Lecture: Deep Learning and Differential Programming

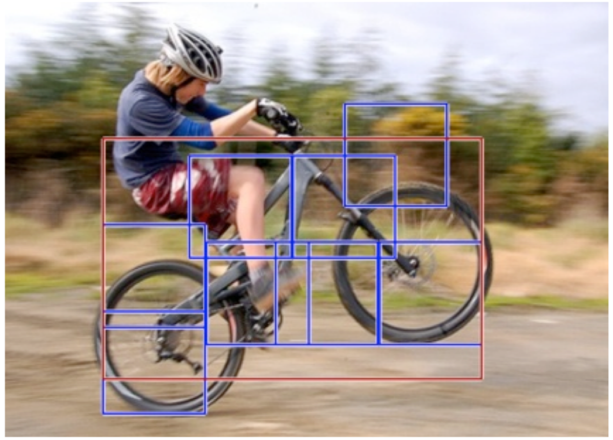
4.2 Attention in Computer Vision

https://liris.cnrs.fr/christian.wolf/teaching

INSA LYON Christian Wolf

Previously at vision conferences: Deformable parts models

- Model an object/human/activity as a collection of local parts
- Optimize over (latent) local part positions



[Felzenszwalb et al., PAMI 2010]

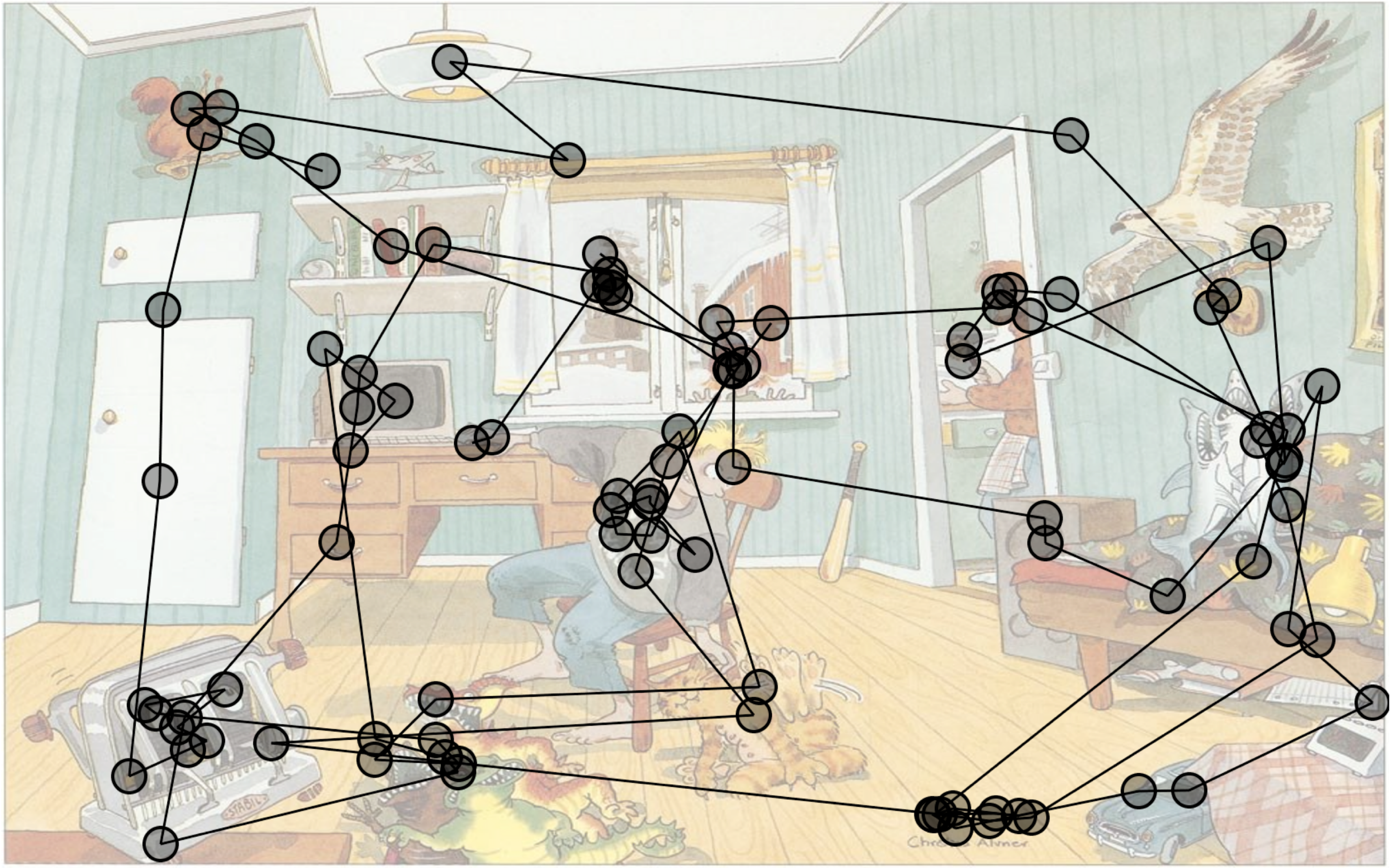
$$\underbrace{\sum_{i=0}^{n} F_i' \cdot \phi(H, p_i)}_{\text{Local appearance}} - \underbrace{\sum_{i=1}^{n} d_i \cdot \phi_d(dx_i, dy_i)}_{\text{Deformation}} + b,$$

?

Is it really necessary to calculate all possible position of parts of searched objects in order to recognize them?

How do humans perform these tasks?

Human attention: gaze patterns



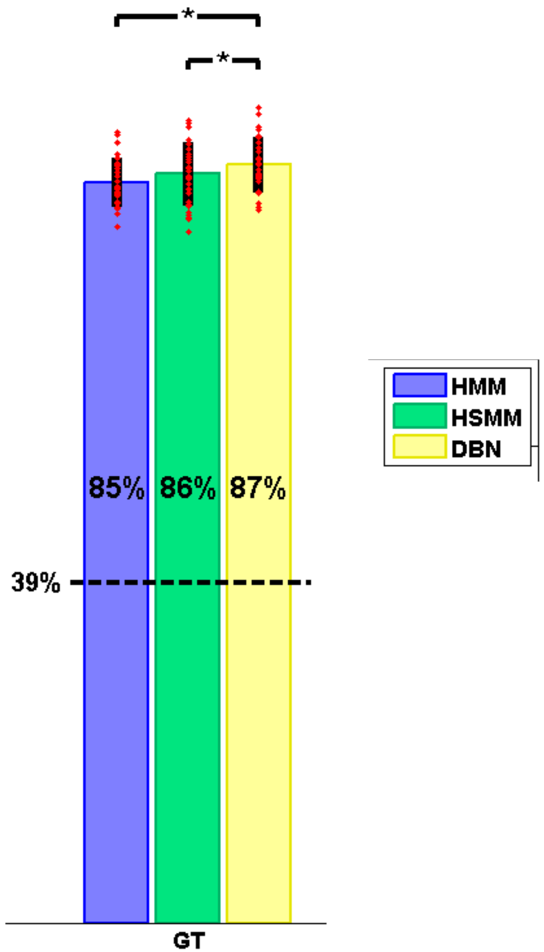
[Johansson, Holsanova, Dewhurst, Holmqvist, 2012]

Gaze can be predicted

Gaze has a several functions:

- scene analysis
- social signals

Attention models can learn to predict gaze



[Mihoub, Bailly, Wolf, Elisei, *PRL 2016*]

Attention in vision

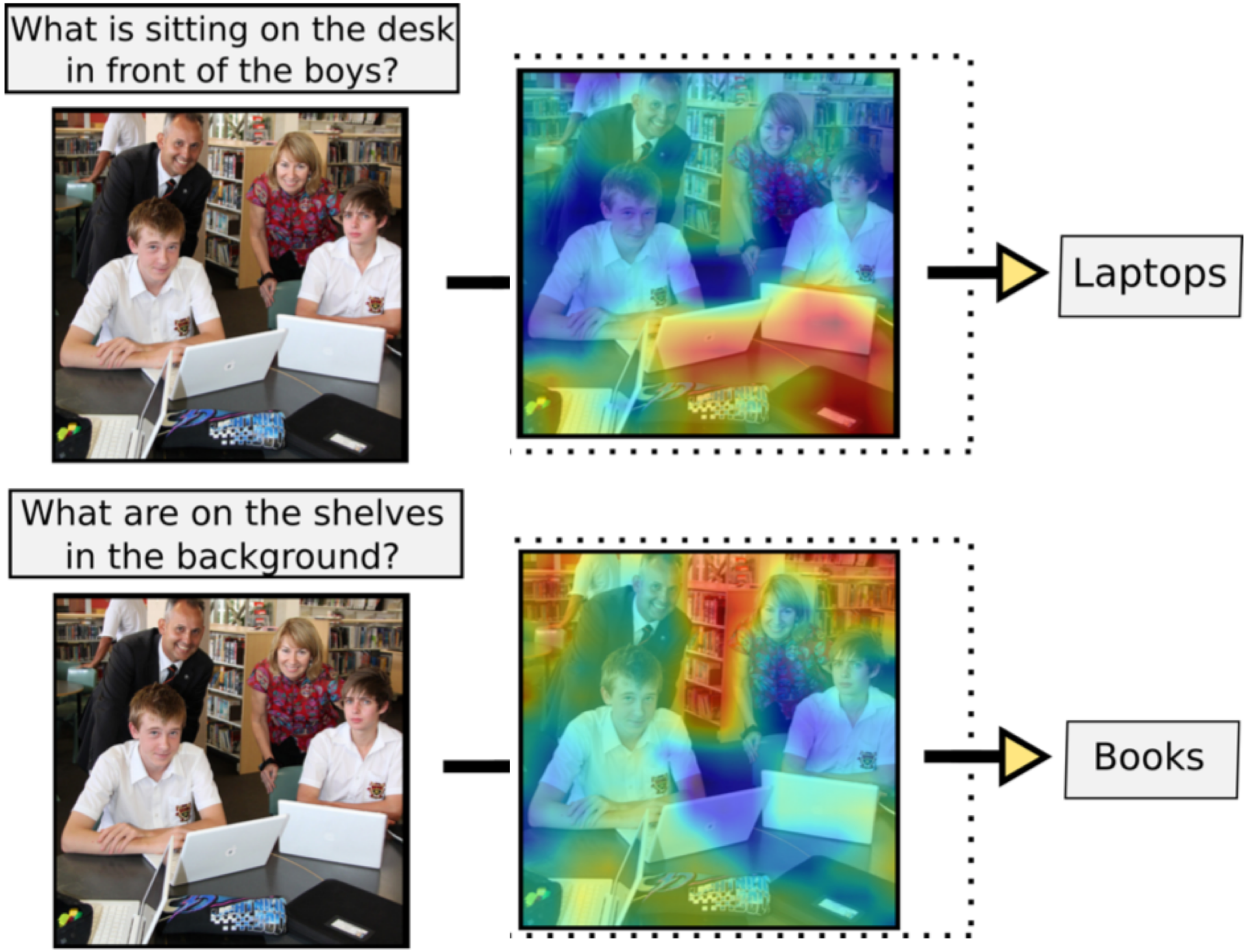
[Durand, Mordan, Thome, Cord, CVPR 2017]

Attention based mechanisms

Can we jointly predict gaze … and the scrutinzed object?

Loss: recognitition performance

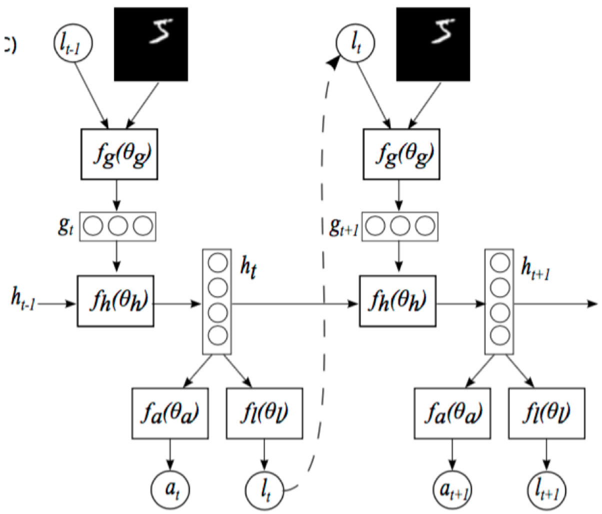
Soft attention: example



[Ben-Younes, Cadene, Thome, Cord, ICCV 2017]

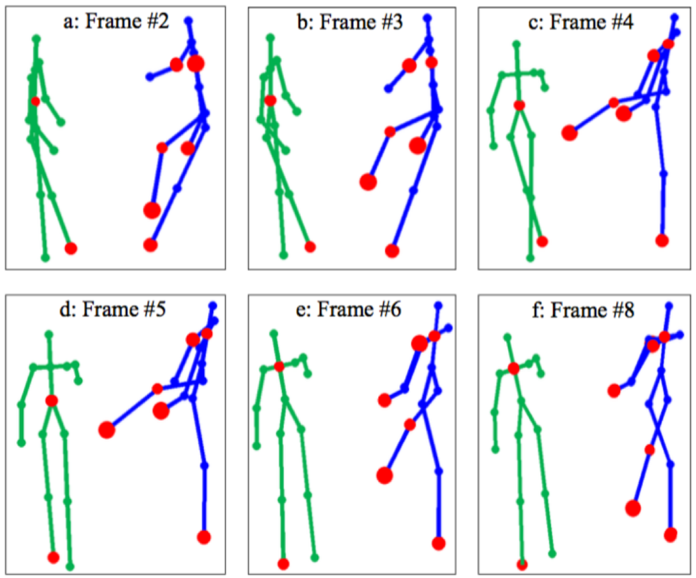
Soft attention vs. hard attention

Hard attention



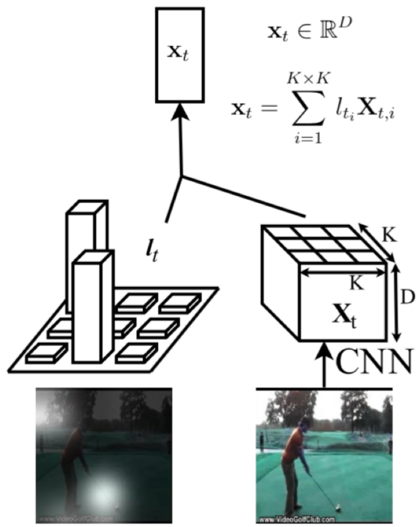
[Mnih et al., NIPS 2015]

Attention on joints



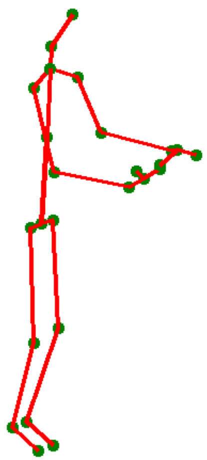
[Song et al., AAAI 2016]

Soft attention in feature maps



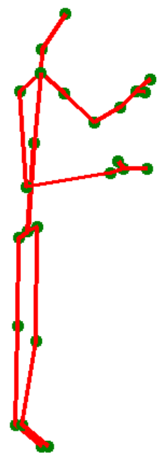
[Sharma et al., ICLR 2016]

Articulated pose alone is not sufficient



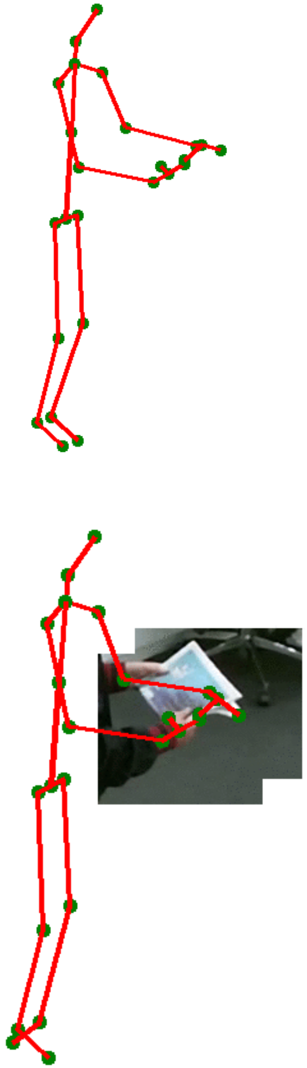
Reading

Same class?!

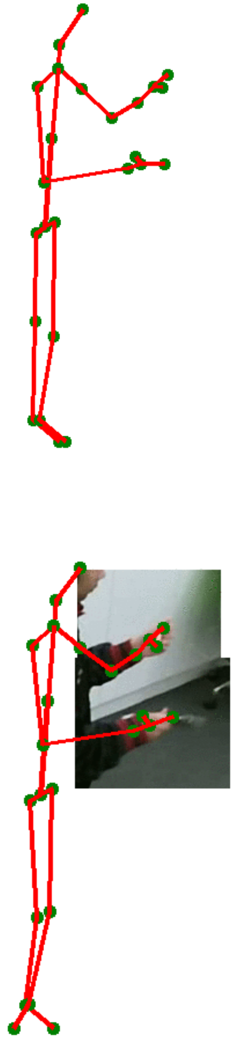


Writing

Articulated pose alone is not sufficient



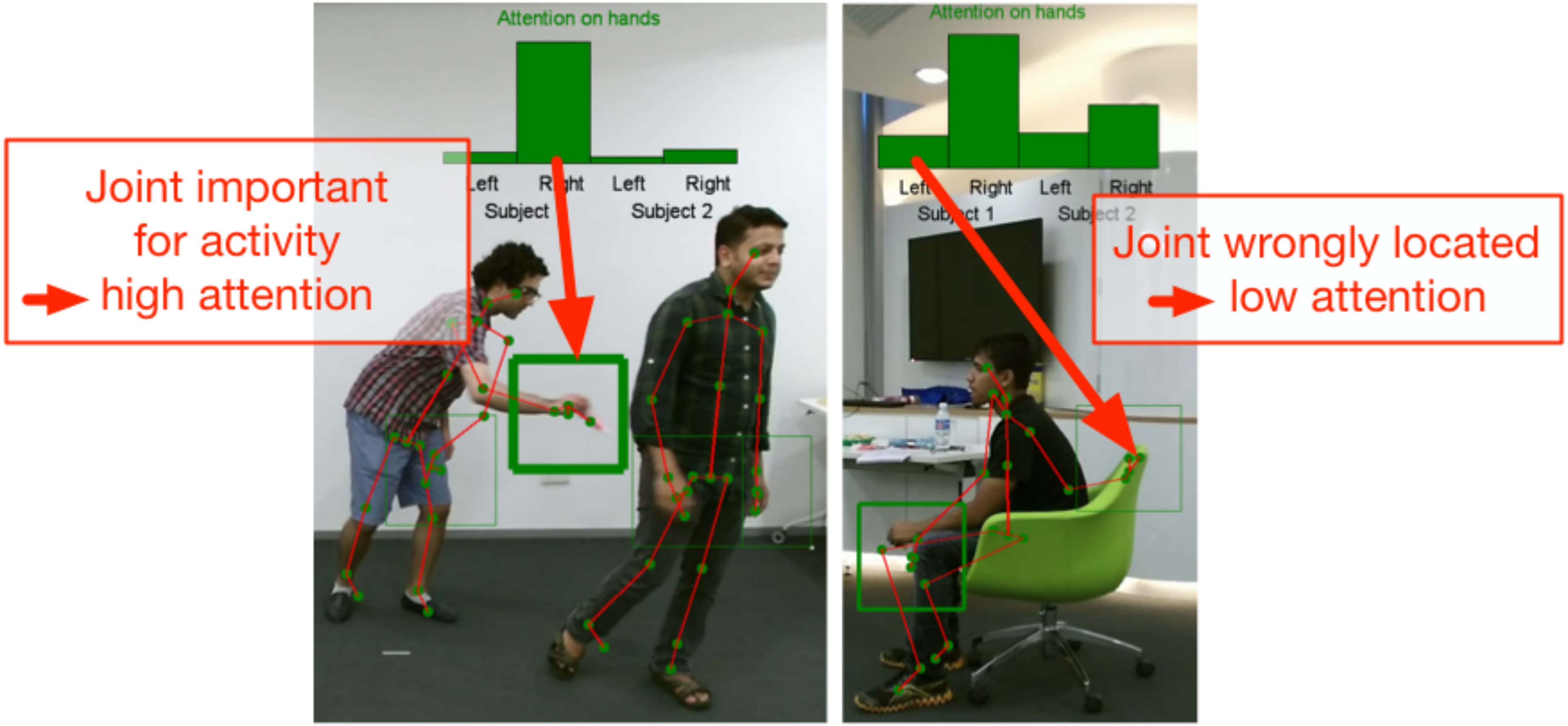
RGB is helpful...



Reading

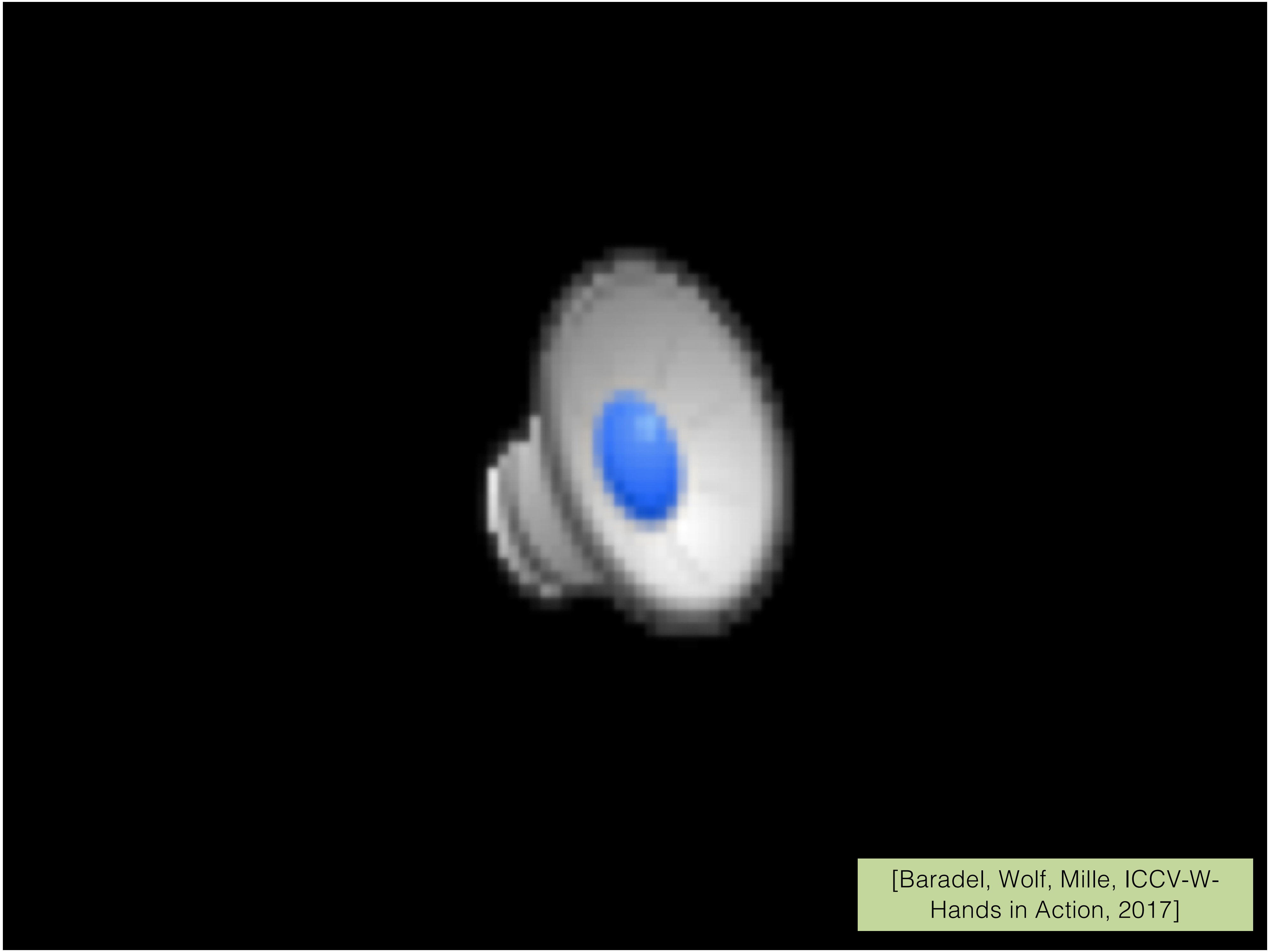
Writing

Attention on relevant parts

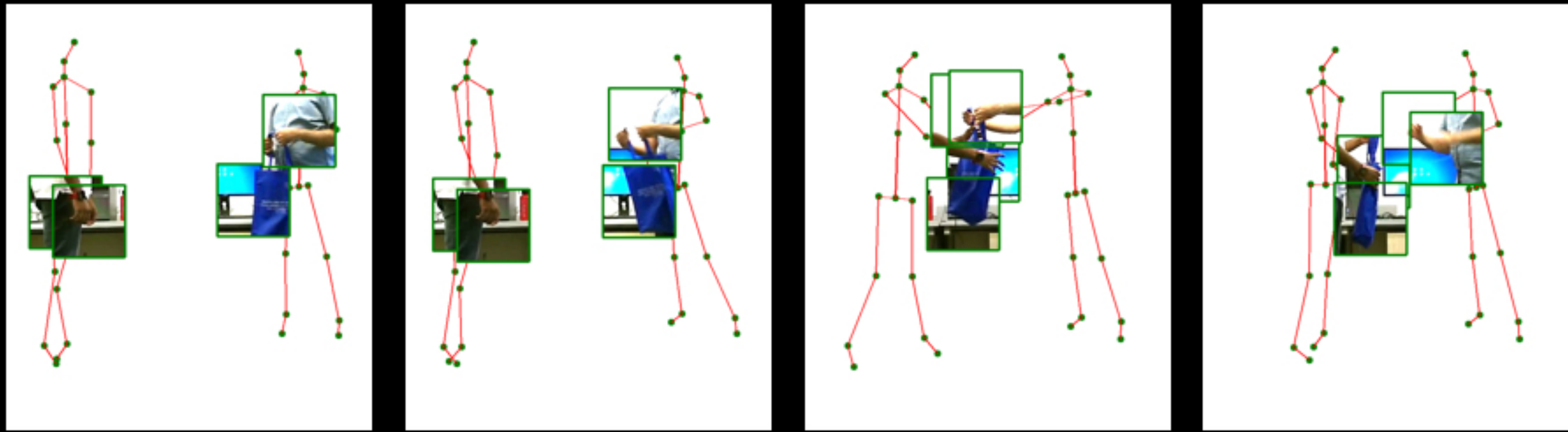


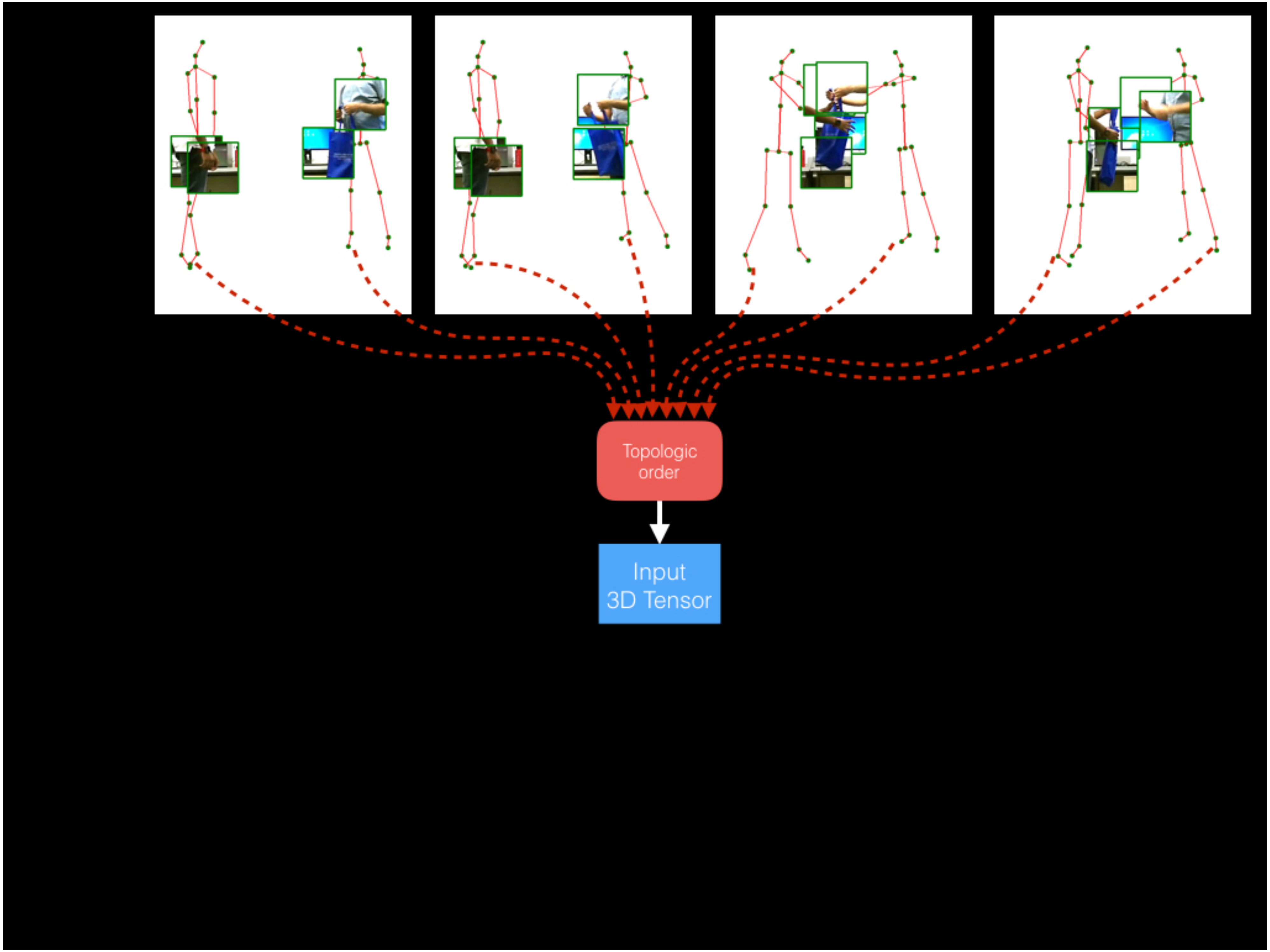
Work of Fabien Baradel,
Phd @ LIRIS

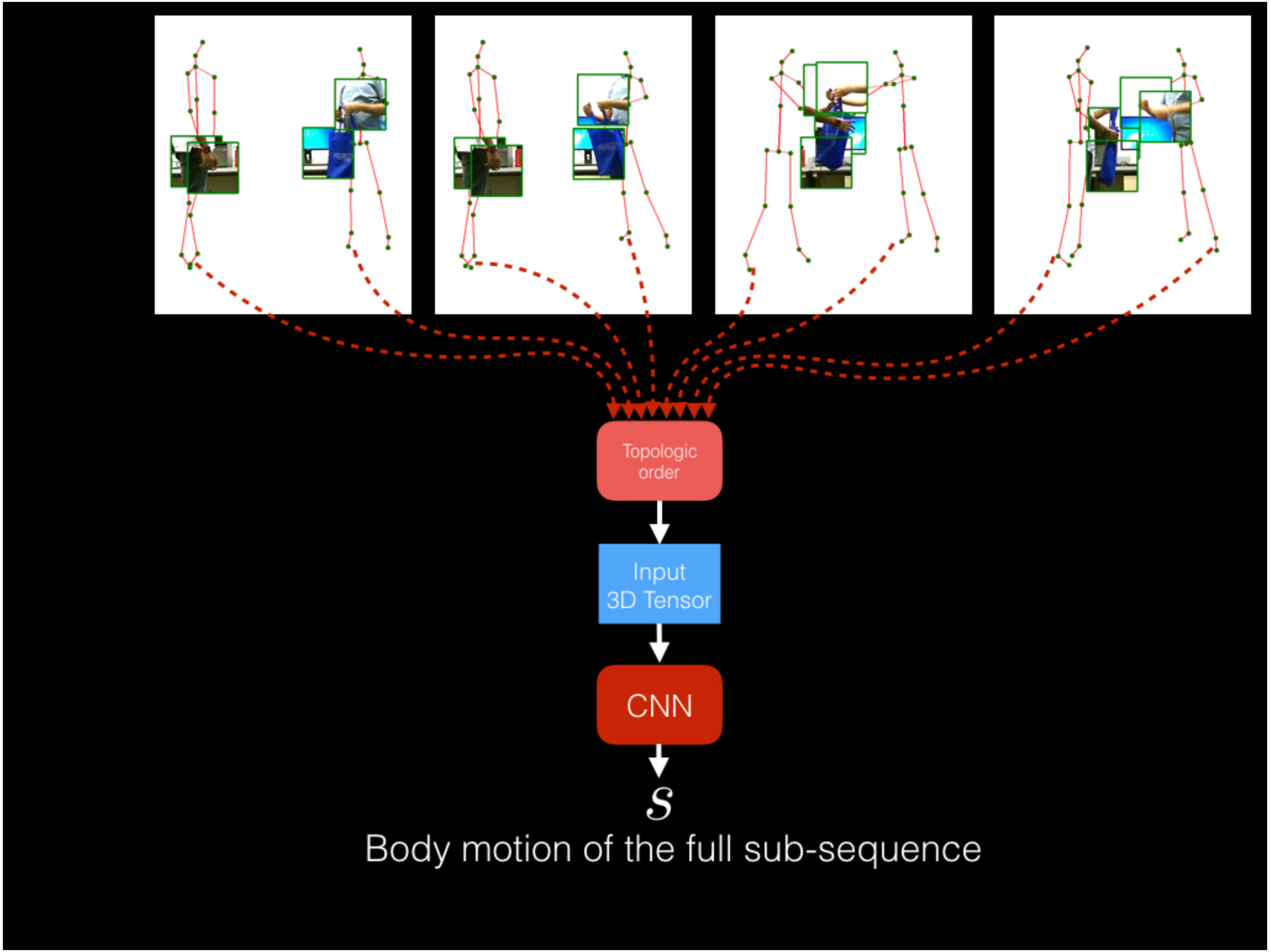
With Julien Mille
(INSA Val de Loire)



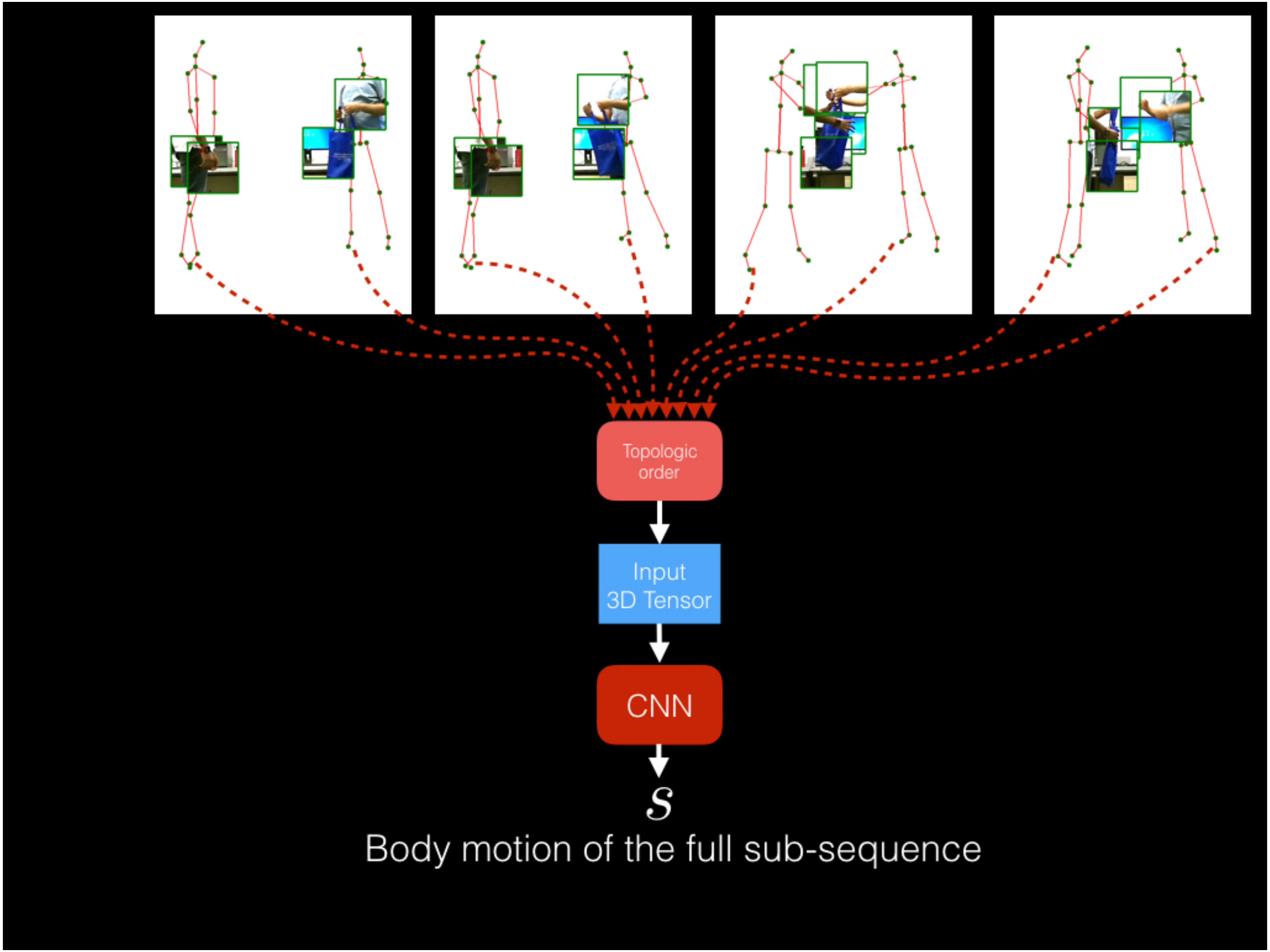
[Baradel, Wolf, Mille, ICCV-W-Hands in Action, 2017]

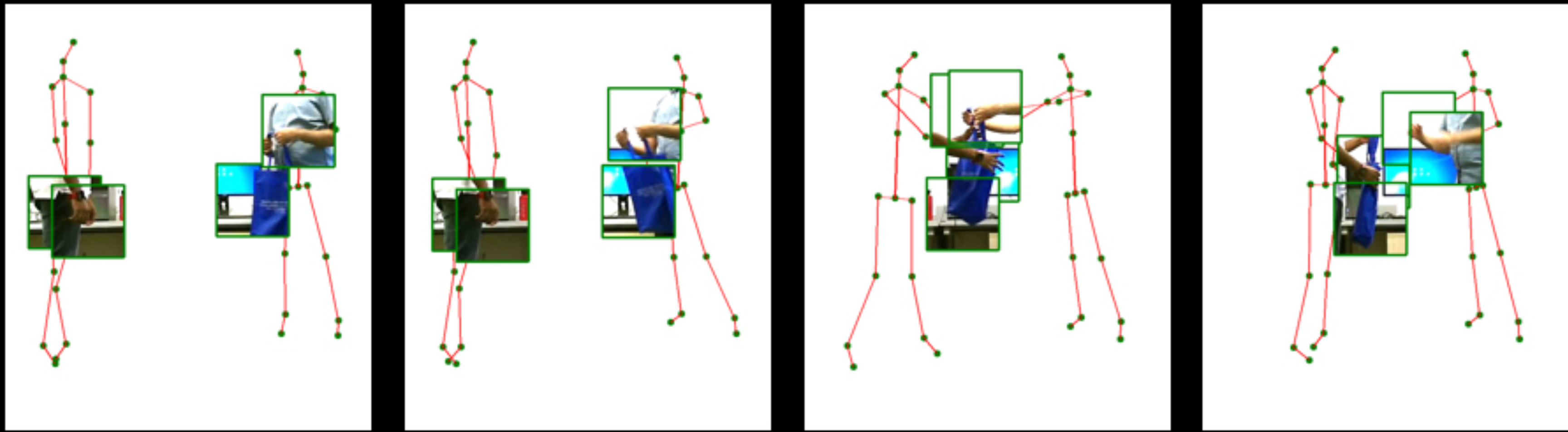



Topologic order
Input 3D Tensor



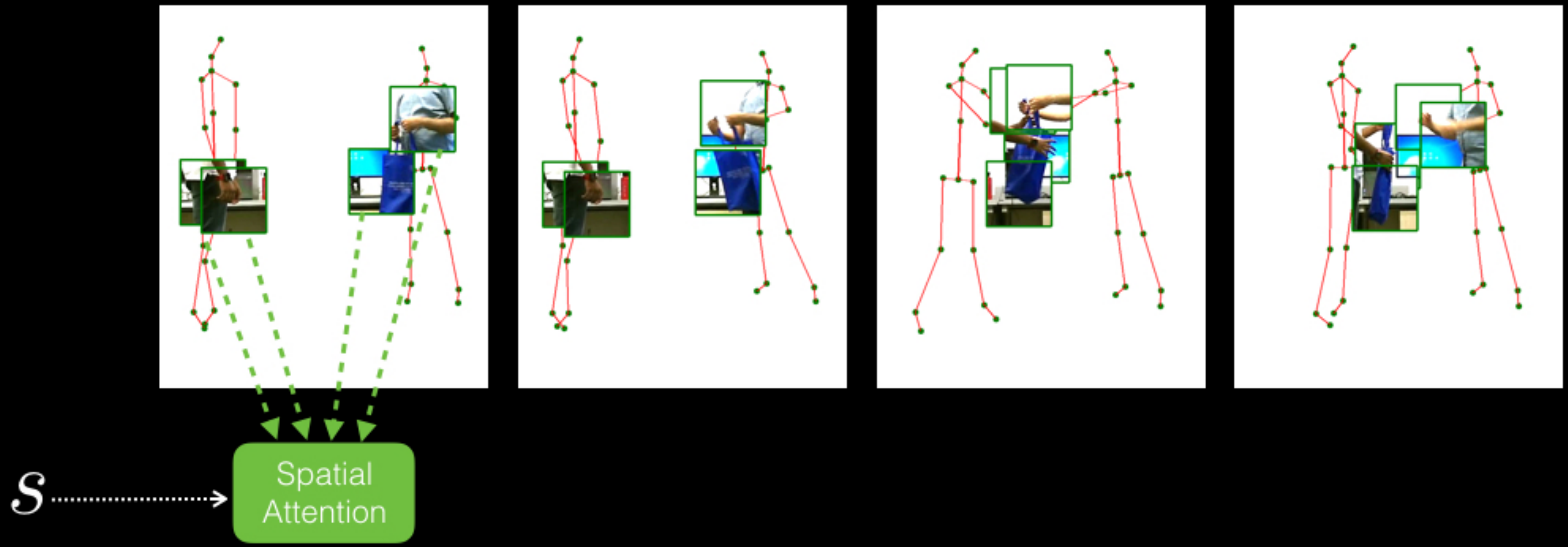
Topologic order
Input 3D Tensor
CNN
S
Body motion of the full sub-sequence


Topologic order
Input 3D Tensor
CNN
S
Body motion of the full sub-sequence

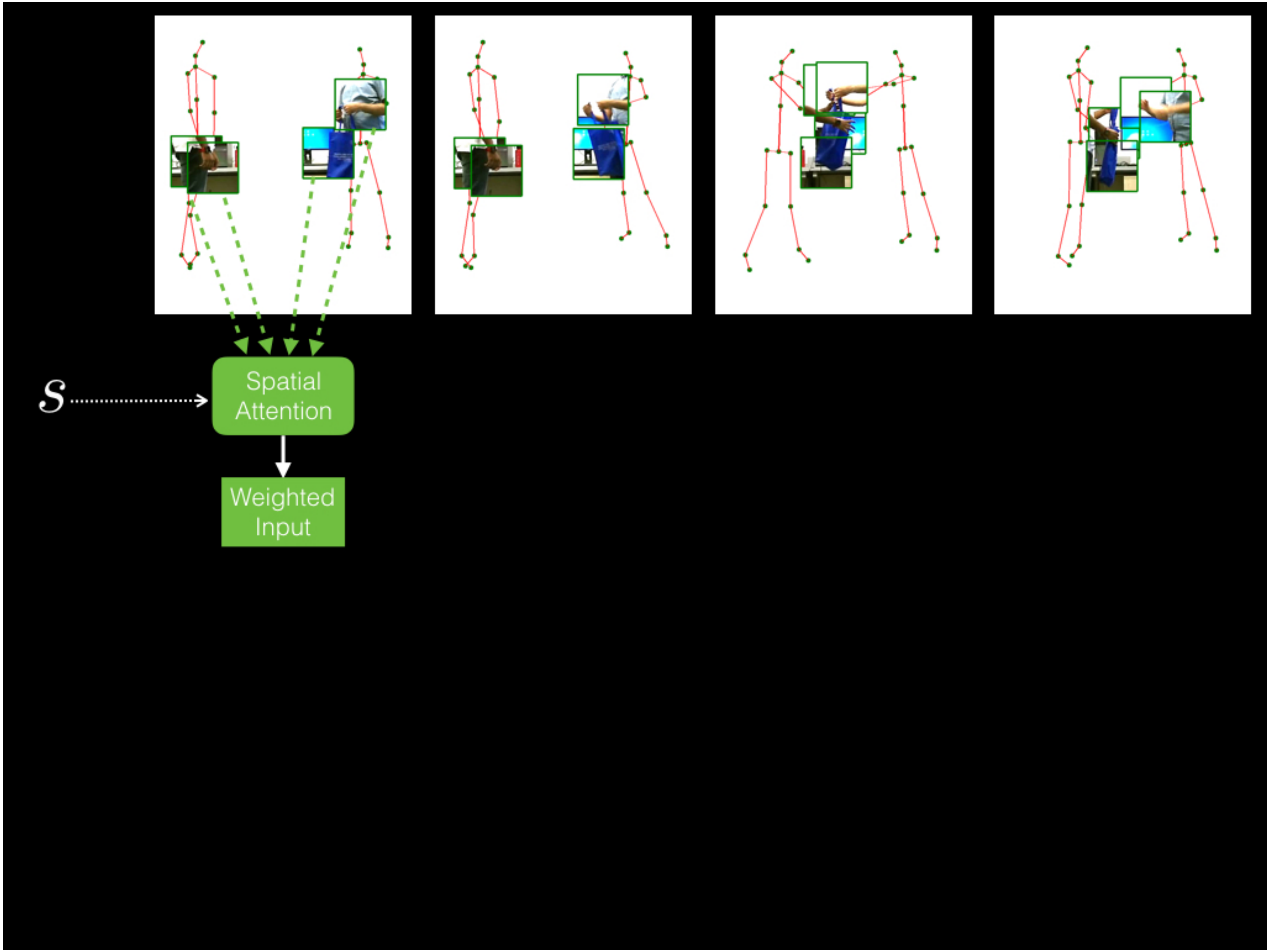


$\boldsymbol{S}$

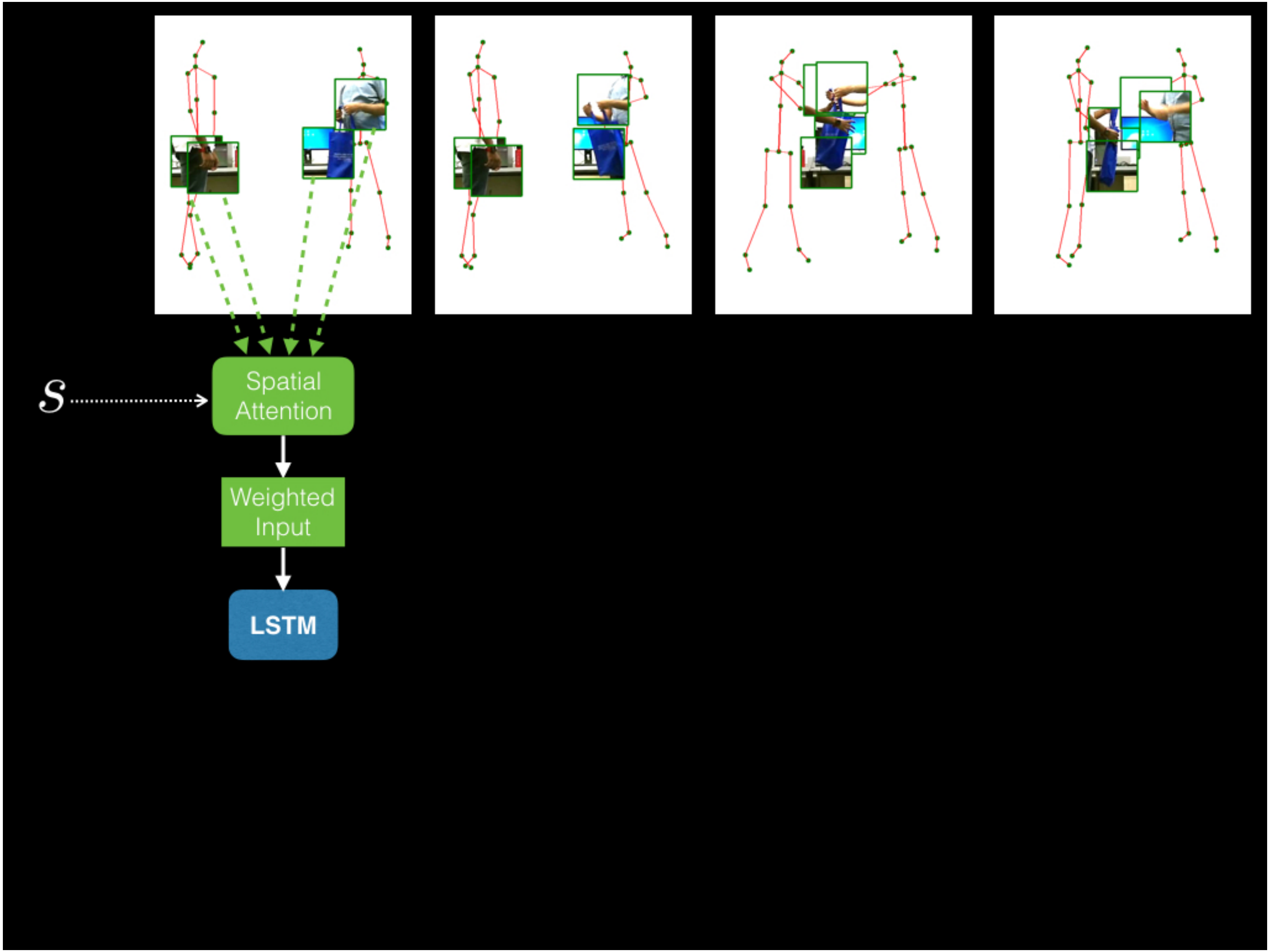
Body motion of the full sub-sequence



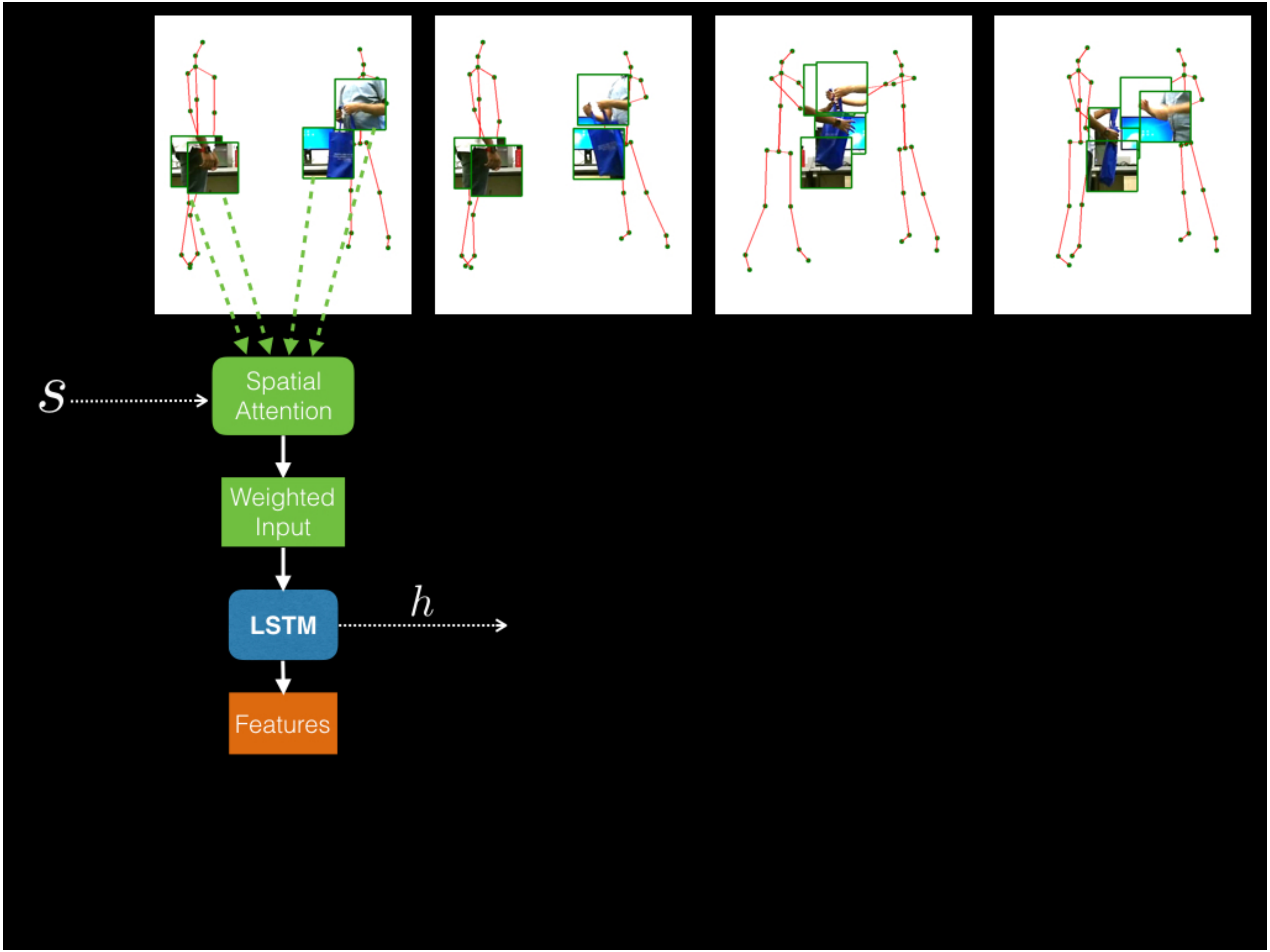
S
Spatial Attention



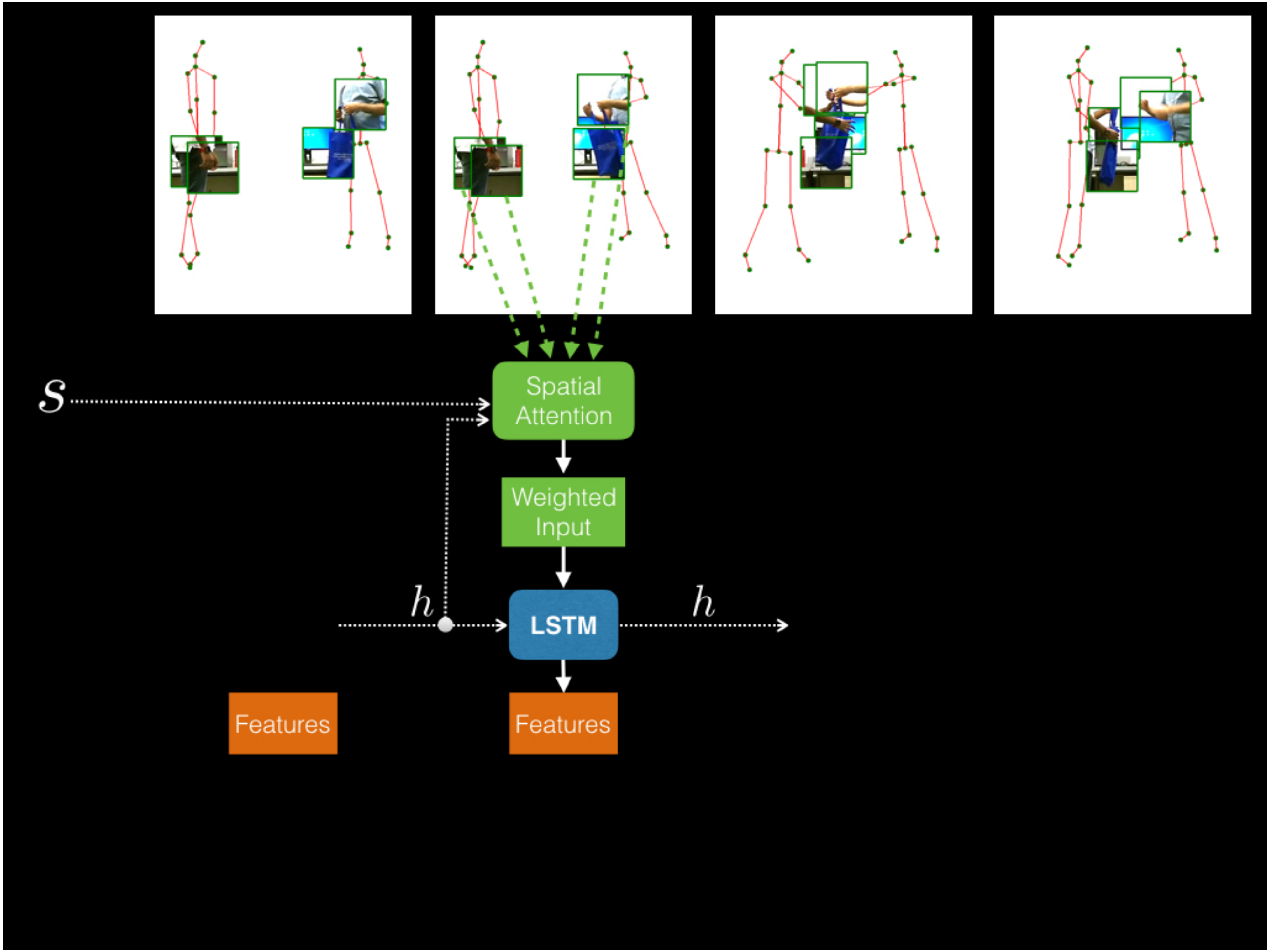
S
Spatial Attention
Weighted Input



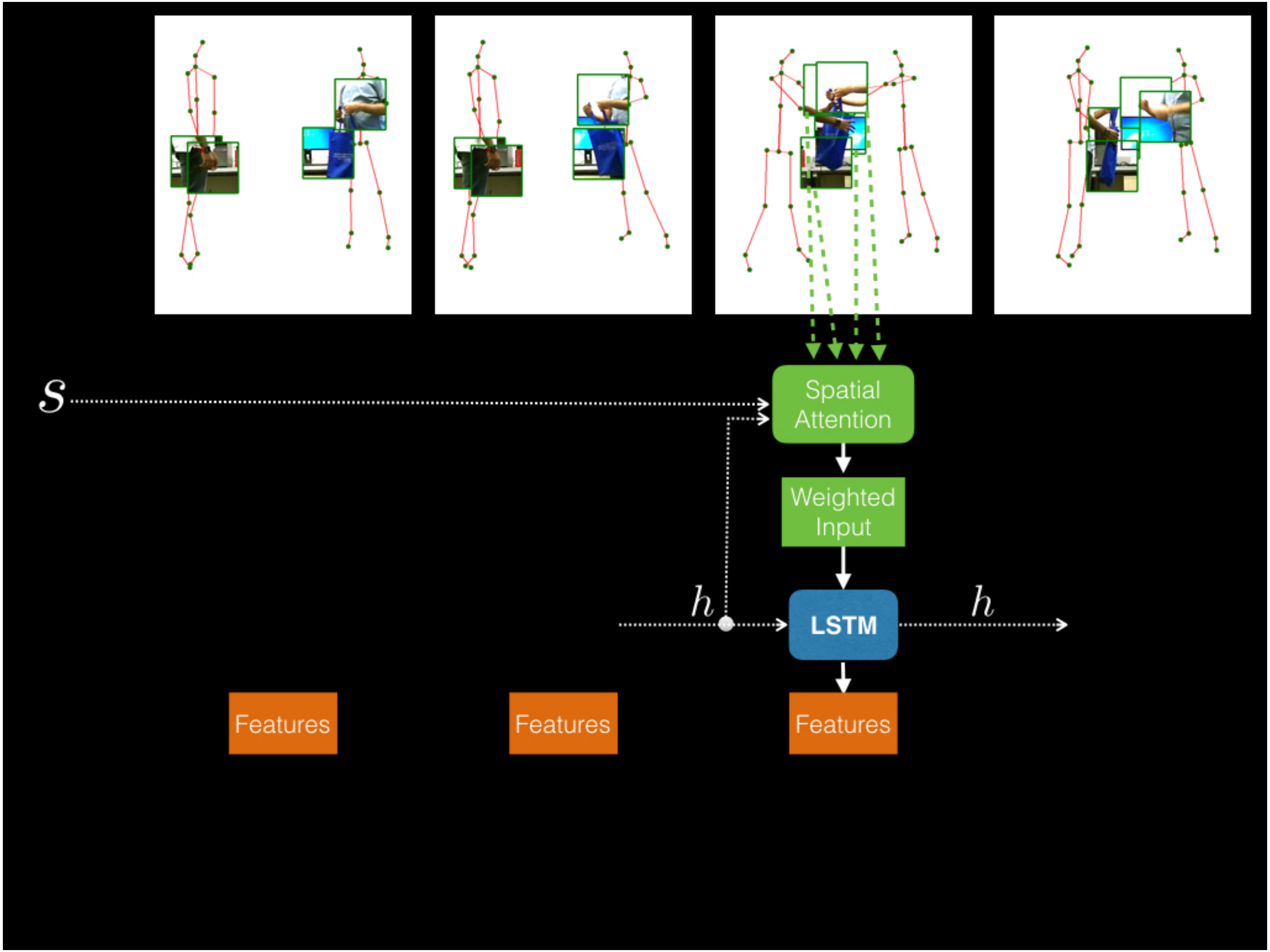
Spatial Attention
Weighted Input
LSTM



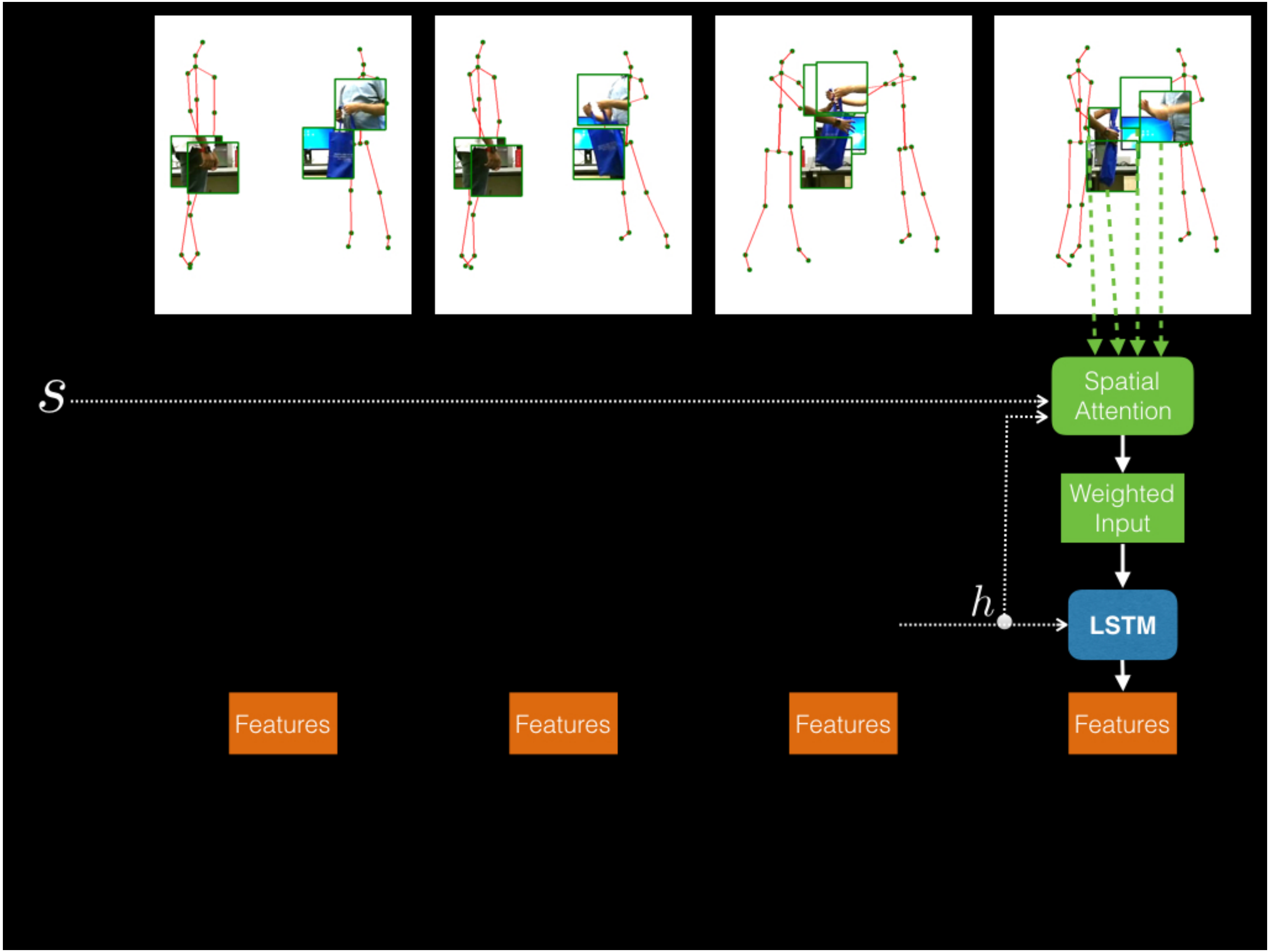
s
Spatial Attention
Weighted Input
LSTM
h
Features

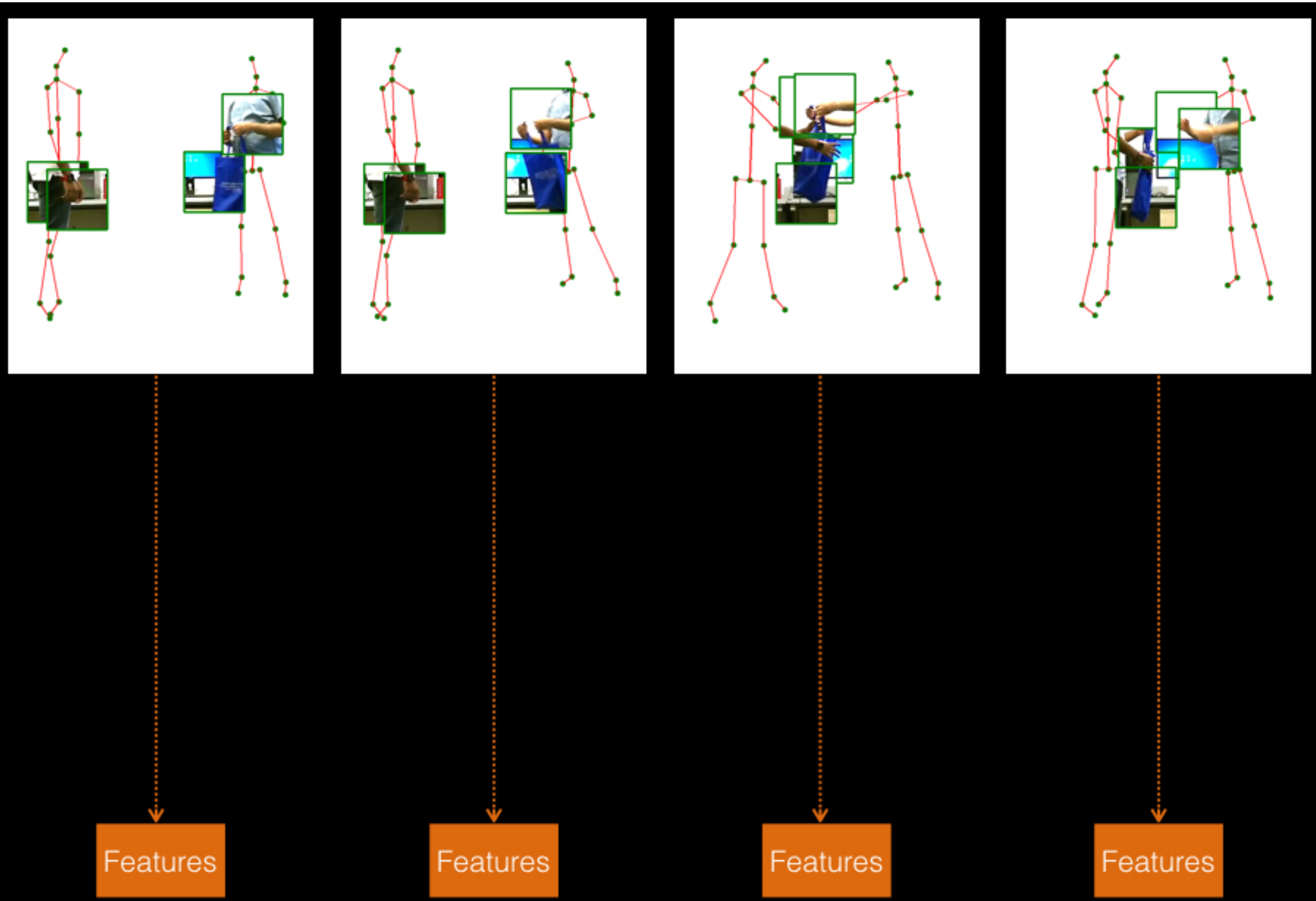


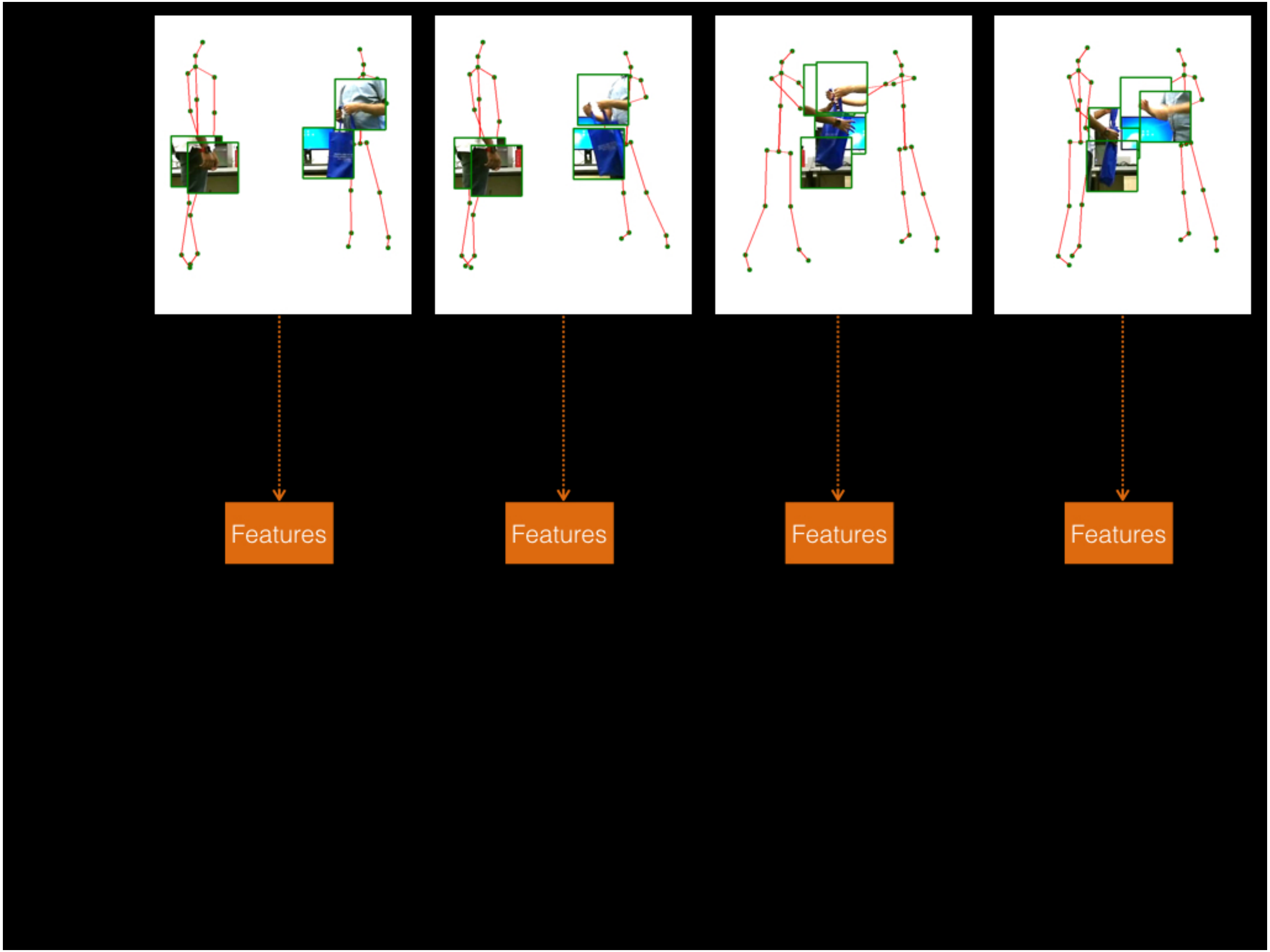
s
Spatial Attention
Weighted Input
h
LSTM
h
Features
Features

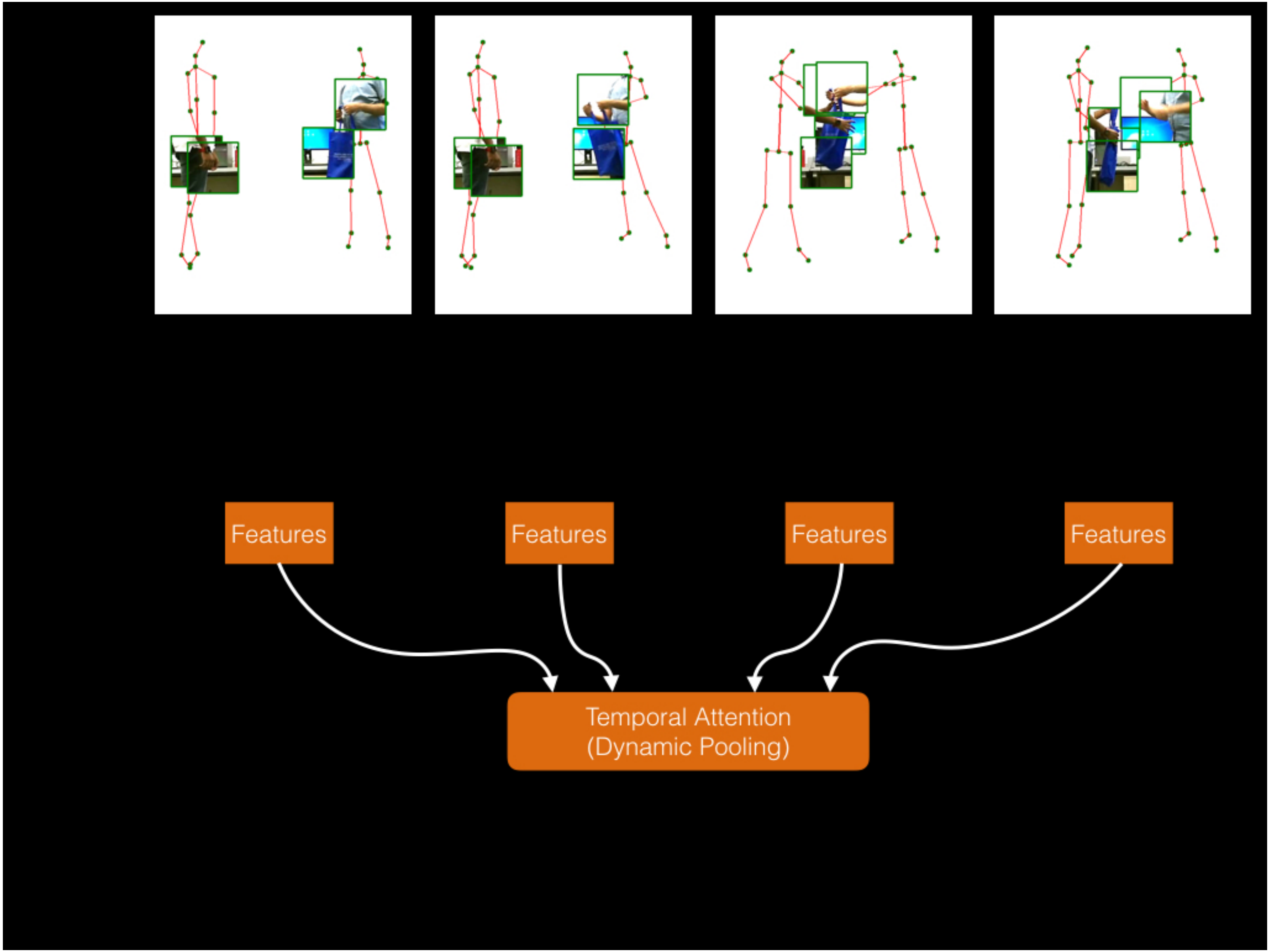


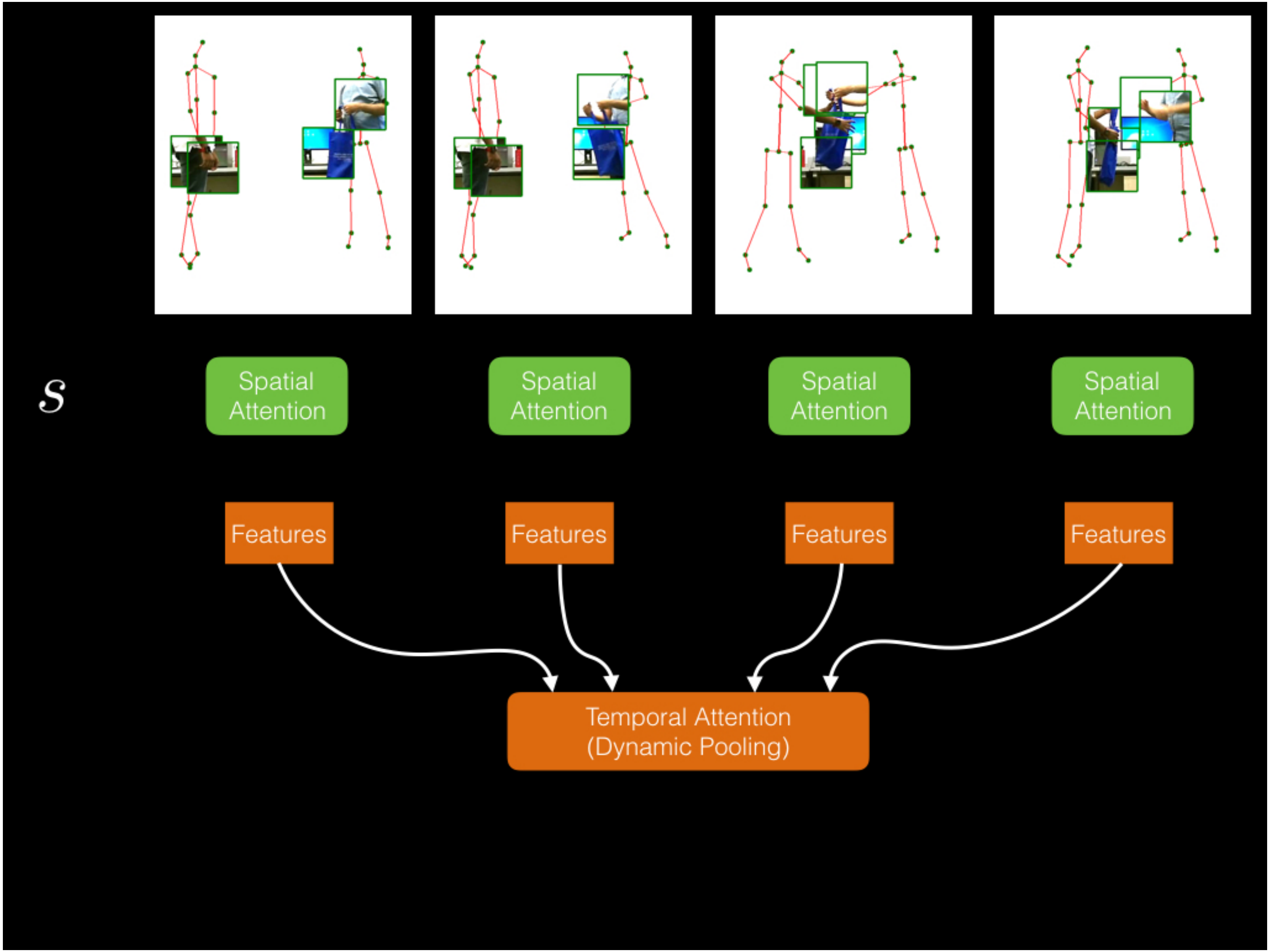
S
Spatial Attention
Weighted Input
h
LSTM
h
Features
Features
Features


s
Spatial Attention
Weighted Input
h
LSTM
Features
Features
Features
Features

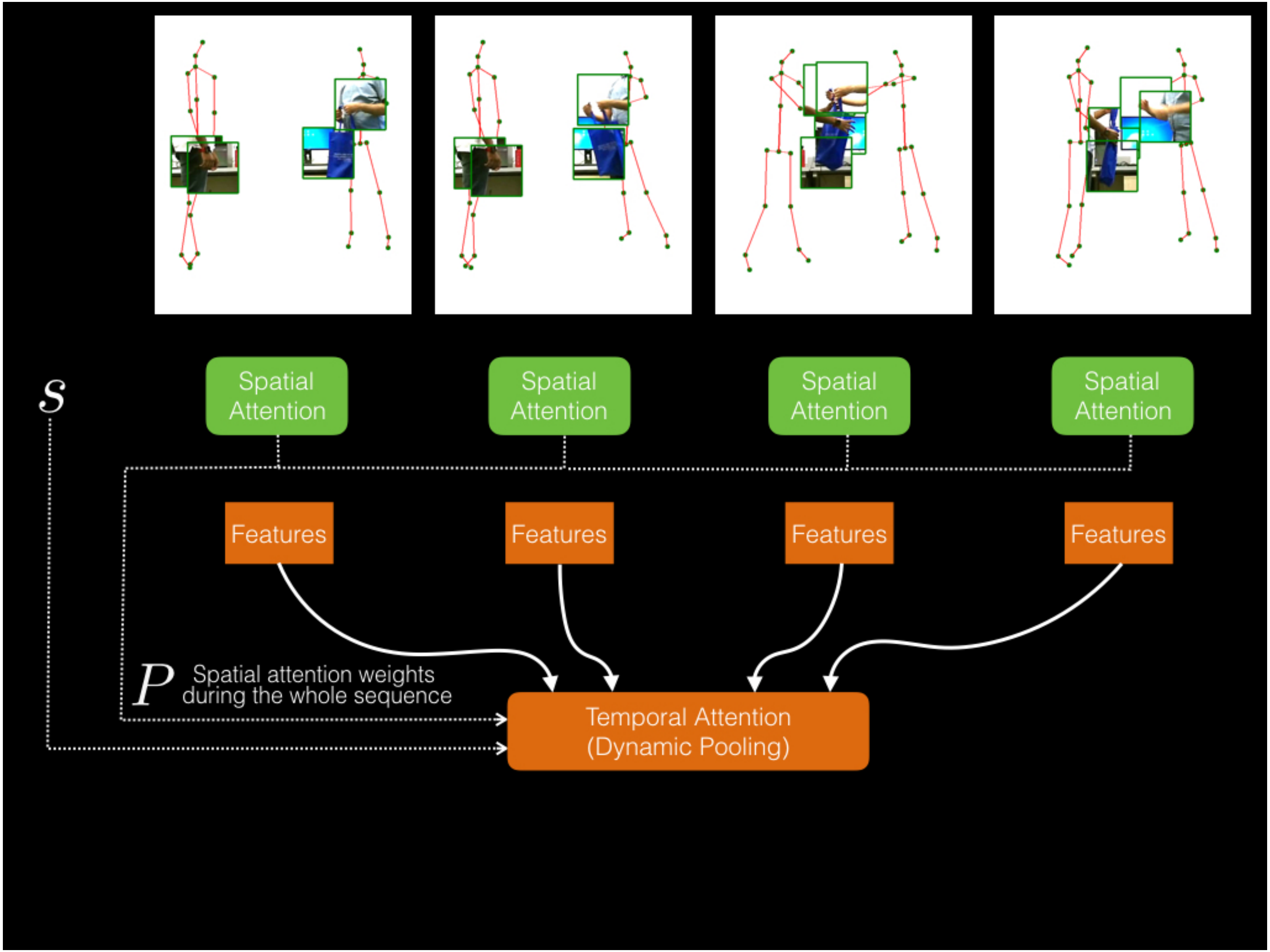

Features
Features
Features
Features

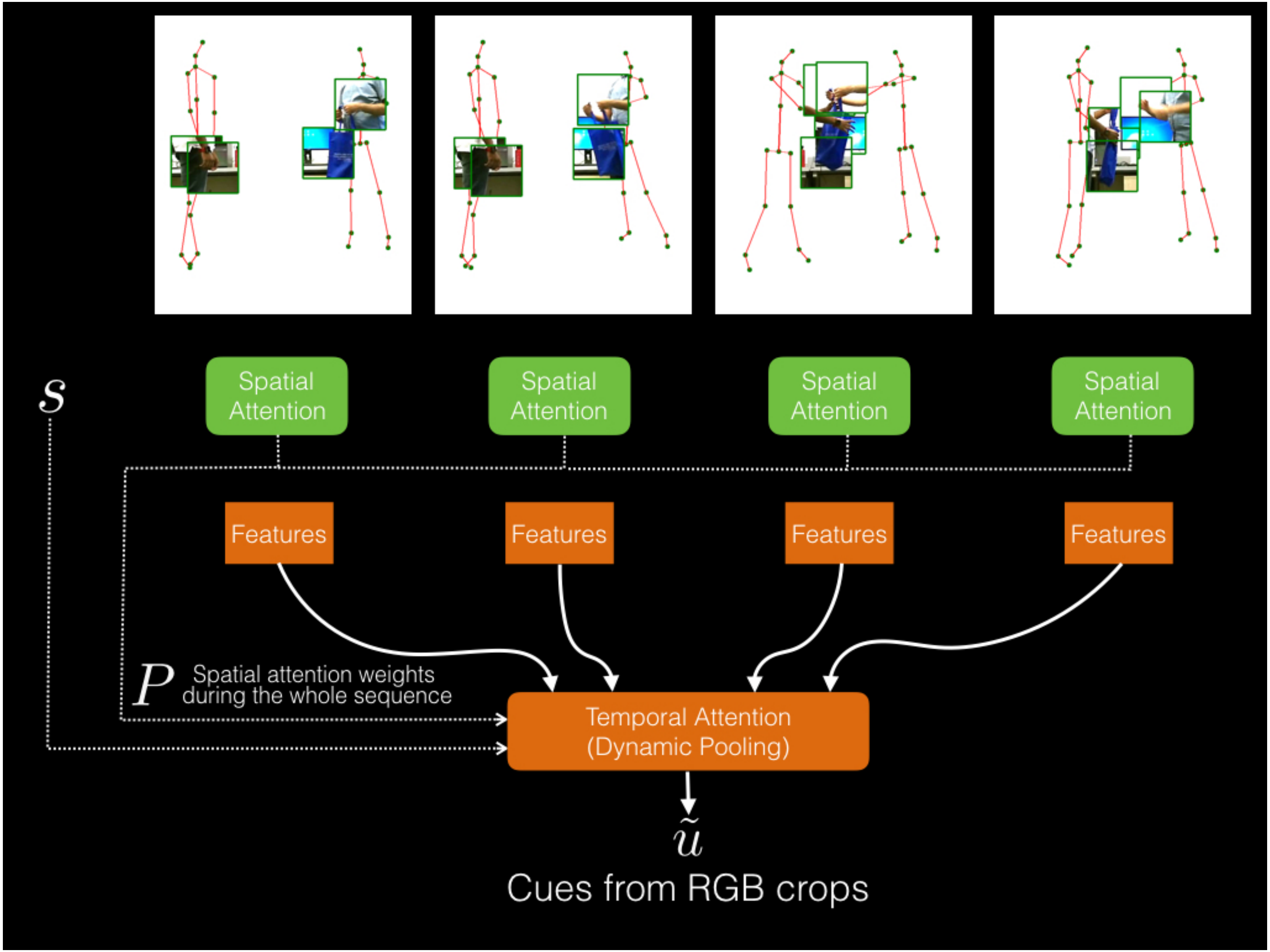

Features
Features
Features
Features


Features
Features
Features
Features
Temporal Attention
(Dynamic Pooling)



S
Spatial Attention
Spatial Attention
Spatial Attention
Spatial Attention
Features
Features
Features
Features
Temporal Attention (Dynamic Pooling)


Spatial Attention
Spatial Attention
Spatial Attention
Spatial Attention
Features
Features
Features
Features
s
P Spatial attention weights during the whole sequence
Temporal Attention (Dynamic Pooling)



Spatial Attention
Spatial Attention
Spatial Attention
Spatial Attention
s
Features
Features
Features
Features
P Spatial attention weights during the whole sequence
Temporal Attention (Dynamic Pooling)
ũ
Cues from RGB crops

Results: comparison w. state of the art

Methods	Pose	RGB	CS	CV	Avg
Lie Group [37]	X	-	50.1	52.8	51.5
Skeleton Quads [9]	X	-	38.6	41.4	40.0
Dynamic Skeletons [13]	X	-	60.2	65.2	62.7
HBRNN [8]	X	-	59.1	64.0	61.6
Deep LSTM [30]	X	-	60.7	67.3	64.0
Part-aware LSTM [30]	X	-	62.9	70.3	66.6
ST-LSTM + TrustG. [23]	X	-	69.2	77.7	73.5
STA-LSTM [34]	X	-	73.2	81.2	77.2
JTM [39]	X	-	76.3	81.1	78.7
DSSCA - SSLM [31]	X	X	74.9	-	-
Ours (pose only)	X	-	**77.1**	**84.5**	**80.8**
Ours (RGB only)	-	X	**75.6**	**80.5**	**78.1**
Ours (pose +RGB)	X	X	**84.8**	**90.6**	**87.7**

Table 1: Results on the NTU RGB+D dataset with Cross-Subject (CS) and Cross-View (CV) settings (accuracies in %)

Transfer learning

Methods	Pose	RGB	Depth	Acc.
Raw skeleton [45]	X	-	-	49.7
Joint feature [45]	X	-	-	80.3
Raw skeleton [46]	X	-	-	79.4
Joint feature [46]	X	-	-	86.9
HBRNN [8]	X	-	-	80.35
Co-occurence RNN [47]	X	-	-	90.4
STA-LSTM [34]	X	-	-	91.5
ST-LSTM + Trust Gate [23]	X	-	-	93.3
DSPM [22]	-	X	X	93.4
Ours (Pose only)	X	-	-	90.5
Ours (RGB only)	-	X	-	72.0
Ours (Pose + RGB)	X	X	-	94.1

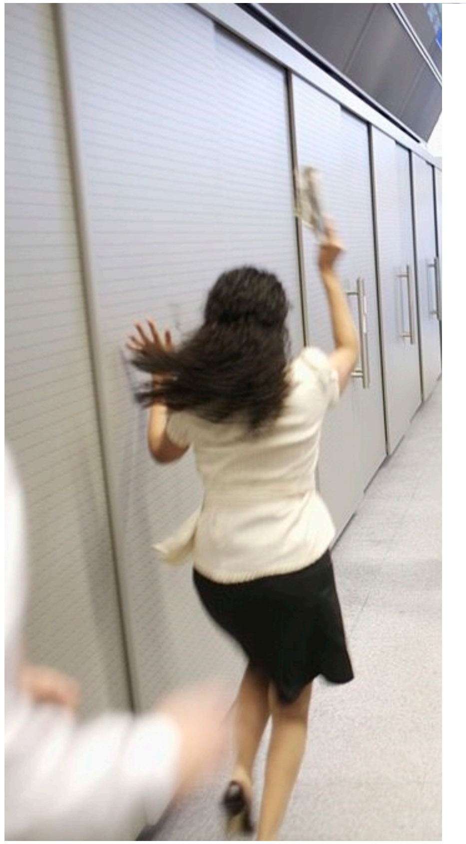
Table 2: Results on SBU Kinect Interaction dataset (accuracies in %)

Methods	Pose	RGB	Depth	Acc.
Action Ensemble [38]	X	-	-	68.0
Efficient Pose-Based [10]	X	-	-	73.1
Moving Pose [47]	X	-	-	73.8
Moving Poselets [36]	X	-	-	74.5
Depth Fusion [48]	-	-	X	88.8
MMMP [32]	X	-	X	91.3
DL-GSGC [24]	X	-	X	95.0
DSSCA - SSLM [31]	-	X	X	97.5
Ours (Pose only)	X	-	-	74.6
Ours (RGB only)	-	X	-	75.3
Ours (Pose + RGB)	X	X	-	90.0

Table 3: Results on MSR Daily Activity 3D dataset (accuracies in %)

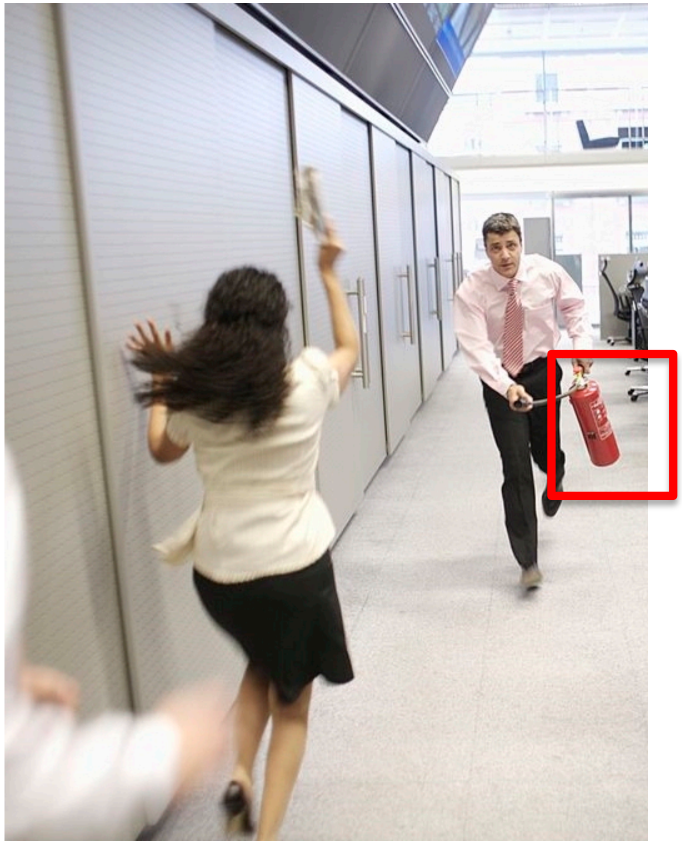
[Baradel, Wolf, Mille, BMVC 2018]

Context



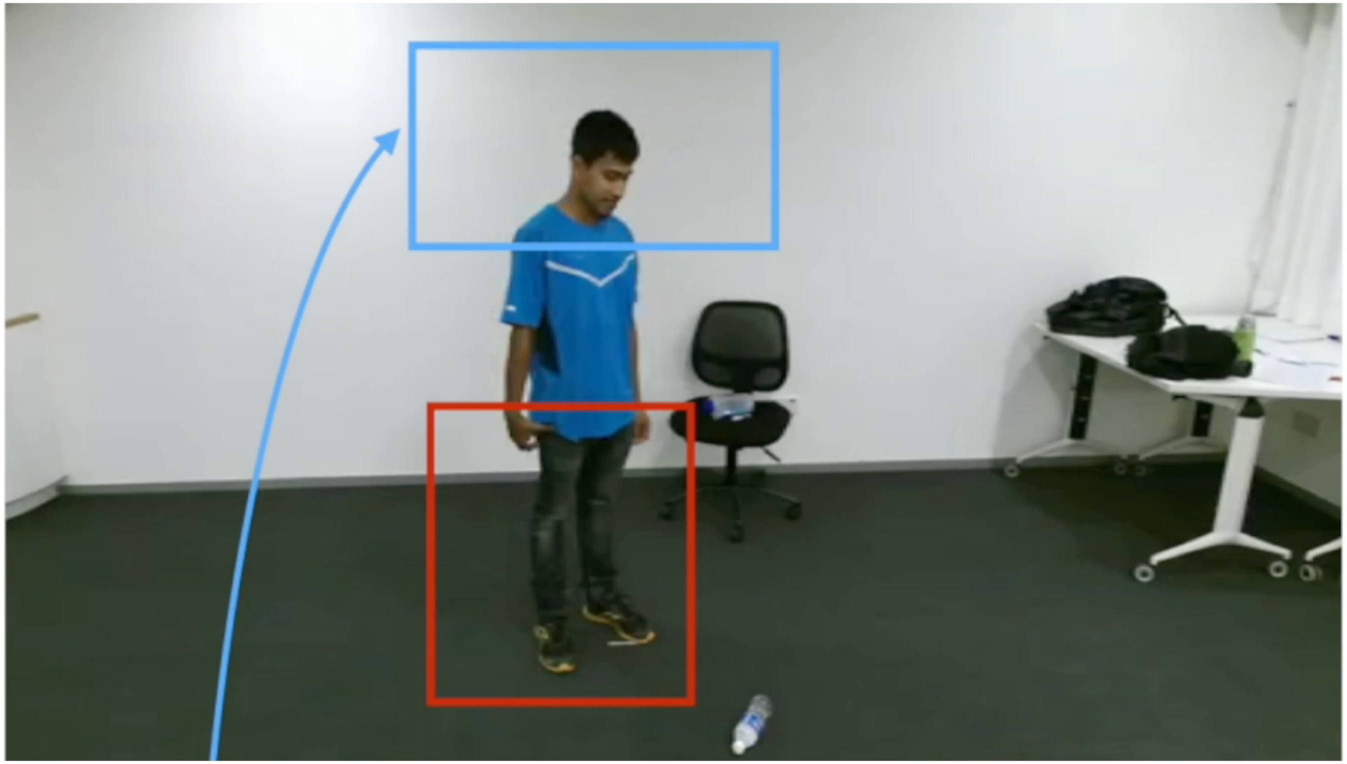
We need to put attention to places which are not always determined by pose

Context



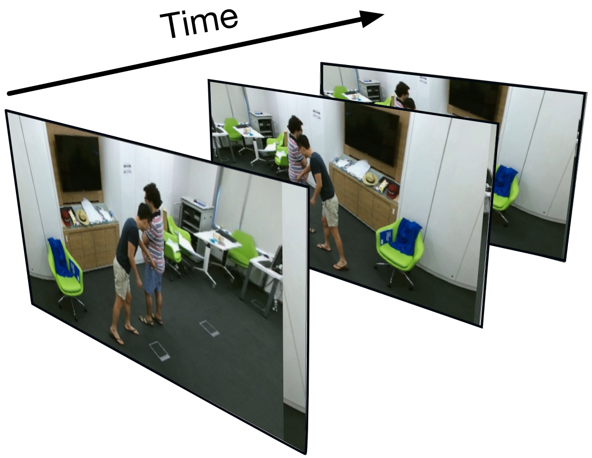
We need to put attention to places which are not always determined by pose

Dynamic spatio-temporal attention



[Baradel, Wolf, Mille, Taylor, CVPR 2018]

Dynamic visual attention

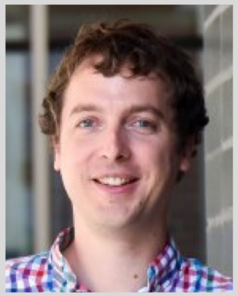


RGB input video

1. Learn where to attend
2. Learn how to track attended glimpse points (assign glimpses to semantic entities)
3. Learn how to recognize activities from a collection of tracked semantic entities

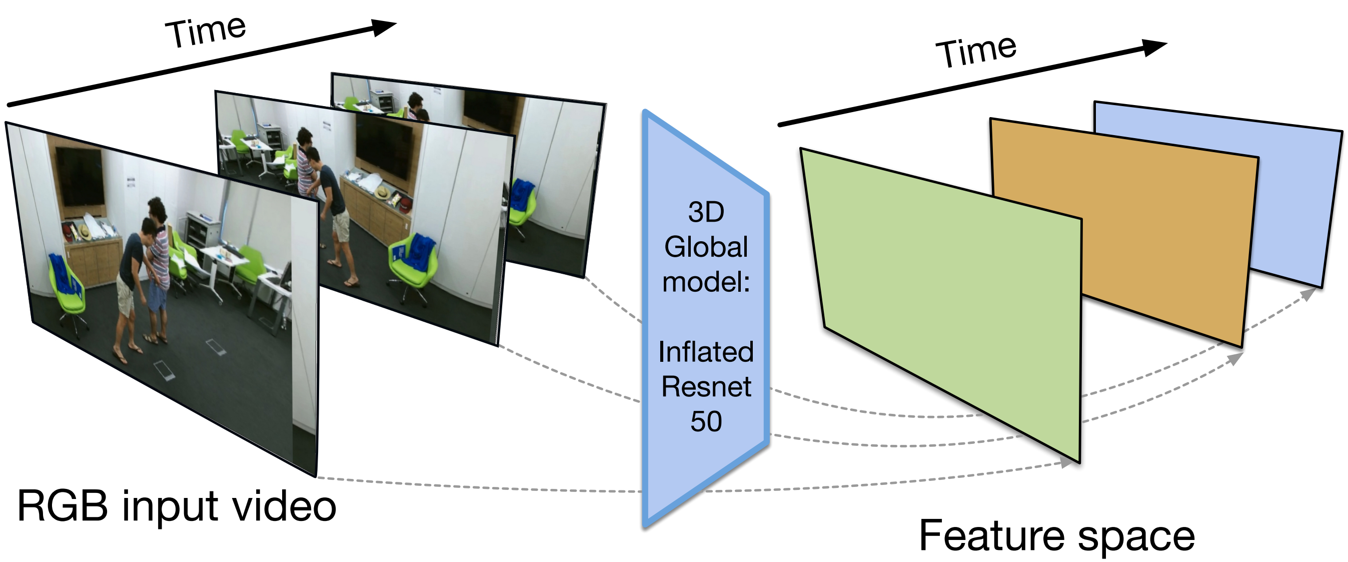
Work of
Fabien Baradel,
Phd @ LIRIS

With Julien
Mille
(INSA VdL)



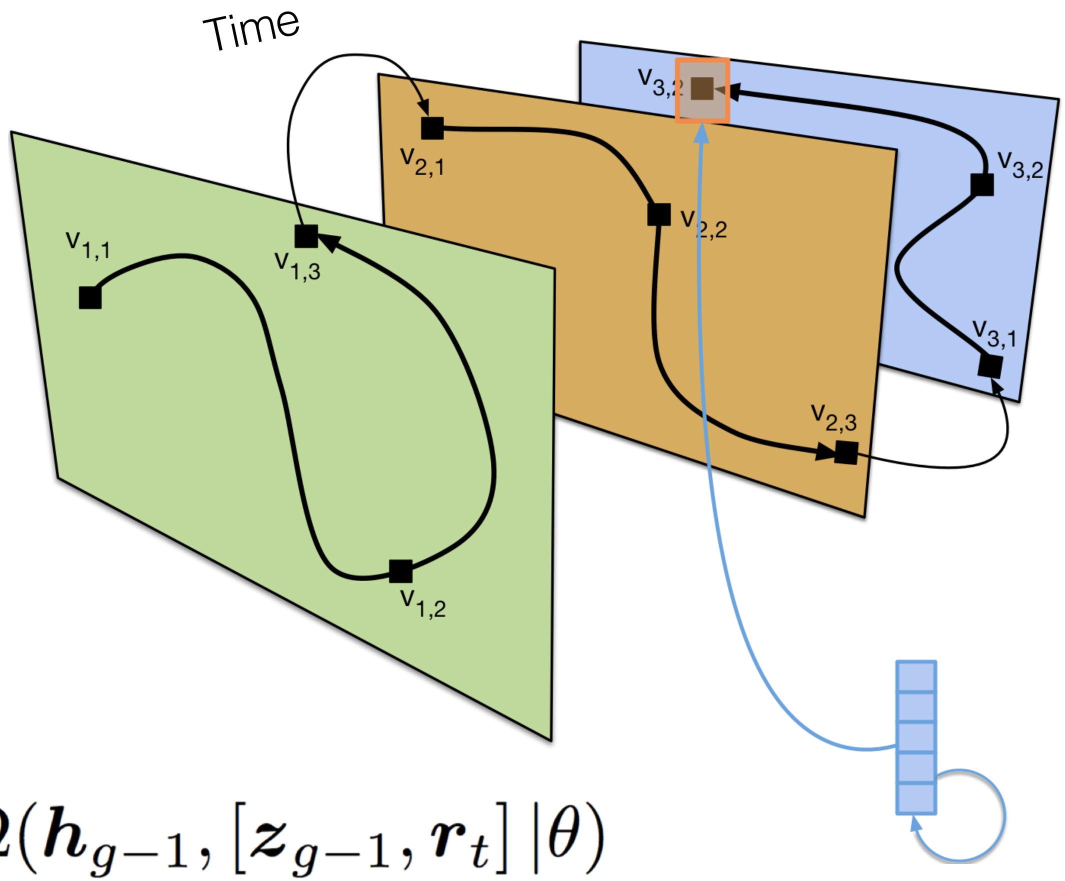
With Graham W. Taylor
(Univ. of Guelph,
Vector Institut)

Attention in feature space



[Baradel, Wolf, Mille, Taylor, CVPR 2018]

Unconstrained differentiable attention



$$\boldsymbol{h}_g = \Omega(\boldsymbol{h}_{g-1}, [\boldsymbol{z}_{g-1}, \boldsymbol{r}_t] \,|\theta)$$

$$\boldsymbol{l}_g = W_l^\top \left[\boldsymbol{h}_g, \boldsymbol{c}_t\right]$$

Hidden state from recurrent recognizers (workers)

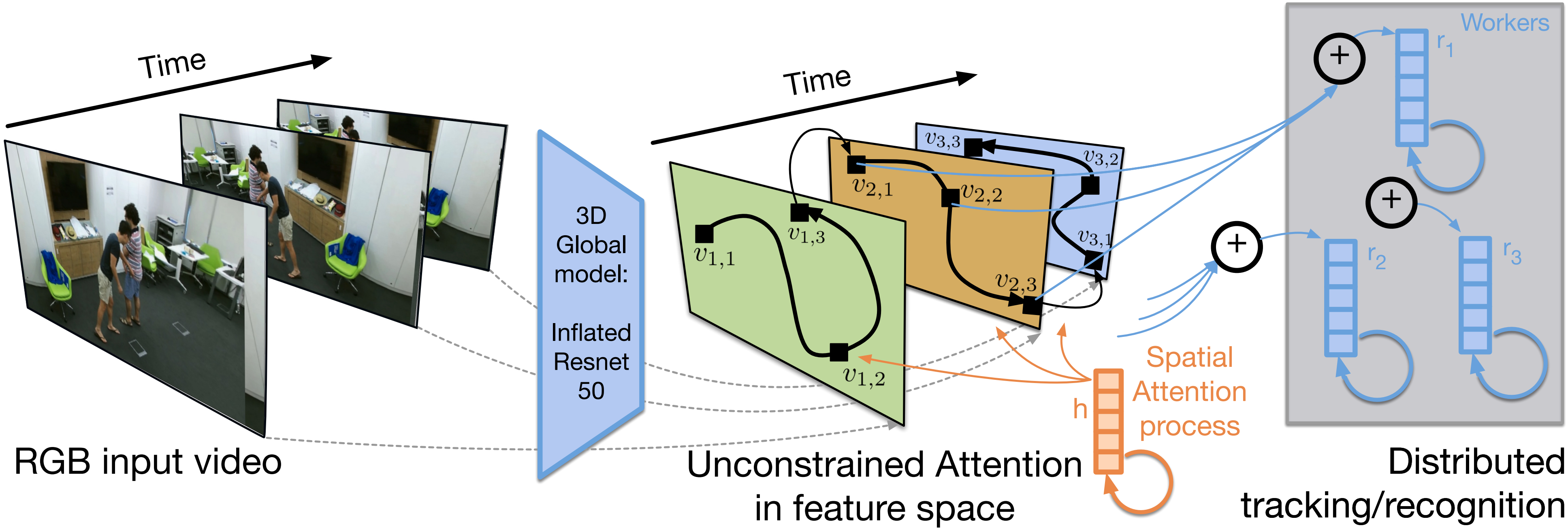
Frame context

[Baradel, Wolf, Mille, Taylor, CVPR 2018]

"Differentiable crop »

(Spatial Transformer Network)

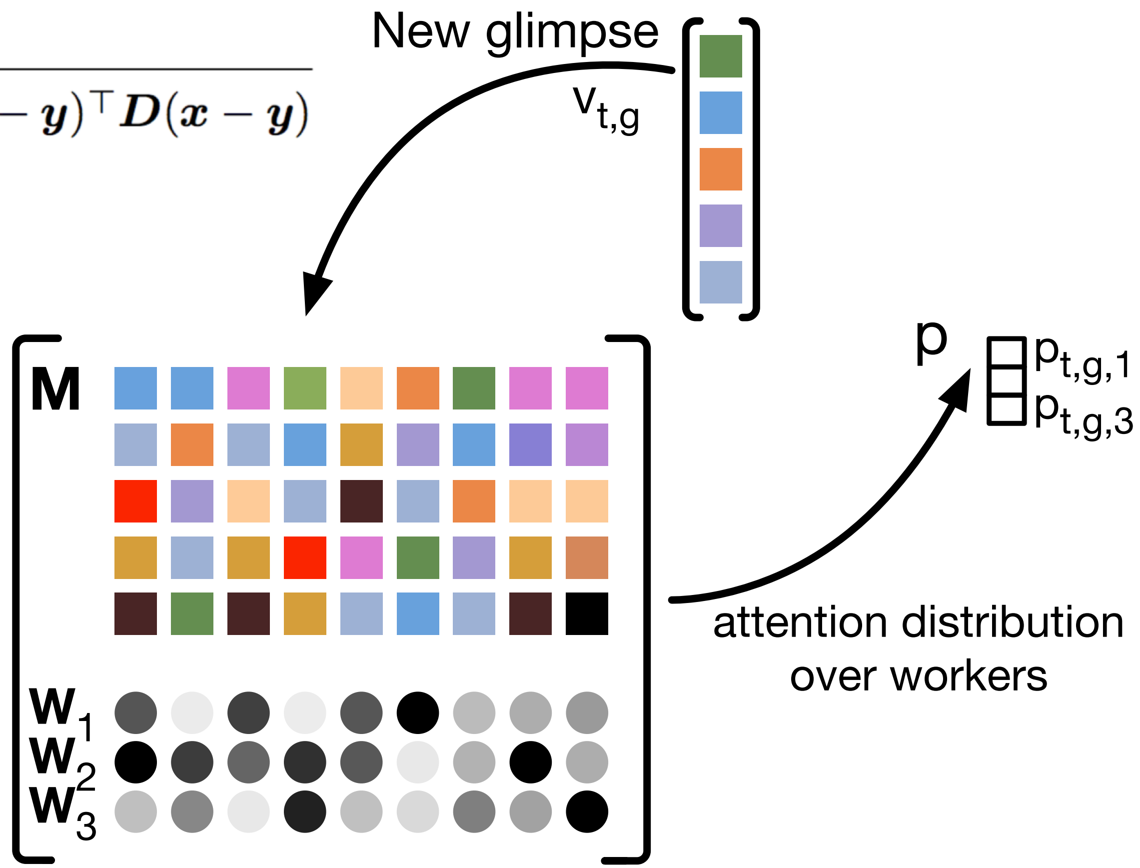
Distributed recognition



[Baradel, Wolf, Mille, Taylor, CVPR 2018]

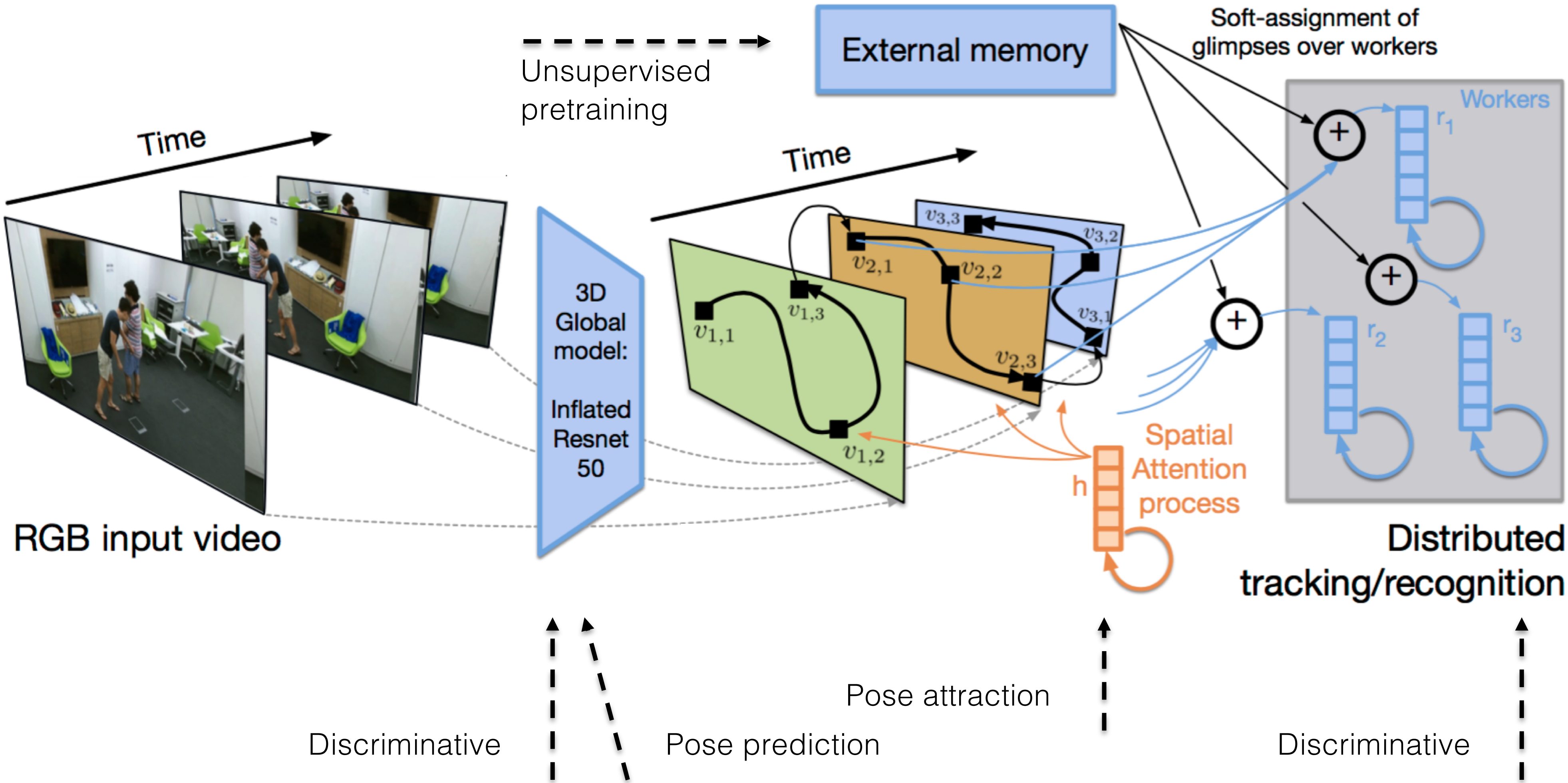
Soft-assignment of glimpses to workers

$$\phi(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^\top \boldsymbol{D}(\boldsymbol{x}-\boldsymbol{y})}$$



[Baradel, Wolf, Mille, Taylor, CVPR 2018]

Intermediate supervision



[Baradel, Wolf, Mille, Taylor, CVPR 2018]

State-of-the-art comparaison

Methods	Pose	RGB	CS	CV	Avg
Lie Group [40]	✓	-	50.1	52.8	51.5
Skeleton Quads [10]	✓	-	38.6	41.4	40.0
Dynamic Skeletons [14]	✓	-	60.2	65.2	62.7
HBRNN [9]	✓	-	59.1	64.0	61.6
Deep LSTM [32]	✓	-	60.7	67.3	64.0
Part-aware LSTM [32]	✓	-	62.9	70.3	66.6
ST-LSTM + TrustG. [26]	✓	-	69.2	77.7	73.5
STA-LSTM [35]	✓	-	73.2	81.2	77.2
Ensemble TS-LSTM [24]	✓	-	74.6	81.3	78.0
GCA-LSTM [27]	✓	-	74.4	82.8	78.6
JTM [41]	✓	-	76.3	81.1	78.7
MTLN [18]	✓	-	79.6	84.8	82.2
VA-LSTM [47]	✓	-	79.4	87.6	83.5
View-invariant [28]	✓	-	80.0	87.2	83.6
DSSCA - SSLM [33]	✓	✓	74.9	-	-
Hands Attention [5]	✓	✓	84.8	90.6	87.7
C3D†	-	✓	63.5	70.3	66.9
Resnet50+LSTM†	-	✓	71.3	80.2	75.8
Glimpse Clouds	-	✓	**86.6**	**93.2**	**89.9**

Table 1. Results on the NTU RGB+D dataset with Cross-Subject and Cross-View settings (accuracies in %); († indicates method has been re-implemented).

Figure 1. Results on Northwestern-UCLA Multiview Action 3D, Cross-View (accuracy in %). V=Visual(RGB), D=Depth, P=Pose.

Methods	Data	$V_{1,2}^{3}$	$V_{1,3}^{2}$	$V_{2,3}^{1}$	Avg
DVV [5]	D	58.5	55.2	39.3	51.0
CVP [11]	D	60.6	55.8	39.5	52.0
AOG [10]	D	45.2	-	-	-
HPM+TM [8]	D	**91.9**	75.2	71.9	79.7
Lie group [9]	P	74.2	-	-	-
HBRNN-L [1]	P	78.5	-	-	-
Enhanced viz. [6]	P	86.1	-	-	-
Ensemble TS-LSTM [3]	P	89.2	-	-	-
Hankelets [4]	V	45.2	-	-	-
nCTE [2]	V	68.6	68.3	52.1	63.0
NKTM [7]	V	75.8	73.3	59.1	69.4
Global model	V	85.6	84.7	79.2	83.2
Glimpse Clouds	V	90.1	**89.5**	**83.4**	**87.6**

SOTA results on two datasets NTU and N-UCLA

Larger difference between Glimpse clouds and global model on N-UCLA

[Baradel, Wolf, Mille, Taylor, CVPR 2018]