SFU

Analyzing and Predicting Human Activities in Video

Greg Mori
Professor
School of Computing Science
Simon Fraser University

Research Director
Borealis AI Vancouver

What does activity recognition involve?

Detection: are there people?

Objects and scenes: where are they?

Action recognition: what are they doing?

Intention/social role: why are they doing this?

Group activity recognition: what is the overall situation?

These are inter-related problems:
model structures

Desiderata for Activity Recognition Models

Label structure

Hu et al., CVPR 16
Deng et al., CVPR 16
Nauata et al., CVPRW 17
Deng et al., CVPR 17

Temporal structure

Yeung et al., CVPR 16
Yeung et al., IJCV 17
He et al., WACV 18
Chen et al., ICCVW 17

Group structure

Ibrahim et al., CVPR 16
Mehrasa et al., SLOAN 18
Khodabandeh et al., arXiv 17
Lan et al. CVPR 12
Zhong et al., 2018

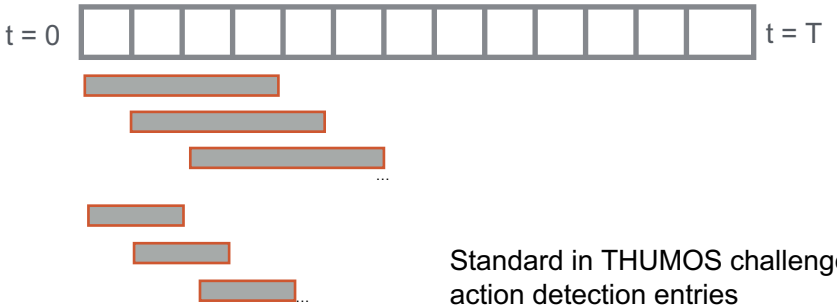
Task: action detection

Input

Output

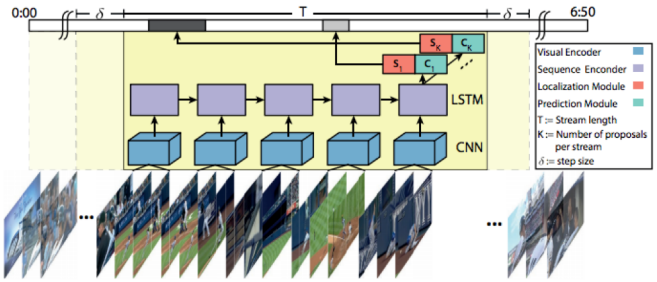
Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Dominant paradigm: Dense processing



Standard in THUMOS challenge action detection entries
Oneata et al. 2014
Wang et al. 2014
Oneata et al. 2014
Yuan et al. 2015

Sliding windows

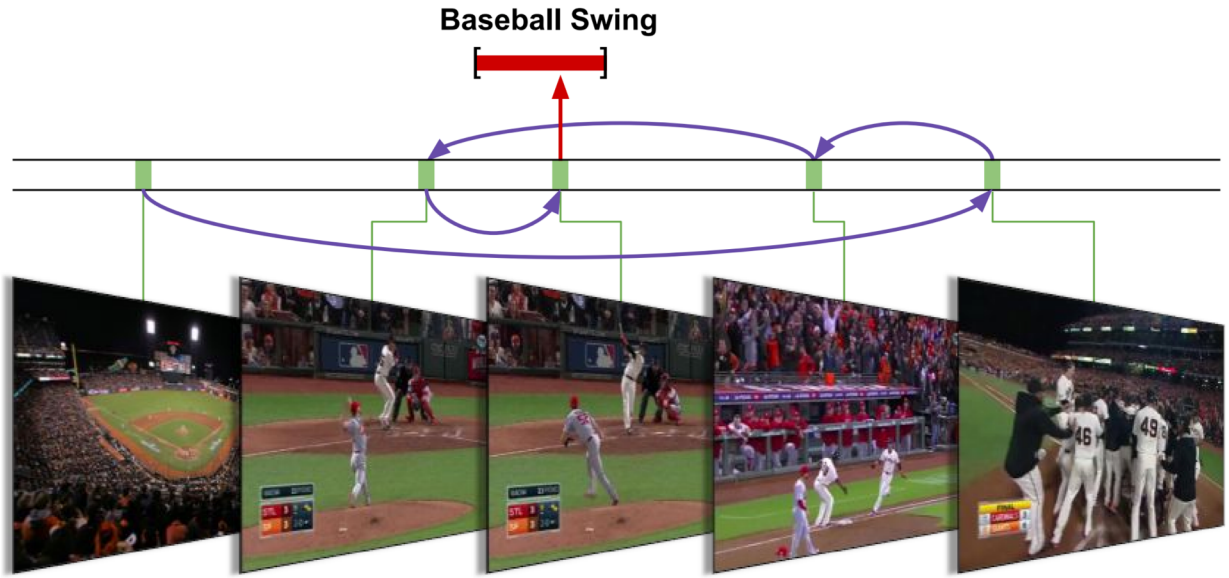


Gkioxari and Malik 2015
Yu et al. 2015
Escorcia et al. 2016
Peng and Schmid 2016
He et al. 2018

Action proposals

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Efficiently detecting actions



Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection

Detected actions

Video

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection

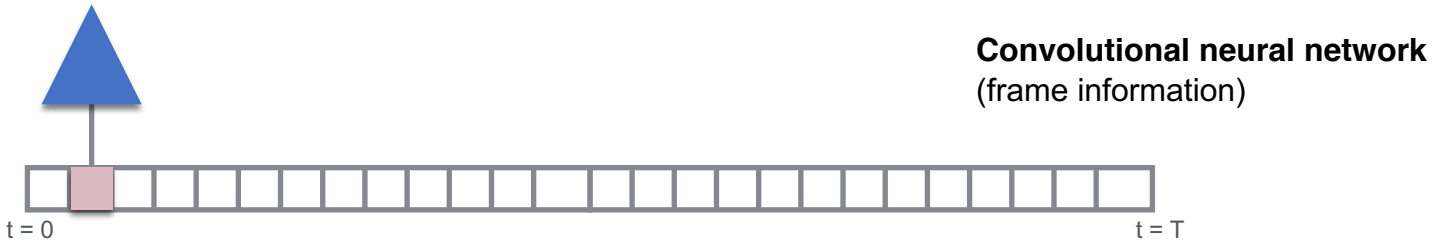
Detected actions

Video

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection

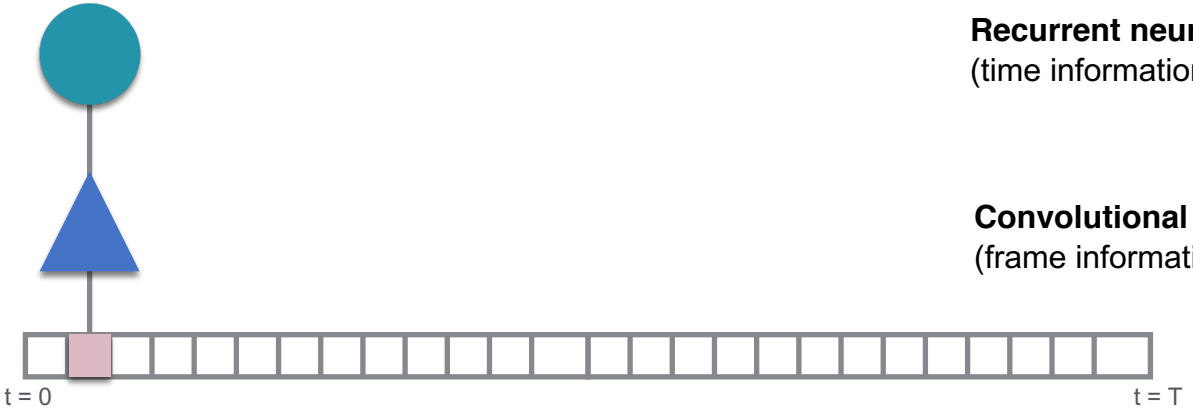
Detected actions



Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection

Detected actions



Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection

Detected actions

Outputs:
Detection **instance hypothesis** [start, end]

Recurrent neural network
(time information)

Convolutional neural network
(frame information)

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection

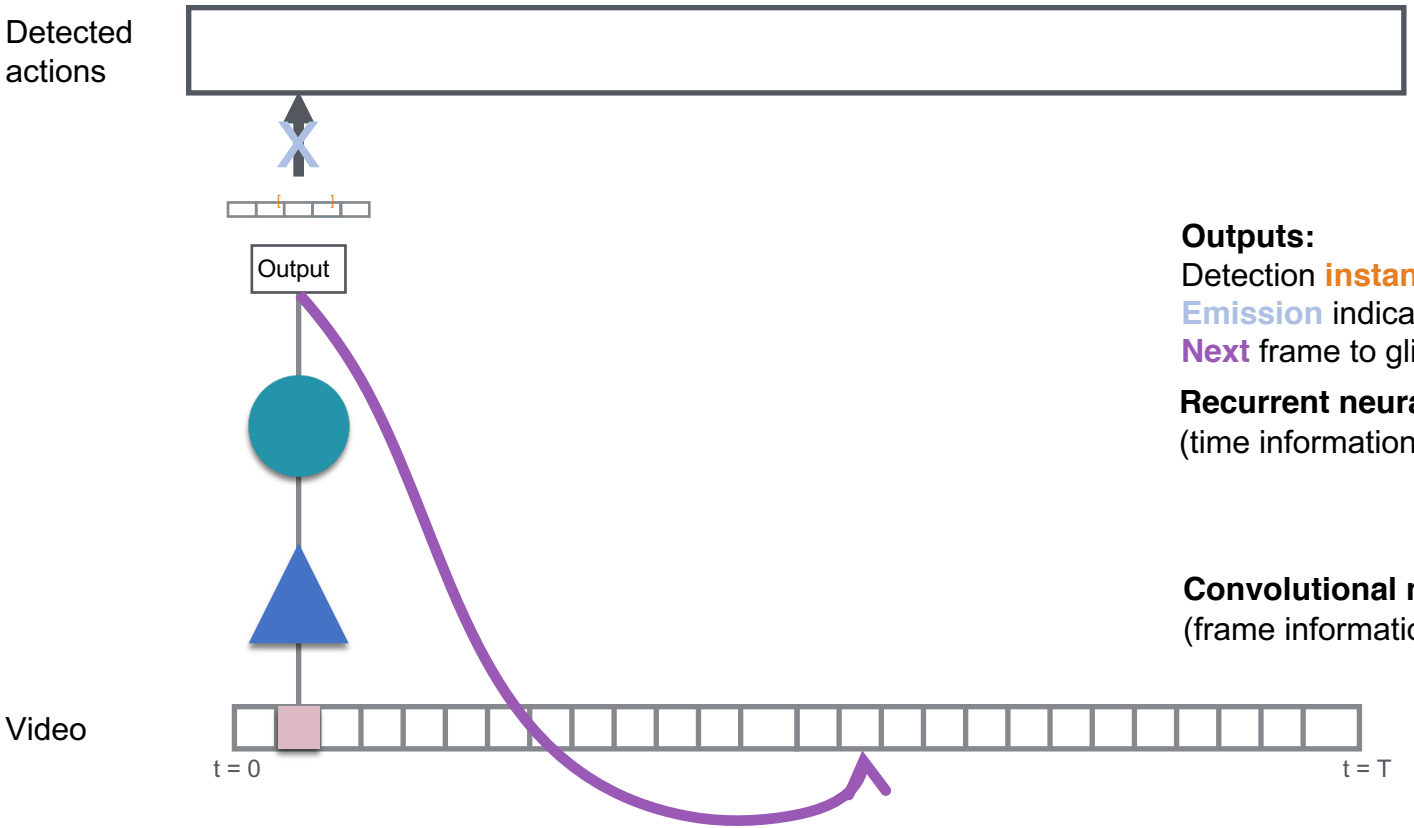
Outputs:
Detection **instance hypothesis** [start, end]
Emission indicator

Recurrent neural network
(time information)

Convolutional neural network
(frame information)

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection



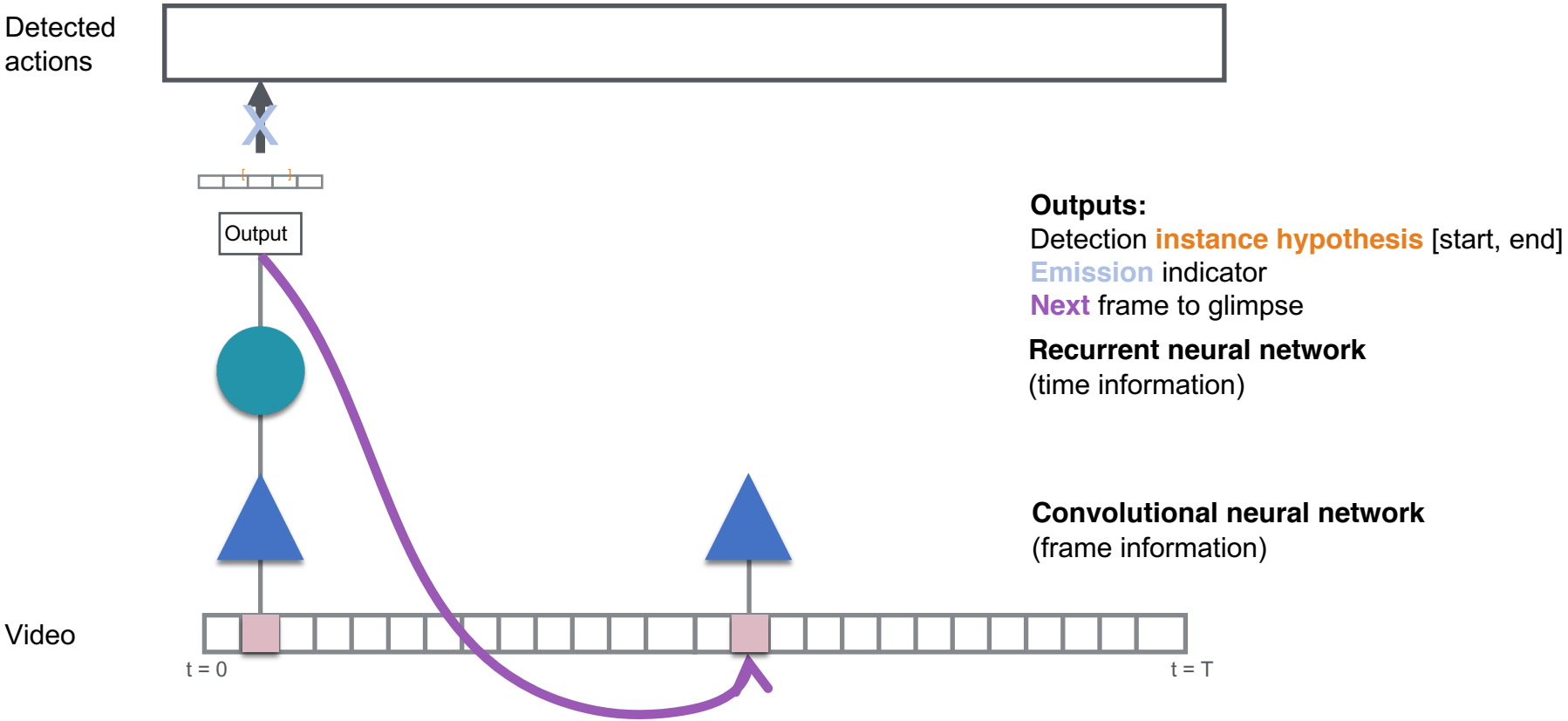
Outputs:
Detection **instance hypothesis** [start, end]
Emission indicator
Next frame to glimpse

Recurrent neural network
(time information)

Convolutional neural network
(frame information)

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection



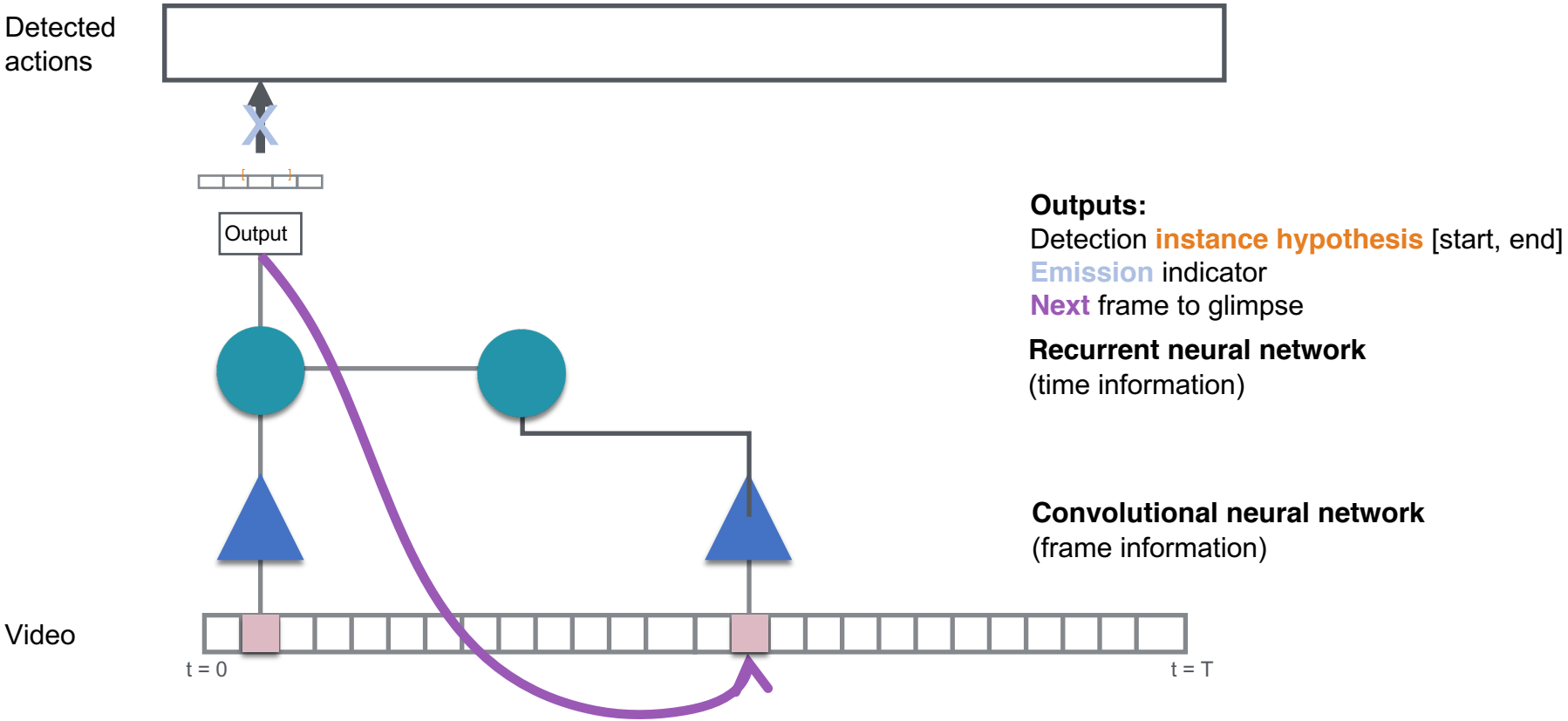
Outputs:
Detection **instance hypothesis** [start, end]
Emission indicator
Next frame to glimpse

Recurrent neural network
(time information)

Convolutional neural network
(frame information)

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection



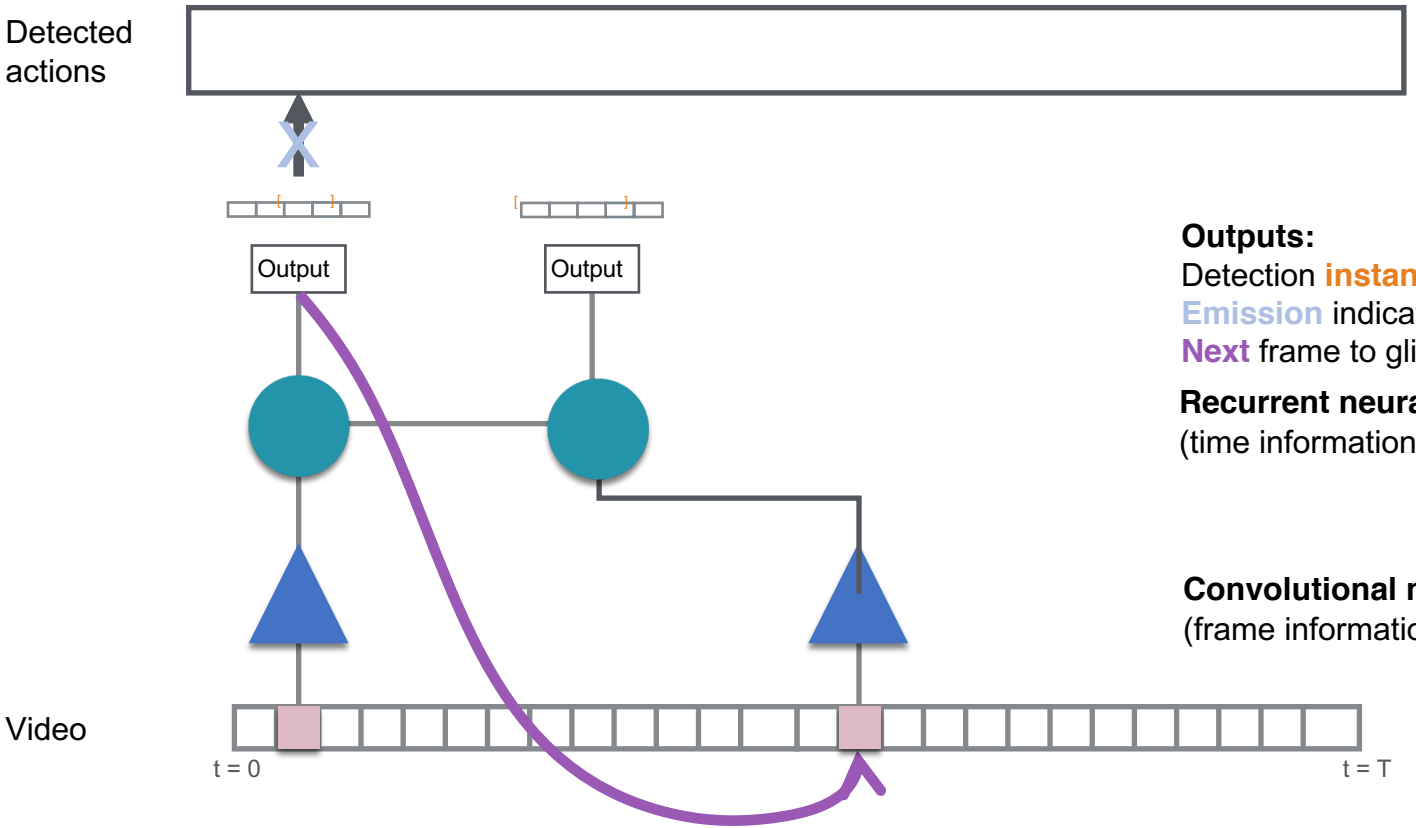
Outputs:
Detection **instance hypothesis** [start, end]
Emission indicator
Next frame to glimpse

Recurrent neural network
(time information)

Convolutional neural network
(frame information)

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection



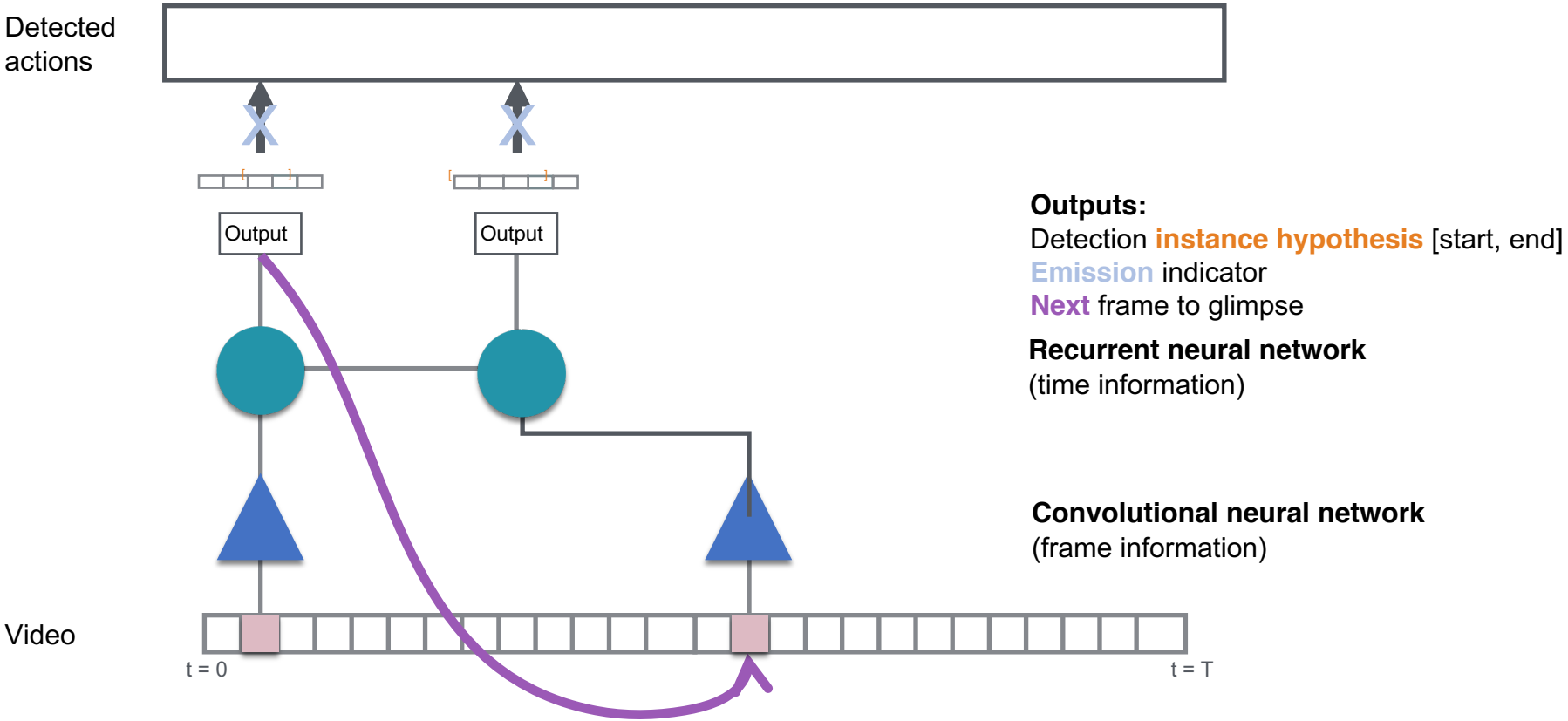
Outputs:
Detection **instance hypothesis** [start, end]
Emission indicator
Next frame to glimpse

Recurrent neural network
(time information)

Convolutional neural network
(frame information)

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection



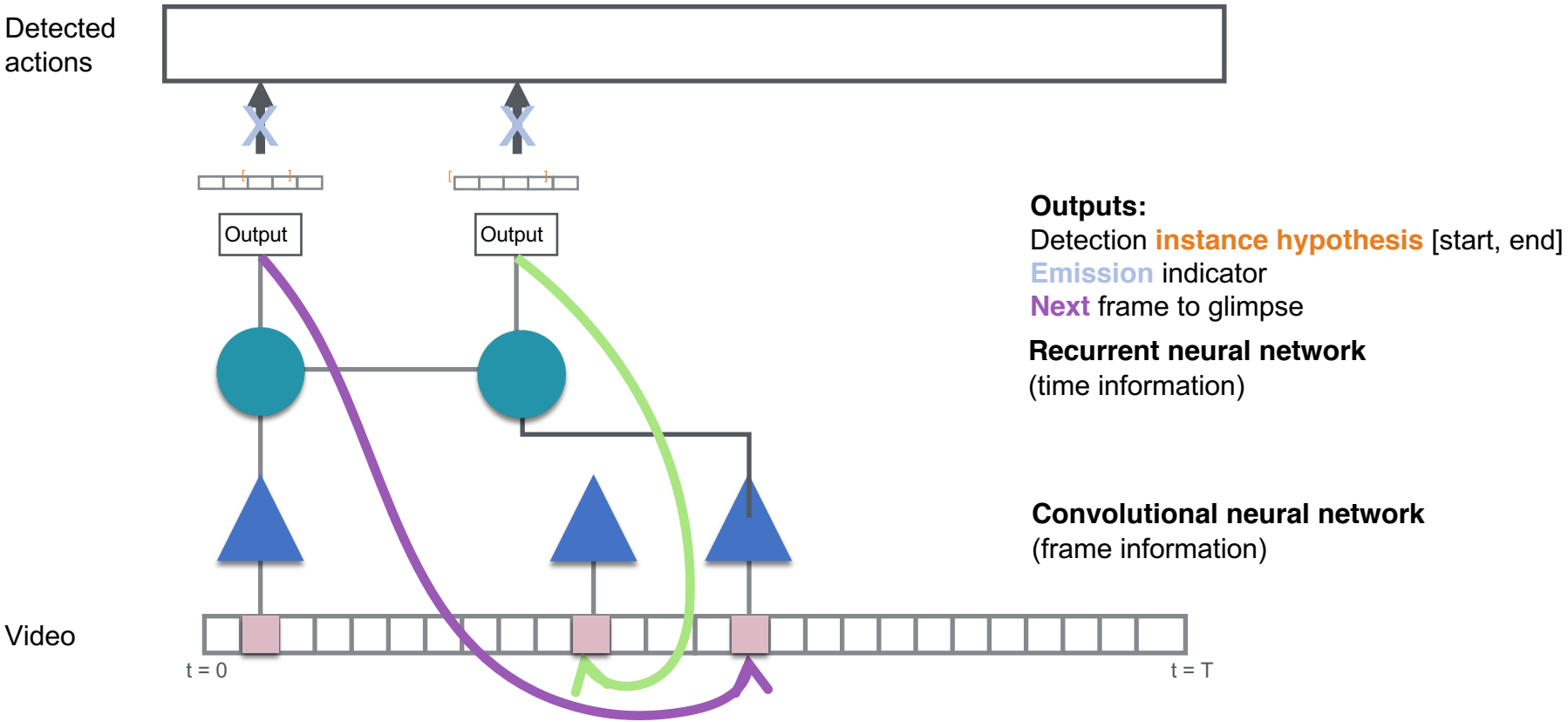
Outputs:
Detection **instance hypothesis** [start, end]
Emission indicator
Next frame to glimpse

Recurrent neural network
(time information)

Convolutional neural network
(frame information)

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection



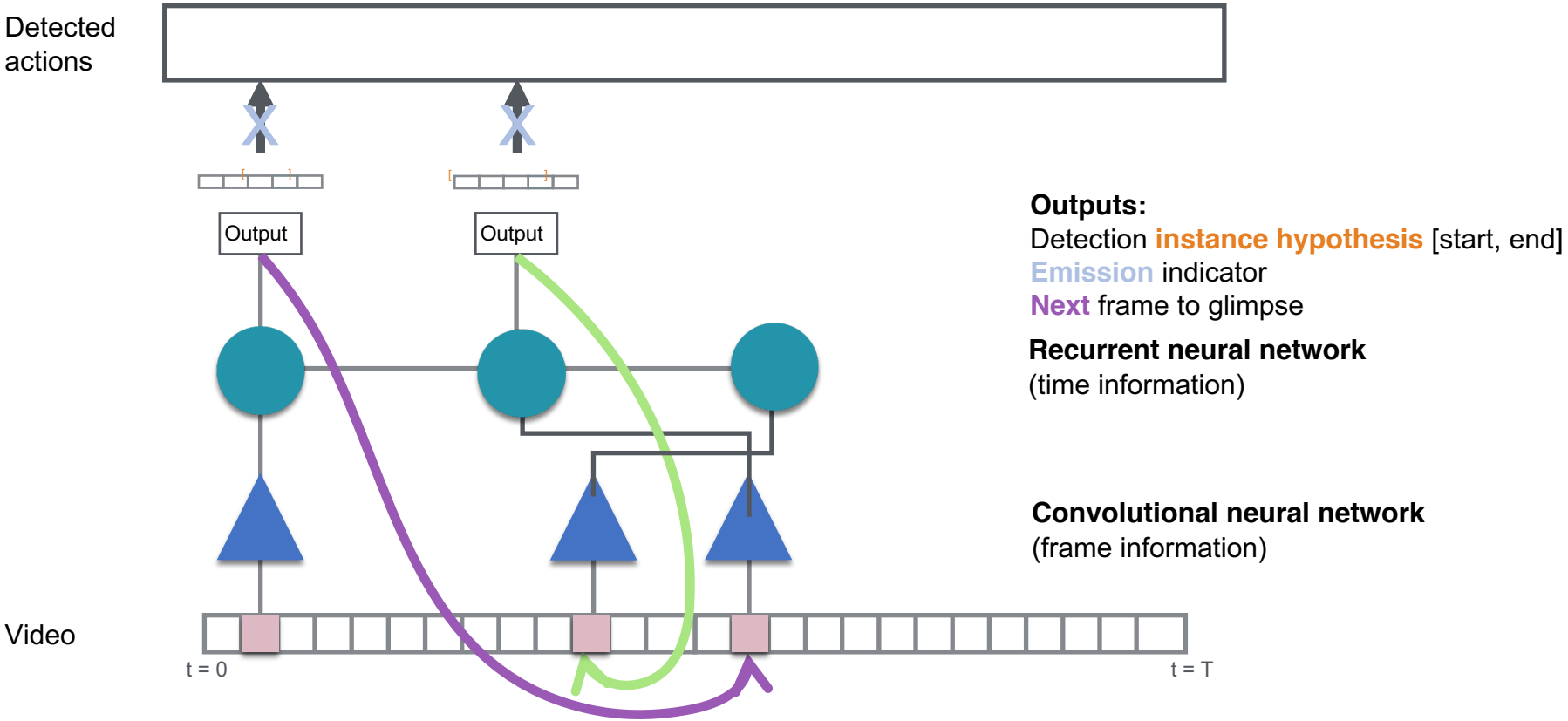
Outputs:
Detection **instance hypothesis** [start, end]
Emission indicator
Next frame to glimpse

Recurrent neural network
(time information)

Convolutional neural network
(frame information)

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection



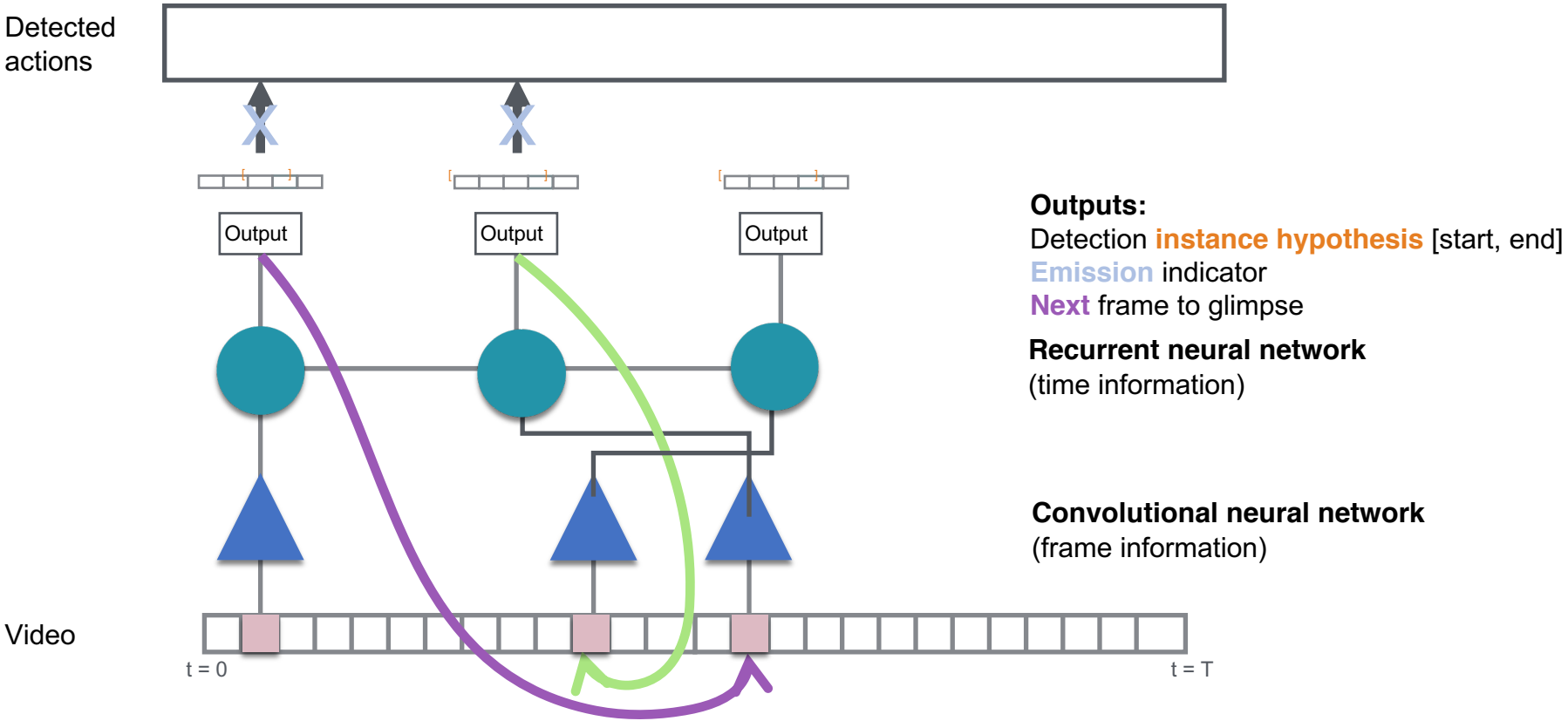
Outputs:
Detection **instance hypothesis** [start, end]
Emission indicator
Next frame to glimpse

Recurrent neural network
(time information)

Convolutional neural network
(frame information)

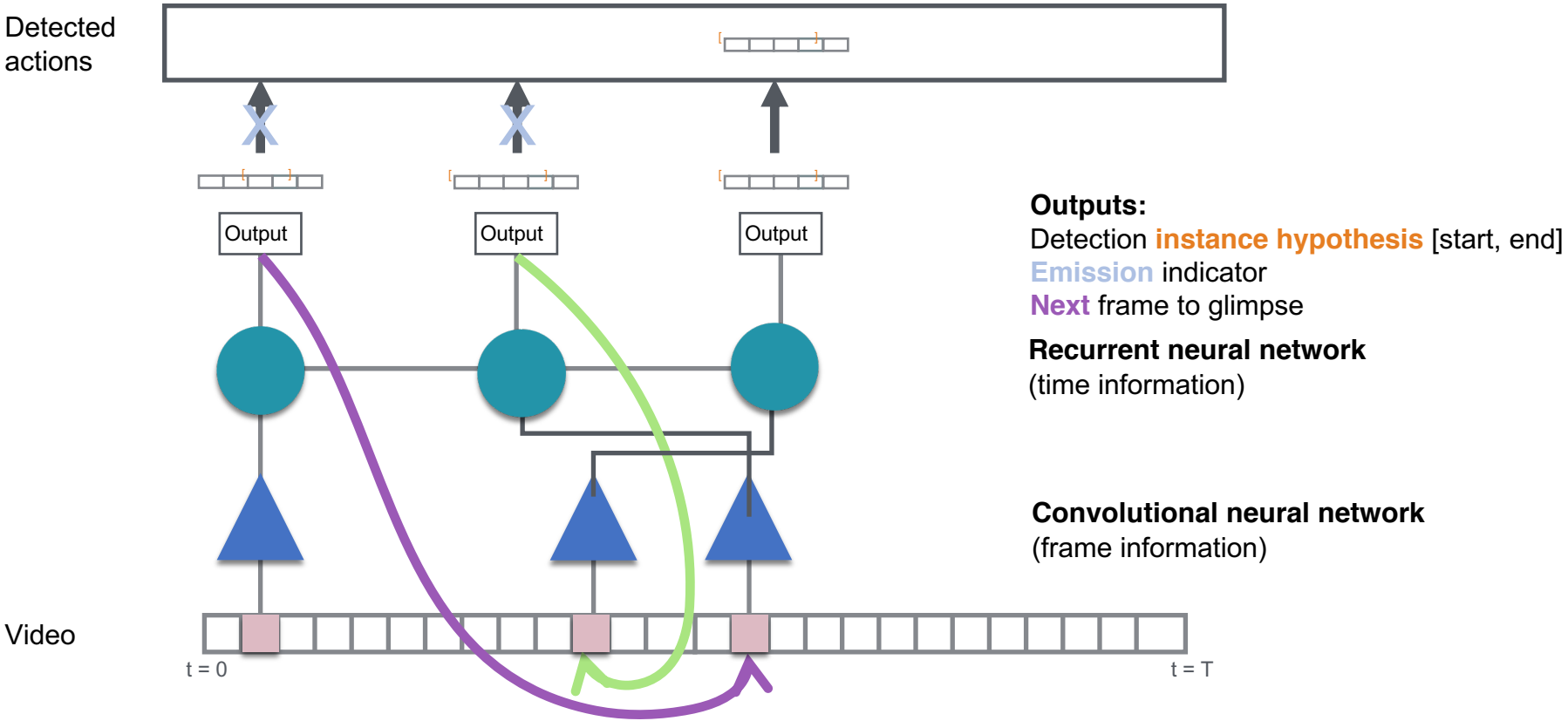
Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection



Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection



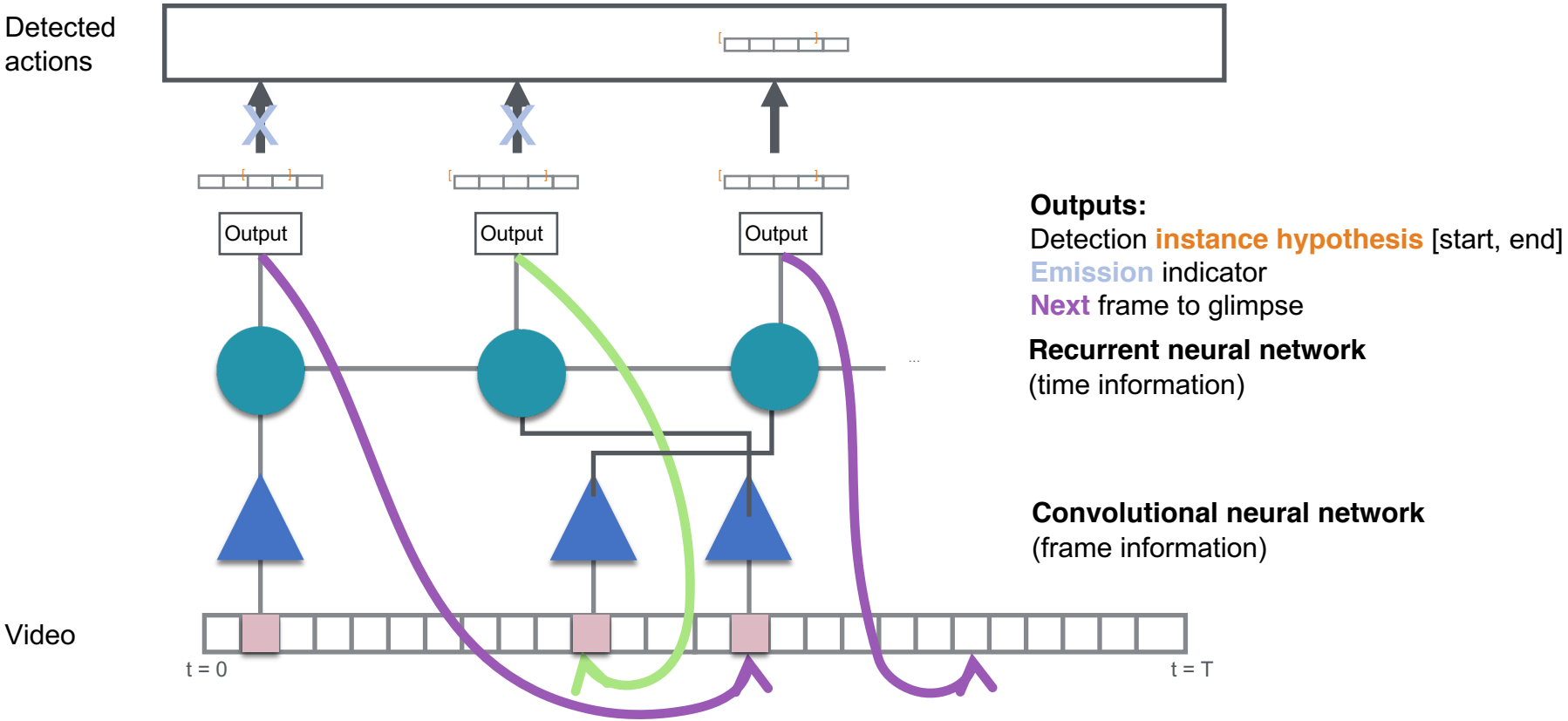
Outputs:
Detection **instance hypothesis** [start, end]
Emission indicator
Next frame to glimpse

Recurrent neural network
(time information)

Convolutional neural network
(frame information)

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Our model for efficient action detection



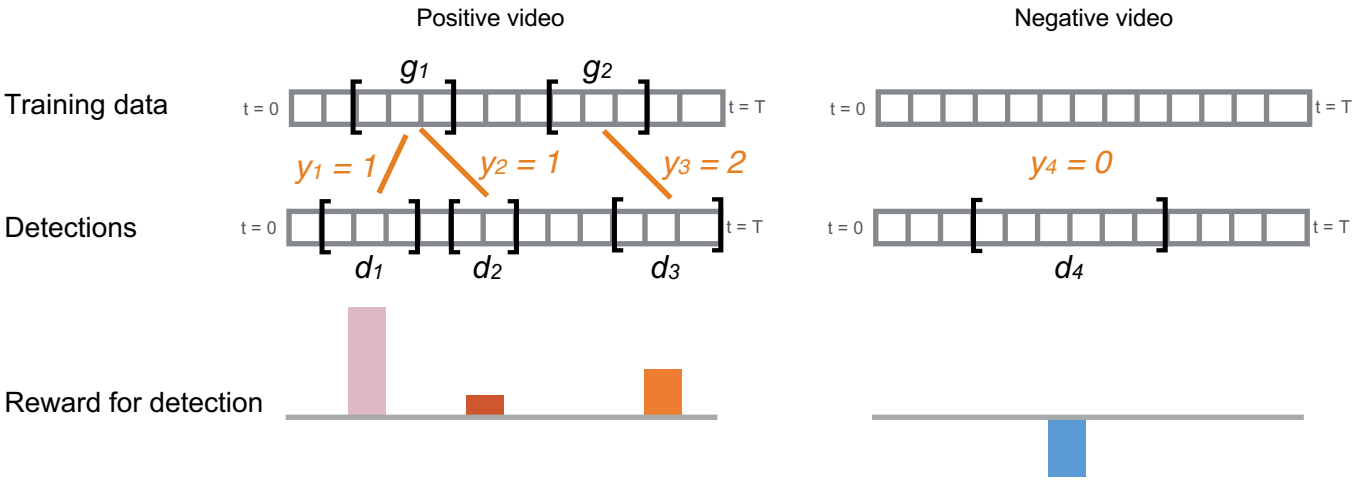
Outputs:
Detection **instance hypothesis** [start, end]
Emission indicator
Next frame to glimpse

Recurrent neural network
(time information)

Convolutional neural network
(frame information)

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Training the detection instance output



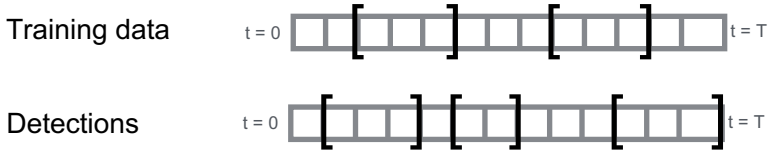
$$\mathcal{L}(D,G) = \sum_i \mathcal{L}_{cls}(d_i, y_i > 0) + \gamma \sum_{i:y_i>0} \mathcal{L}_{loc}(d_i, g_{y_i})$$

cross-entropy classification loss

L_2 distance localization loss

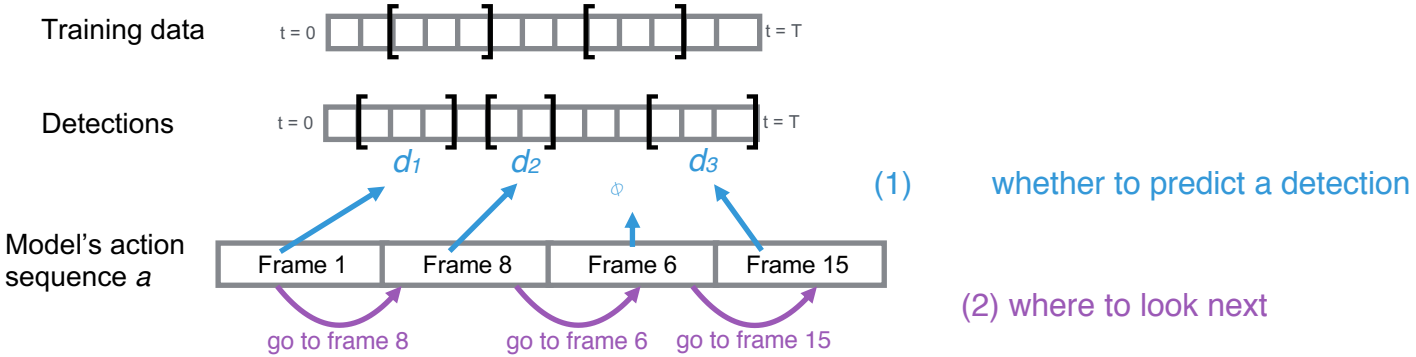
Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Training the non-differentiable outputs



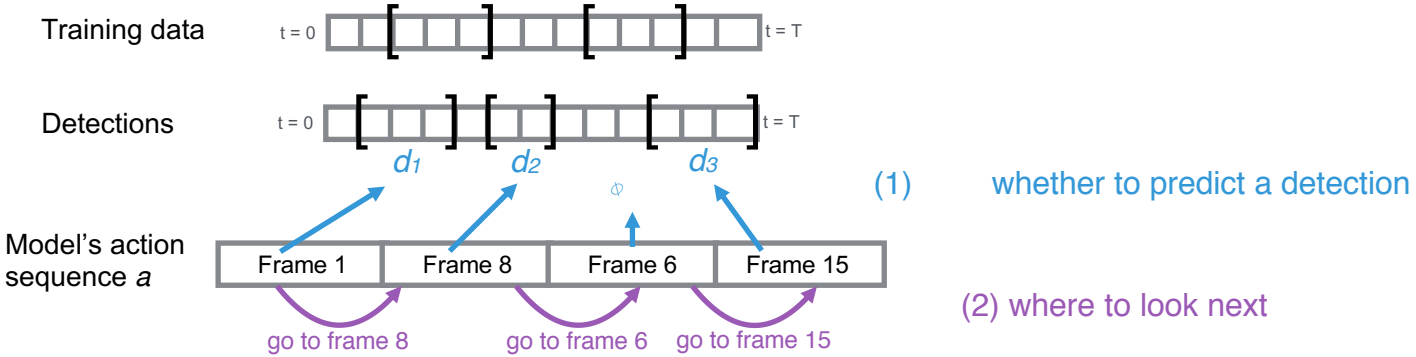
Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Training the non-differentiable outputs



Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

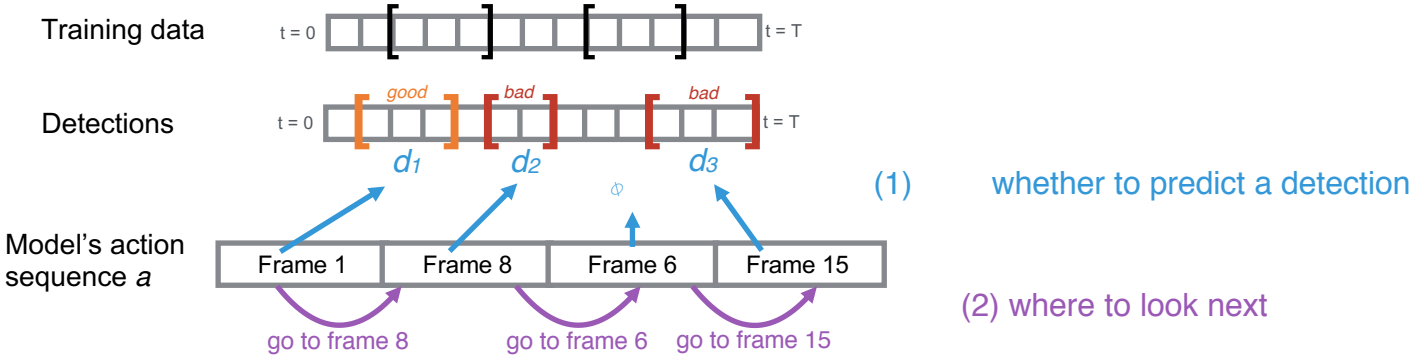
Training the non-differentiable outputs



Train an policy π_θ for actions (1) and (2) using REINFORCE [Williams 1992]

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Training the non-differentiable outputs

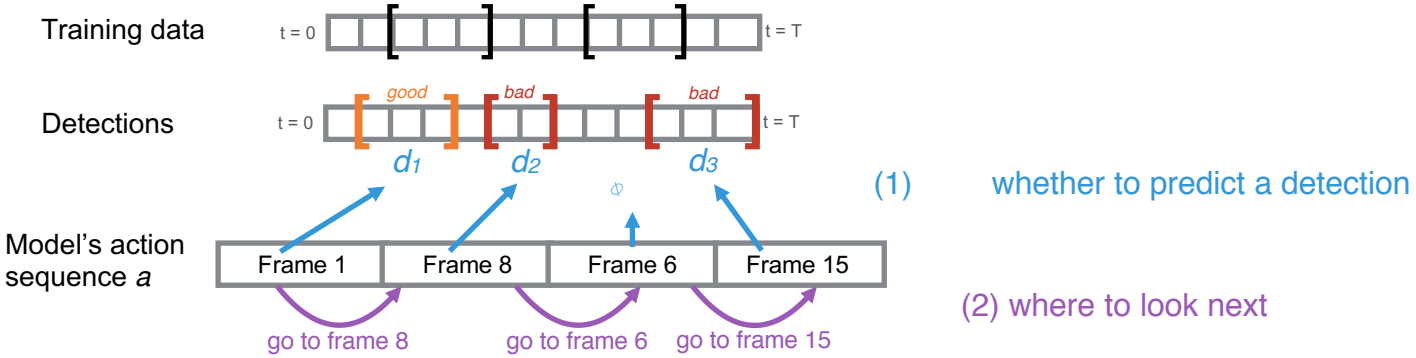


Train an policy π_θ for actions (1) and (2) using REINFORCE [Williams 1992]

Reward for an action sequence a: $r(a) = \mathbf{N}^+ - \alpha\mathbf{N}^-$

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Training the non-differentiable outputs



Train an policy π_θ for actions (1) and (2) using REINFORCE [Williams 1992]

Reward for an action sequence a: $r(a) = \mathbf{N}^+ - \alpha\mathbf{N}^-$

Objective: $J(\theta) = \sum_a p_\theta(a) r(a)$

Gradient: $\nabla J(\theta) = \sum_a p_\theta(a) r(a) \nabla \log p_\theta(a)$

Monte-Carlo approximation: $\nabla J(\theta) \approx \frac{1}{K} \sum_{k=1}^{K} r(a^k) \sum_{t=1}^{T} \nabla \log \pi_\theta(a_t^k | M_t^k)$

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

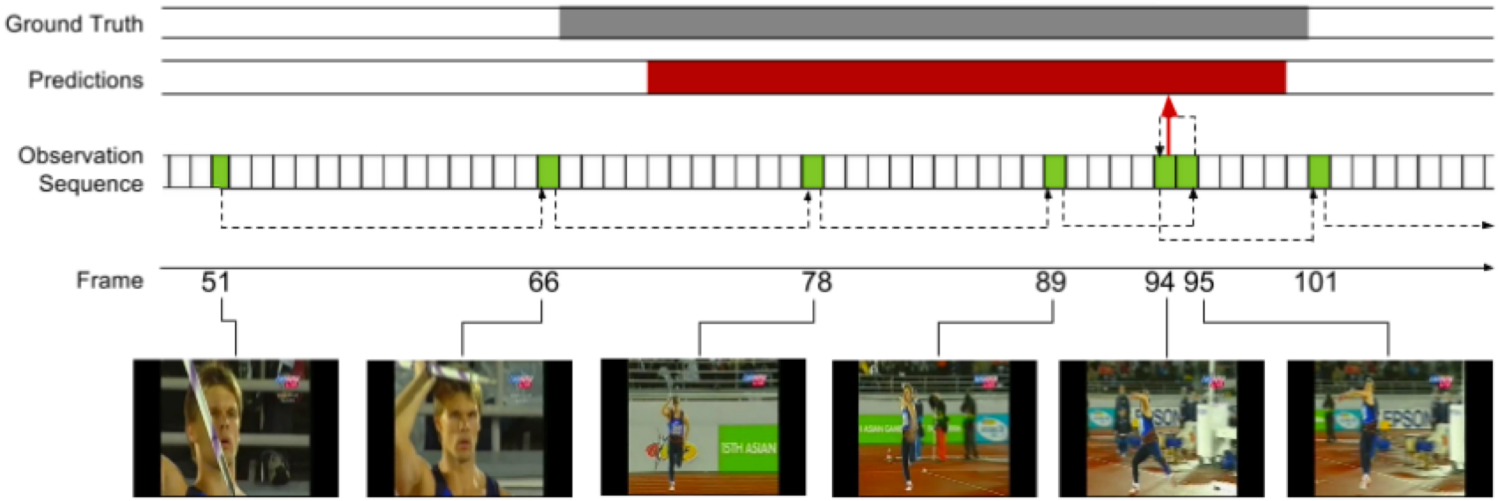
Action detection results

Dataset	Detection AP at IOU 0.5	
	State-of-the-art	Our result
THUMOS 2014	14.4	**17.1**
ActivityNet sports	33.2	**36.7**
ActivityNet work	31.1	**39.9**

While glimpsing only 2% of frames

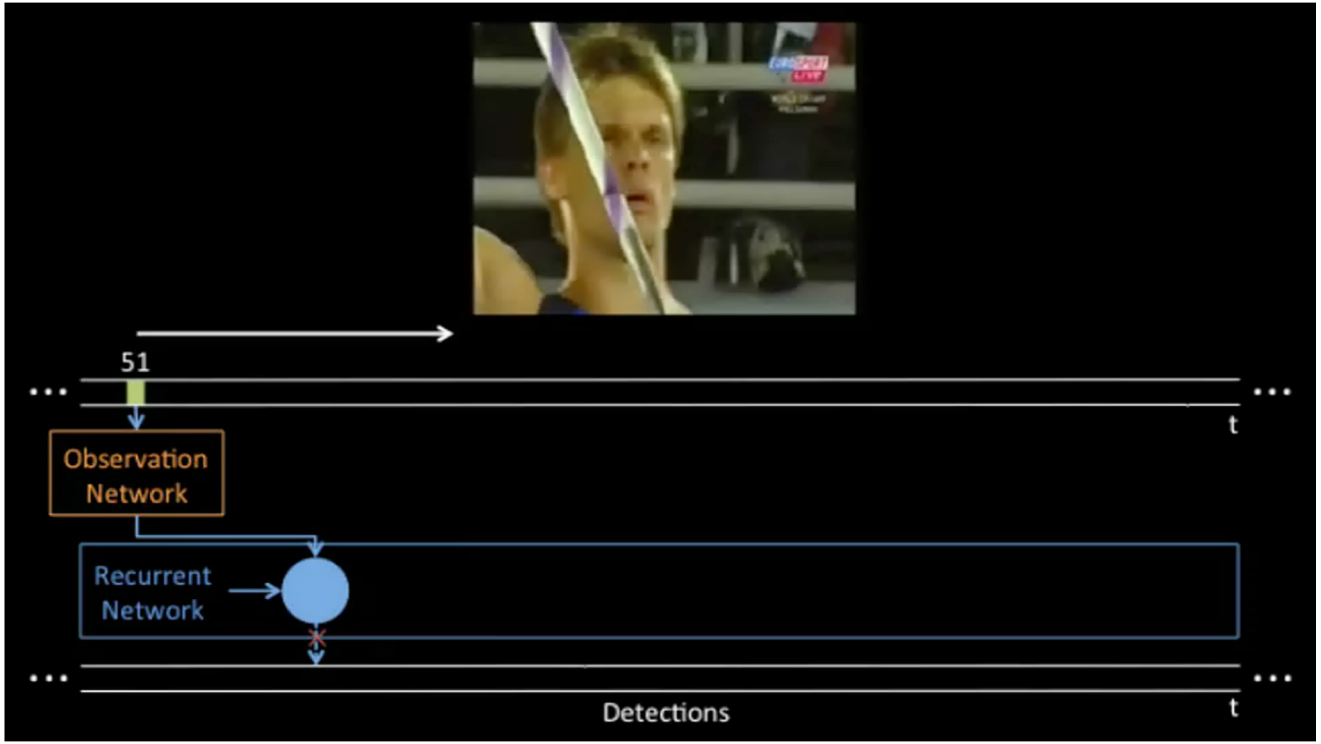
Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Learned policies



Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Learned policies



Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Importance of prediction indicator output

	mAP (IOU = 0.5)
Ours (full model)	**17.1**
Ours w/o prediction indicator output (always predict)	**12.4**

Deciding when to output a prediction (learning to do non-maximum suppression) matters.

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

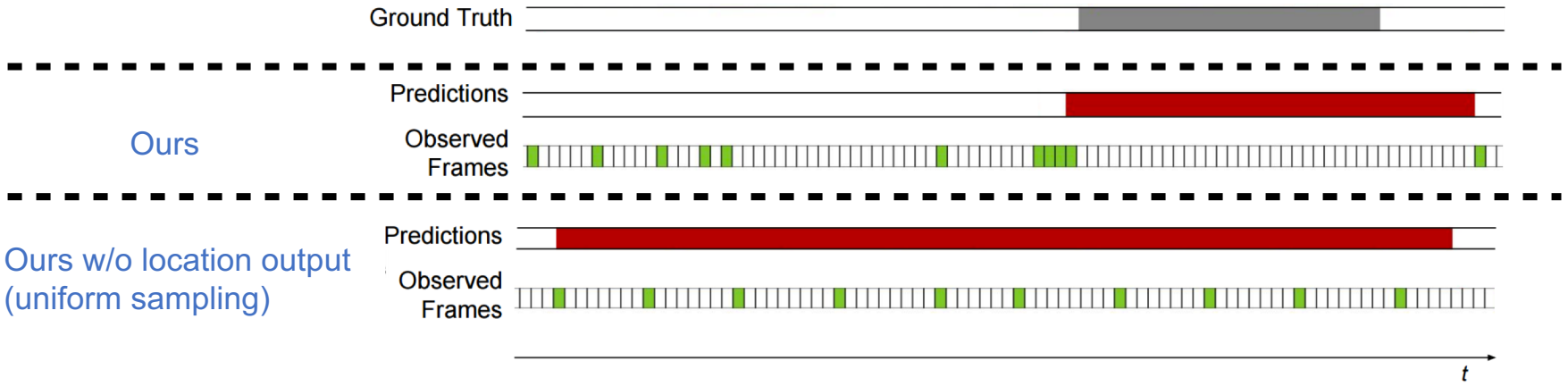
Importance of location output

	mAP (IOU = 0.5)
Ours (full model)	**17.1**
Ours w/o prediction indicator output (always predict)	12.4
Ours w/o location output (uniform sampling)	**9.3**

Deciding where to look next (location output) has even greater effect.

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Importance of location output



Uniform sampling does not always have sufficient temporal resolution where it's needed.

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Removing both prediction indicator and location outputs

	mAP (IOU = 0.5)
Ours (full model)	**17.1**
Ours w/o prediction indicator output (always predict)	12.4
Ours w/o location output (uniform sampling)	9.3
Ours w/o prediction indicator w/o location output (always predict, with uniform sampling)	**8.6**

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Importance of location regression

	mAP (IOU = 0.5)
Ours (full model)	**17.1**
Ours w/o prediction indicator output (always predict)	12.4
Ours w/o location output (uniform sampling)	9.3
Ours w/o prediction indicator w/o location output (always predict, with uniform sampling)	8.6
Ours w/o location regression (always output mean action duration)	**5.5**

Simply outputting mean action duration gives significantly worse performance.

Yeung, Russakovsky, Mori, Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. CVPR 2016.

Desiderata for Activity Recognition Models

Label structure

Hu et al., CVPR 16
Deng et al., CVPR 16
Nauata et al., CVPRW 17
Deng et al., CVPR 17

Temporal structure

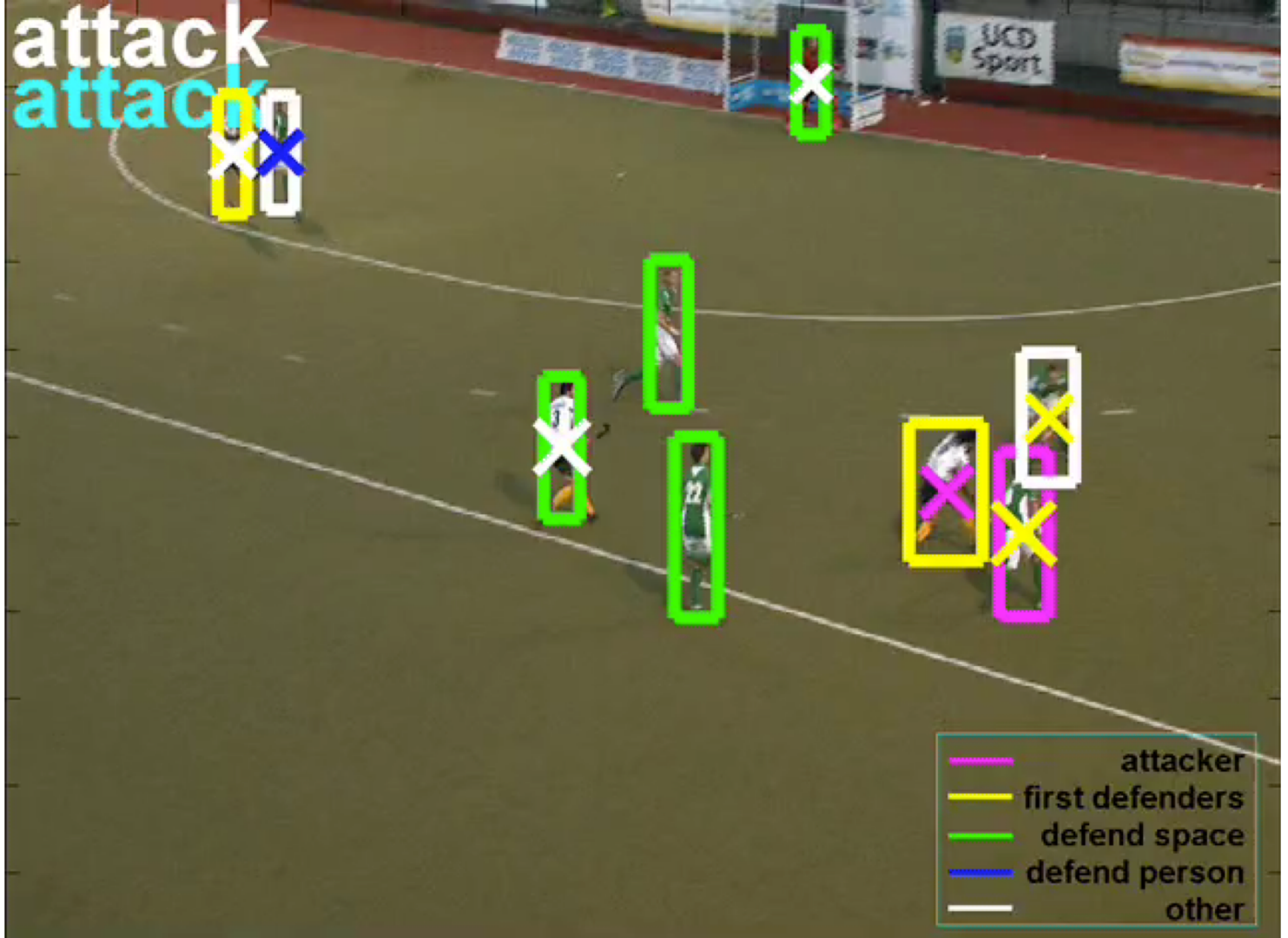
Yeung et al., CVPR 16
Yeung et al., IJCV 17
He et al., WACV 18
Chen et al., ICCVW 17

Group structure

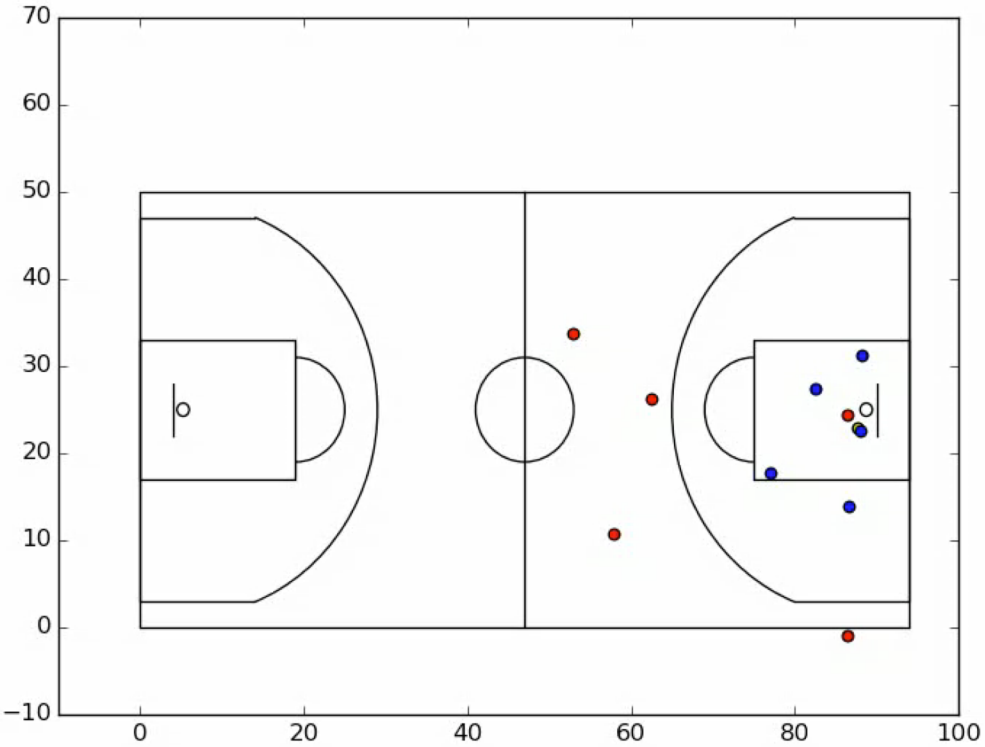
Ibrahim et al., CVPR 16
Mehrasa et al., SLOAN 18
Khodabandeh et al., arXiv 17
Lan et al. CVPR 12
Zhong et al., 2018

Role of Context in Actions

Who has the puck?


attack
attack
attacker
first defenders
defend space
defend person
other

Analyzing Human Trajectories to Recognize Actions



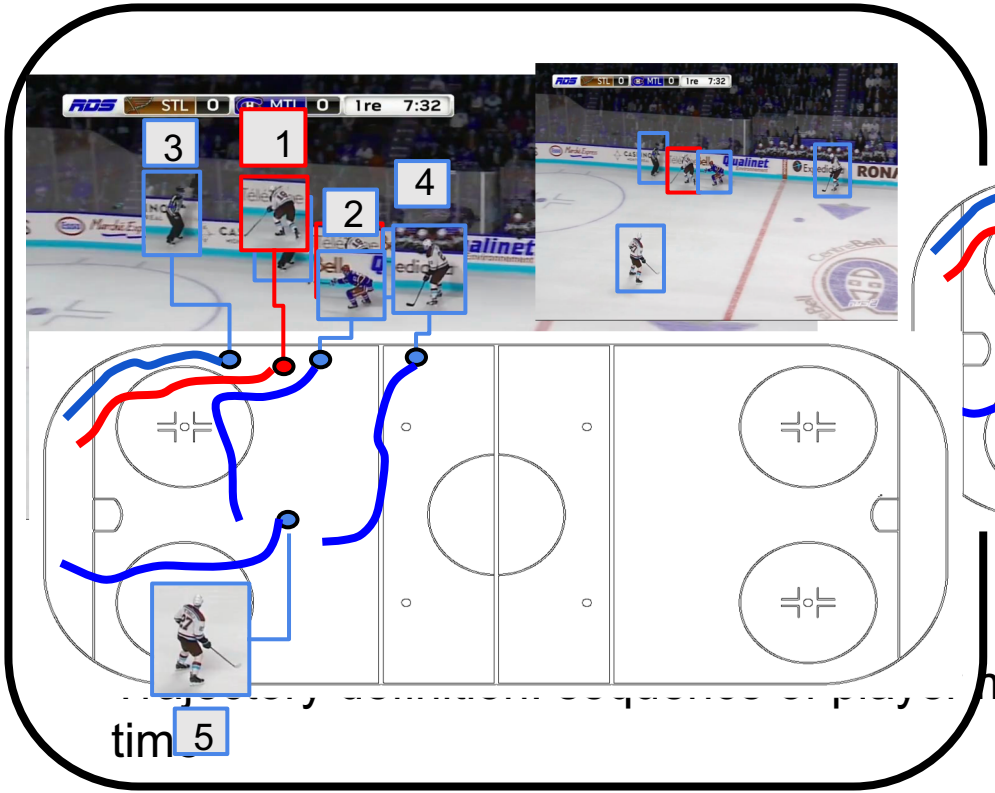
Which team is it?

Who was player X?

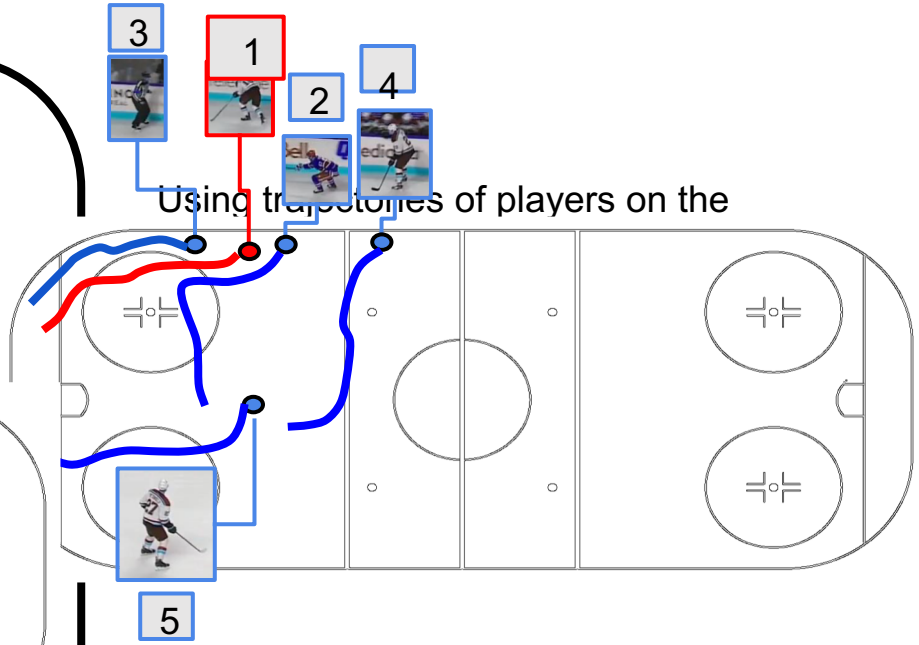
Will the shot be successful?

Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Motivation



Using trajectories of players on the

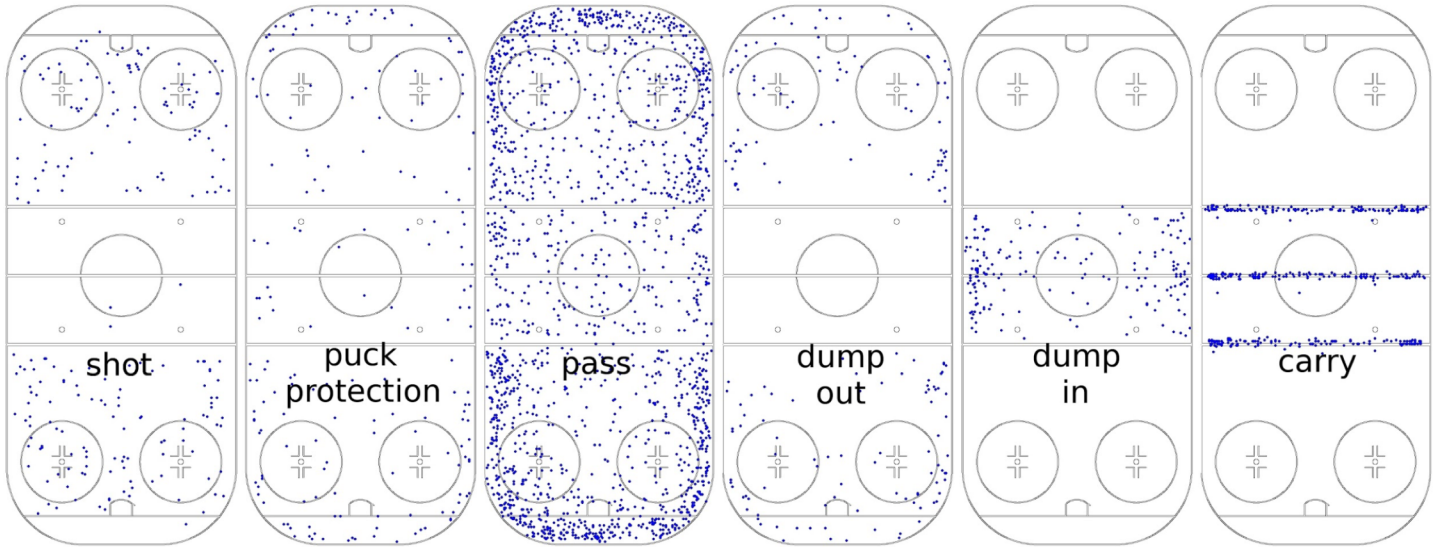


movements across space over

tim

Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

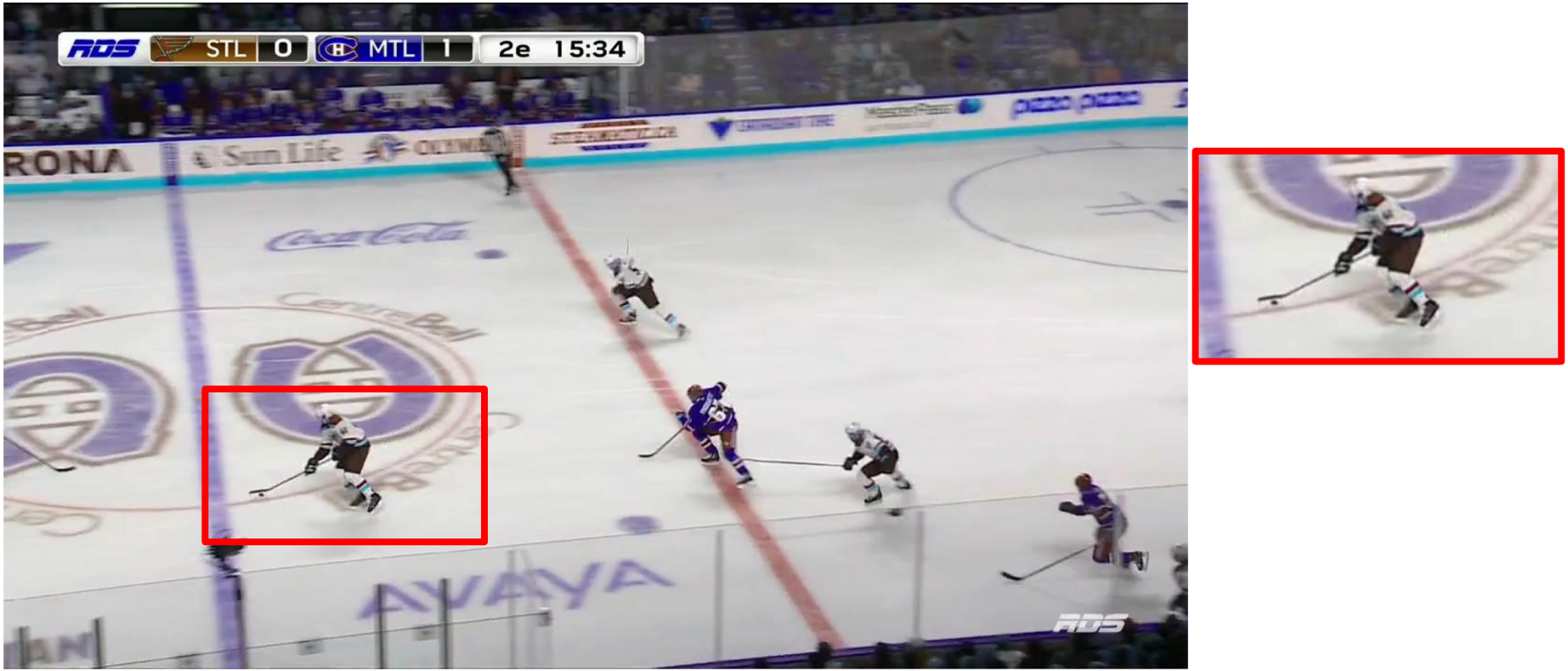
Motivation



locations matter!

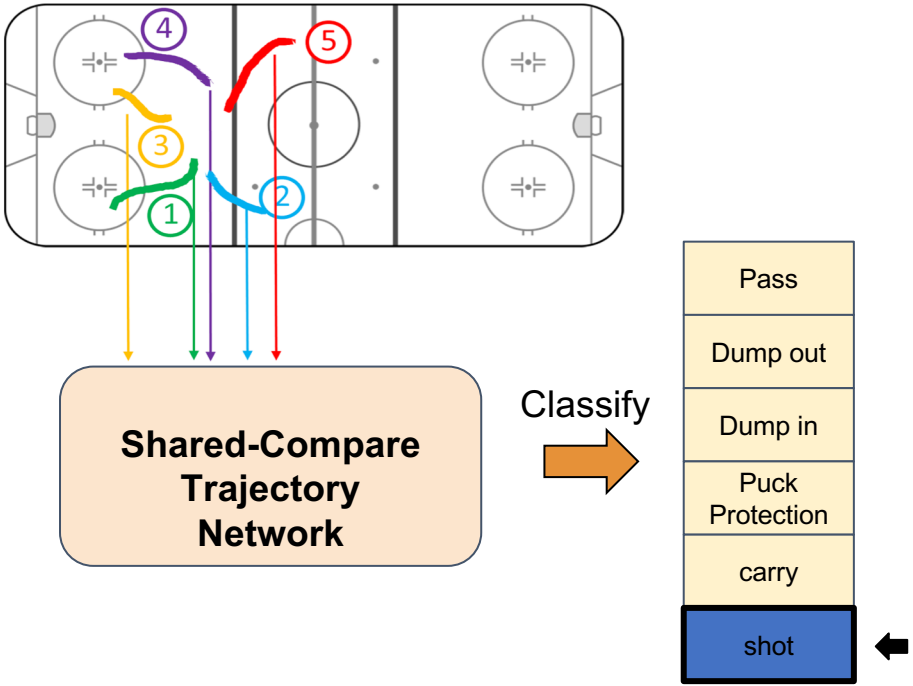
Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Key Player Definition



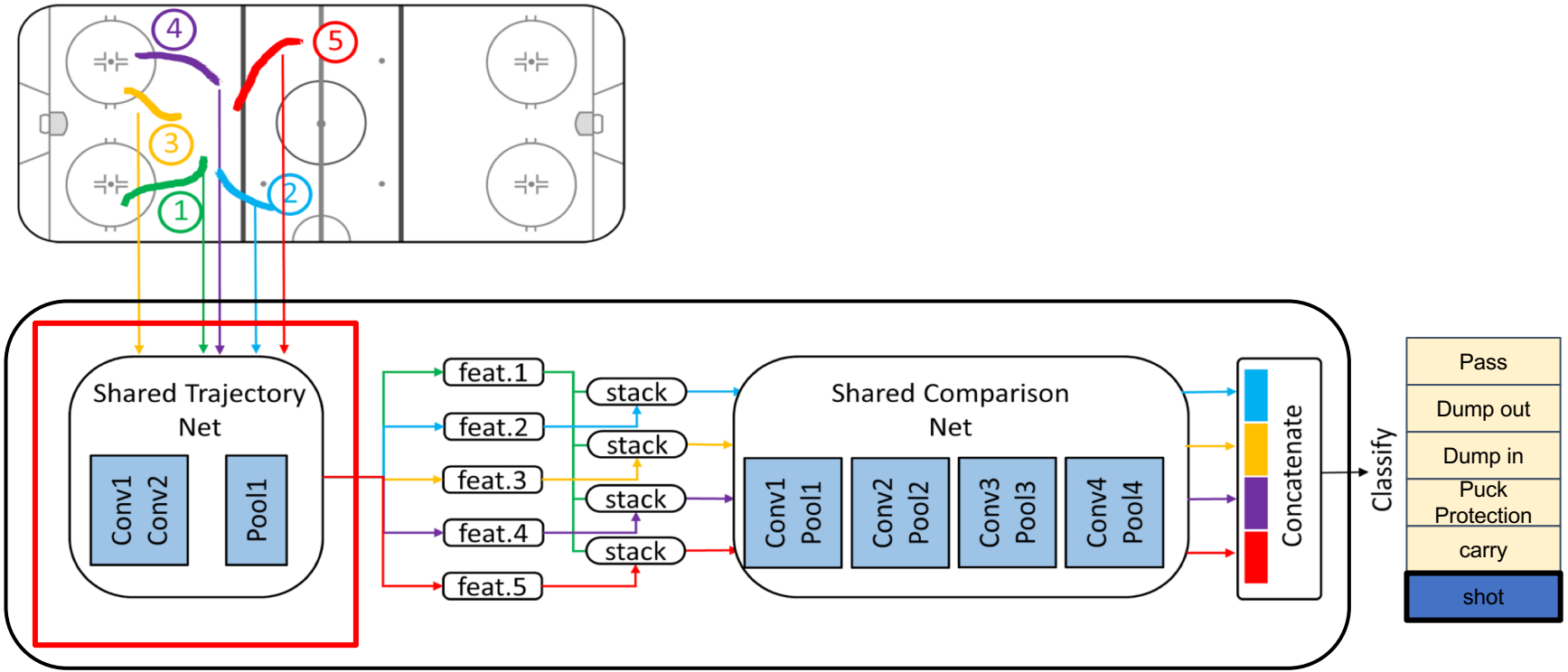
Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Shared-Compare Trajectory Network



Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

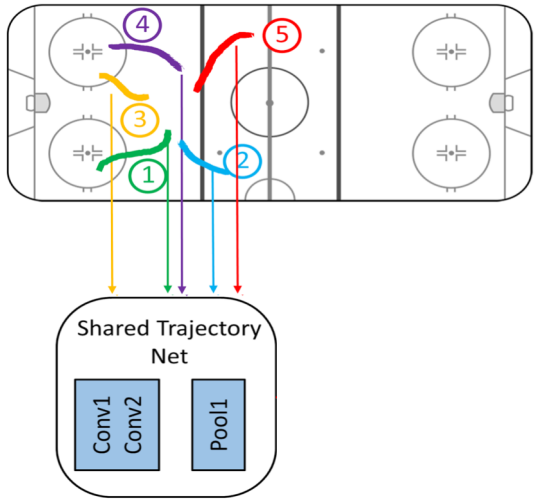
Shared-Compare Trajectory Network



Shared-Compare Trajectory Network

Shared Trajectory Network

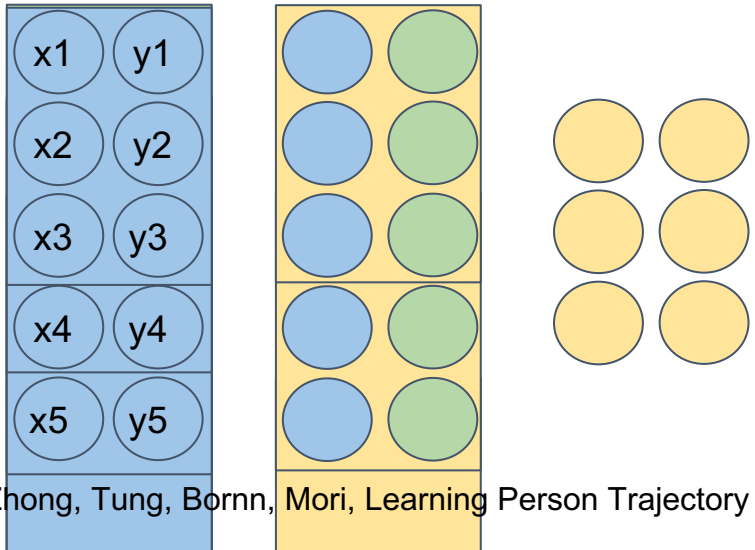
- Consists of 1D convolution and max-pooling
- Learning generic representation for each ind



1D max-pooling layer

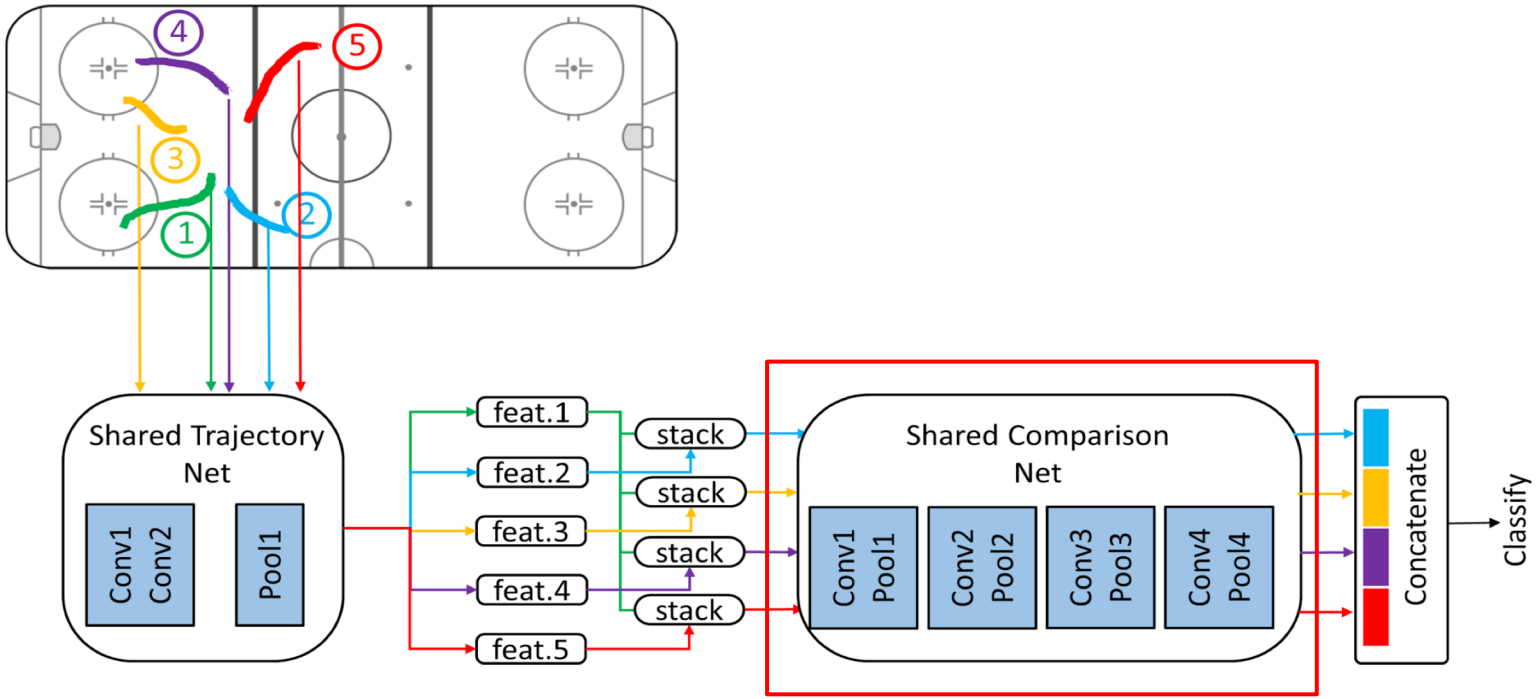
Kernel Size =C * K * M

Pooling stride =2



Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Shared-Compare Trajectory Network

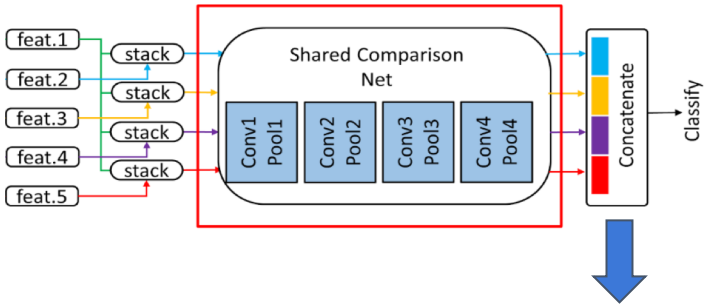


Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Shared Compare Network

Input:

- Pairs of individual trajectory features provided by Shared Trajectory Network
- Pairs are formed relative to a “key player”



Learning:

- The relative motion patterns of pairs
- Interaction cues of players

Enforce an ordering among the players

Output: relative motion pattern representation of each pair

Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Experiments

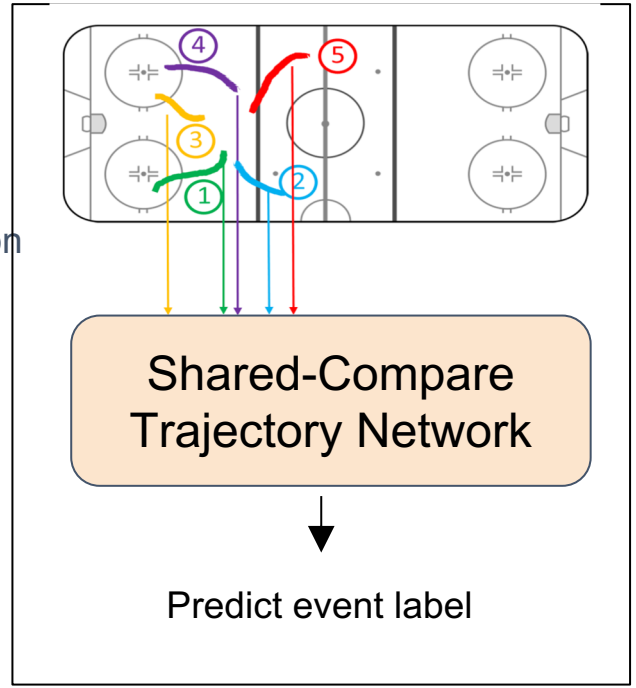
- Event Recognition on the Sportlogiq Dataset
- Team Identification on the NBA Dataset

Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Event recognition using Sportlogiq dataset

Task Definition

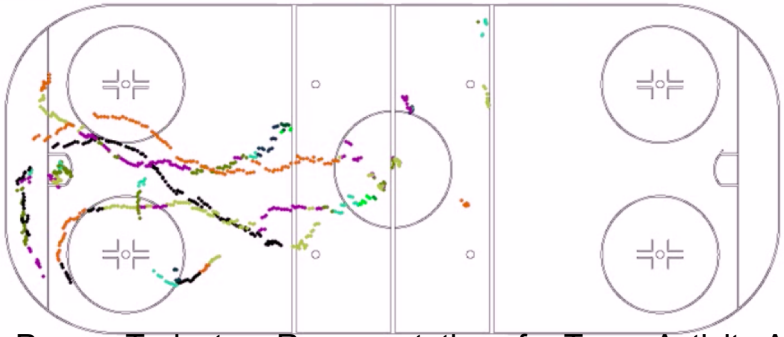
- Event classification
- 6 event classes
  - pass, dump in, dump out, shot, carry, puck protection
- Dataset: Sportlogiq hockey dataset



Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Event recognition using Sportlogiq dataset

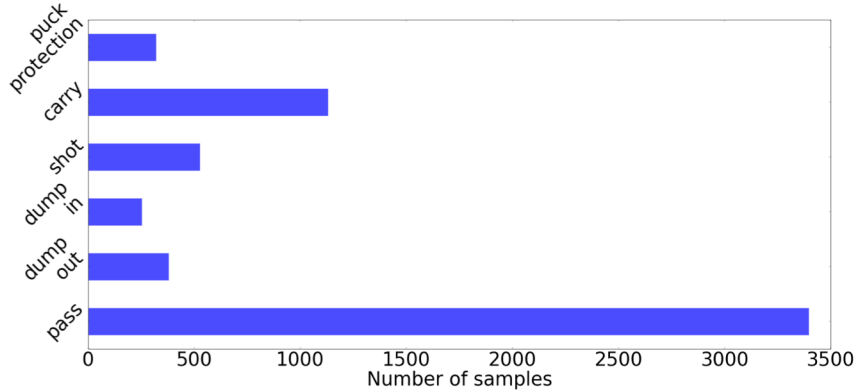
How the Sportlogiq dataset looks



Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Event recognition using Sportlogiq dataset

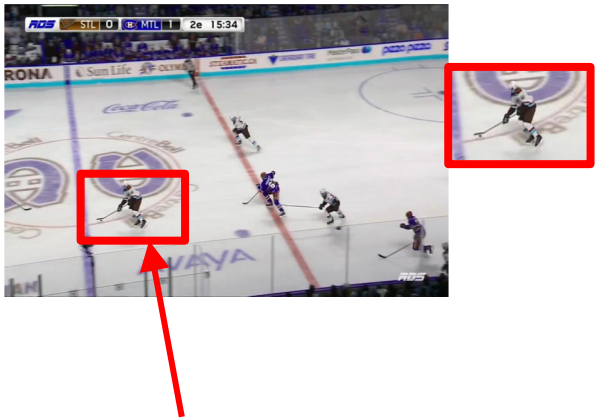
- Sportlogiq Dataset Information
  - State of the art algorithms are used to automatically detect and track players in raw broadcast video
  - Trajectory data are estimated using homography
  - Trajectory length: 16 frames
  - # players used is fixed: 5
  - # of samples of each event
  - 4 games for training, 2 games for validation, and 2 games for testing



Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Event recognition using Sportlogiq dataset

- Training phase:
  - Key player is provided
  - Remaining players are ranked by proximity to the key player
- Test phase:
  - Both cases of known and unknown key player
  - Average pooling strategy for the case of unknown key player



Event recognition on Sportlogiq dataset

Unknown Key Player

	IDT	C3D	Fine-tuned C3D	Shared-Cmp
pass	72.86%	71.10%	77.45%	78.13%
dump out	13.75%	11.66%	18.15%	22.14%
dump in	6.35%	7.58%	19.04%	26.63%
shot	13.05%	23.37%	38.96%	40.52%
carry	45.66%	64.75%	65.65%	61.10%
puck protection	6.28%	6.50%	7.98%	8.72%
mAP	26.32%	30.83%	37.87%	39.54%

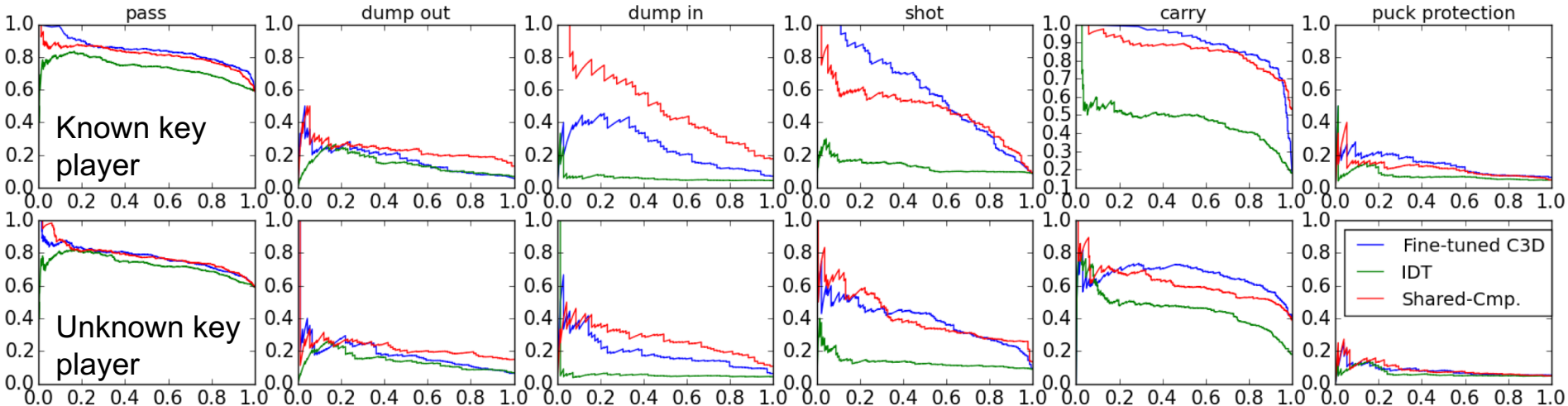
Known Key Player

	IDT	C3D	Fine-tuned C3D	Shared-Cmp
pass	73.35%	77.30%	84.34%	81.33%
dump out	14.34%	10.17%	17.10%	23.11%
dump in	5.77%	10.25%	24.83%	50.04%
shot	13.07%	34.17%	58.88%	48.51%
carry	47.38%	86.37%	90.10%	85.96%
puck protection	7.28%	11.83%	13.99%	11.54%
mAP	26.86%	38.35%	48.21%	50.08%

- In comparison to IDT 13.2 higher mAP
- In comparison to C3D trained from scratch 8.7 higher mAP
- In comparison to fine-tuned C3D 1.7 higher mAP

Event recognition on Sportlogiq dataset

Precision-recall curve



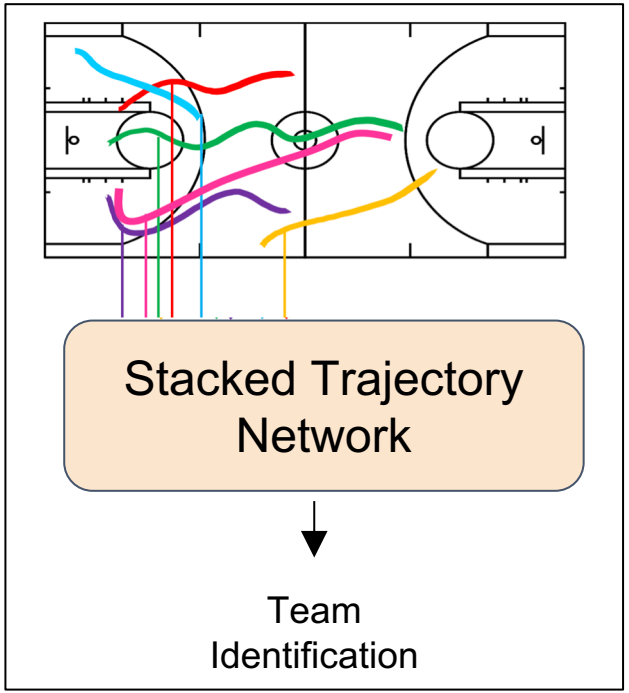
Experiments

- Event Recognition on the Sportlogiq Dataset
- Team Identification on the NBA Dataset

Team Identification on the NBA Dataset

Task Definition

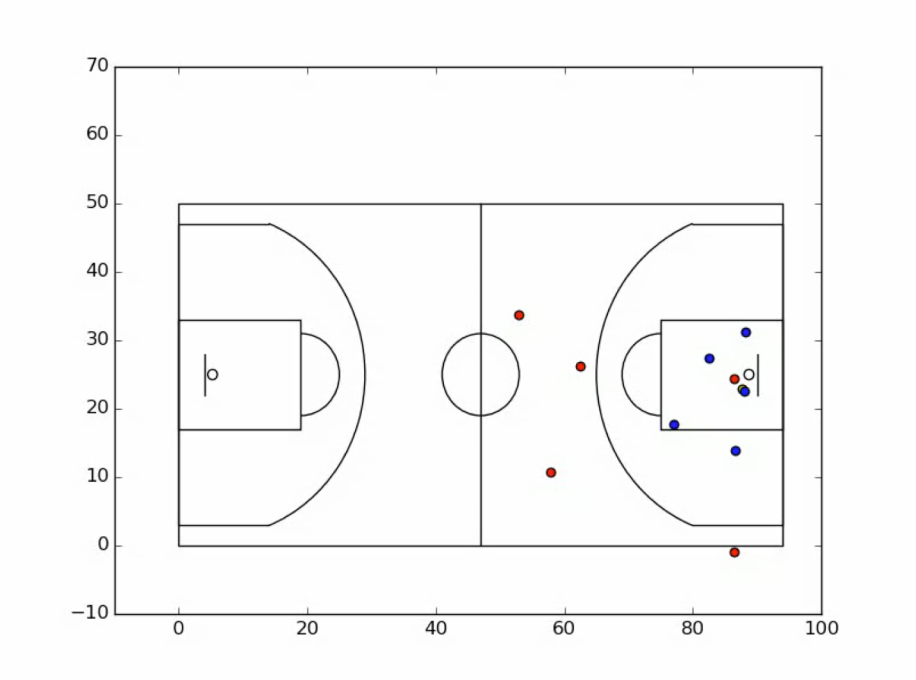
- Team Identification
- Stacked Trajectory Network
- 30 NBA teams
- Dataset: NBA basketball dataset



Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Team Identification on the NBA Dataset

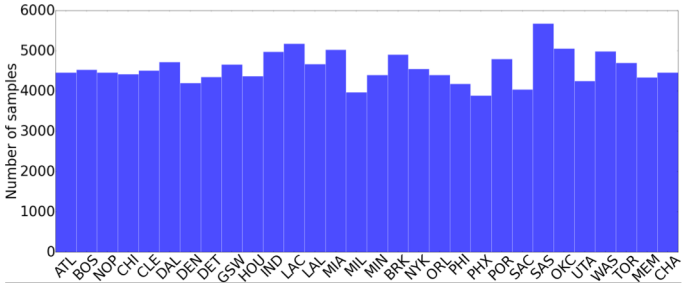
How the NBA dataset looks like



Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Team Identification using NBA dataset

- Dataset Information
  - Trajectory data are acquired by a multi-camera system
  - Sampling rate: 25Hz
  - Extract 137176 possessions from 1076 games
  - 200 frames per possession
  - 82375 poss. for training, 27437 poss. for testing, and 27437 poss. for validation
  - Number of poss. per team



Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

Team Identification on the NBA Dataset

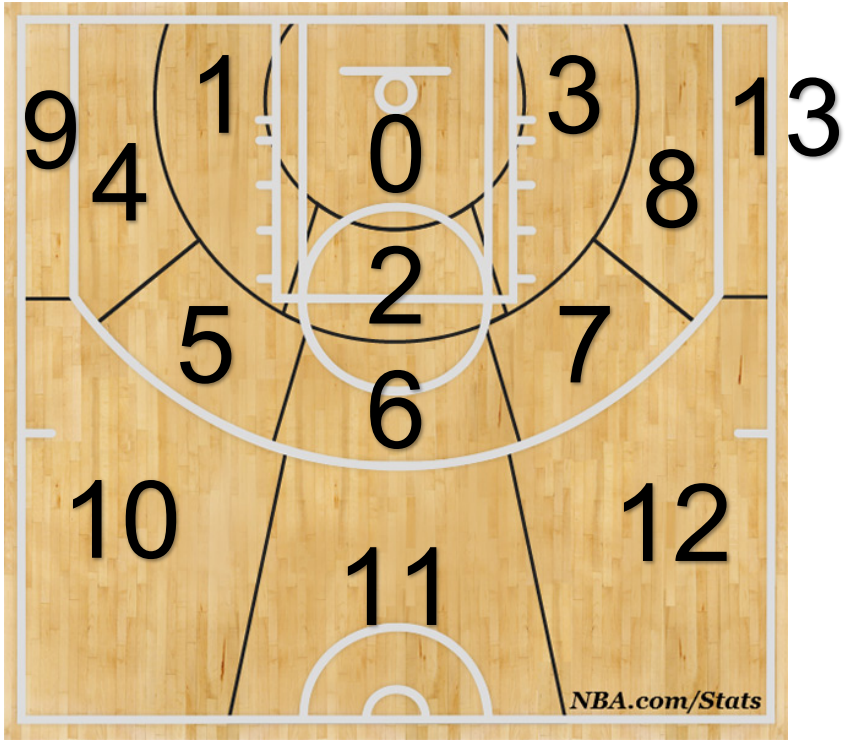
Results

layers	acc	hit@2	hit@3	game acc
2conv	10.68%	18.09%	24.31%	50.00%
3conv	18.86%	28.89%	36.47%	87.05%
4conv	22.34%	33.03%	40.47%	93.41%
5conv	24.78%	35.61%	42.95%	95.91%
5conv+2fc	25.08%	35.83%	42.85%	94.32%

Mehrasa, Zhong, Tung, Bornn, Mori, Learning Person Trajectory Representations for Team Activity Analysis, SLOAN 2018

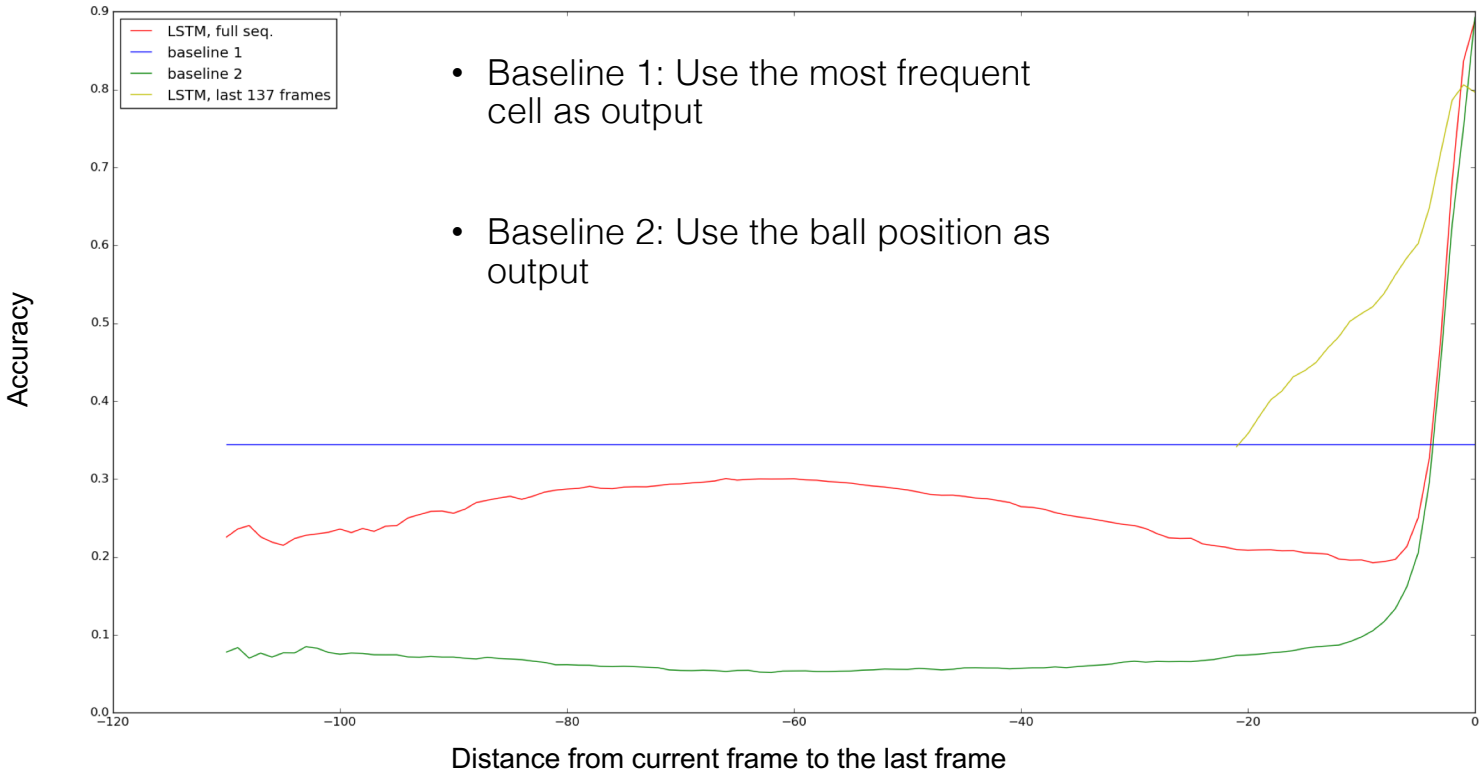
Shot location Prediction

- Task: Predict where the next shot will take place
- Input: A sequence of 2D positions of 10 players and the ball in the court coordinates.
- Output: A distribution over shooting zones; A cell where the next shot will most likely take place
- This discretization is commonly used for analyzing hot shooting zones



Reference: http://www.nba.com/bucks/features/boeder-130923

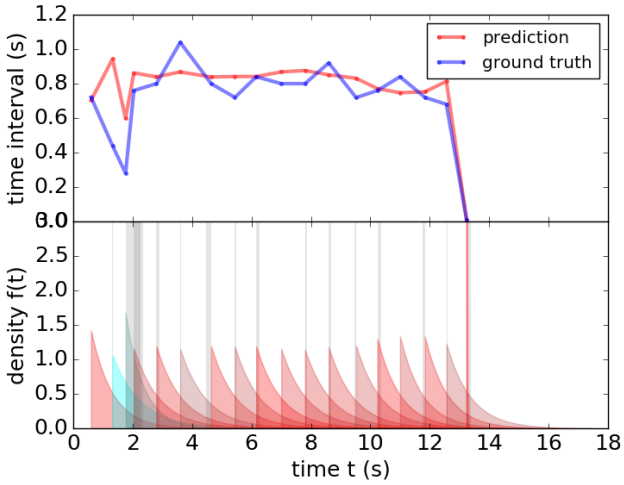
Result

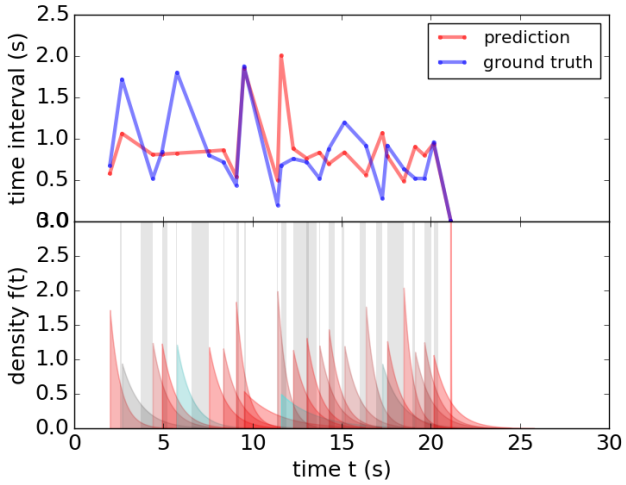


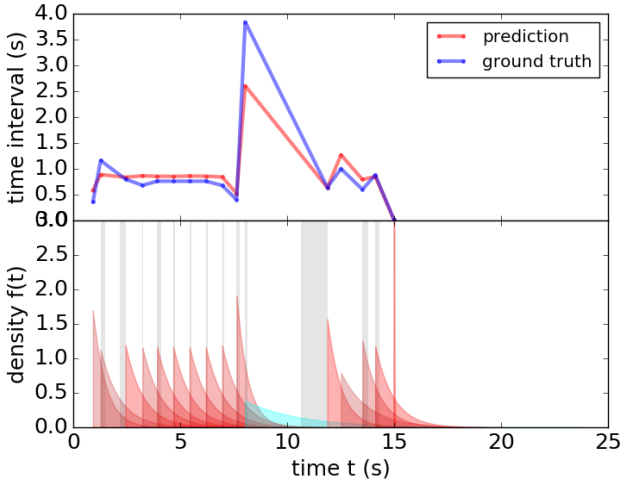
- Baseline 1: Use the most frequent cell as output
- Baseline 2: Use the ball position as output

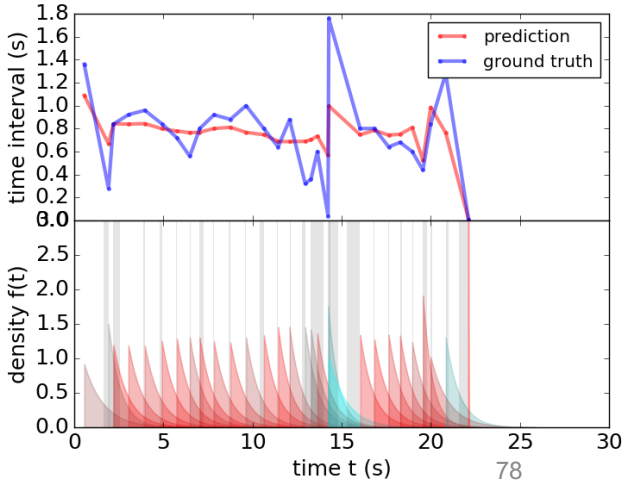
Show video

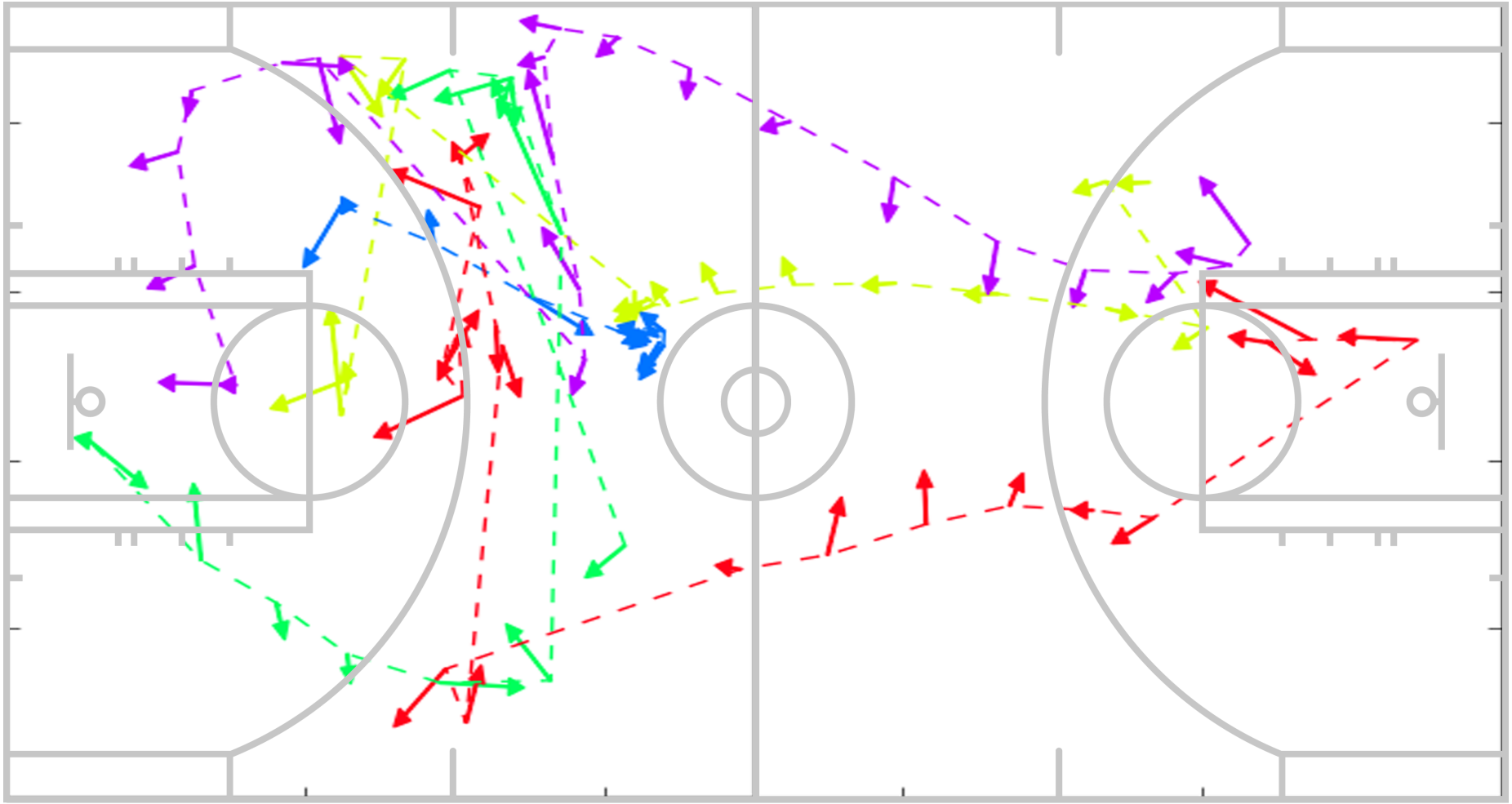
- Predict next activity
  - When
  - Where
  - What











Conclusion

Methods for handling *structures* in deep networks

Label structure: message passing algorithms for multi-level image/video labeling; purely from image data or with partial labels

Temporal structure: action detection in time; efficient glimpsing of video frames

Group structure: network structures to connect related people, gating functions or modules for reasoning about relations