Relational Action Forecasting

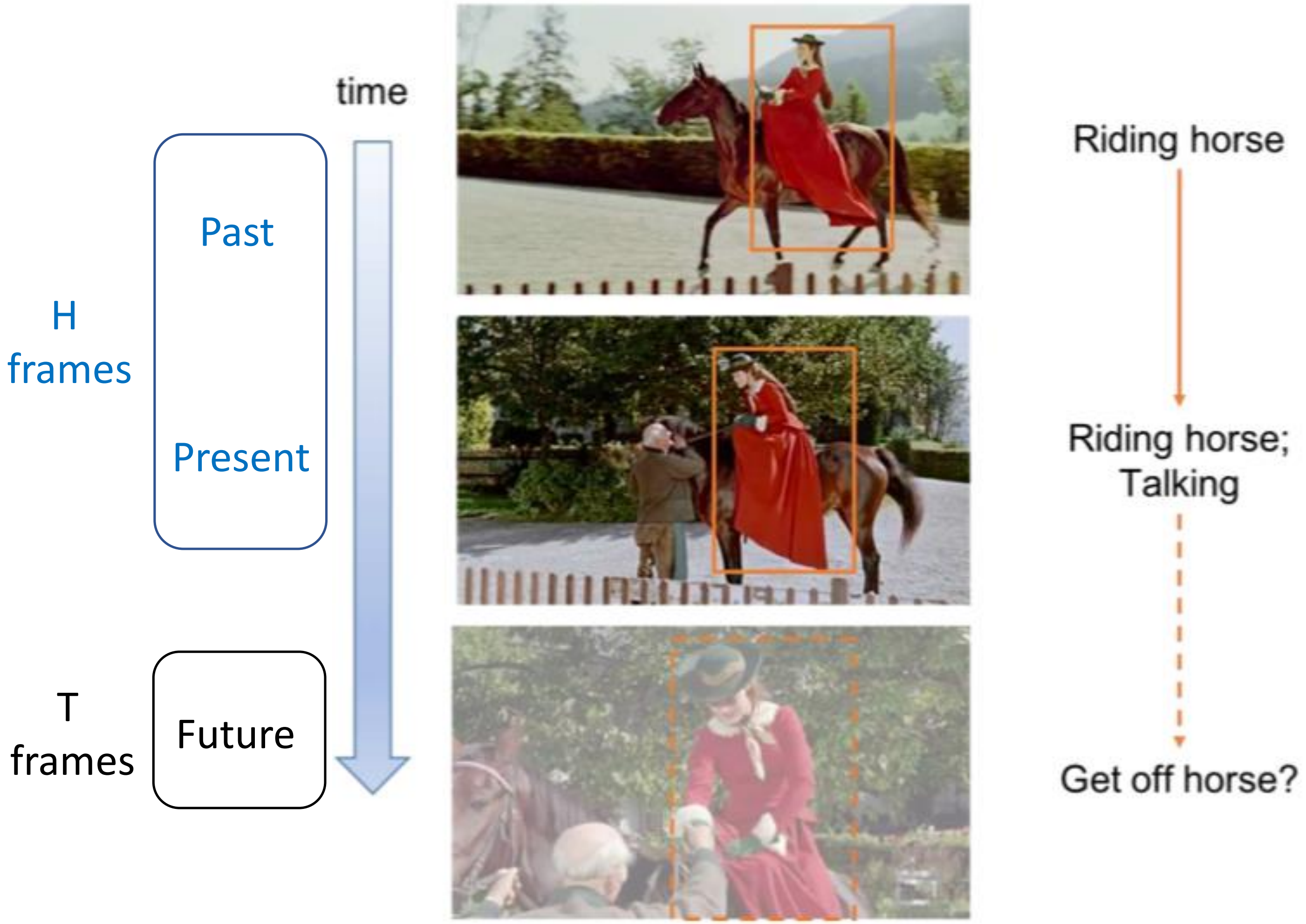
Chen Sun; Abhinav Shrivastava et al., CVPR 2019

Presenters: Wanda Wang, Vincent Chung, Shreyas Hosamane

UCDAVIS
UNIVERSITY OF CALIFORNIA

Problem Statement

Given a history of H previous frames, the goal is to detect actors and to predict their future actions for the next T frames.



Why is action forecasting important?

Self-driving cars

Human interaction robot

Related work

- Action recognition
- Future prediction
- Relational reasoning

Action recognition

- Action classification
- Temporal action localization
- Spatio-temporal action detection

Action recognition: Action classification

$$p(a^{0:T}|V^{0:T})$$

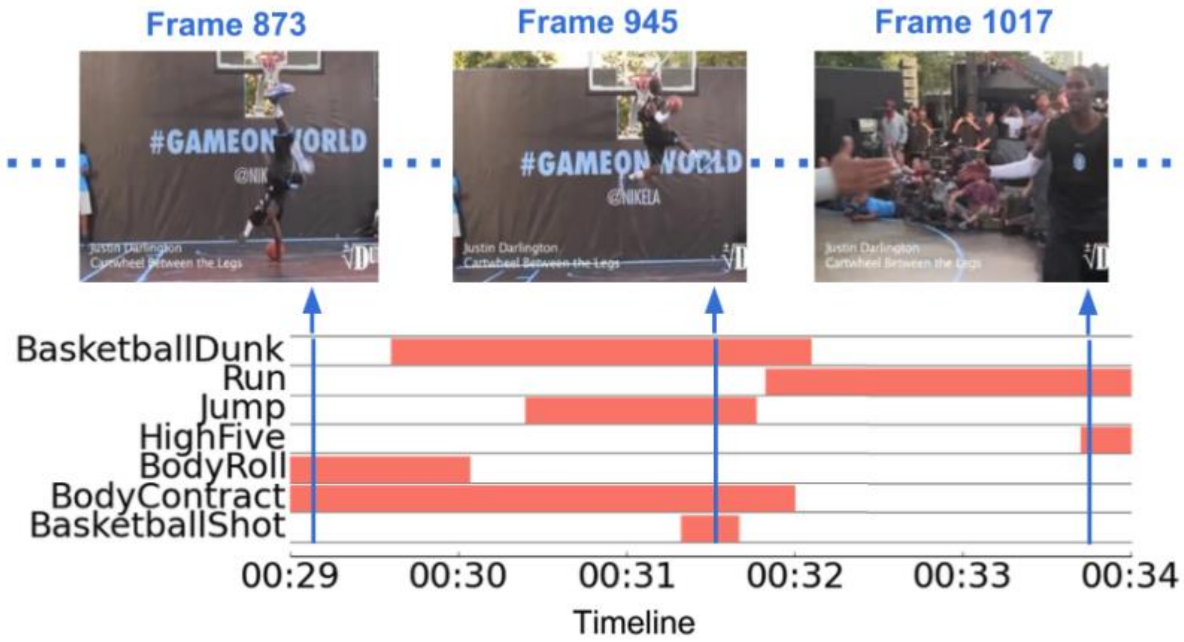
$V^{0:T}$: frames of the entire video
$a^{0:T}$: predicted action labels

M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009
S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *IJCV*, 2017
Song, L., Zhang, S., Yu, G., & Sun, H. (2019). TACNet: Transition-Aware Context Network for Spatio-Temporal Action Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 11987-11995).

Action recognition: Temporal action localization

$$p(a^{0:T}, b^{0:T} | V^{0:T})$$

$V^{0:T}$: frames of the entire video
$a^{0:T}$: predicted action labels
$b^{0:T}$: predicted locations of actions

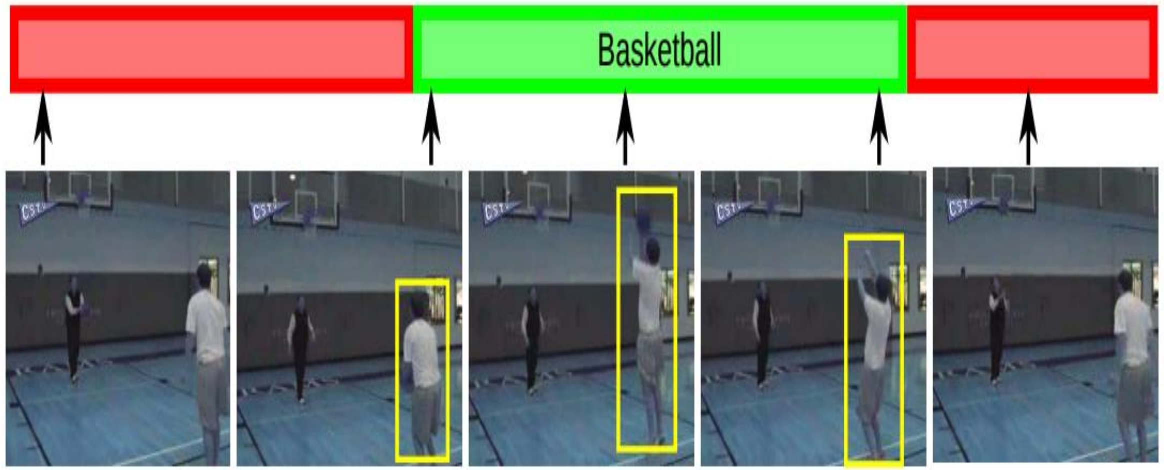


M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009
S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *IJCV*, 2017
Song, L., Zhang, S., Yu, G., & Sun, H. (2019). TACNet: Transition-Aware Context Network for Spatio-Temporal Action Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 11987-11995).

Action recognition: Spatio-temporal action detection

$$p(a_{1:N}^{0:T}, b_{1:N}^{0:T} | V^{0:T})$$

N : number of predicted actors
$V^{0:T}$: frames of the entire video
$a_{1:N}^{0:T}$: predicted action labels of N actors
$b_{1:N}^{0:T}$: predicted locations of N actors



M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009
S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *IJCV*, 2017
Song, L., Zhang, S., Yu, G., & Sun, H. (2019). TACNet: Transition-Aware Context Network for Spatio-Temporal Action Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 11987-11995).

Future prediction

Pixel prediction

Petrovic CVPR 2006, Vondrick NIPS 2016,
Walker ECCV 2016, Xue NIPS 2016,
Villegas ICML 2017, Denton ICML 2018,

Trajectory prediction

Kitani ECCV 2012, Alahi CVPR 2016,
Robicquet ECCV 2016, Lee CVPR 2017,
Gupta CVPR 2018, Sun ICLR 2019,

Relational reasoning

- This work aims to capture human-human relationships to reason about future actions

A.Alahi,K.Goel,V.Ramanathan,A.Robicquet,L.Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 2016.

Set of actions for last observed frame

Action forecasting

Sitting
Talking

Next:
Clink glass

Holding
Standing

Next:
Serving

Hugging
Standing

Next:
Kissing

CVPR 2019 Oral Session 1-1C: Action & Video : https://www.youtube.com/watch?v=JwaBi_2JFeU&t=1105s

Proposed Approach

Mathematical representation

$$p(N, b_{1:N}^{0}, a_{1:N}^{0:T} | V^{-H:0})$$

$V^{-H:0}$ → visual history of H previous frames

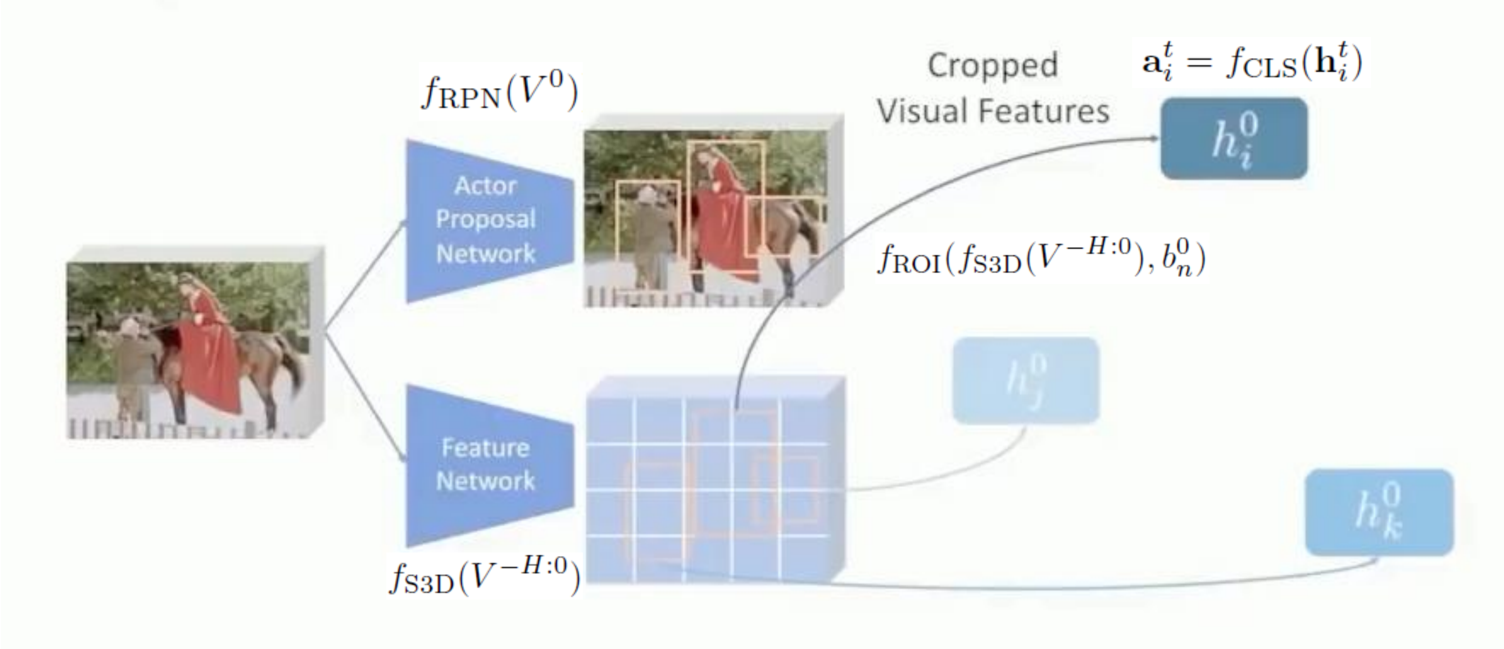
N→ number of predicted actors

$b_{1:N}^{0}$ → predicted locations (bounding boxes) of N actors at time t = 0

$a_{1:N}^{0:T}$ → predicted action labels for N actors for time t = 0:T

Proposed Approach

Creating the nodes in the graph

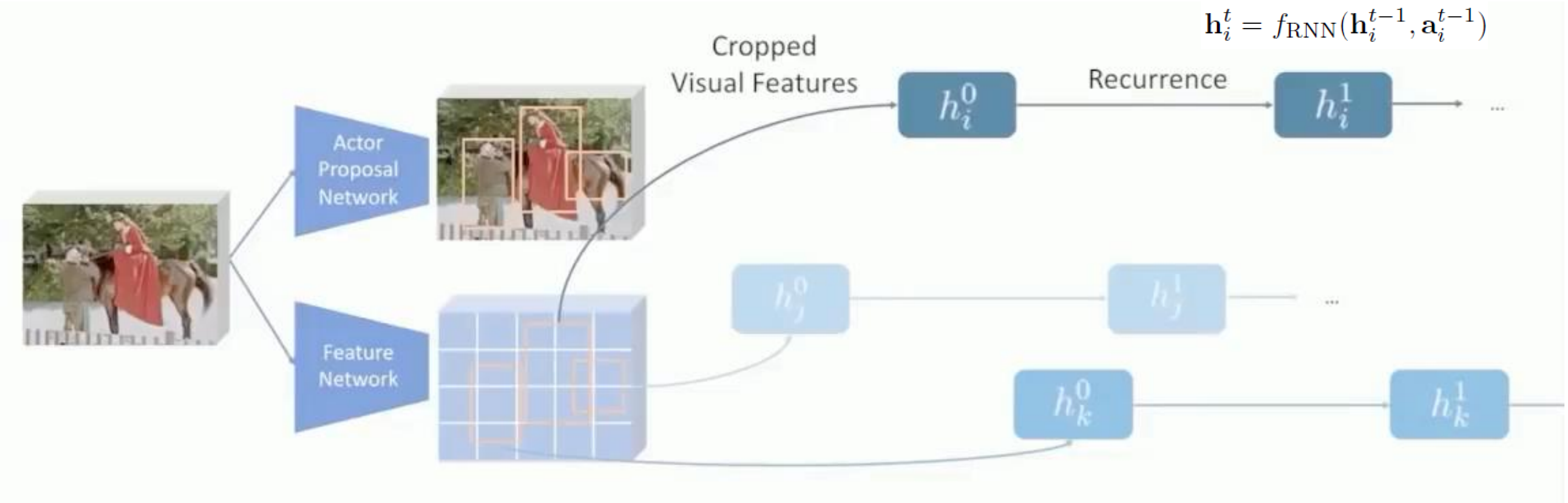


$$p(N, b_{1:N}^0, a_{1:N}^{0:T} | V^{-H:0}) = \delta(N, b_{1:N}^0 | f_{\text{RPN}}(V^0)) \underline{p(a_{1:N}^0, h_{1:N}^0 | b_{1:N}^0, V^{-H:0})} \prod_{t=1}^{T} p(a_{1:N}^t, h_{1:N}^t | a_{1:N}^{t-1}, h_{1:N}^{t-1})$$

$$\prod_{n=1}^{N} \text{Cat}(a_n^0 | f_{\text{CLS}}(h_n^0)) \delta(h_n^0 | f_{\text{ROI}}(f_{\text{S3D}}(V^{-H:0}), b_n^0))$$

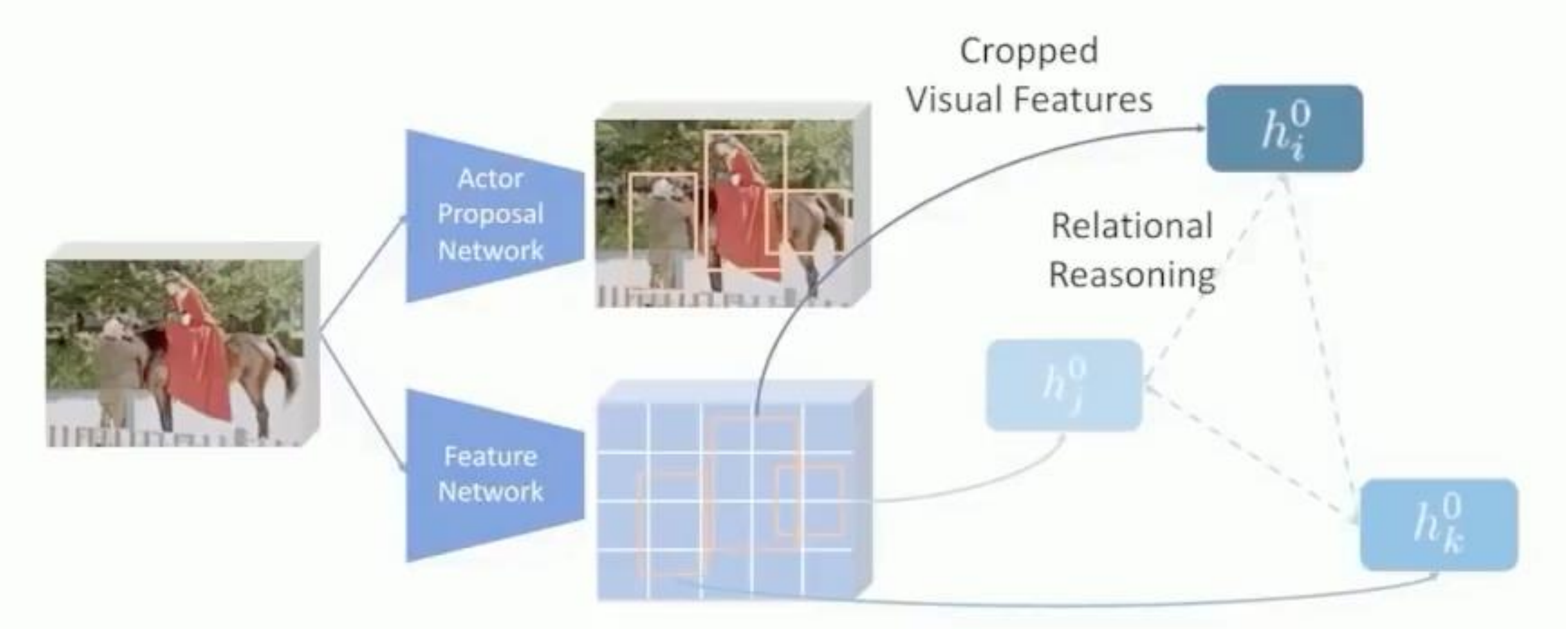
Proposed Approach

Modeling node dynamics



Proposed Approach

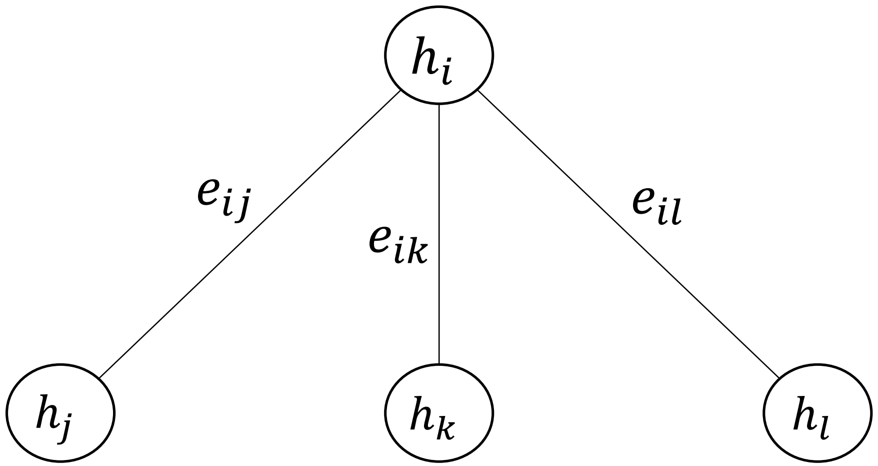
Modeling the edges



$$p(a^t_{1:N}, h^t_{1:N} | a^{t-1}_{1:N}, h^{t-1}_{1:N}) = \prod_{n=1}^{N} \mathrm{Cat}(a^t_n | f_{\mathrm{CLS}}(h^t_n)) \delta(h^t_n | f_{\mathrm{RNN}}(\tilde{h}^{t-1}_n, a^{t-1}_n)) \delta(\tilde{h}^{t-1}_n | f_{\mathrm{GNN}}(h^{t-1}_{1:N}))$$

Proposed Approach

Modeling the edges



$$\mathbf{e}_{ij} = f_{\text{edge}}(\mathbf{h}_i, \mathbf{h}_j)$$

$$\tilde{\mathbf{h}}_i = f_{\text{node}}\left(\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\mathbf{e}_{ij}\right)$$

$$\mathbf{h}_i^t = f_{\text{RNN}}(\tilde{\mathbf{h}}_i^{t-1}, \mathbf{a}_i^{t-1})$$

sensitive to noisy nodes

Proposed Approach

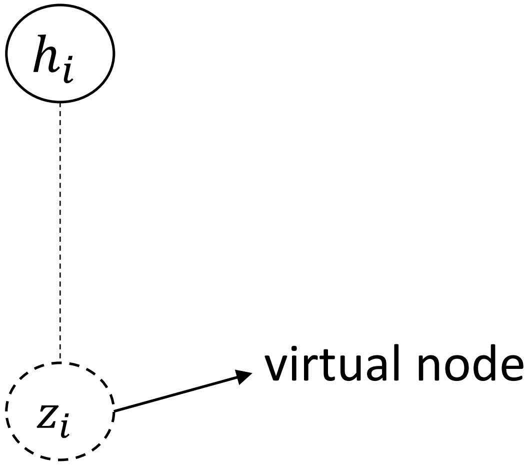
Modeling the edges

$$\mathbf{z}_i = \sum_j \alpha_{ij}\mathbf{h}_j$$

$$\alpha_{ij} = \mathrm{softmax}(f_{\mathrm{attn}}(\mathbf{e}_{ij}))$$

$$\tilde{\mathbf{h}}_i = f_{\mathrm{node}}(\mathbf{z}_i)$$

$$\boxed{\tilde{\mathbf{h}}_i = f_{\mathrm{node}}([\mathbf{h}_i; \mathbf{z}_i])}$$

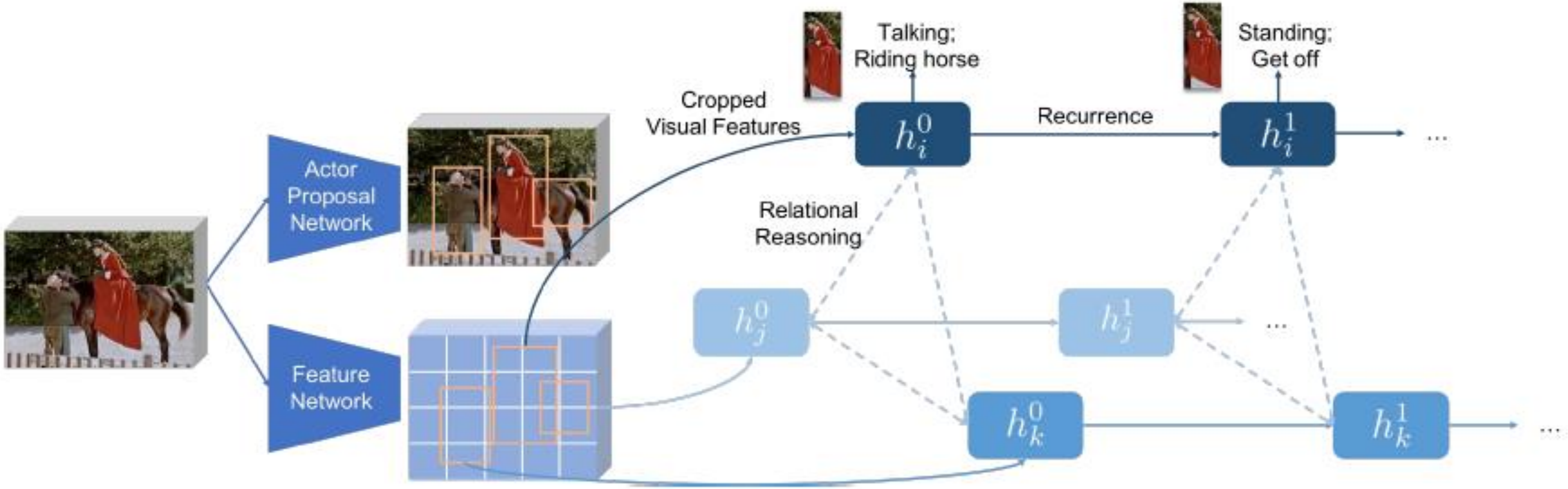


difficulty distinguishing node features from neighbor features

$$\tilde{h}_{1:N} = f_{\mathrm{GNN}}(h_{1:N}) = \prod_{i=1}^{N} \delta(\tilde{h}_i | f_{\mathrm{node}}([h_i, z_i]))\delta(z_i | \sum_j \alpha_{ij} h_j) \prod_{j=1}^{N} \delta(\alpha_{ij} | \mathcal{S}_j(f_{\mathrm{attn}}(e_{i,1:N}))\delta(e_{ij} | f_{\mathrm{edge}}(h_i, h_j))$$

Proposed Approach

Overall DR^{2}N model



Proposed Approach

Training

$$\mathcal{L}^{\text{total}} = \alpha \mathcal{L}^{\text{loc}} + \sum_{t=0}^{T} \beta_t \mathcal{L}_t^{\text{cls}}$$

$\mathcal{L}^{\text{loc}} \rightarrow$ bounding box localization loss

$\mathcal{L}_t^{\text{cls}} \rightarrow$ action classification loss at time t

$\alpha, \beta_t \rightarrow$ scalar weights

$\alpha = 1$

$\beta_0 = 1$, linearly decrease such that $\beta_t = 0.5$

Proposed Approach

Implementation details

- RPN: ResNet-50 initialized with pre-trained ImageNet weights
- Feature network: S3D-G weights pre-trained from Kinetics-400

- GRU: RNN architecture to model action dynamics

- Synchronous SGD with batch size 4 per GPU
- 10 GPUs in total

- Warm-start, cosine learning rate decay
- Gradient multiplier

Experimental setup

AVA (Atomic Visual Actions)

- Large-scale action detection dataset (430 15-min clips)
- 80 atomic visual actions
- Multiple actors with multiple action labels
- If IoU > 0.5 and action label is correct → true positive

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

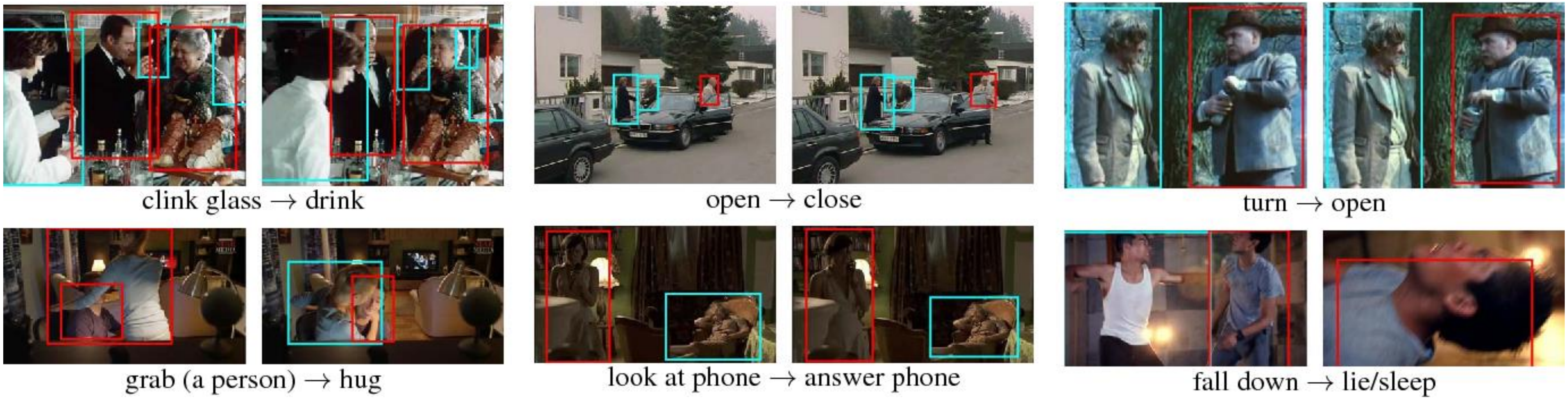
Left: Sit, Ride, Talk to; Right: Sit, Drive, Listen to

Left: Stand, Watch; Middle: Stand, Play instrument; Right: Sit, Play instrument

Orange: 1 pose action
Red: 0-3 interactions with objects
Blue: 0-3 interactions with other people

C. Gu, C. Sun, S. Vijayanarasimhan, C. Pantofaru, D. A. Ross, G. Toderici, Y. Li, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In CVPR, 2018.

Atomic Actions



The text shows pairs of atomic actions for the people in red bounding boxes.

C. Gu, C. Sun, S. Vijayanarasimhan, C. Pantofaru, D. A. Ross, G. Toderici, Y. Li, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In CVPR, 2018.

J-HMDB (Joint-annotated Human Motion Database)

- 21 categories (1.4 seconds/clip)
- One actor with a single action label
- Test for an early action classification problem

JHMDB Dataset : http://jhmdb.is.tue.mpg.de/dataset

Methods for ablation study

Single-head

- Train a separate model for each future time step

Multi-head

- Train one model for all future t

GRU

- Predicted from hidden states of GRU
- Without edges (relations)

Graph-GRU

- With edges

RN: equal weights

GAT: weighted differently but is not sure if it should focus on features from itself or features from neighbors

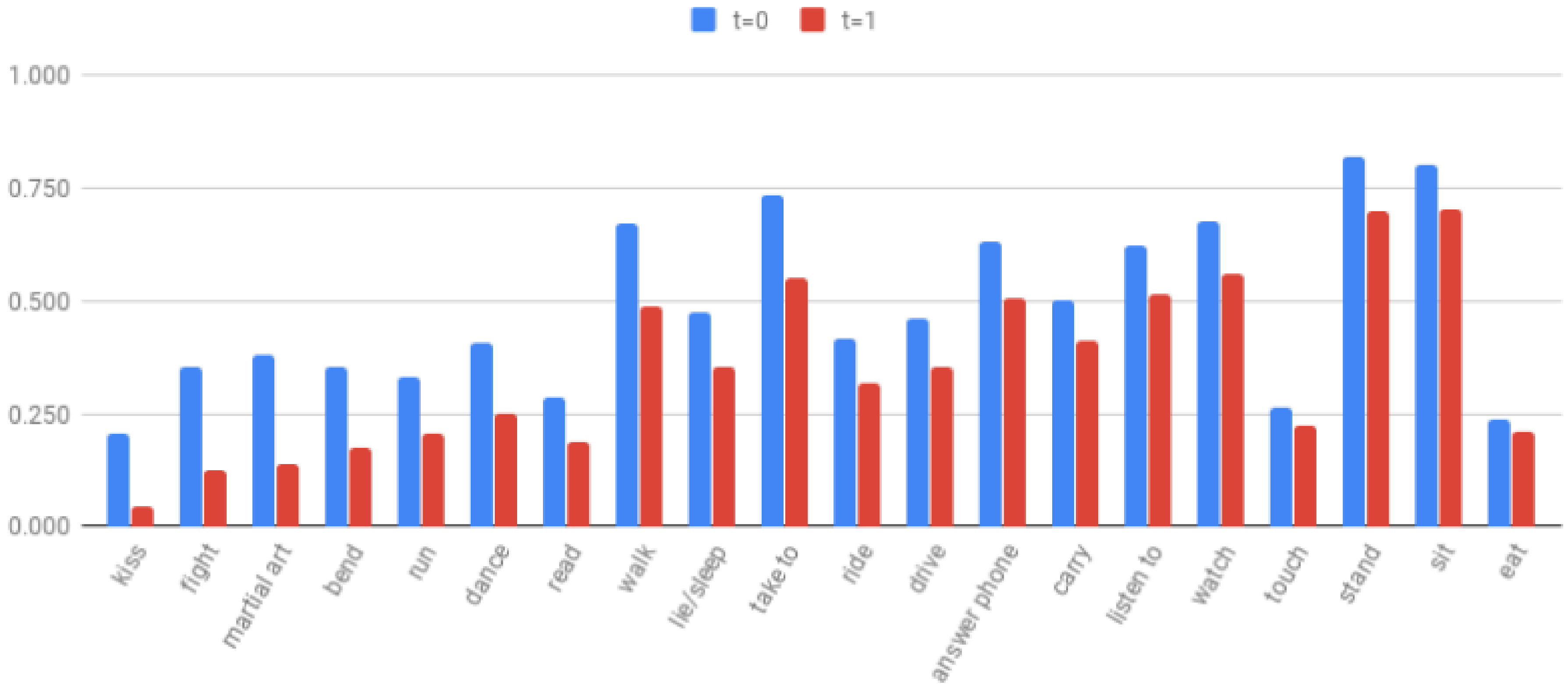
$\mathbf{DR^2N}$**:** proposed method

Action forecasting results on AVA dataset

Method	Dynamics Model	Relation Model	$t=0$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Single-head	-	-	19.1	7.8	5.3	4.2	2.6	1.8
Multi-head	-	-	16.0	9.4	6.8	5.4	4.3	3.6
GRU	GRU	-	18.7	13.1	10.3	8.0	6.7	5.7
Graph-GRU	GRU	RN [51]	17.3	12.3	9.9	7.7	6.5	5.3
Graph-GRU	GRU	GAT [61]	16.4	12.3	9.3	7.3	6.2	5.2
Graph-GRU	GRU	DR^2N (Us)	**20.4**	**14.4**	**11.2**	**9.3**	**7.5**	**6.8**

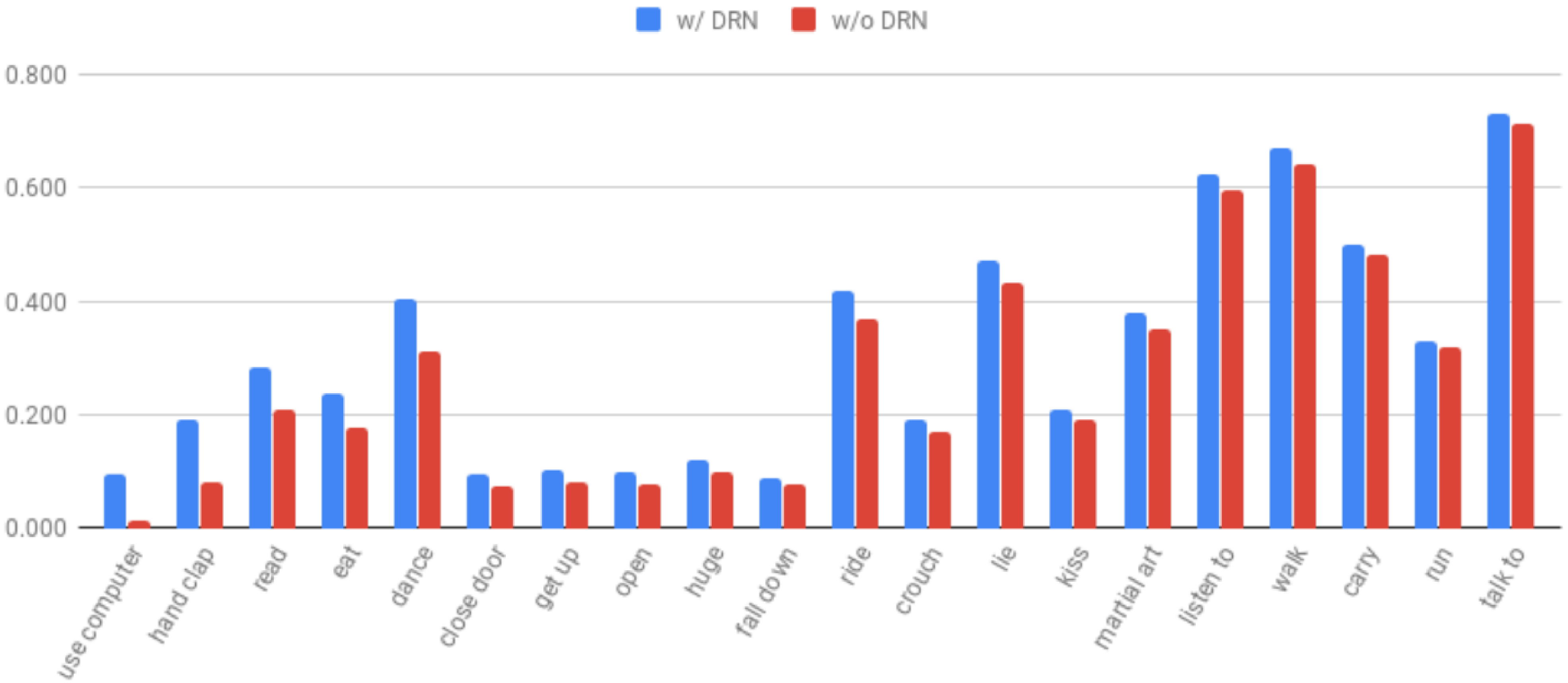
4. DR^2N outperforms the other two methods

Short duration action is hard to predict



Change in AP performance from t = 0 to t = 1

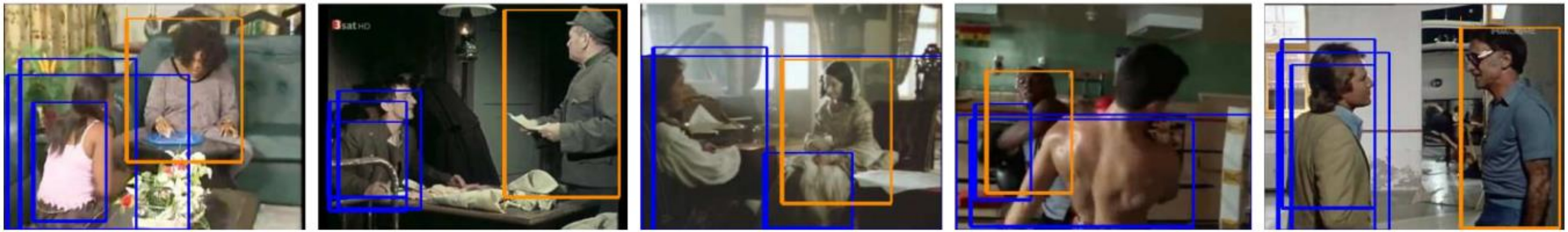
Most gains for actions with explicit interactions or where other actors provide useful context



Change in AP performance from adding graph connections at t = 0

Visualization of discriminative relations

- Orange boxes: query actor (whose actions are to be predicted)
- Blue boxes: top 3 actor proposals with highest attention weights



Results Visualization

Prediction: Eat
Ground truth: Eat

Prediction: Get Up
Ground truth: Get Up

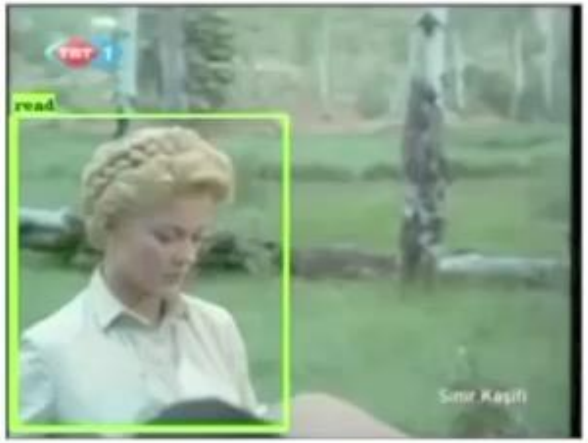
Results Visualization

Prediction: Put down
Ground truth: Put down

Prediction: Smoke
Ground truth: Smoke

Results Visualization

Prediction: Read
Ground truth: Read



Prediction: Open (door)
Ground truth: Open (door)

Results Visualization (Failure modes)

Prediction: Get up
Ground truth: Keep kneeling
Reason: Wrong duration

Prediction: Close
Ground truth: Open
Reason: Multiple futures

Early action classification on J-HMDB

- Take the first K% of frames and predict the label of the clip

Model	10%	20%	30%	40%	50%
Soomro *et al.* [55]	≈ 5	≈ 12	≈ 21	≈ 25	≈ 30
Singh *et al.* [54]	≈ 48	≈ 59	≈ 62	≈ 66	≈ 66
GRU	52.5	56.2	61.1	65.2	65.9
GAT [61]	58.1	61.8	64.4	68.7	68.8
DR^2N	**60.6**	**65.8**	**68.1**	**71.4**	**71.8**

Strengths & Weaknesses

Strengths

- Proposed method outperforms existing methods on a complex dataset like AVA
- DR^2N also improves performance on task of early action classification from previous SOTA of 48% to 60%

Weaknesses

- In the relations visualization, duplicate detections exist that might unnecessarily increase graph size and also affect the weights
- Struggles with actions having multiple possible futures and when duration is hard to predict

Possible extensions

- Implement NMS to suppress duplicate detections and obtain better graph representations
- Conduct study with different values of H to demonstrate effect of temporal context duration
- Model human-object relationships in addition to human-human

Conclusion

- This paper proposed a multi-person action forecasting model (DR^2N) that considers temporal and spatial interactions among actors
- AVA and J-HMDB datasets used for evaluation demonstrate effectiveness