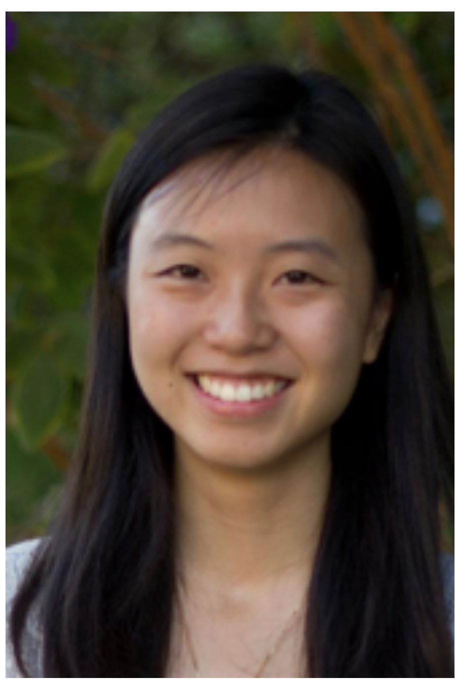
Towards web-scale video understanding

Olga Russakovsky



Serena Yeung
(Stanford)

Achal Dave
(CMU)

PRINCETON UNIVERSITY

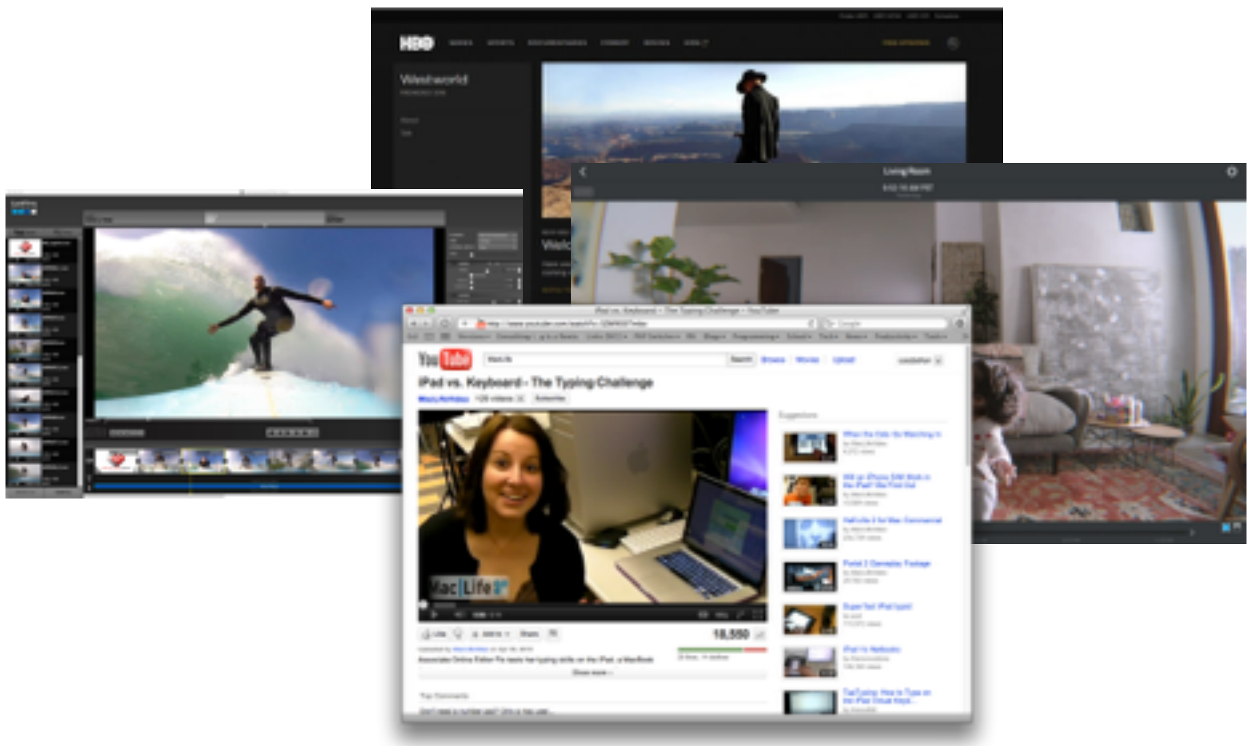
Stanford University

Carnegie Mellon University

400 hours of video are uploaded to YouTube every minute

70% of Internet traffic was videos in 2016, will be over 80% by 2020

[1]http://________________________________
[2]White paper: Cisco VNI Forecast and Methodology, 2015-2020



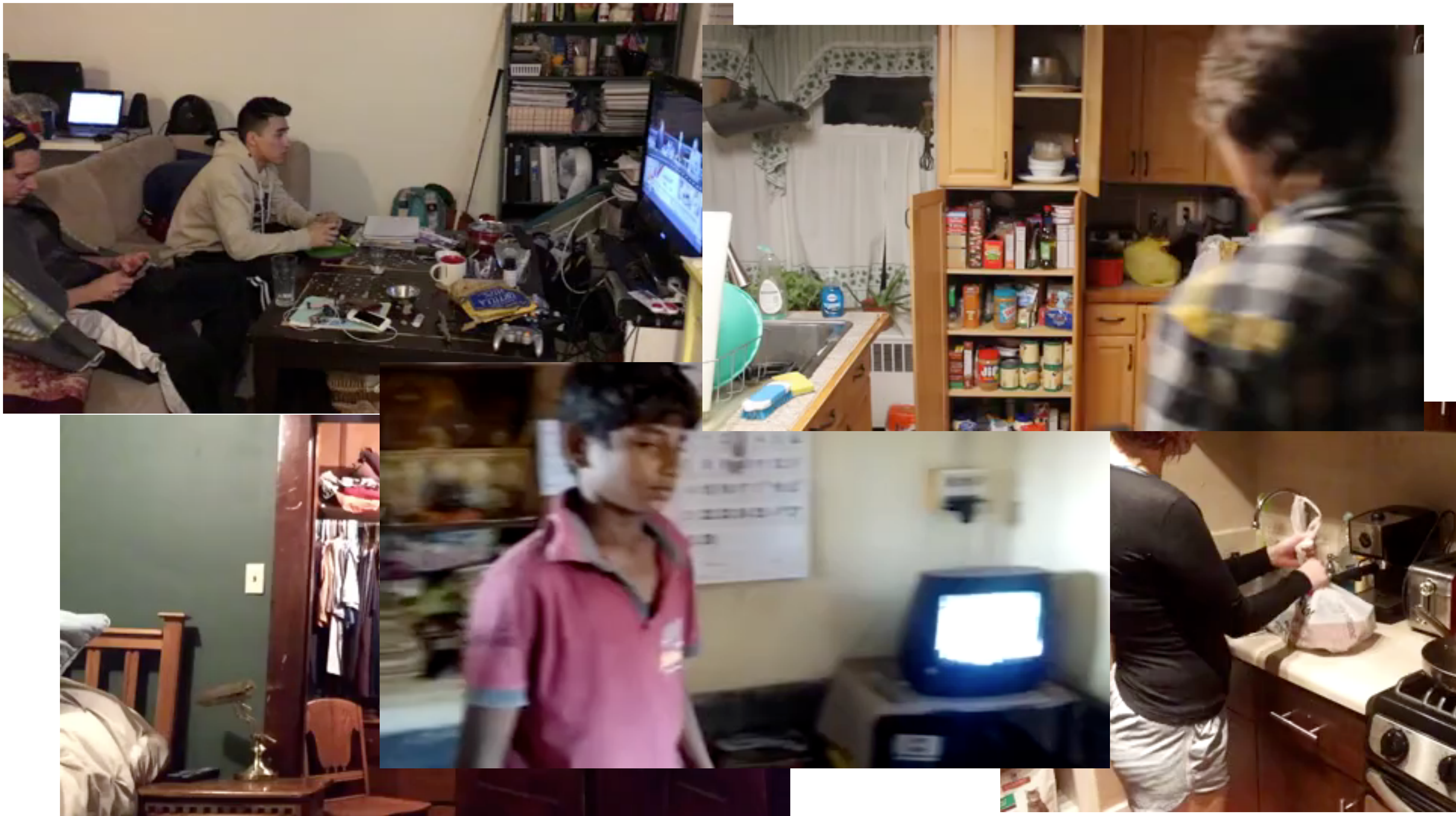
Videos

Knowledge of the dynamic visual world

Capture temporal cues

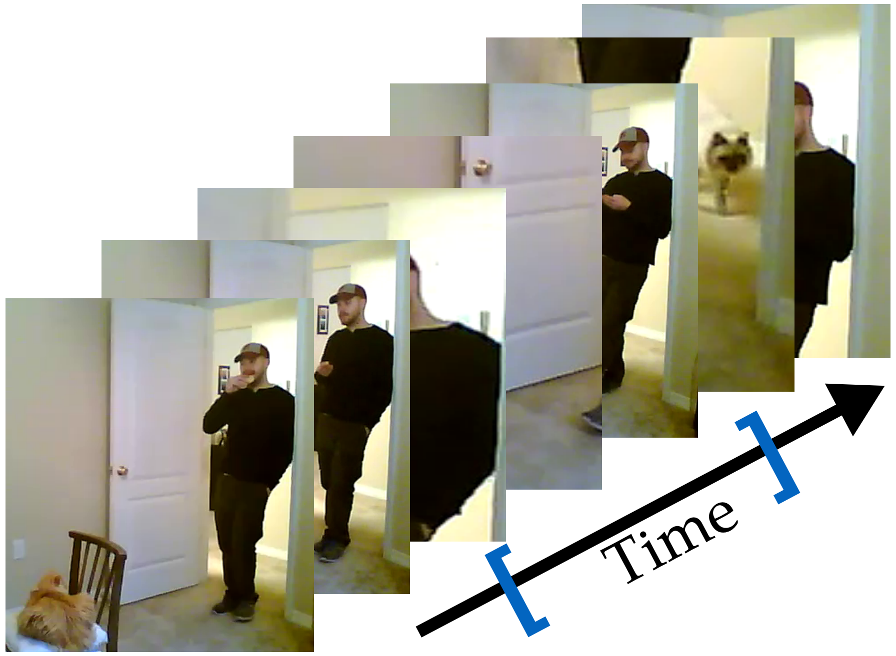
(while handling correlations)

Allocate computation



Forego expensive annotation

(while embracing ambiguity)



Agreement over
<u>spatial</u> boundaries in <u>images</u>:
96-98% above 0.5 IOU
[Papadopoulos et al. ICCV 2017]

Agreement over
<u>temporal</u> boundaries in <u>videos</u>:
76% above 0.5 IOU
[Sigurdsson et al. ICCV 2017]

Challenges of videos @ scale

Modeling

Capture temporal cues while handling correlations

Learning

Learn new concepts cheaply and while embracing ambiguity

Inference

Allocate computation to enable large-scale processing

Challenges of videos @ scale

Modeling

Capture temporal cues while handling correlations

Learning

Learn new concepts cheaply and while embracing ambiguity

Inference

Allocate computation to enable large-scale processing

Groundtruth

BodyBend

BodyContract

ClapHands

FistPump

HammerThrow

HammerThrowRelease

HammerThrowSpin

HammerThrowWindUp

PickUp

Run

Sit

Squat

Stand

Throw

Walk

Some desired modeling properties

- Capture temporal cues
- Effectively handle correlated examples
- Provide an interpretable notion of memory
- Operate in an online manner

Current approaches

- **Two-stream networks** [Simonyan et al. NIPS 2014]: incorporates motion through optical flow
  - Computationally intensive!
- **C3D** [Tran et al. ICCV 2015]: Operates via 3D convolutions on groups of video frames
  - Memory intensive
  - Tends to oversmooth
- **Recurrent networks, e.g., Clockwork RNNs** [Koutnik et al. ICML 2014]: Maintain memory of “entire” history of video
  - History not easily interpretable
  - Training requires SGD on correlated data

Predictive-corrective networks

- Key idea: Inspired by Kalman Filtering
- Suppose our images and action scores evolve smoothly, as with a linear dynamical system:

Actions $\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + noise$

Frames $\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + noise$

- Can create improved estimates of action scores by:

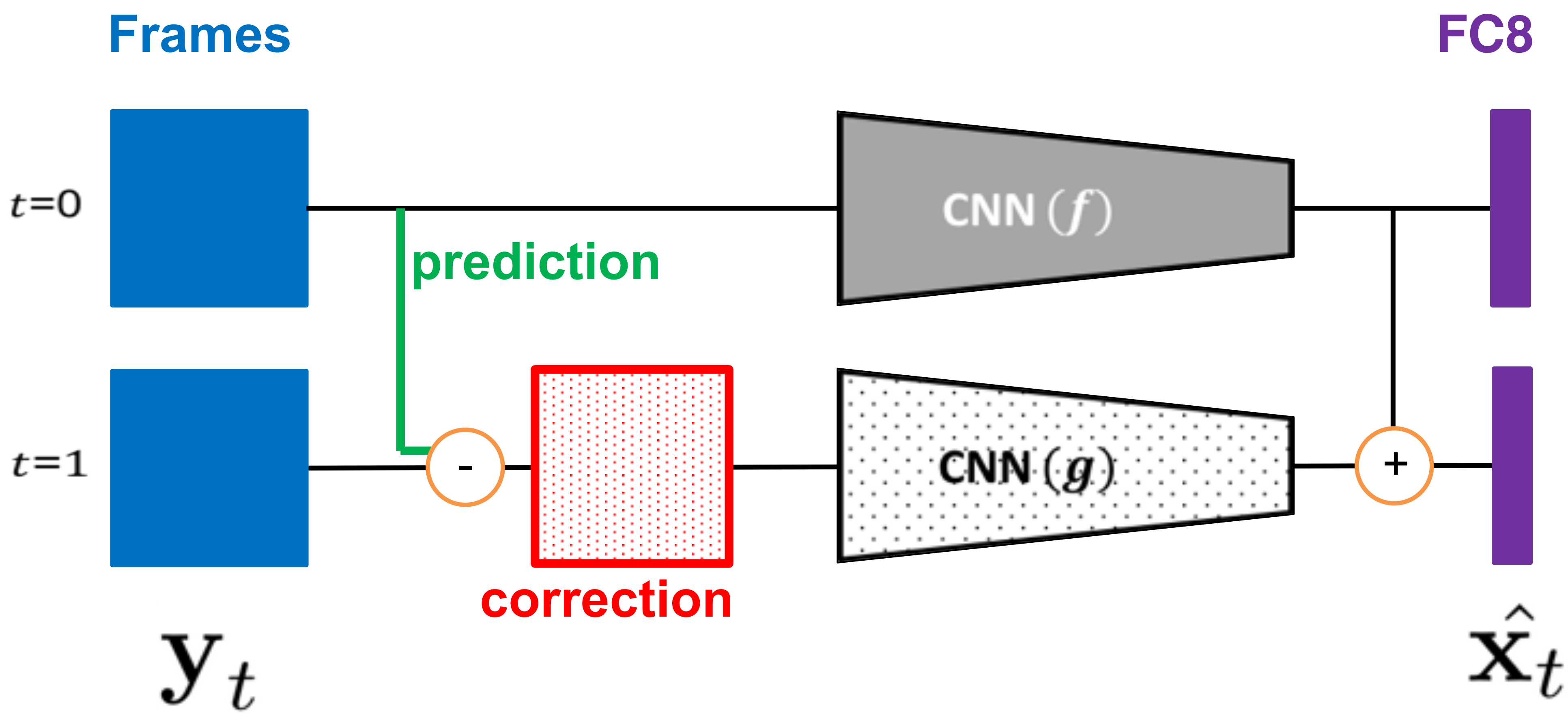
$$\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_{t-1} + g(\mathbf{y}_t - \hat{\mathbf{y}}_t)$$

Prediction

Correction

[Dave, Russakovsky, Ramanan. "Predictive-Corrective Networks for Action Detection." CVPR 2017]

Predictive-corrective instantiation



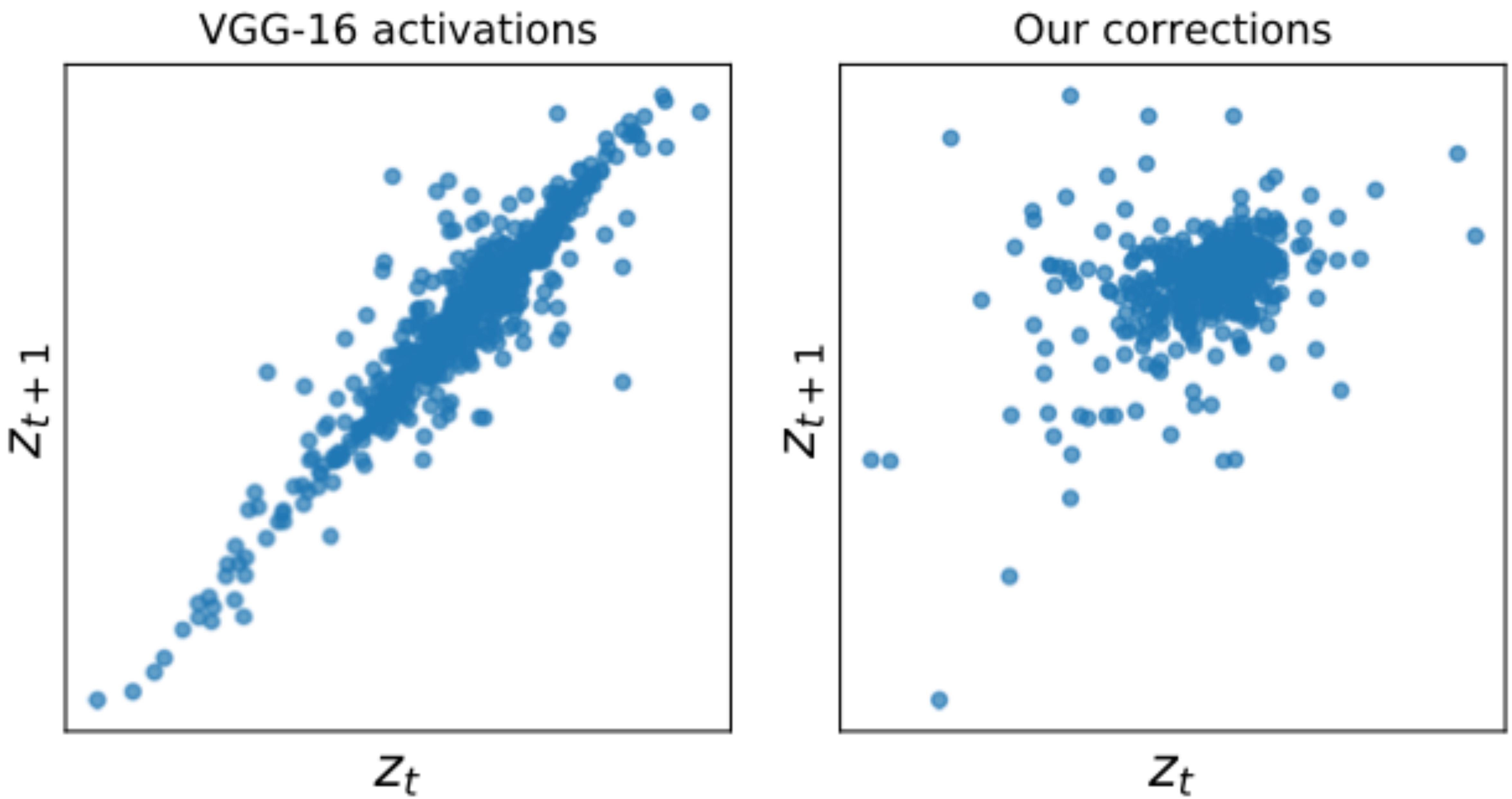
$$\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_{t-1} + g(\mathbf{y}_t - \hat{\mathbf{y}}_t)$$

Prediction

Correction

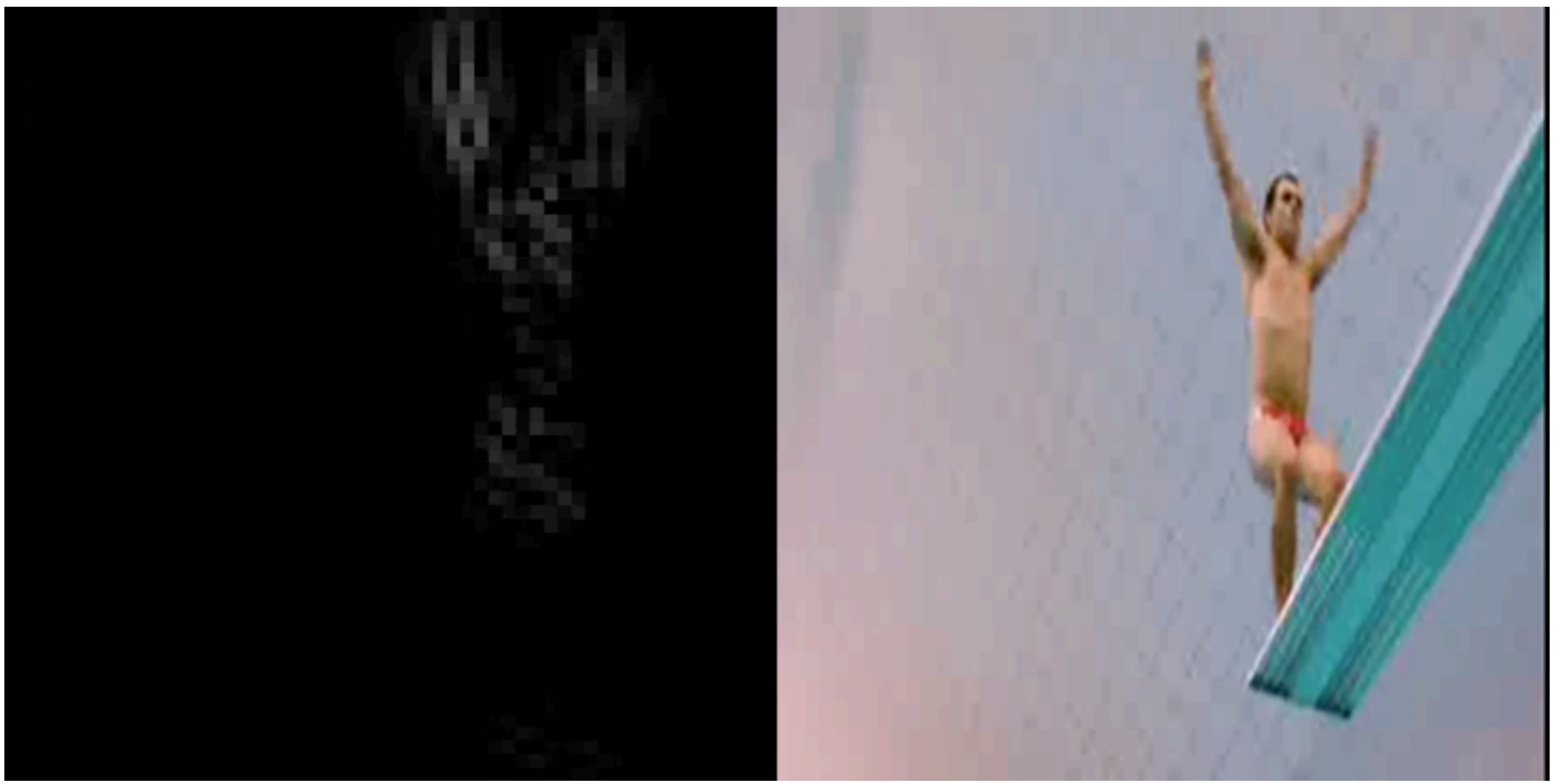
[Dave, Russakovsky, Ramanan. "Predictive-Corrective Networks for Action Detection." CVPR 2017]

De-correlate data (conv4-3 layer)



[Dave, Russakovsky, Ramanan. “Predictive-Corrective Networks for Action Detection.” CVPR 2017]

Visualizing the corrections



[Dave, Russakovsky, Ramanan. “Predictive-Corrective Networks for Action Detection.” CVPR 2017]

To summarize

Observe t=0

Predict t=1

Observe t=1

Correct

[Dave, Russakovsky, Ramanan. “Predictive-Corrective Networks for Action Detection.” CVPR 2017]

Results

Per-frame classification (mAP)

	THUMOS	MultiTHUMOS	Charades
Single-frame	34.7	25.4	7.9
Two-stream	36.2	27.6	**8.9**
LSTM (RGB)	**39.3**	28.1	7.7
Predictive-Corrective	38.9	**29.7**	**8.9**

[Dave, Russakovsky, Ramanan. “Predictive-Corrective Networks for Action Detection.” CVPR 2017]

Results

Per-frame classification (mAP)

	THUMOS	MultiTHUMOS	Charades
Single-frame	34.7	25.4	7.9
Two-stream	36.2	27.6	**8.9**
LSTM (RGB)	**39.3**	28.1	7.7
Predictive-Corrective	38.9	**29.7**	**8.9**

[Dave, Russakovsky, Ramanan. "Predictive-Corrective Networks for Action Detection." CVPR 2017]

Results

Per-frame classification (mAP)

	THUMOS	MultiTHUMOS	Charades
Single-frame	34.7	25.4	7.9
Two-stream	36.2	27.6	**8.9**
LSTM (RGB)	**39.3**	28.1	7.7
Predictive-Corrective	38.9	**29.7**	**8.9**

[Dave, Russakovsky, Ramanan. “Predictive-Corrective Networks for Action Detection.” CVPR 2017]

Challenges of videos @ scale

Modeling

Capture temporal cues using a Kalman filter

- Competitive with two-stream without optical flow
- Simplifies learning by decorrelating the input

[Dave, Russakovsky, Ramanan. CVPR 2017]

Learning

Learn new concepts cheaply and while embracing ambiguity

Inference

Allocate computation to enable large-scale processing

Challenges of videos @ scale

Modeling

Capture temporal cues using a Kalman filter

- Competitive with two-stream without optical flow
- Simplifies learning by decorrelating the input

[Dave, Russakovsky, Ramanan. CVPR 2017]

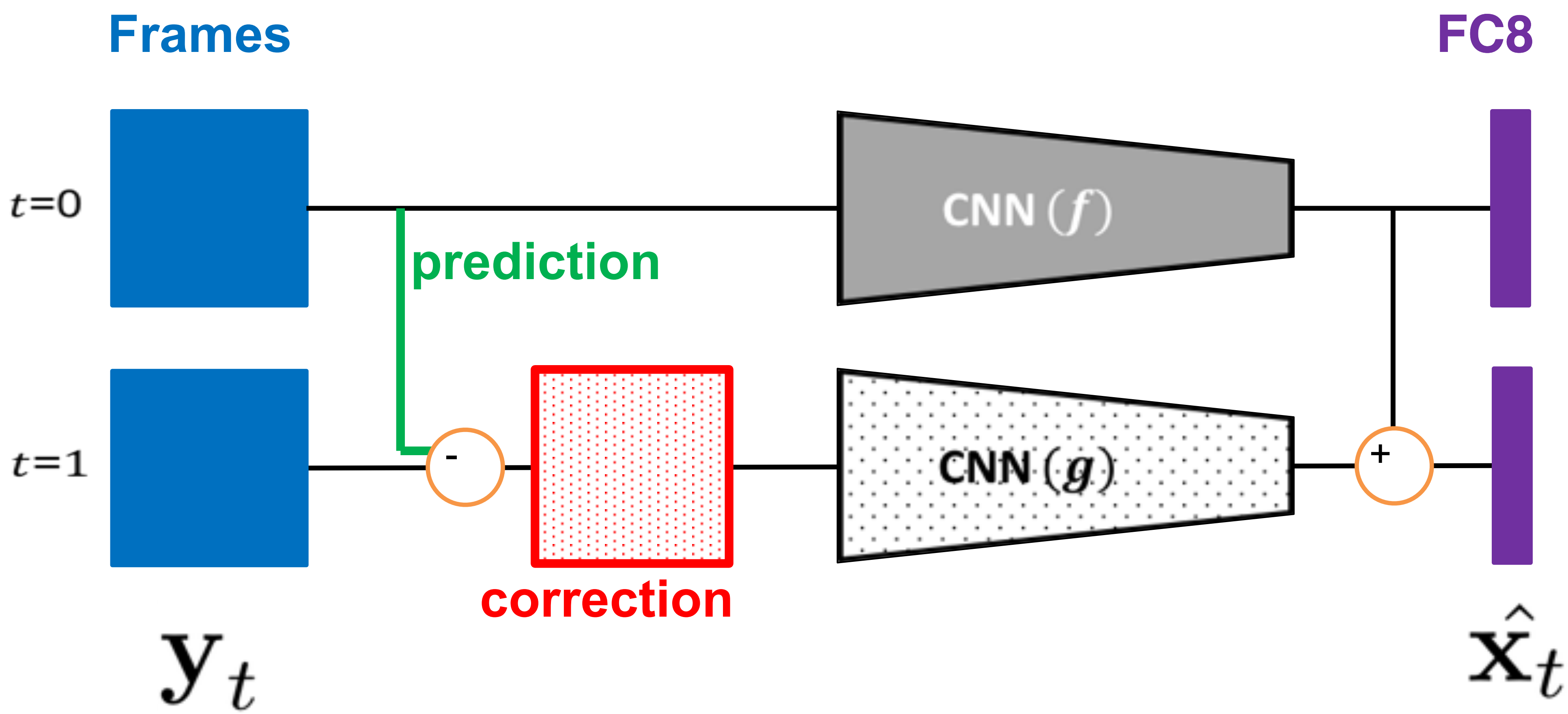
Inference

Allocate computation to enable large-scale processing

Learning

Learn new concepts cheaply and while embracing ambiguity

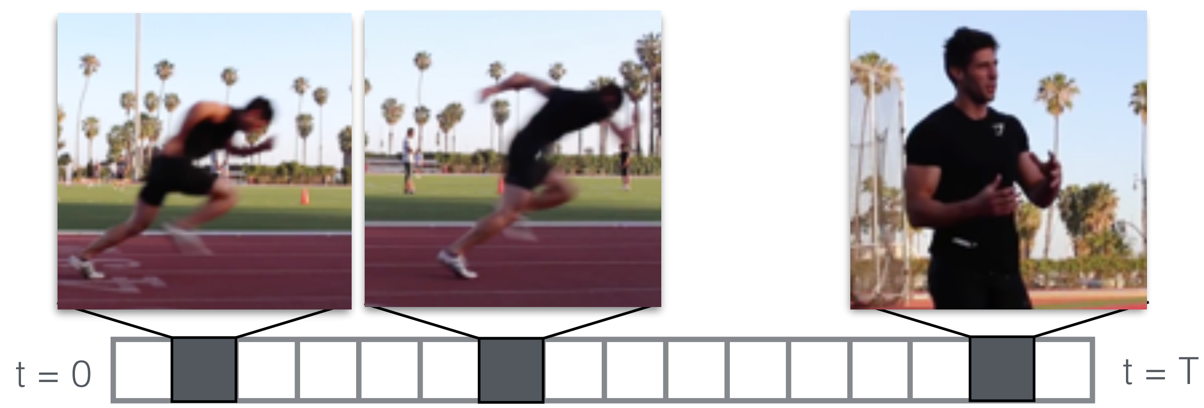
Back to predictive-corrective



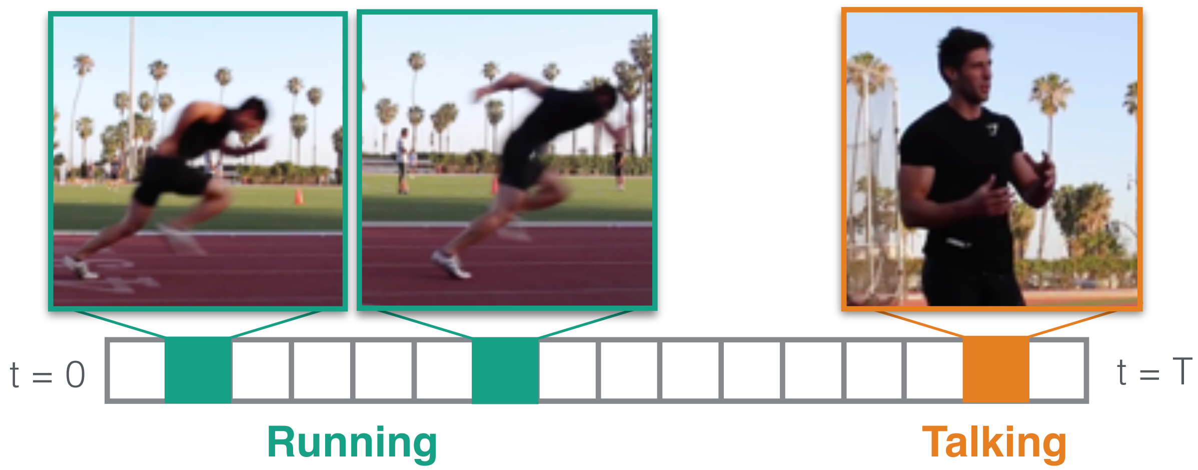
- Can save computation by ignoring the frame if correction is too small (~2x savings)
  - But still need to look at every frame!

[Dave, Russakovsky, Ramanan. “Predictive-Corrective Networks for Action Detection.” CVPR 2017]

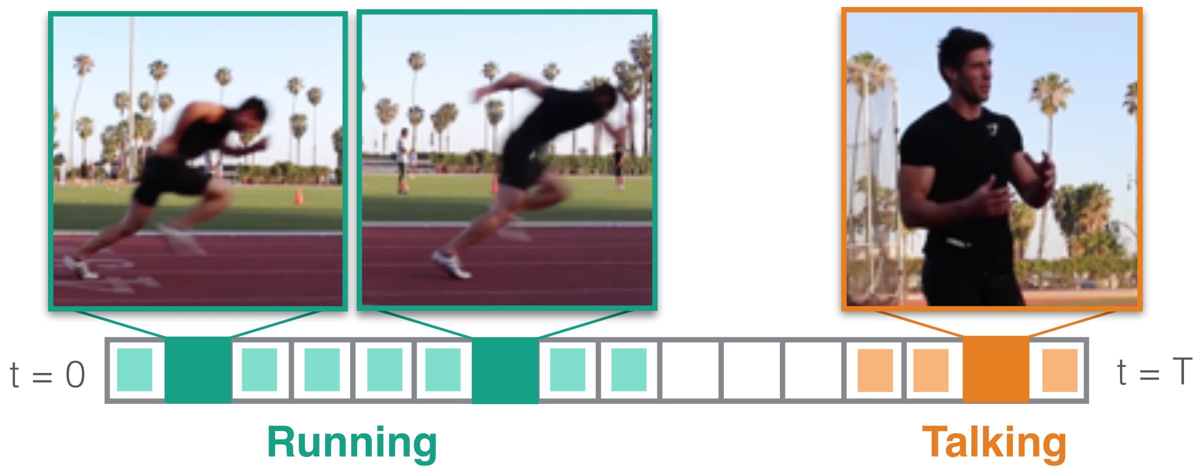
Efficient video processing



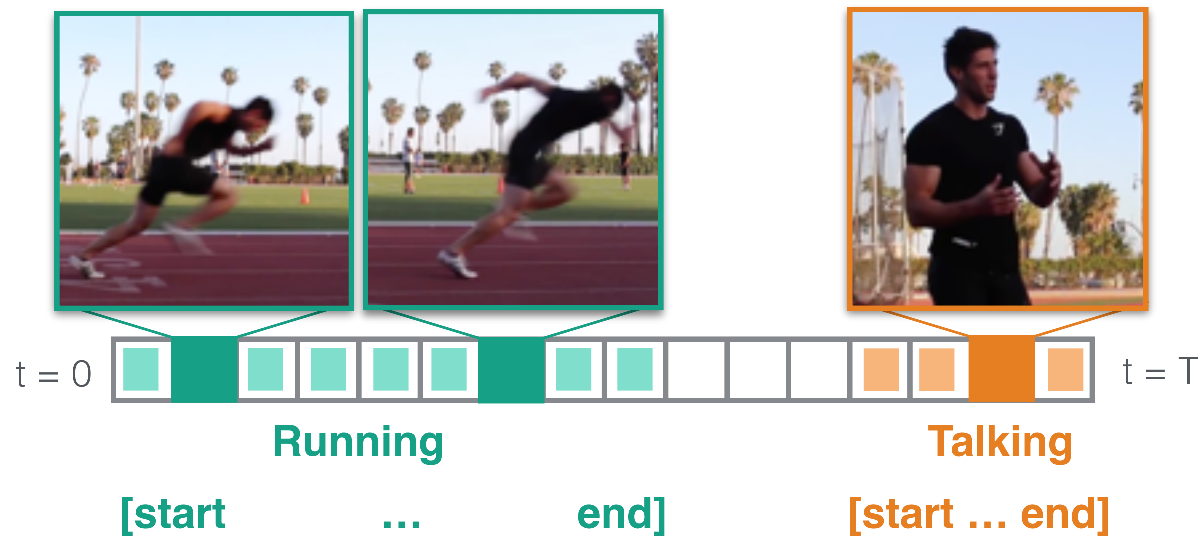
Efficient video processing



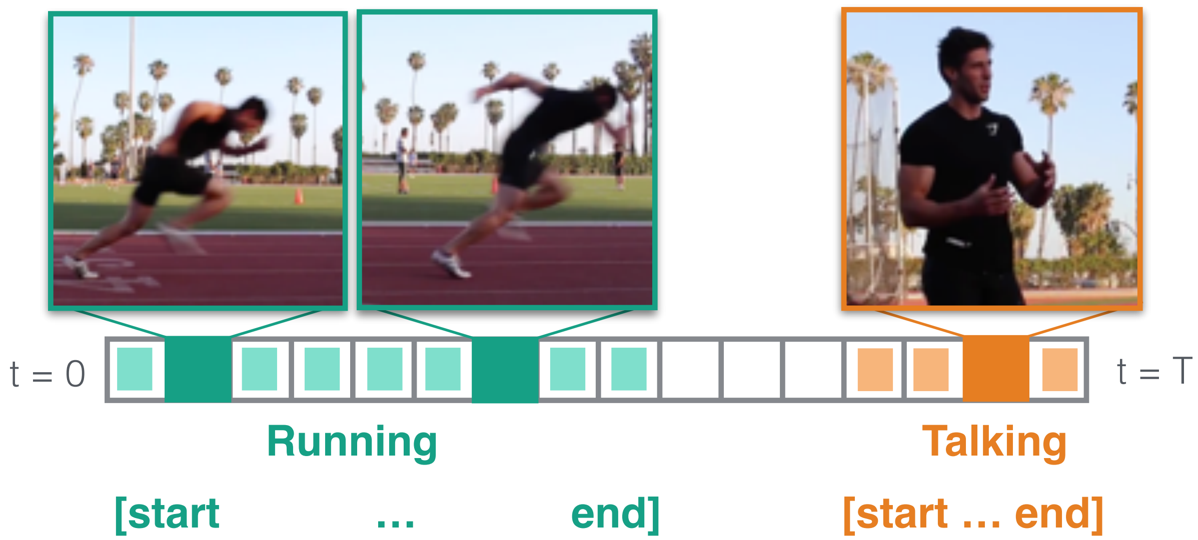
Efficient video processing



Efficient video processing



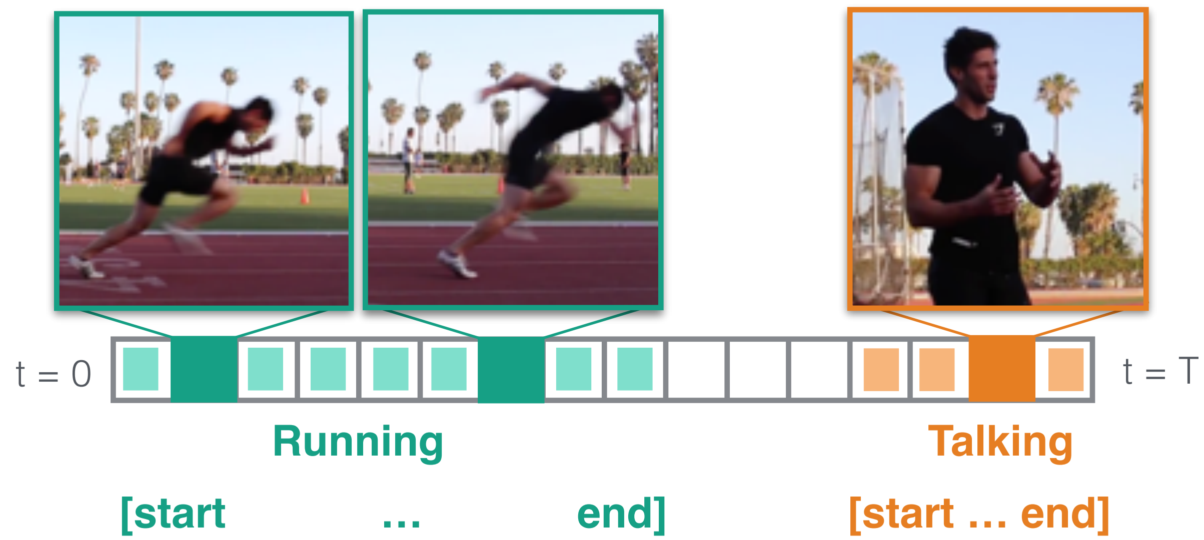
Efficient video processing



“Knowing the output or the final state… there is no need to explicitly store many previous states”

[N. I. Badler. “Temporal Scene Analysis…” **1975**]

Efficient video processing

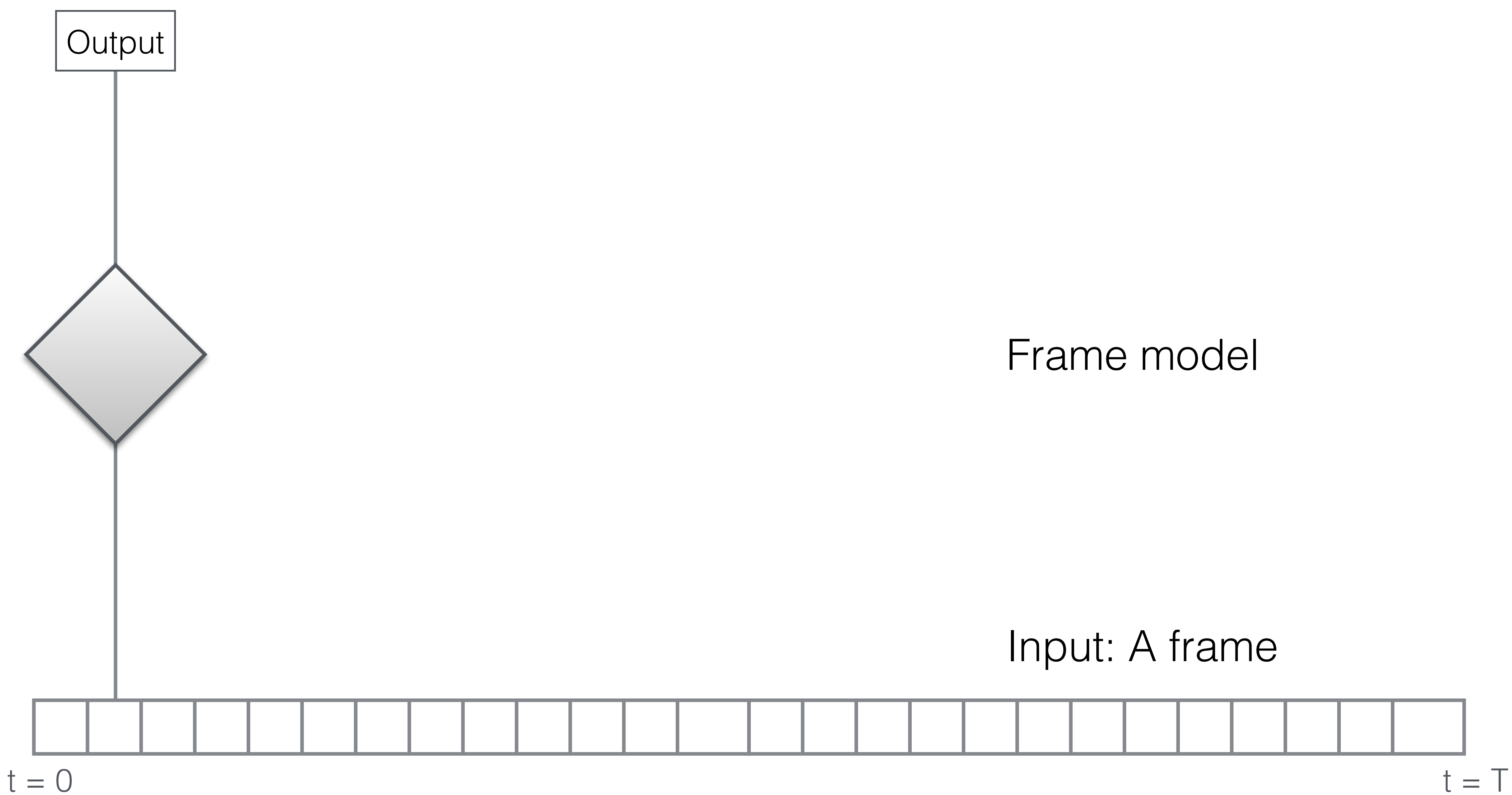


"Knowing the output or the final state… there is no need to explicitly store many previous states"

"Time may be represented in several ways… The intervals between 'pulses' need not be equal."

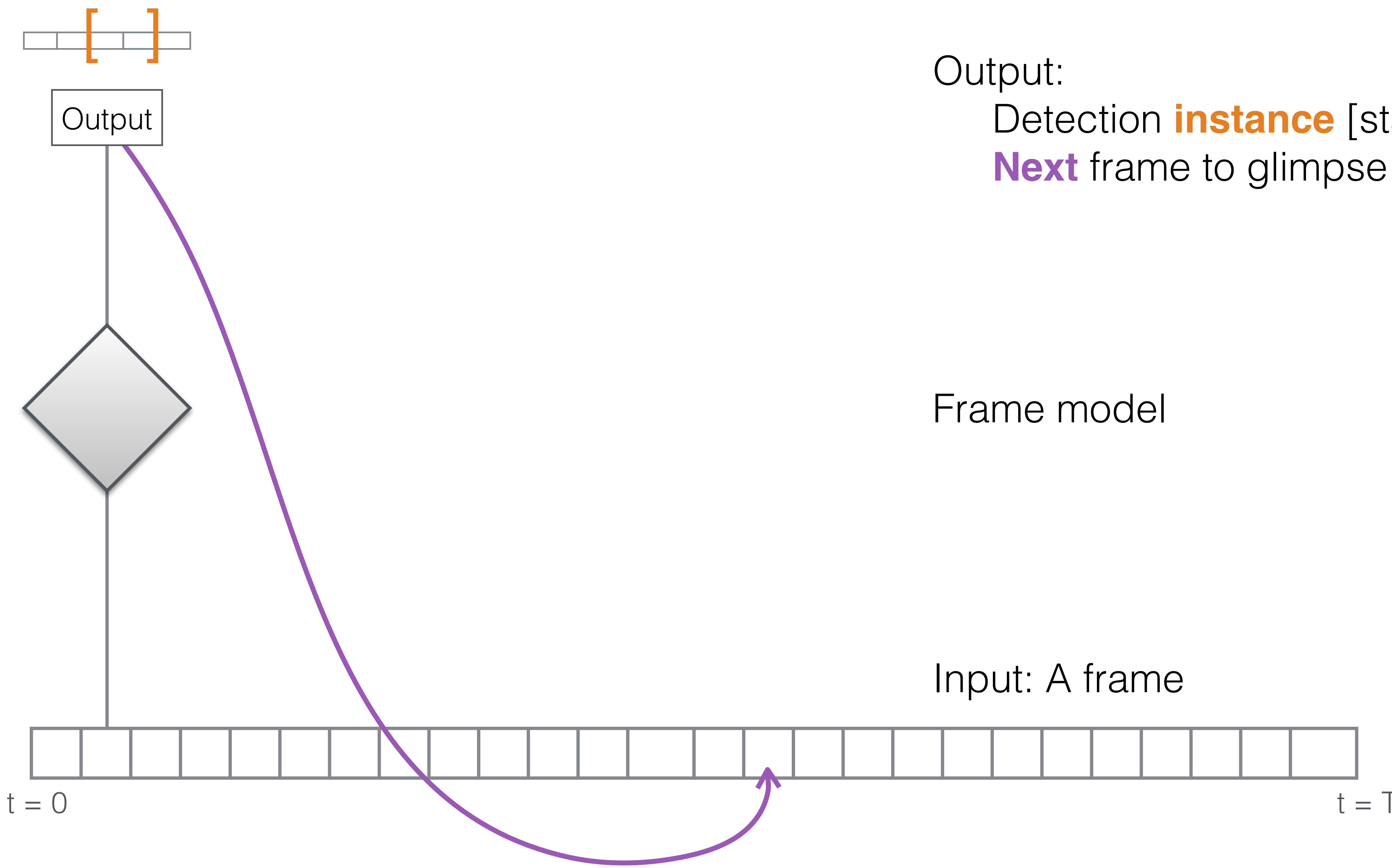
[N. I. Badler. "Temporal Scene Analysis…" **1975**]

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



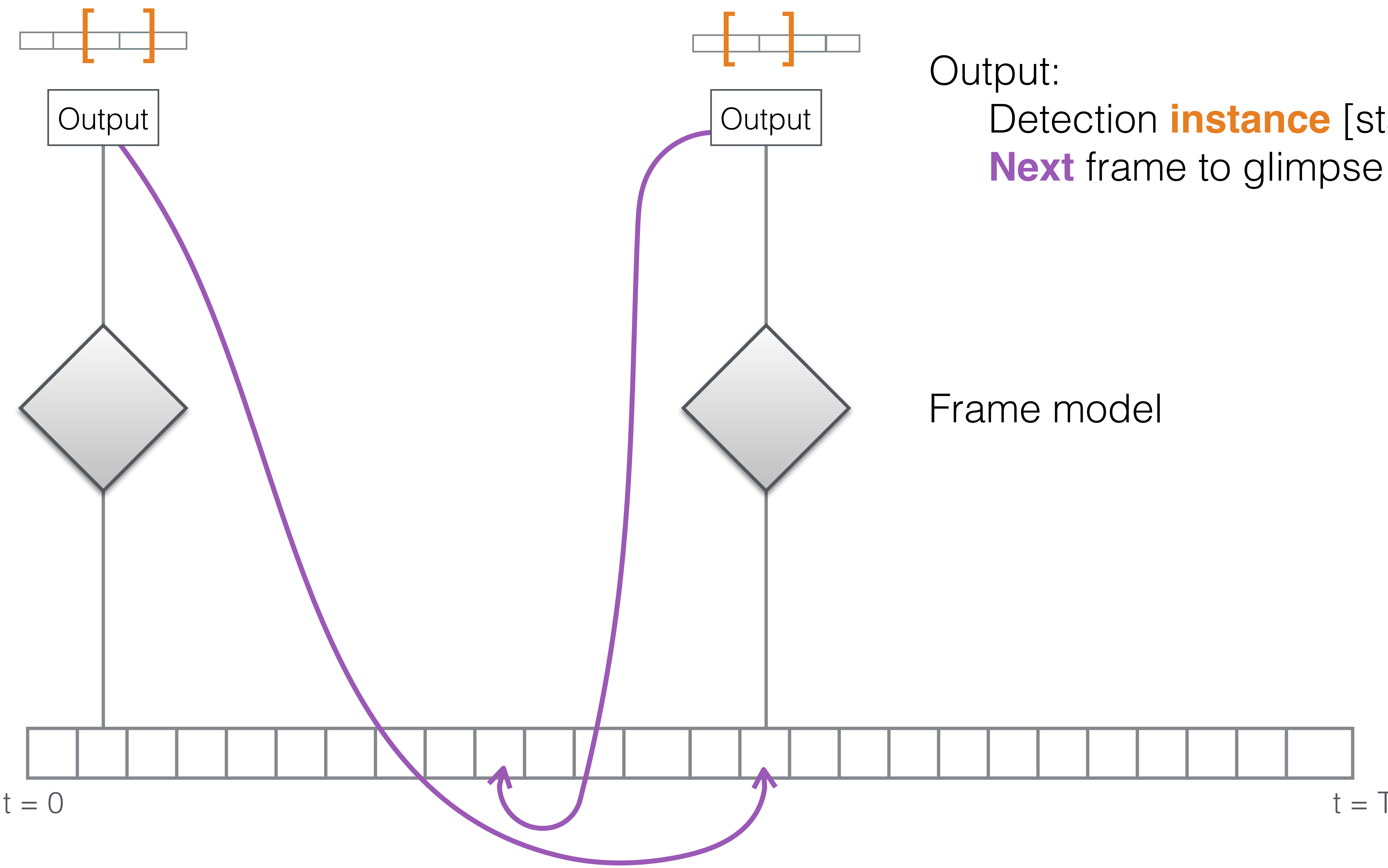
Output:
Detection **instance** [start, end]
Next frame to glimpse

Frame model

Input: A frame

[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

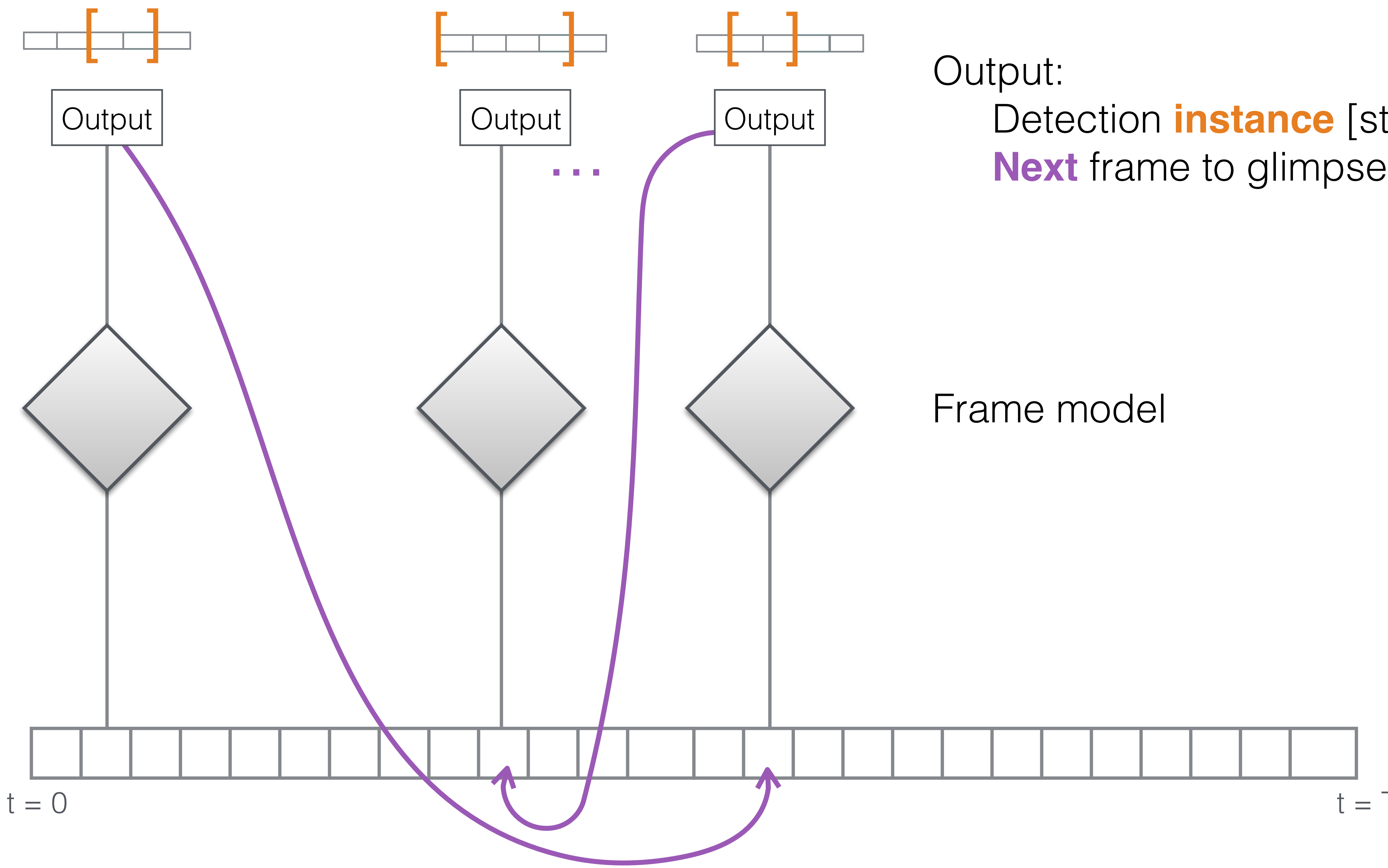


Output:
Detection **instance** [start, end]
Next frame to glimpse

Frame model

[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



Output:

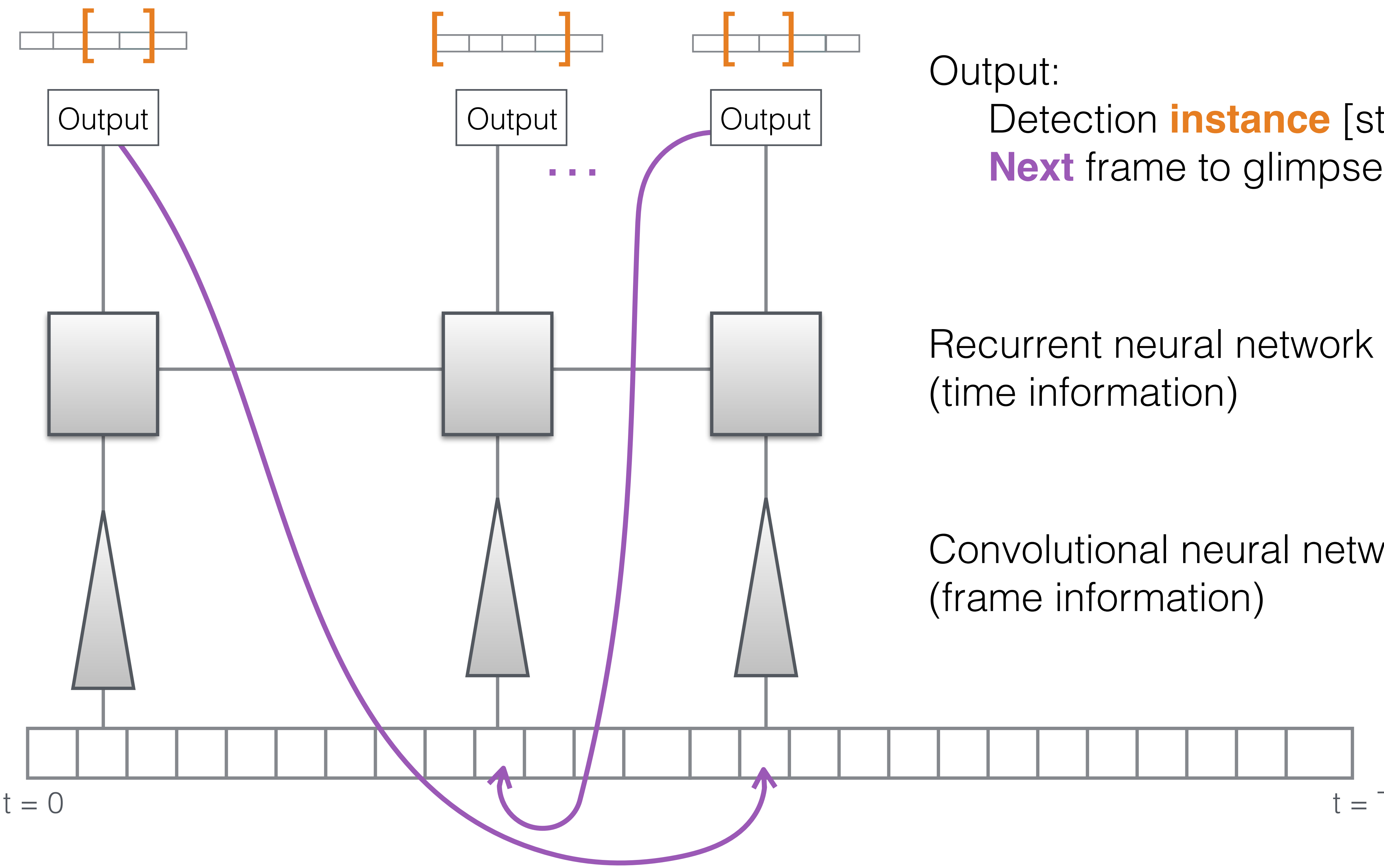
Detection **instance** [start, end]

Next frame to glimpse

Frame model

[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



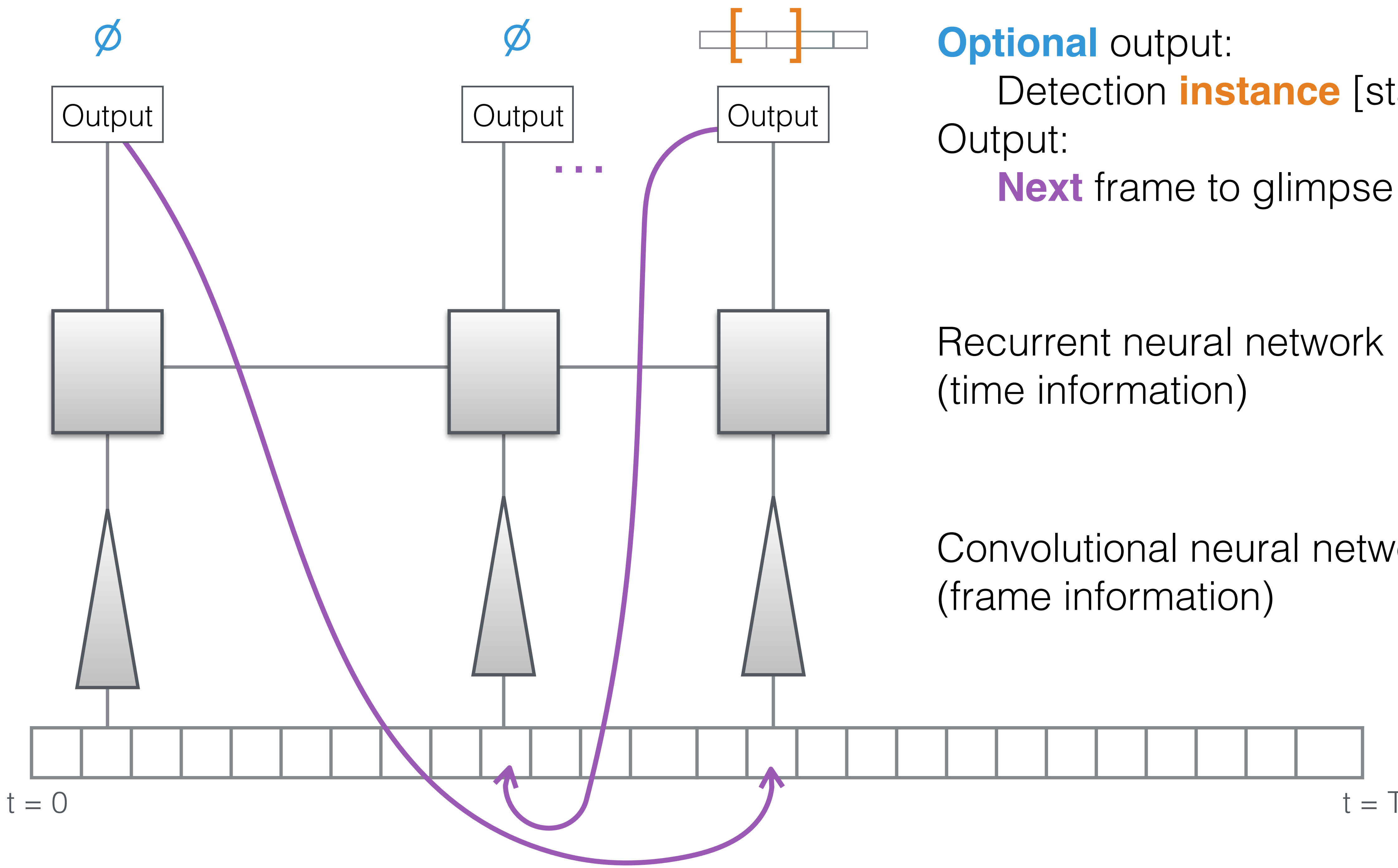
Output:
Detection **instance** [start, end]
Next frame to glimpse

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



Optional output:

Detection **instance** [start, end]

Output:

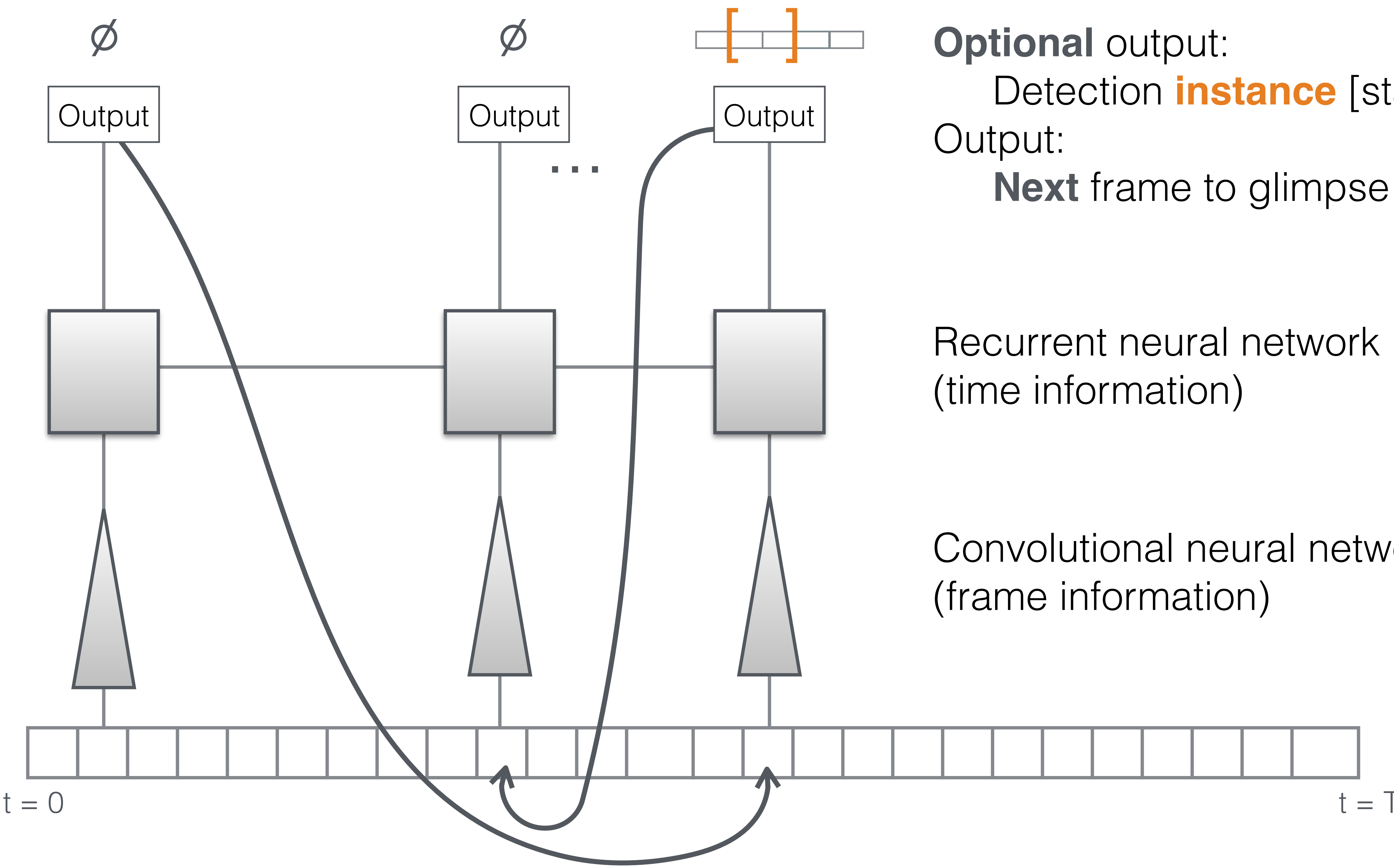
Next frame to glimpse

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]

$$\mathcal{L}(D,G) = \sum_i \underbrace{\mathcal{L}_{cls}(d_i, y_i > 0)}_{\text{cross-entropy classification loss}} + \gamma \sum_{i:y_i>0} \underbrace{\mathcal{L}_{loc}(d_i, g_{y_i})}_{L_2 \text{ distance localization loss}}$$



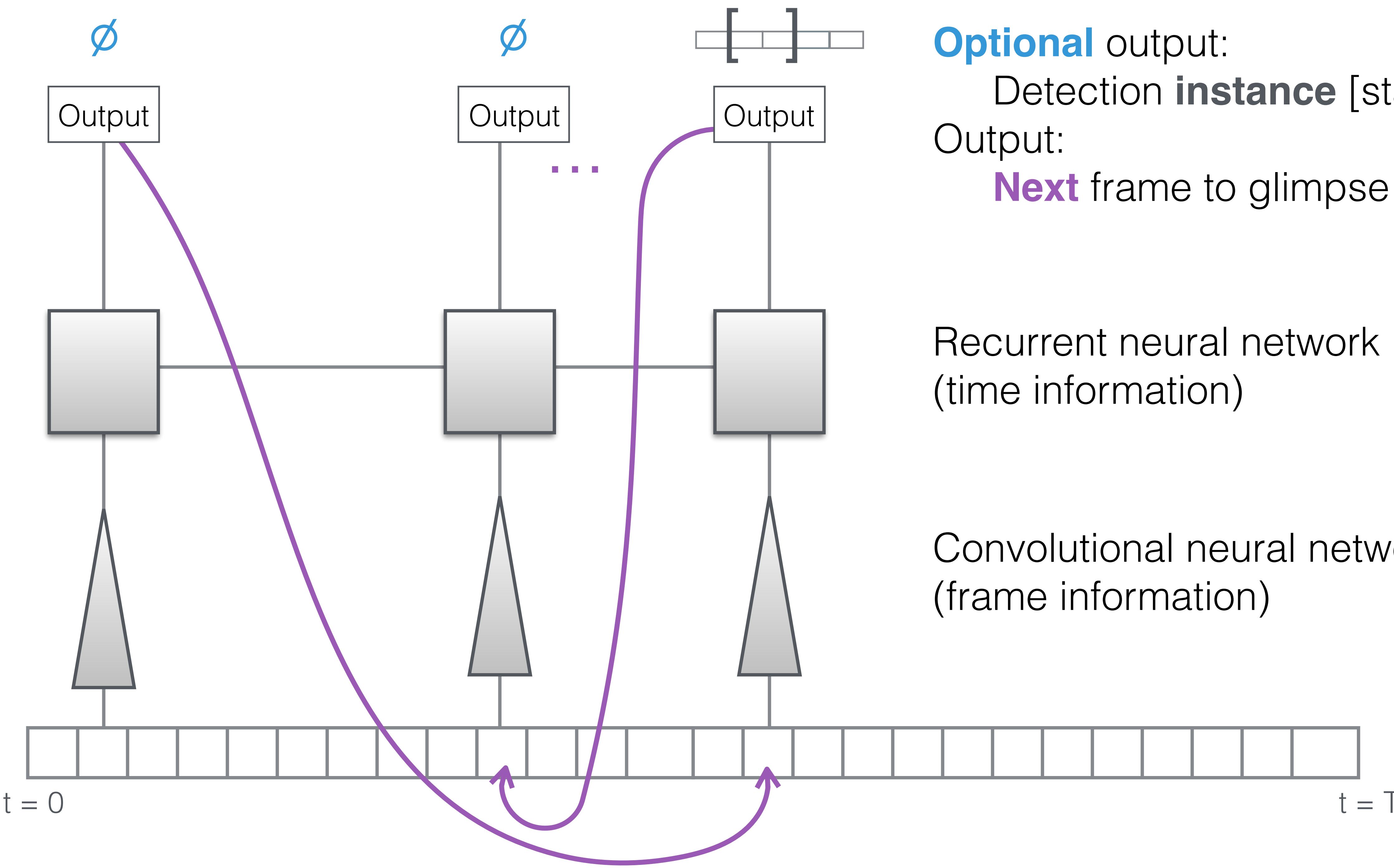
Optional output:
Detection **instance** [start, end]
Output:
Next frame to glimpse

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]

Train a policy using REINFORCE



Optional output:

Detection **instance** [start, end]

Output:

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]

- Accuracy
- Efficiency
- Interpretability

[Yeung, Russakovsky, Mori, Fei-Fei. "End-to-end learning of action detection from frame glimpses in videos." CVPR 2016]

☑ Accuracy

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**

☐ *Efficiency*

☐ *Interpretability*

[Yeung, Russakovsky, Mori, Fei-Fei. "End-to-end learning of action detection from frame glimpses in videos." CVPR 2016]

☑ Accuracy

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**

☐ *Interpretability*

☑ Efficiency

Glimpse only 2% of video frames

[Yeung, Russakovsky, Mori, Fei-Fei. "End-to-end learning of action detection from frame glimpses in videos." CVPR 2016]

Accuracy

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**

Efficiency

Glimpse only 2% of video frames

Samping	Detection AP at IOU 0.5
Uniform	9.3
Our glimpses	**17.1**

Interpretability

[Yeung, Russakovsky, Mori, Fei-Fei. "End-to-end learning of action detection from frame glimpses in videos." CVPR 2016]

☑ Accuracy

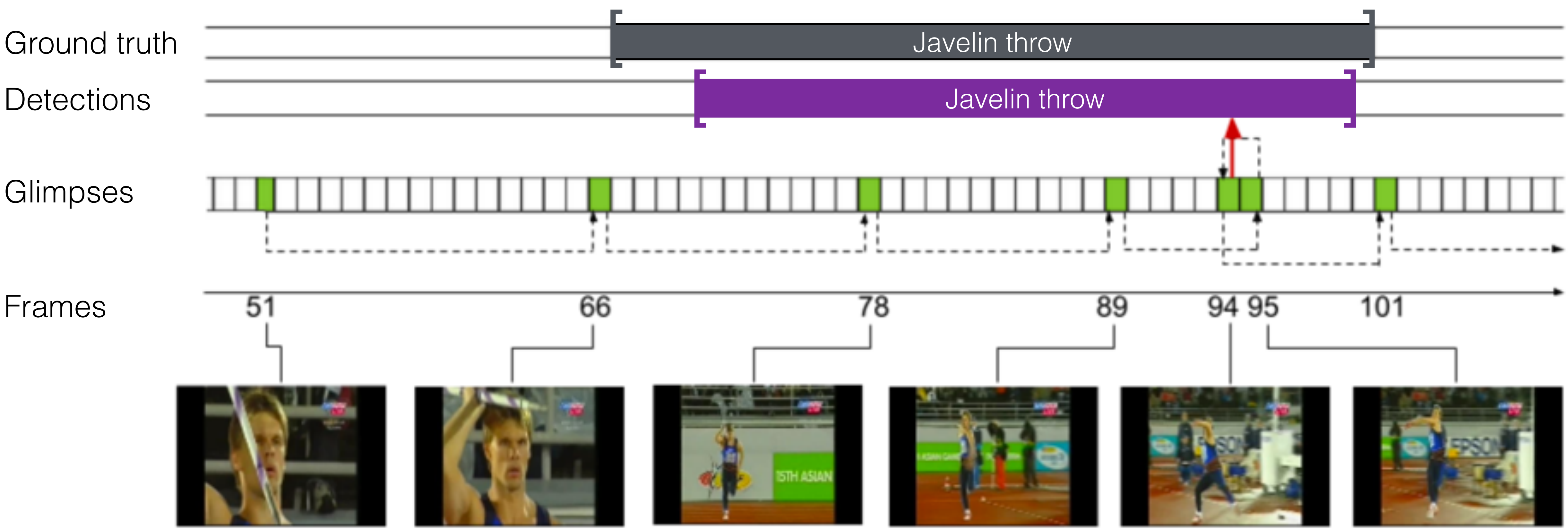
Dataset	Detection AP at IOU 0.5	
	State-of-the-art	Our result
THUMOS 2014	14.4	**17.1**
ActivityNet sports	33.2	**36.7**
ActivityNet work	31.1	**39.9**

☑ Efficiency

Glimpse only 2% of video frames

Samping	Detection AP at IOU 0.5
Uniform	9.3
Our glimpses	**17.1**

☑ Interpretability



[Yeung, Russakovsky, Mori, Fei-Fei. "End-to-end learning of action detection from frame glimpses in videos." CVPR 2016]

Challenges of videos @ scale

Modeling

Capture temporal cues using a Kalman filter

- Competitive with two-stream without optical flow
- Simplifies learning by decorrelating the input

[Dave, Russakovsky, Ramanan. CVPR 2017]

Inference

Focus computation on a small subset of key frames

- Only looks at 2% of frames while maintaining accuracy
- Uses RL to learn where to look and when to output

[Yeung, Russakovsky, Mori, Fei-Fei. CVPR 2016]

Learning

Learn new concepts cheaply and while embracing ambiguity

Challenges of videos @ scale

Modeling

Capture temporal cues using a Kalman filter

- Competitive with two-stream without optical flow
- Simplifies learning by decorrelating the input

[Dave, Russakovsky, Ramanan. CVPR 2017]

Inference

Focus computation on a small subset of key frames

- Only looks at 2% of frames while maintaining accuracy
- Uses RL to learn where to look and when to output

[Yeung, Russakovsky, Mori, Fei-Fei. CVPR 2016]

Learning

Learn new concepts cheaply and while embracing ambiguity

Labeling videos is expensive

- Takes significantly longer to label a video than an image
- Temporal bounds even more expensive — and ambiguous
- How can we practically learn about new concepts in video?

Instructions

Below is a link to a video of one or two people, please watch each video and answer the questions.

- This HIT contains multiple videos, each followed by few questions. *The number of videos and questions is balanced such that the task should take **3 minutes.***
- Make sure you ***fully and carefully watch each*** video so you **do not miss anything**. ***This is important.***
- It is possible that many of the actions in this HIT do not match. It is important to verify an action is indeed ***not*** present in the video.
- **Check all that apply! If there is any doubt, check it anyway for good measure.**
- **Read each and every question carefully. Do not take shortcuts, it will cause you to miss something.**

- [x] Check here if **someone is *Taking a picture of something*** in the video
- [x] Check here if someone is **interacting with *cup/glass/bottle*** in the video

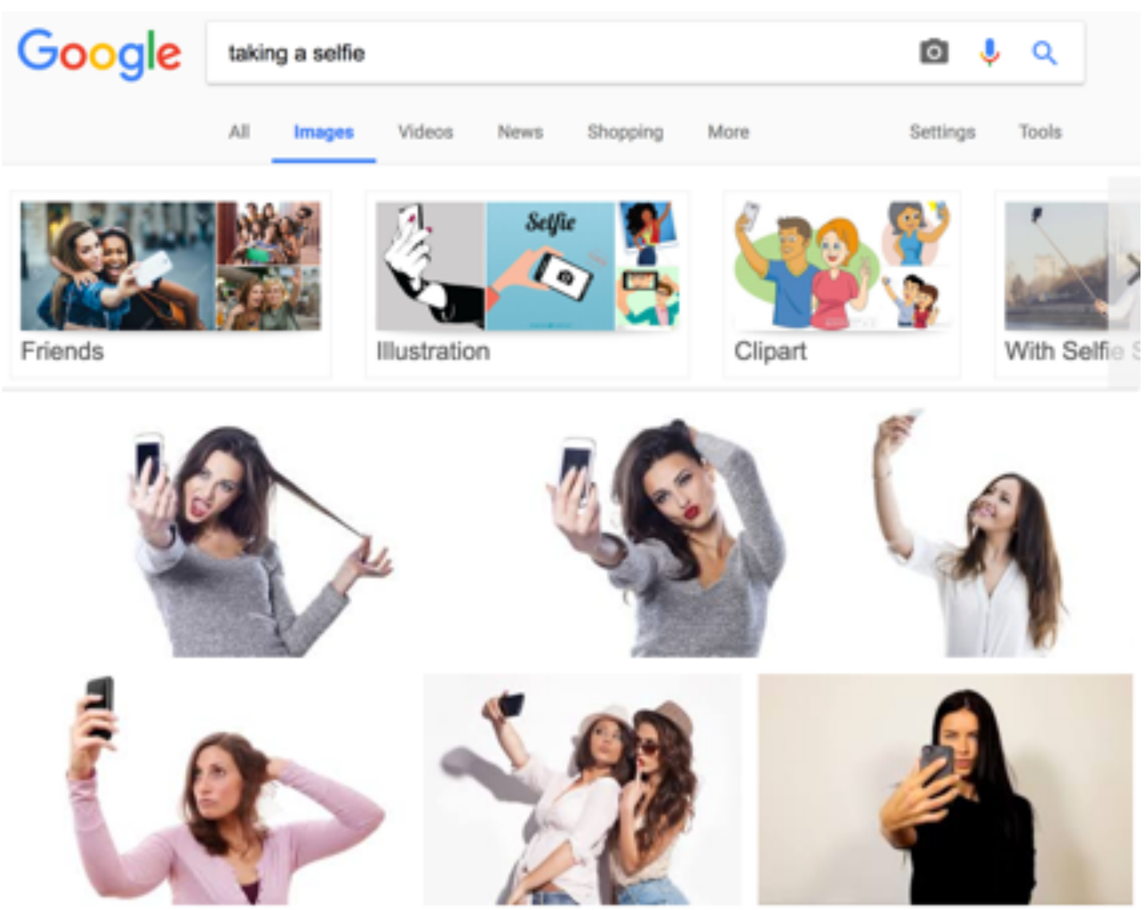
If checked, how? (**Select all that apply**. Use ctrl or cmd to select multiple):

Drinking from a cup/glass/bottle
Holding a cup/glass/bottle of something
Pouring something into a cup/glass/bottle
Putting a cup/glass/bottle somewhere
Taking a cup/glass/bottle from somewhere
Washing a cup/glass/bottle
Other

- [ ] Check here if someone is **interacting with *laptop*** in the video
- [ ] Check here if someone is **interacting with *doorknob*** in the video
- [ ] Check here if someone is **interacting with *table*** in the video
- [ ] Check here if someone is **interacting with *broom*** in the video
- [ ] Check here if someone is **interacting with *picture*** in the video

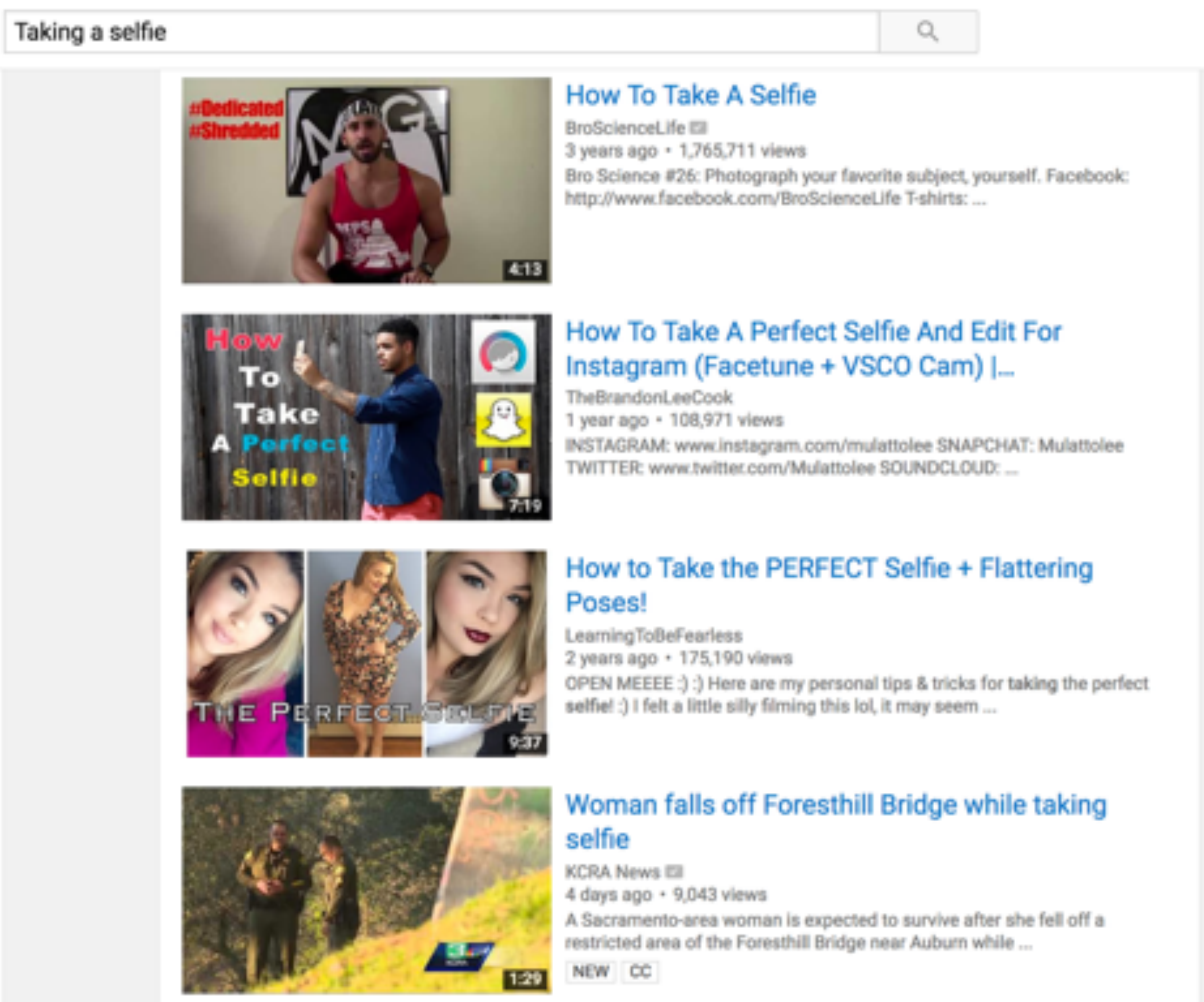
[Sigurdsson, Russakovsky, Farhadi, Laptev, Gupta. "Much Ado About Time: Exhaustive Annotation of Temporal Data." HCOMP 2016]

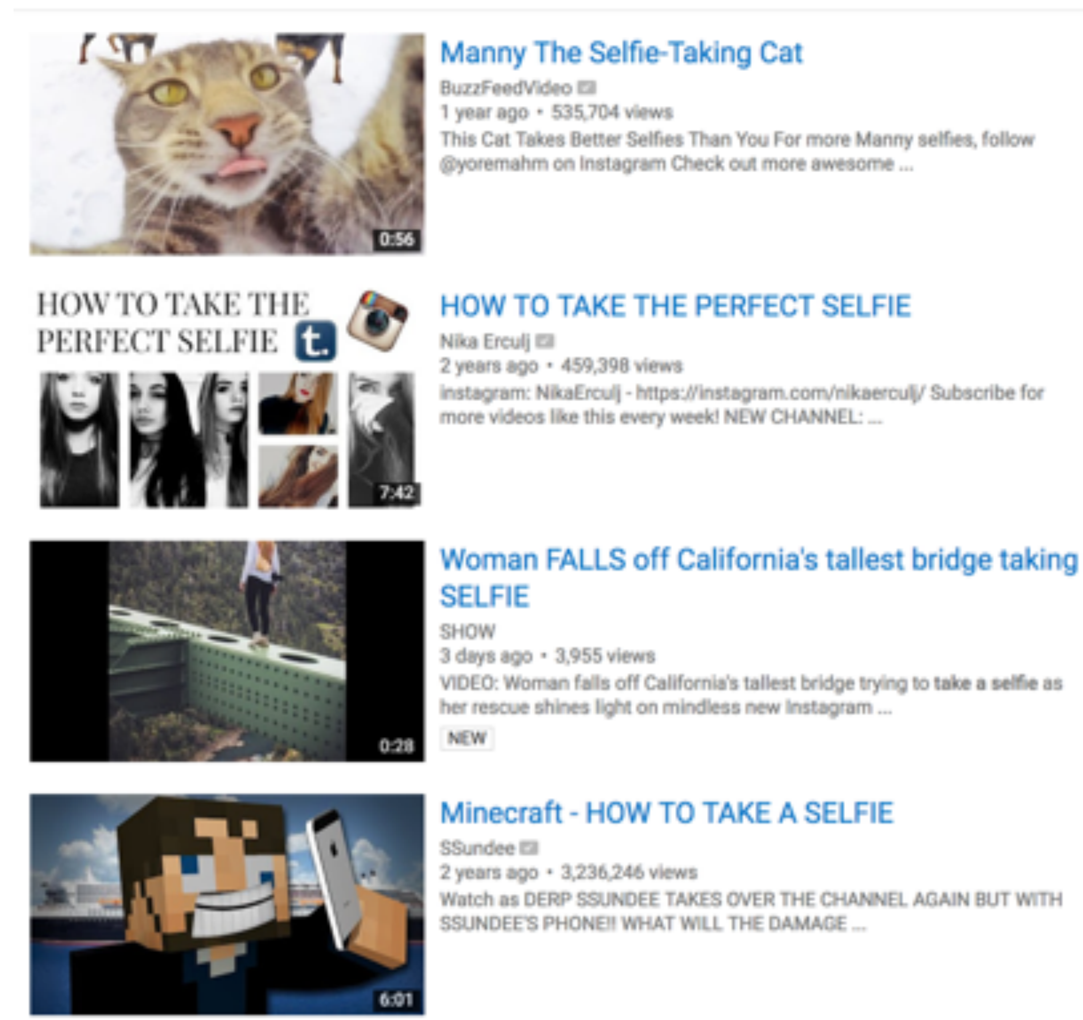
Learning new concepts from <u>image</u> search



Reasonably clean

Learning new concepts from video search





Very very noisy

Balancing diversity vs. semantic drift

- Want diverse training examples
- But too much diversity can also lead to semantic drift

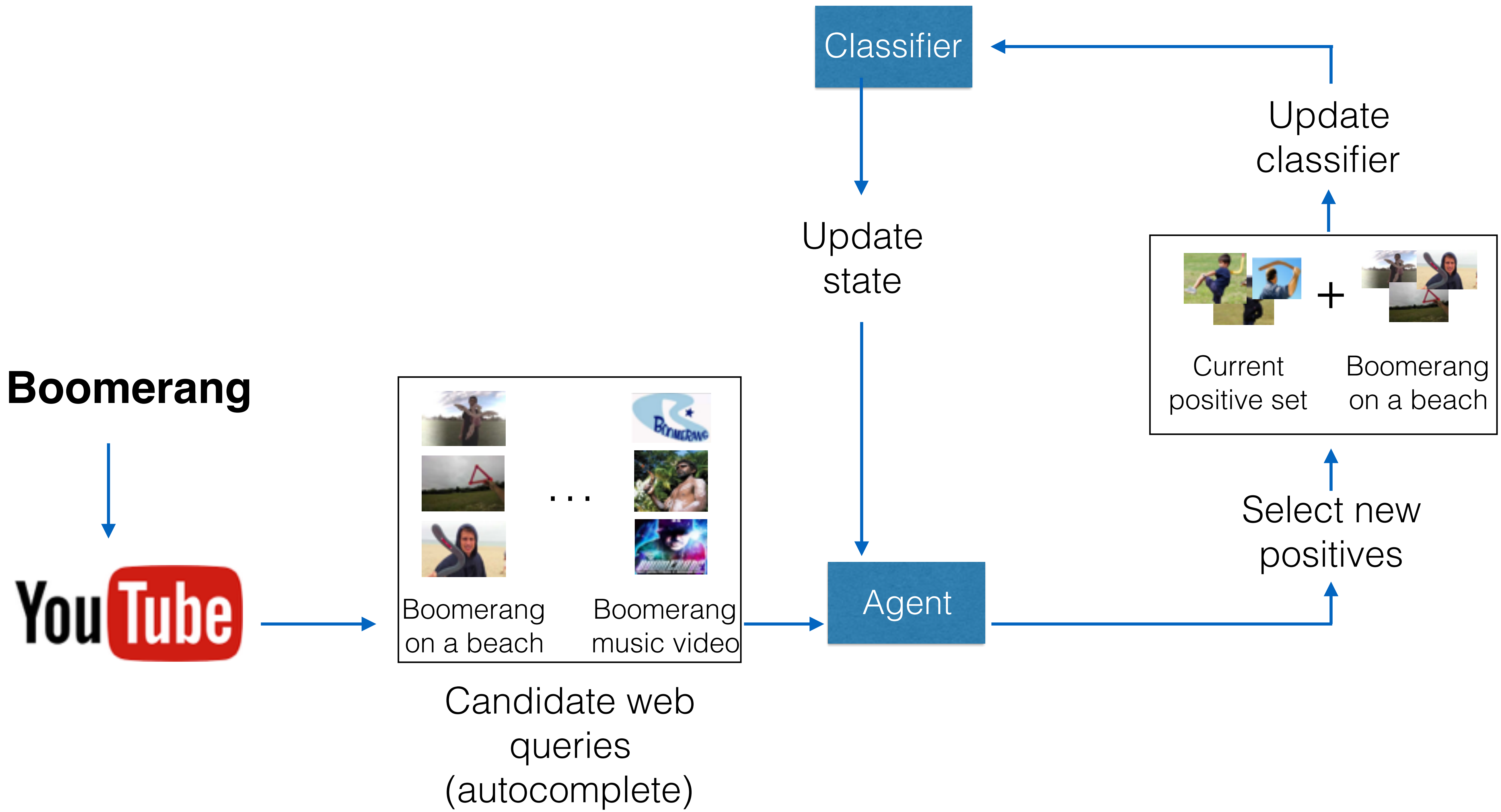
[Yeung, Ramanathan, Russakovsky, Shen, Mori, Fei-Fei. "Learning to learn from noisy web videos." CVPR 2017]

Prior approaches

- **NEIL** [Chen et al. 2013, Chen et al. 2015] incorporate learned relationships between objects
- **OPTIMOL** [Li et al. 2007] uses rule-based heuristics (e.g. entropy)
- **Semi-supervised approaches** (e.g. [Joachims et al. 1999], [Zhu et al. 2002], [Zhou et al. 2004]) optimize globally over a fixed-size dataset

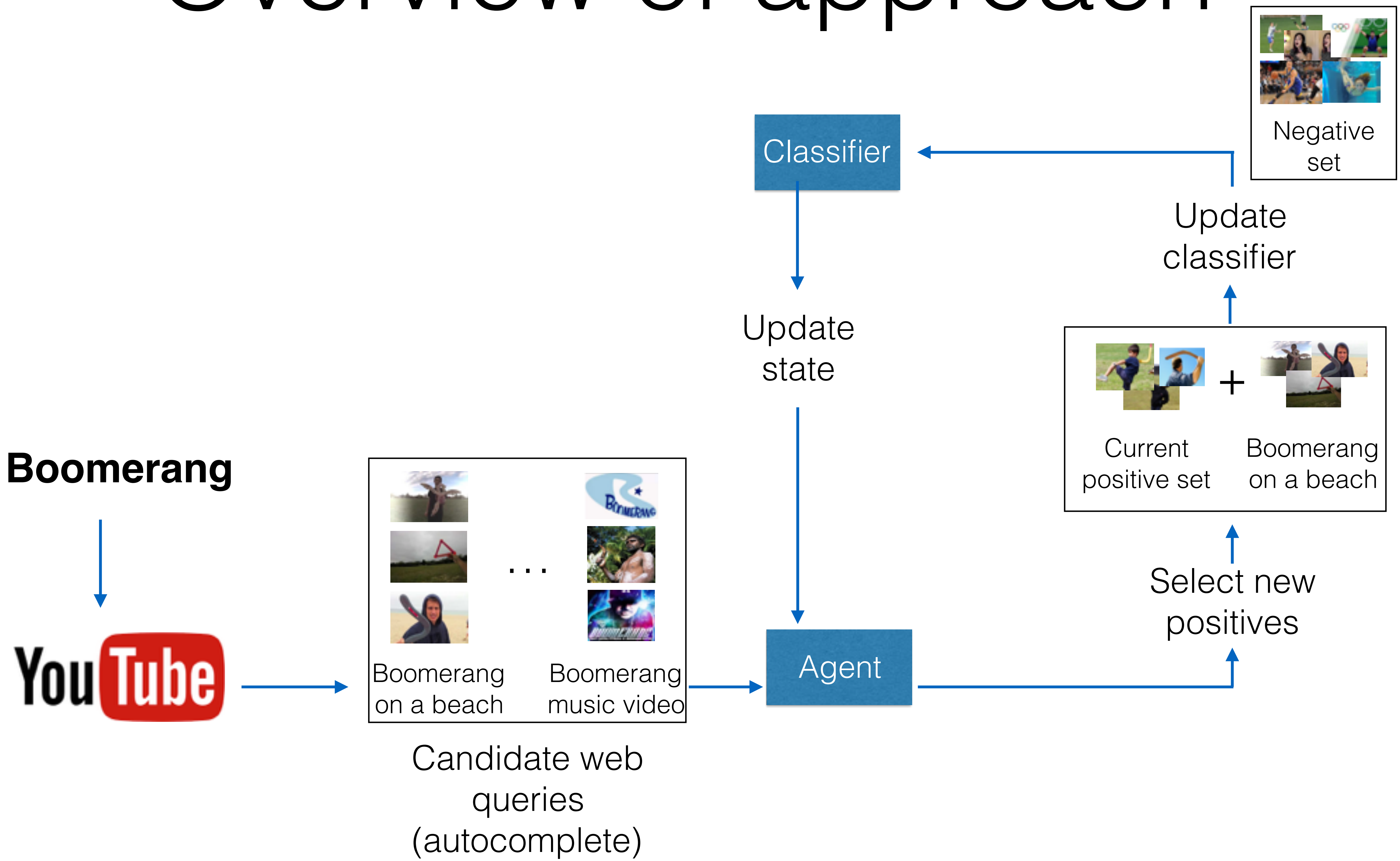
[Yeung, Ramanathan, Russakovsky, Shen, Mori, Fei-Fei. "Learning to learn from noisy web videos." CVPR 2017]

Overview of approach



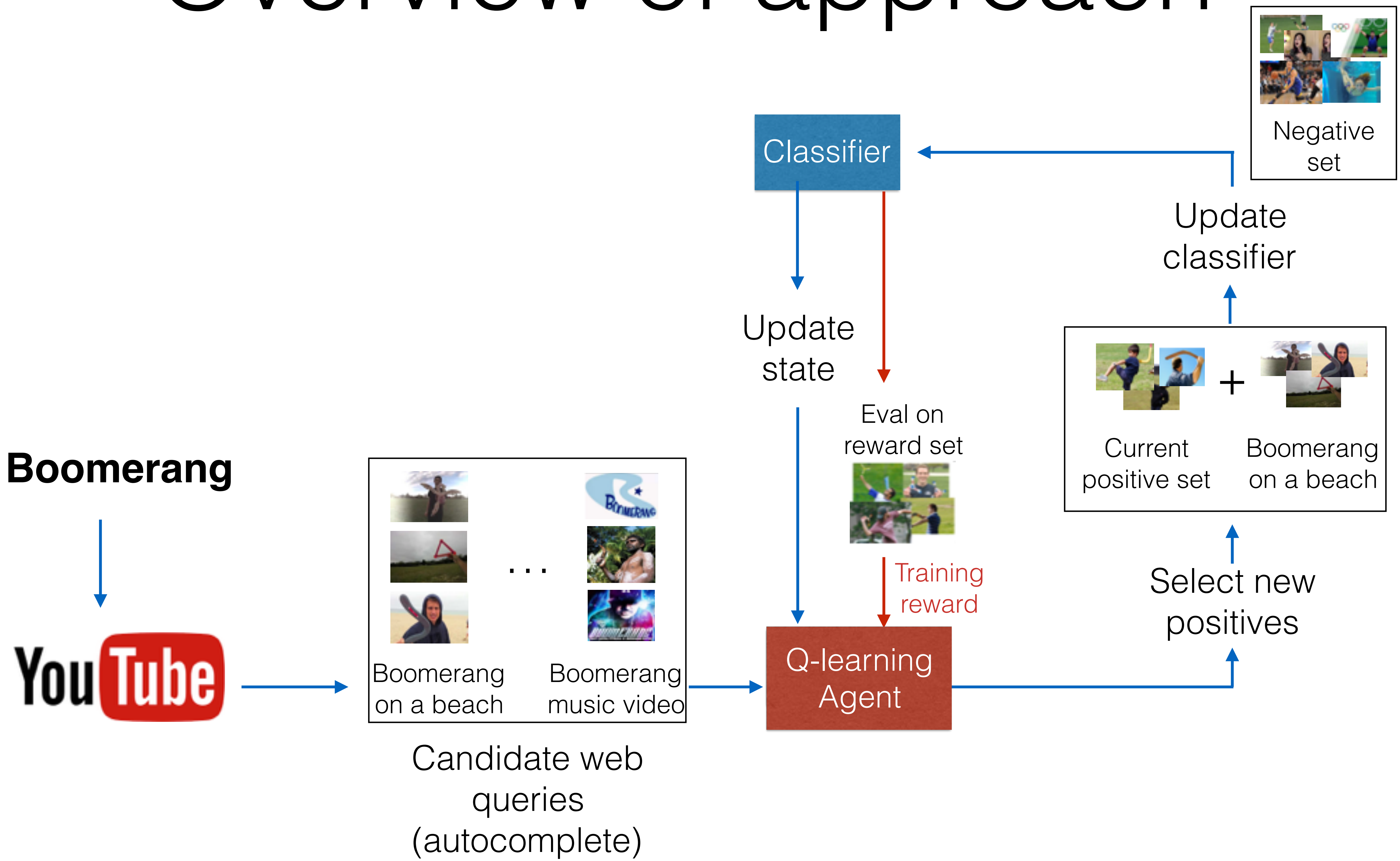
[Yeung, Ramanathan, Russakovsky, Shen, Mori, Fei-Fei. “Learning to learn from noisy web videos.” CVPR 2017]

Overview of approach



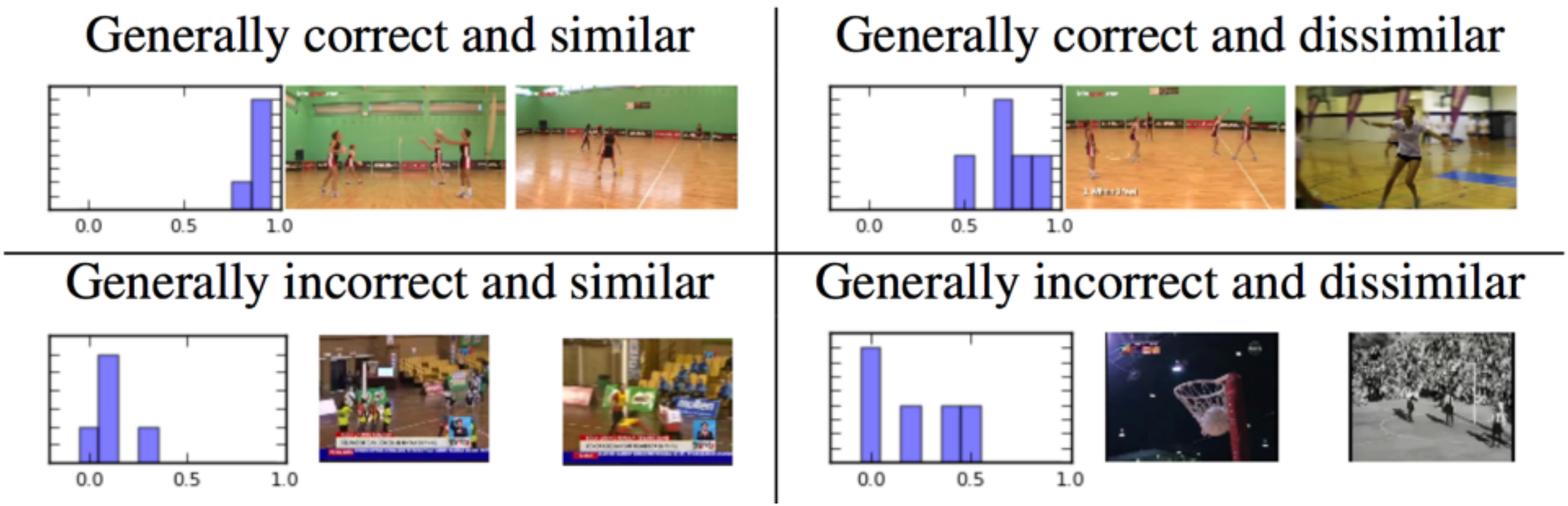
[Yeung, Ramanathan, Russakovsky, Shen, Mori, Fei-Fei. "Learning to learn from noisy web videos." CVPR 2017]

Overview of approach



[Yeung, Ramanathan, Russakovsky, Shen, Mori, Fei-Fei. “Learning to learn from noisy web videos.” CVPR 2017]

Reward incorporates classifier uncertainty



[Yeung, Ramanathan, Russakovsky, Shen, Mori, Fei-Fei. “Learning to learn from noisy web videos.” CVPR 2017]

Testing on Sports1M

Classes: 300 for training, 105 for testing
Videos: YouTube for training, Sport1M-test for testing

Method	Accuracy
Seed	64.3
Label Propagation [Zhu and Ghahramani. ICML 2002]	67.2
Label Spreading [Zhou et al. NIPS 2004]	67.3
TSVM [Joachims ICML 1999]	72.5
Greedy	74.7
Greedy w/ clusters [ala NEIL & OPTIMOL]	74.3
Greedy w/ KL-divergence	74.7
Ours	**77.0**

[Yeung, Ramanathan, Russakovsky, Shen, Mori, Fei-Fei. "Learning to learn from noisy web videos." CVPR 2017]

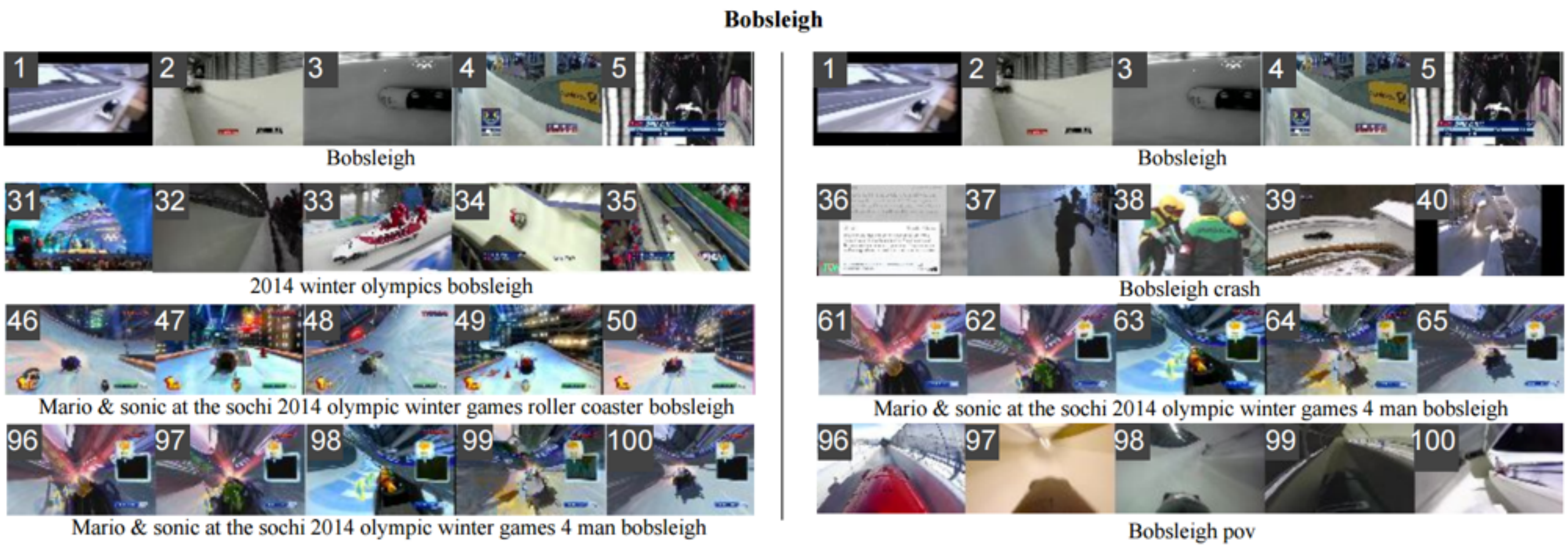
Testing on Sports1M

Greedy classifier

Ours

[Yeung, Ramanathan, Russakovsky, Shen, Mori, Fei-Fei. "Learning to learn from noisy web videos." CVPR 2017]

Testing on Sports1M

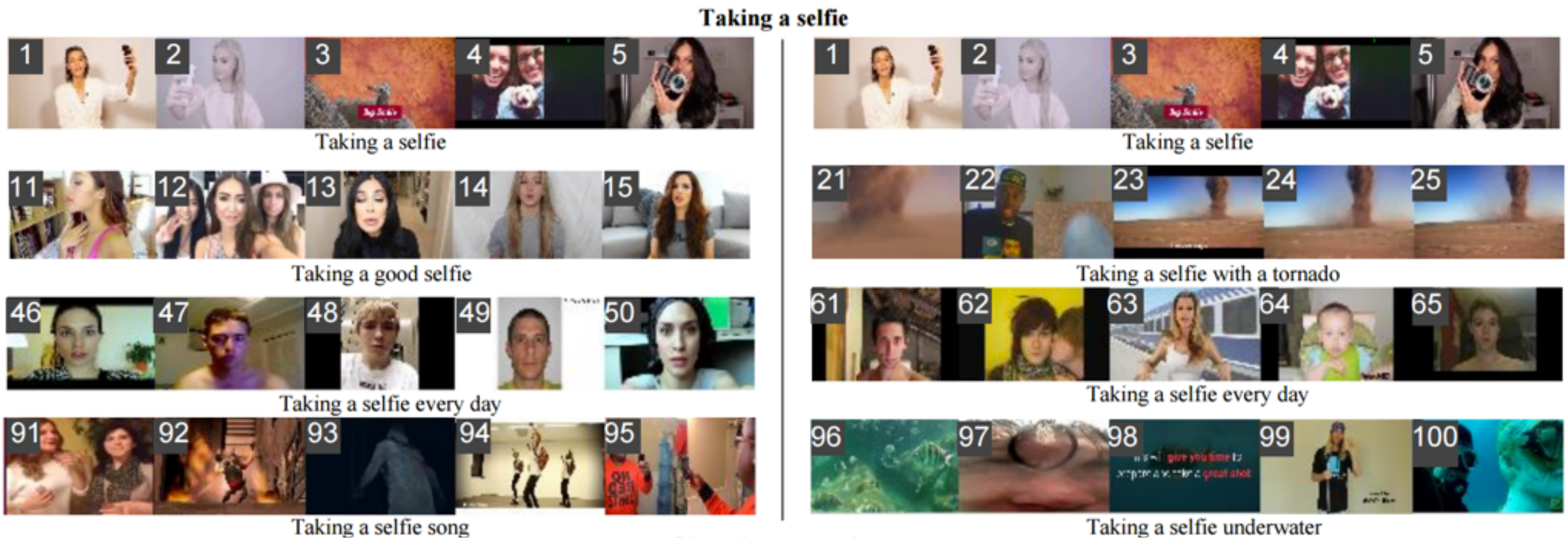


Greedy classifier

Ours

[Yeung, Ramanathan, Russakovsky, Shen, Mori, Fei-Fei. “Learning to learn from noisy web videos.” CVPR 2017]

Novel classes



Greedy classifier

Ours

[Yeung, Ramanathan, Russakovsky, Shen, Mori, Fei-Fei. "Learning to learn from noisy web videos." CVPR 2017]

Challenges of videos @ scale

Modeling

Capture temporal cues using a Kalman filter

- Competitive with two-stream without optical flow
- Simplifies learning by decorrelating the input

[Dave, Russakovsky, Ramanan. CVPR 2017]

Inference

Focus computation on a small subset of key frames

- Only looks at 2% of frames while maintaining accuracy
- Uses RL to learn where to look and when to output

[Yeung, Russakovsky, Mori, Fei-Fei. CVPR 2016]

Learning

Use noisy web search results to learn new concepts

- Determines how to select positive examples with RL
- Avoids expensive annotation

[Yeung, Ramanathan, Russakovsky, Shen, Mori, Fei-Fei. CVPR 2017]

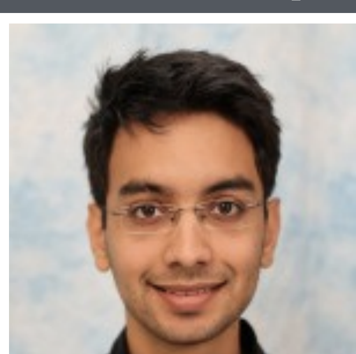
Challenges of videos @ scale

Modeling

Capture temporal cues using a Kalman filter

- Competitive with two-stream without optical flow
- Simplifies learning by decorrelating the input

[Dave, Russakovsky, Ramanan. CVPR 2017]



Inference

Focus computation on a small subset of key frames

- Only looks at 2% of frames while maintaining accuracy
- Uses RL to learn where to look and when to output

[Yeung, Russakovsky, Mori, Fei-Fei. CVPR 2016]

Learning

Use noisy web search results to learn new concepts

- Determines how to select positive examples with RL
- Avoids expensive annotation

[Yeung, Ramanathan, Russakovsky, Shen, Mori, Fei-Fei. CVPR 2017]