Glimpse Clouds: Human Activity Recognition from Unstructured Feature Points

Fabien Baradel, Christian Wolf, Julien Mille,
Graham W. Taylor

- Authorship
- Background
- Proposed Method
- Experiments
- Conclusion

ResNet50

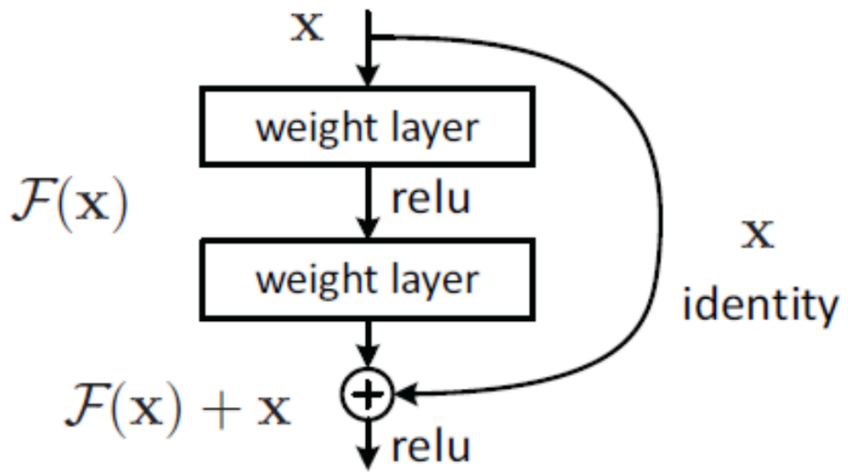
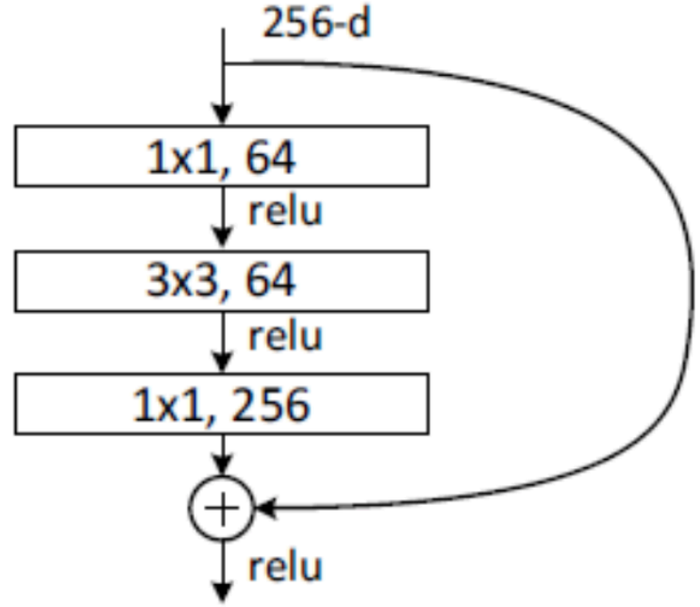


Figure 2. Residual learning: a building block.



- Image classification
- Residual learning & bottleneck
- K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition

Attention and External Memory

- Soft attention and hard attention
  - Soft: weights each part of the observation dynamically
  - Hard: takes hard decisions when choosing parts of the input data
- External memory for extra information

- Authorship
- Background
- Proposed Method
- Experiments
- Conclusion

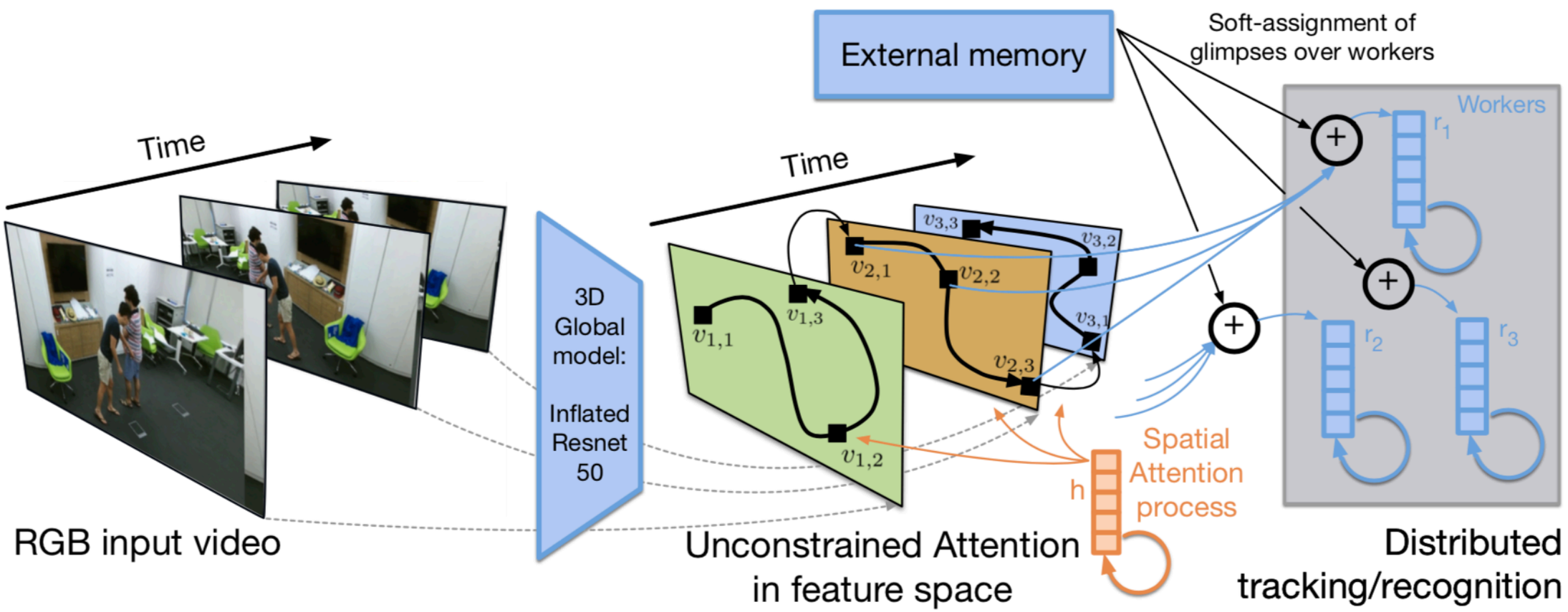
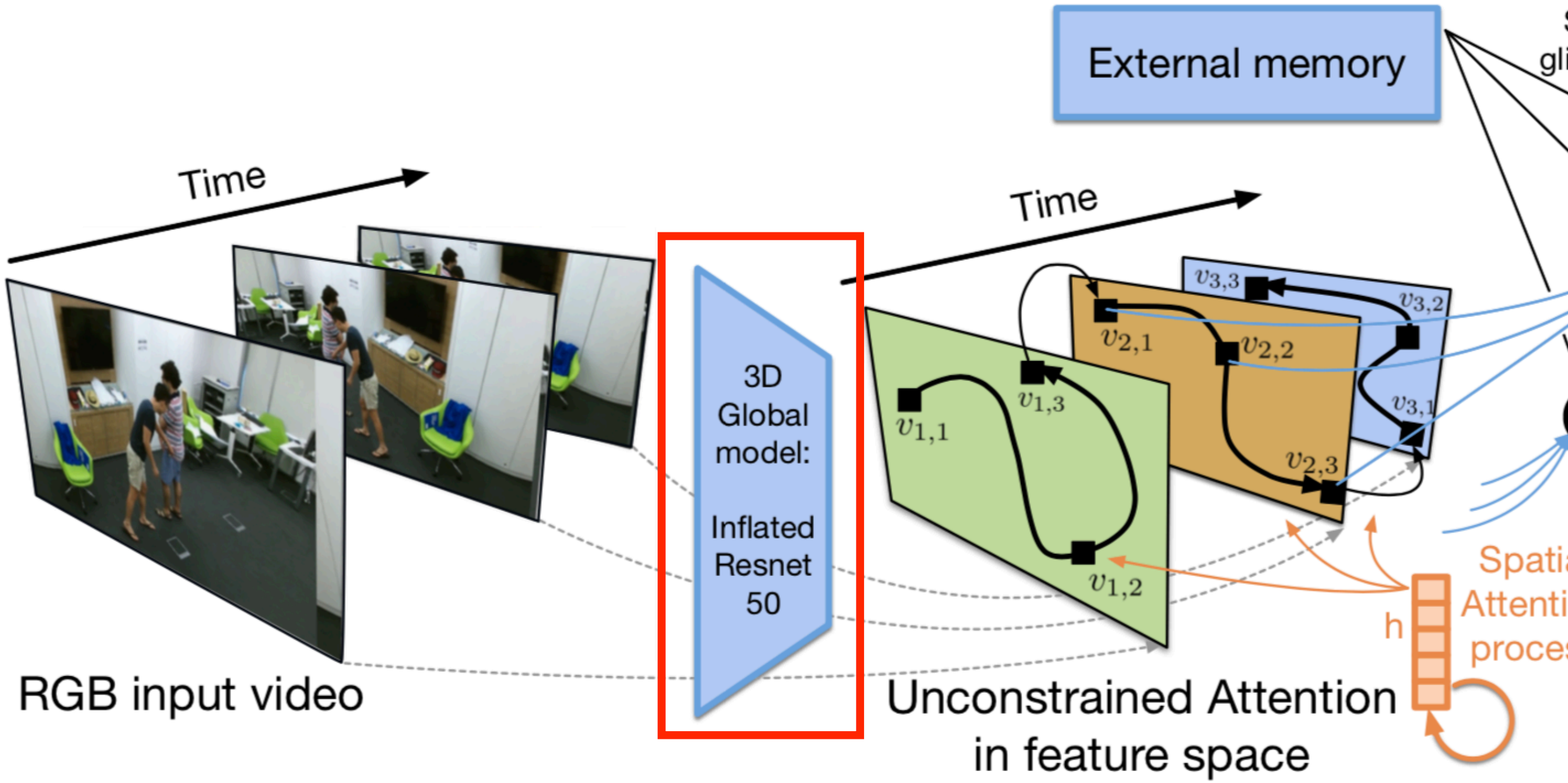


Figure 1. We recognize human activities from unstructured collections of spatio-temporal glimpses with distributed recurrent tracking/recognition and soft-assignment among glimpse points and trackers.



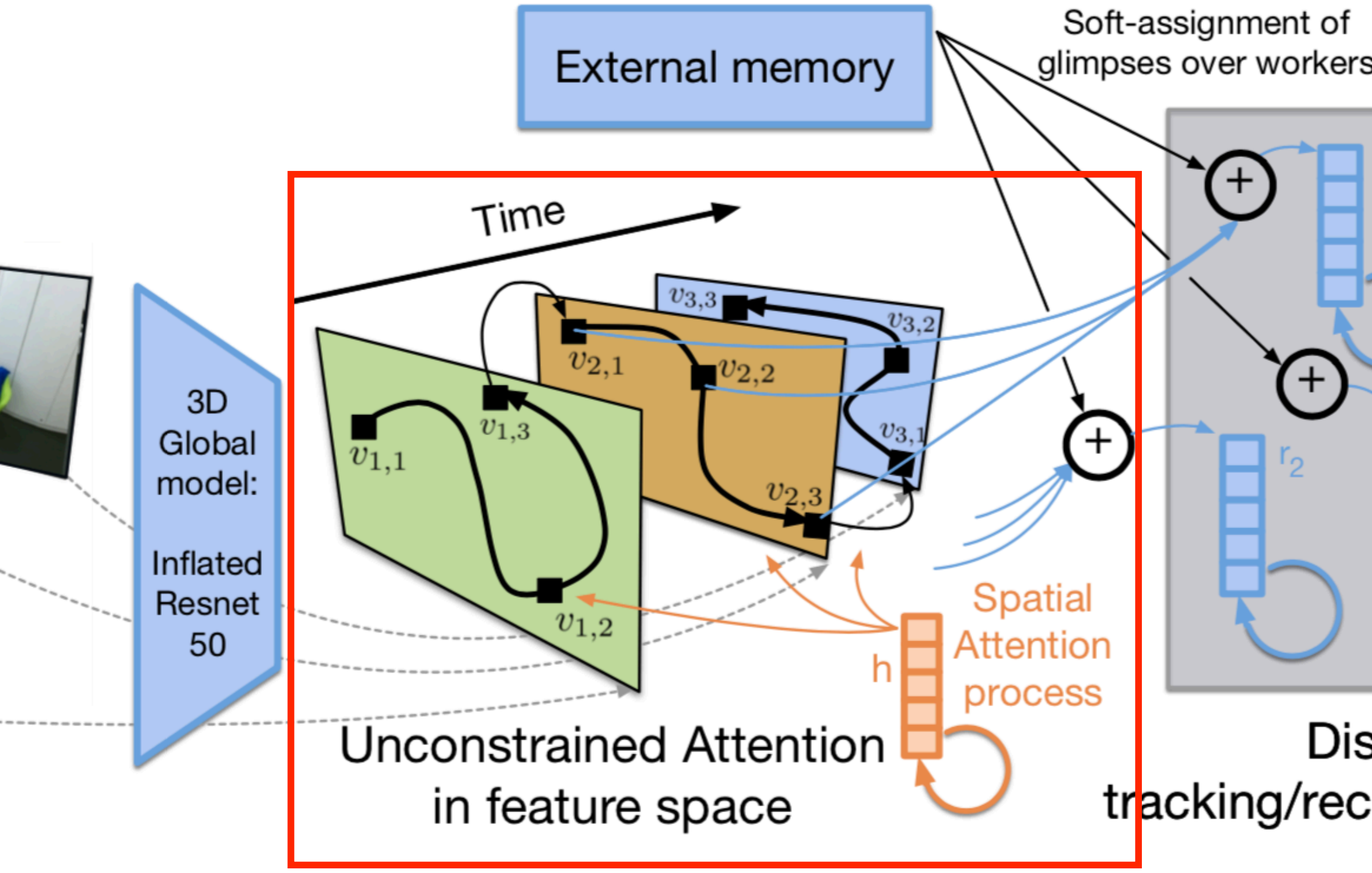
e 1. We recognize human activities from unstructured collections of spatio-temporal glimps
cognition and soft-assignment among glimpse points and trackers.

- Proceed from the ResNet-50 network and inflate the 2D spatial convolutional kernels into 3D kernels

- Input: T * H * W * 3

- Output: T * H’ * W’ * C’

- Temporal receptive field is more than one frame



from unstructured collections of spatio-temporal glimpses with distribu

- Attention representation $\boldsymbol{l}_{t,g} = \left[x_g, y_g, s_g^x, s_g^y\right]_t^\top$
- Attention feature $\boldsymbol{z}_{t,g} = \Gamma(\boldsymbol{Z}_{t,g}) = \frac{1}{H'W'} \sum_m \sum_n \boldsymbol{Z}_{t,g}(m,n)$
- Feature dimension: 1 * C

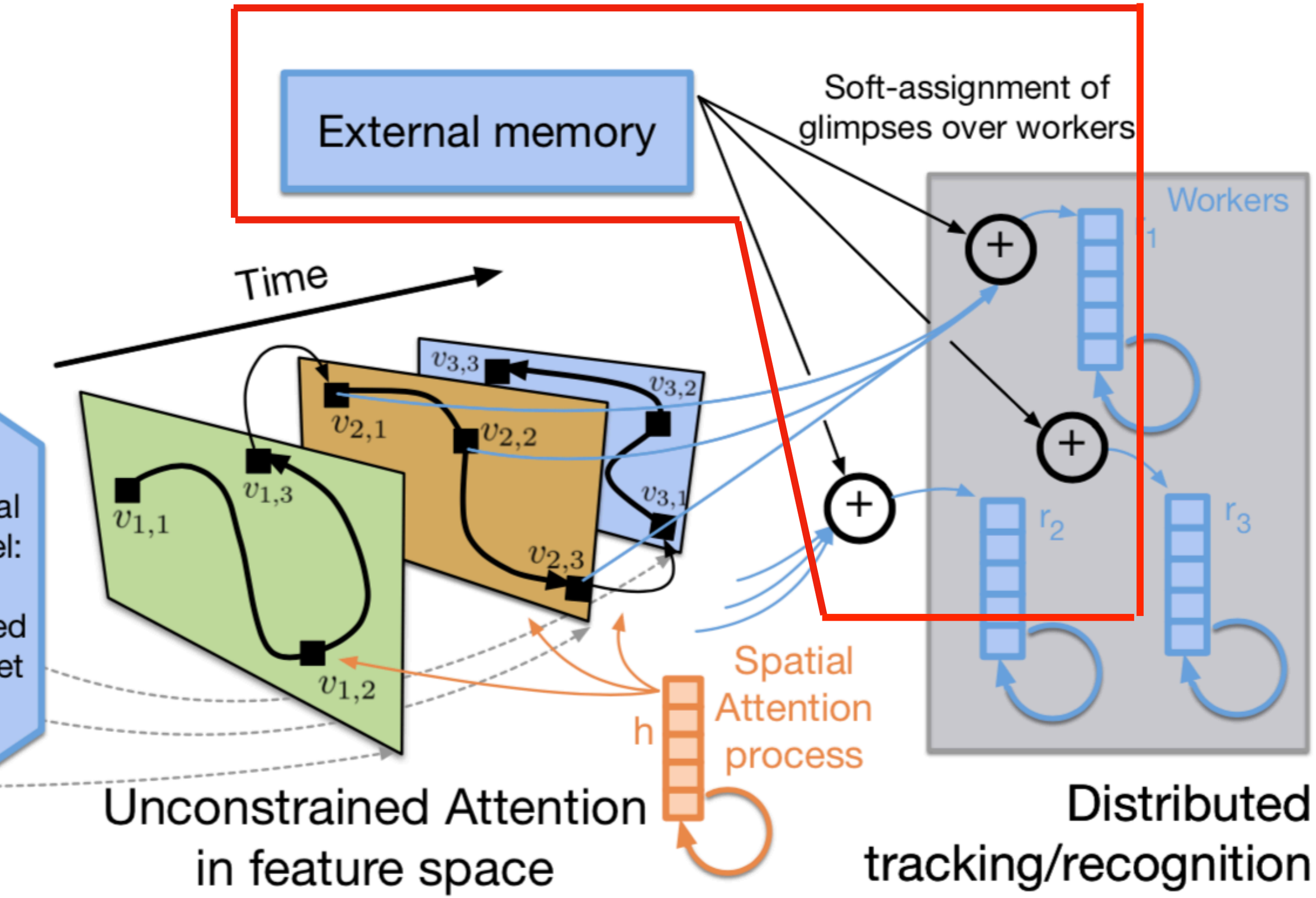
- GRU to predict next:

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

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

To be trained

- P.s. c & r coming soon



ctured collections of spatio-temporal glimpses with distributed recurrent tra

To be trained

- “what” and “where” features: $\boldsymbol{v}_{t,g} = \boldsymbol{z}_{t,g} \otimes \Lambda(\boldsymbol{l}_{t,g}|\theta_\Lambda)$
- “workers”: GRUs, take the “important” features
- Input to the workers, linear combination: $\tilde{\boldsymbol{v}}_{t,c} = \boldsymbol{V}_t \boldsymbol{p}_{t,c}$

- Higher weights to glimpses assigned to this worker in the past
- Using external memory to get the combination weight

To be trained

- Definition of distance: $\phi(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^\top \boldsymbol{D}(\boldsymbol{x}-\boldsymbol{y})}$
- External memory bank: $\boldsymbol{M} = \{\boldsymbol{m}_k\}$
- Each worker's own bank: $\boldsymbol{W}_c = \{w_{c,k}\}$
- Distance pre-trained as similarity
- p sums to 1: $p_{t,c,g} = \sigma_\alpha \left(\sum_k e^{-t^{m_k}} \times w_{c,k} \left[1 - \phi(\boldsymbol{v}_{t,g}, \boldsymbol{m}_k)\right] \right)$
- Memory bank delete the oldest one when eviction

External memory

Soft-assignment of
glimpses over workers

Workers

r_1

r_2

r_3

Time

$v_{3,3}$

$v_{3,2}$

$v_{2,1}$

$v_{2,2}$

$v_{3,1}$

$v_{1,3}$

$v_{2,3}$

$v_{1,2}$

h

Spatial
Attention
process

nconstrained Attention
in feature space

Distributed
tracking/recognition

- “workers”: GRUs

To be trained

$$\boldsymbol{r}_{t,c} = \Psi_c(\boldsymbol{r}_{t-1,c}, \tilde{\boldsymbol{v}}_{t,c}|\theta_{\Psi_c})$$

$$\boldsymbol{r}_t = \sum_c \boldsymbol{r}_{t,c}$$

- Recognition:

$$\boldsymbol{q}_c = W_c \cdot \boldsymbol{r}_c$$

$$\hat{\boldsymbol{y}} = \text{softmax}\left(\sum_c^C \boldsymbol{q}_c\right)$$

- $\boldsymbol{c}_t$ obtained by global average pooling over the spatial domain of the penultimate feature maps of a given timestep

- Pose estimation: $\boldsymbol{y}_t^p = W_p^\top \boldsymbol{c}_t$

- Recall(P16): $\boldsymbol{h}_g = \Omega(\boldsymbol{h}_{g-1}, [\boldsymbol{z}_{g-1}, \boldsymbol{r}_t] \,|\theta)$

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

Loss Function

- 3 parts: $\mathcal{L} = \mathcal{L}_D(\hat{\boldsymbol{y}}, \boldsymbol{y}) + \mathcal{L}_P(\hat{\boldsymbol{y}}^p, \boldsymbol{y}^p) + \mathcal{L}_G(\boldsymbol{l}, \boldsymbol{y}^p)$
- 1 classification, cross-entropy loss
- 2 pose estimation

$$\mathcal{L}_{G_1}^t(\boldsymbol{l}, \boldsymbol{y}^p) = \frac{1}{1 + \sum_{g_1}^{G} \sum_{g_2}^{G} ||\boldsymbol{l}_{t,g_1}, \boldsymbol{l}_{t,g_2}||}$$

- 3 glimpse

$$\mathcal{L}_{G_2}^t(\boldsymbol{l}, \boldsymbol{y}^p) = \sum_{g}^{G} \min_{j} ||\boldsymbol{l}_{t,g}, \boldsymbol{y}_j^p||$$

$$\boldsymbol{l}_{t,g} = \left[x_{t,g}, y_{t,g}, s_{t,g}^x, s_{t,g}^y\right]^T$$

$$\mathcal{L}_G(\boldsymbol{l}, \boldsymbol{y}^p) = \sum_{t}^{T} \left(\mathcal{L}_{G_1}^t(\boldsymbol{l}, \boldsymbol{y}^p) + \mathcal{L}_{G_2}^t(\boldsymbol{l}, \boldsymbol{y}^p)\right)$$

- Authorship
- Background
- Proposed Method
- Experiments
- Conclusion

Two Datasets

- NTU RGB+D Dataset
- N-UCLA: Northwestern-UCLA Multiview Action 3D Dataset
  - Use the model trained on NTU as a pretrained model and fine-tune it on N-UCLA

Methods	Pose	RGB	CS	CV	Avg
Lie Group [51]	✓	-	50.1	52.8	51.5
Skeleton Quads [13]	✓	-	38.6	41.4	40.0
Dynamic Skeletons [18]	✓	-	60.2	65.2	62.7
HBRNN [11]	✓	-	59.1	64.0	61.6
Deep LSTM [43]	✓	-	60.7	67.3	64.0
Part-aware LSTM [43]	✓	-	62.9	70.3	66.6
ST-LSTM + TrustG. [33]	✓	-	69.2	77.7	73.5
STA-LSTM [46]	✓	-	73.2	81.2	77.2
Ensemble TS-LSTM [29]	✓	-	74.6	81.3	78.0
GCA-LSTM [34]	✓	-	74.4	82.8	78.6
JTM [53]	✓	-	76.3	81.1	78.7
MTLN [23]	✓	-	79.6	84.8	82.2
VA-LSTM [59]	✓	-	79.4	87.6	83.5
View-invariant [35]	✓	-	80.0	87.2	83.6
DSSCA - SSLM [44]	✓	✓	74.9	-	-
STA-Hands [5]	X	X	82.5	88.6	85.6
Hands Attention [6]	✓	✓	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 2. Results on the NTU RGB+D dataset with Cross-Subject and Cross-View settings (accuracies in %); († indicates method has been re-implemented).

Methods	Spatial Attention	Soft Workers	L_D	L_P	L_G	CS	CV	Avg
Global model	-	-	✓	-	-	84.5	91.5	88.0
Global model	-	-	✓	✓	-	85.5	92.1	88.8
Global model+$\sum$ Glimpses + GRU	-	-	✓	✓	-	85.8	92.4	89.1
Glimpse Clouds	✓	✓	✓	-	-	85.7	92.5	89.1
Glimpse Clouds	✓	✓	✓	✓	-	86.4	93.0	89.7
Glimpse Clouds	✓	✓	✓	-	✓	86.1	92.9	89.5
Glimpse Clouds	✓	✓	✓	✓	✓	**86.6**	**93.2**	**89.9**
Glimpse Clouds + Global model	✓	✓	✓	✓	✓	86.6	93.2	89.9

Table 3. Results on NTU: ablation study

Glimpse	Type of attention	CS	CV	Avg
3D tubes	Attention	85.8	92.7	89.2
Seq. 2D	Random sampling	80.3	87.8	84.0
Seq. 2D	Saliency	86.2	92.9	89.5
Seq. 2D	**Attention**	**86.6**	**93.2**	**89.9**

Table 4. Results on the NTU: different attention and alternative strategies.

Predict glimpse location through a location network. No past information

- Other experiments determines other parameters like the worker number or the glimpse number

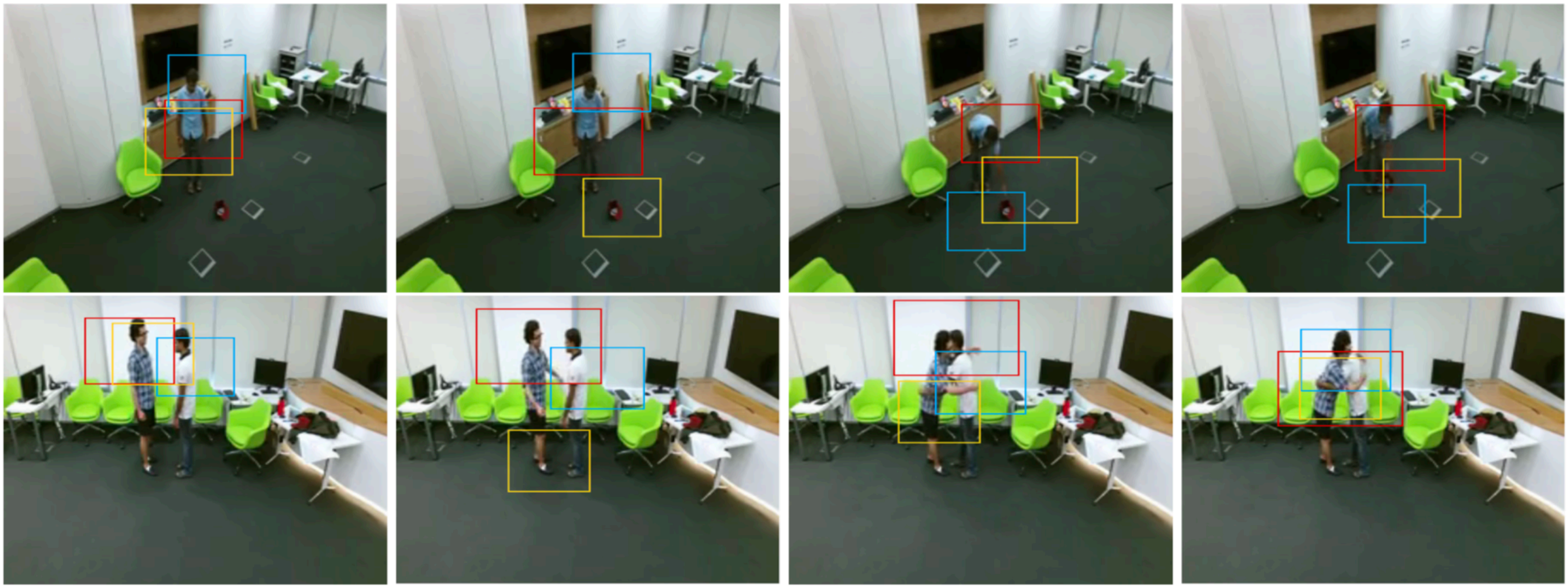


Figure 3. An illustration of the glimpse distribution for several sequences of the NTU dataset. Here we set 3 glimpses per frame (G=3, Red: first glimpse, Blue: second glimpse, Yellow: third one).

- Authorship
- Background
- Proposed Method
- Experiments
- Conclusion

- Only RGB when testing, though pose information when training

- Attention process, soft-assigned

- A little bit complex but code released