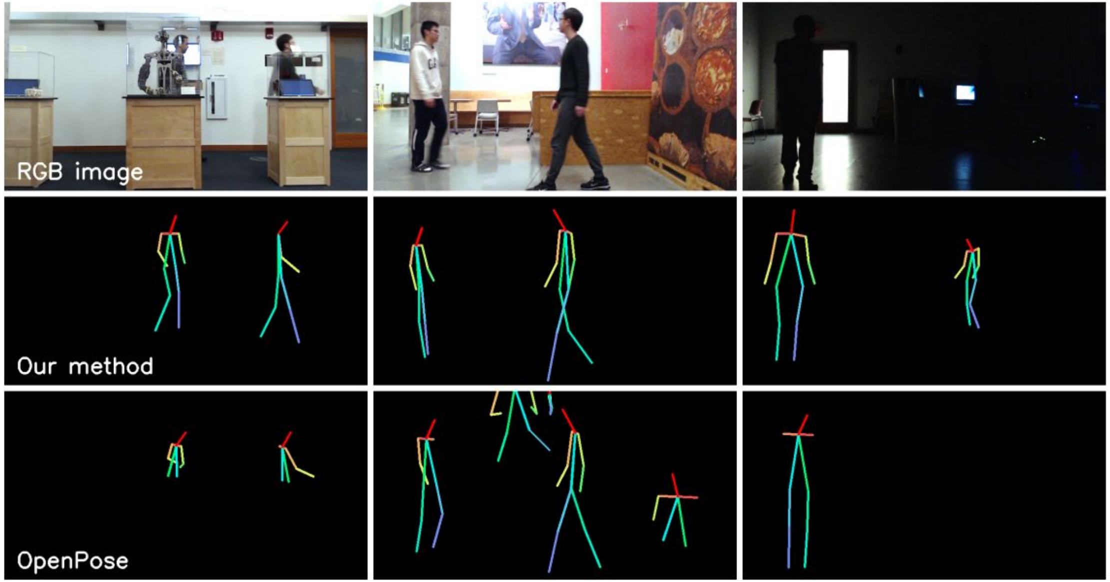
$$OKS = \frac{\sum_i e^{-\frac{d_i^2}{2s^2k_i^2}} \delta(v_i > 0)}{\sum_i \delta(v_i > 0)}$$

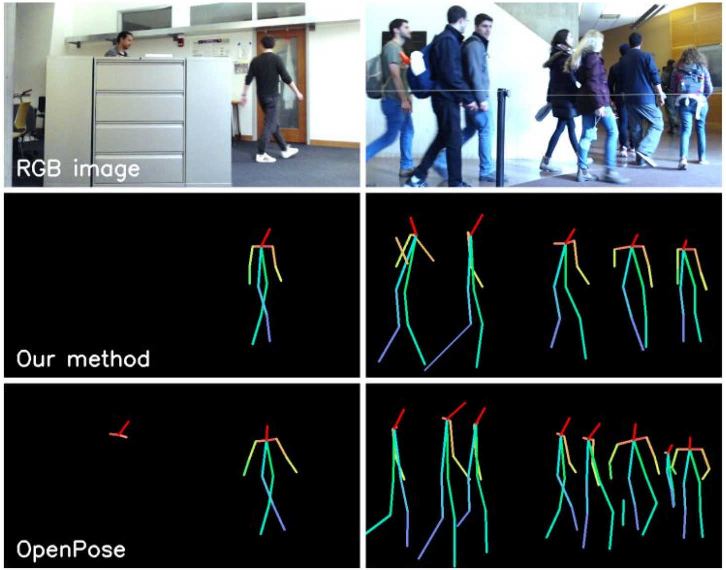


Well lit and occlusion free environments



Not so well lit, with occlusion and even reflectors





Importance of considering multiple windows over time

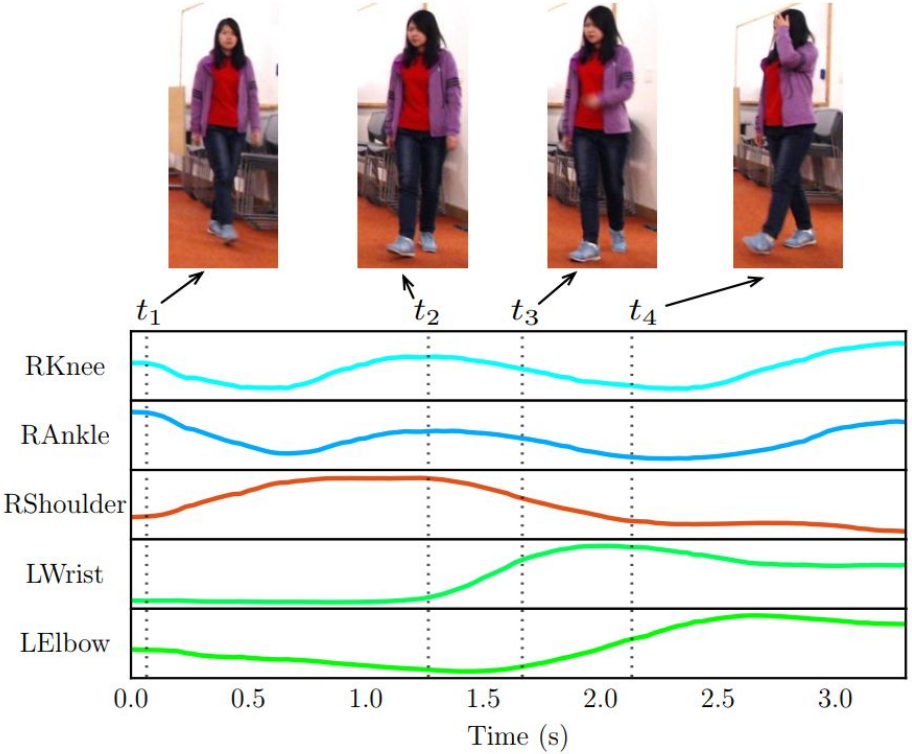


Figure 9: Activation of different keypoints over time.

# RF frames	AP
6	30.8
20	50.8
50	59.1
100	**62.4**

Table 3: Average precision of pose estimation trained on varying lengths of input frames.

Person Identification

Based on the gait of a person one can identify a person

Over 100 different persons:

Method	Visible scenes		Through-walls	
	Top1	Top3	Top1	Top3
RF-Pose	83.4	96.1	84.4	96.3