

# Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity- Representativeness Reward

Kaiyang Zhou, Yu Qiao, Tao Xiang



中国科学院深圳先进技术研究院  
SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY  
CHINESE ACADEMY OF SCIENCES

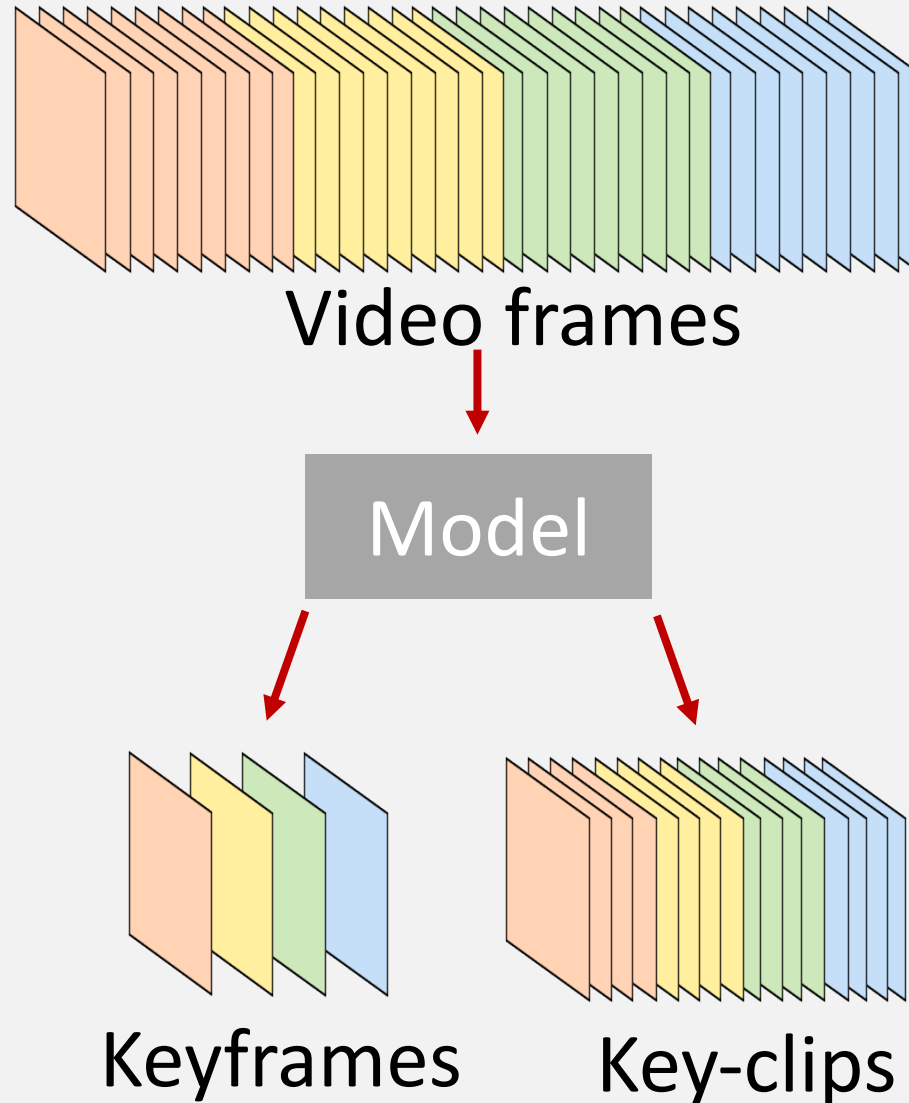


Queen Mary  
University of London

AAAI 2018

# What is video summarization?

**Goal:** to automatically summarize videos into keyframes or key-clips.




We want


- Diverse
- Representative

# Application of video summarization

e.g. YouTube video preview




### Trending




**Melania Trump Gives Her Own State Of The Union**

The Late Show with Stephen C...  
766K views • 17 hours ago



**Michelle Obama Talks with Birthday Girl Ellen About**

TheEllenShow ✓  
1M views • 12 hours ago



**Bruno M Finesse**

Vevo ✓  
13M view

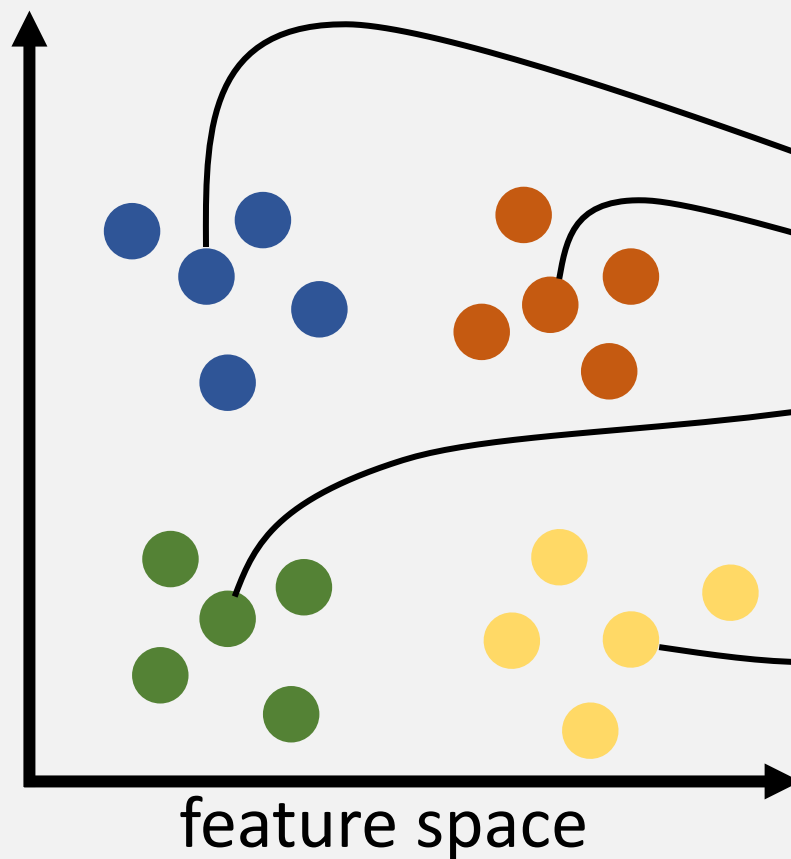
# Unsupervised video summarization

Idea: to analyze correlations between frames in feature space

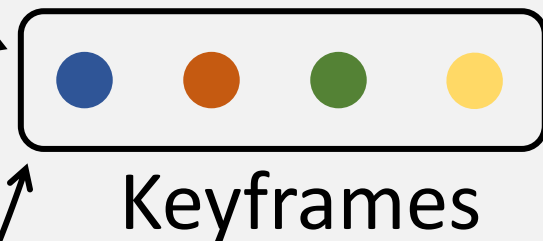
1. Feature extraction



2. Clustering



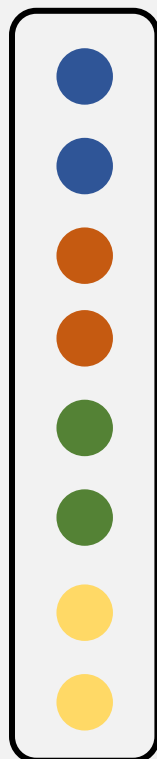
3. Keyframes extraction



# Supervised video summarization

Idea: to exploit human labels

scores:  $y = \{0.1, 0.8, 1.0, 0.2, \dots\}$  keyframes:  $y = \{0, 1, 1, 0, \dots\}$



Training

$$\longrightarrow \text{loss} = (y - w^T X)^2$$

Inference

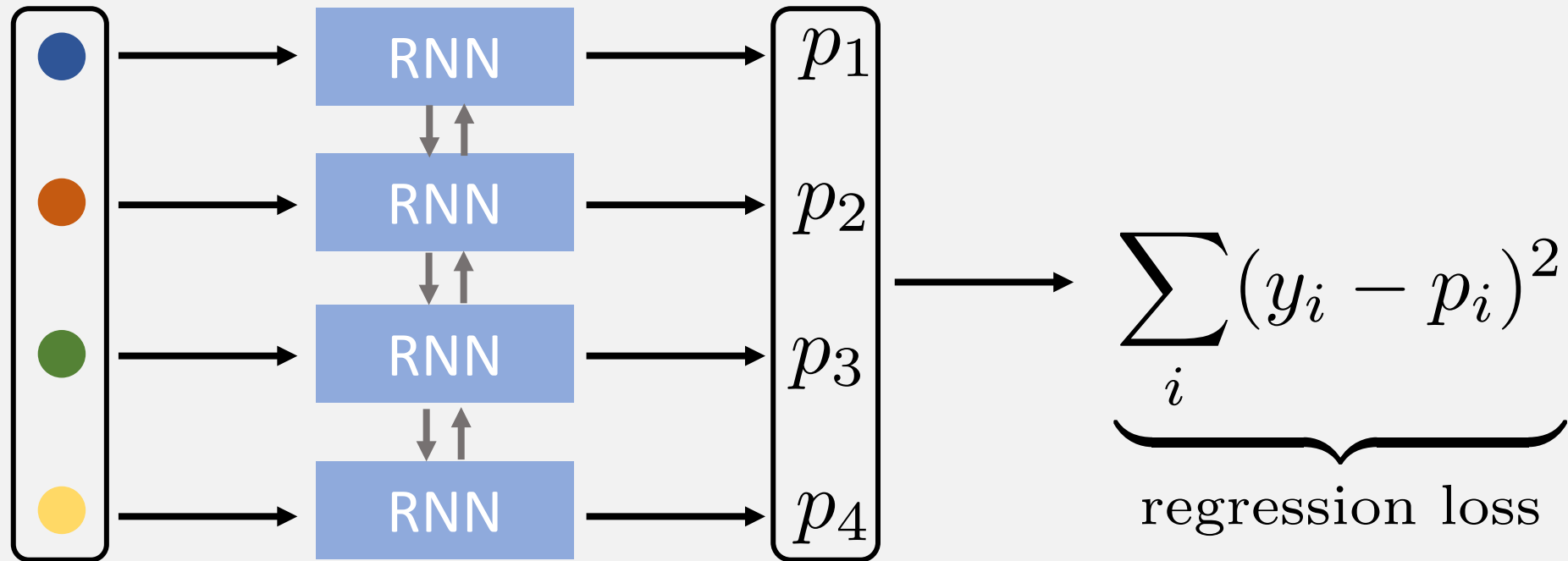
$$\longrightarrow p = w^T X'$$

feature vectors

- Temporal relations are hard to capture by linear models.

# Recurrent neural network with supervised learning

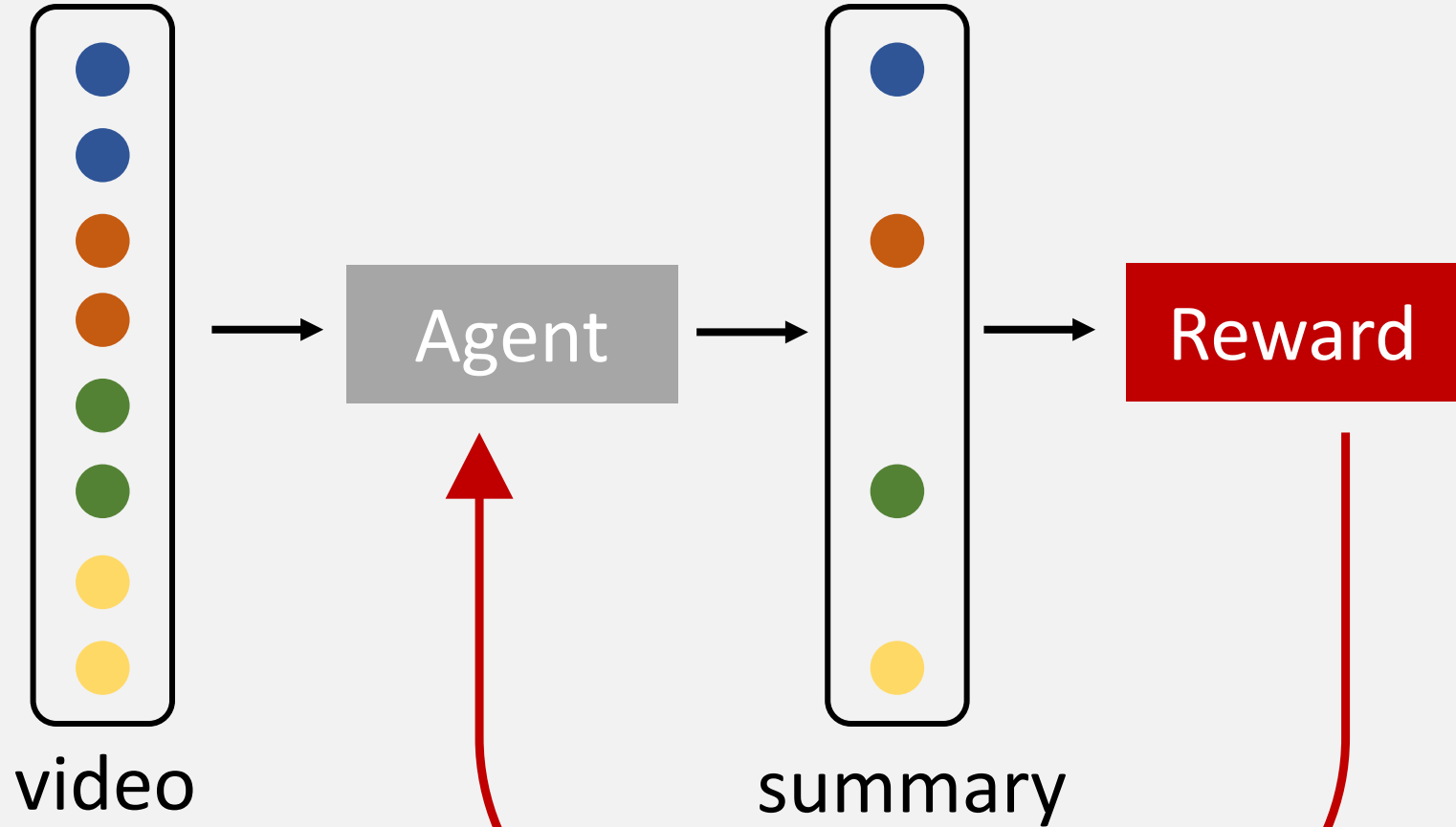
Idea: use RNN to capture temporal relations



- Collecting labels here is much more expensive than that of other tasks.
- Labels may not provide good supervision signals. (b/c labels are subjective)

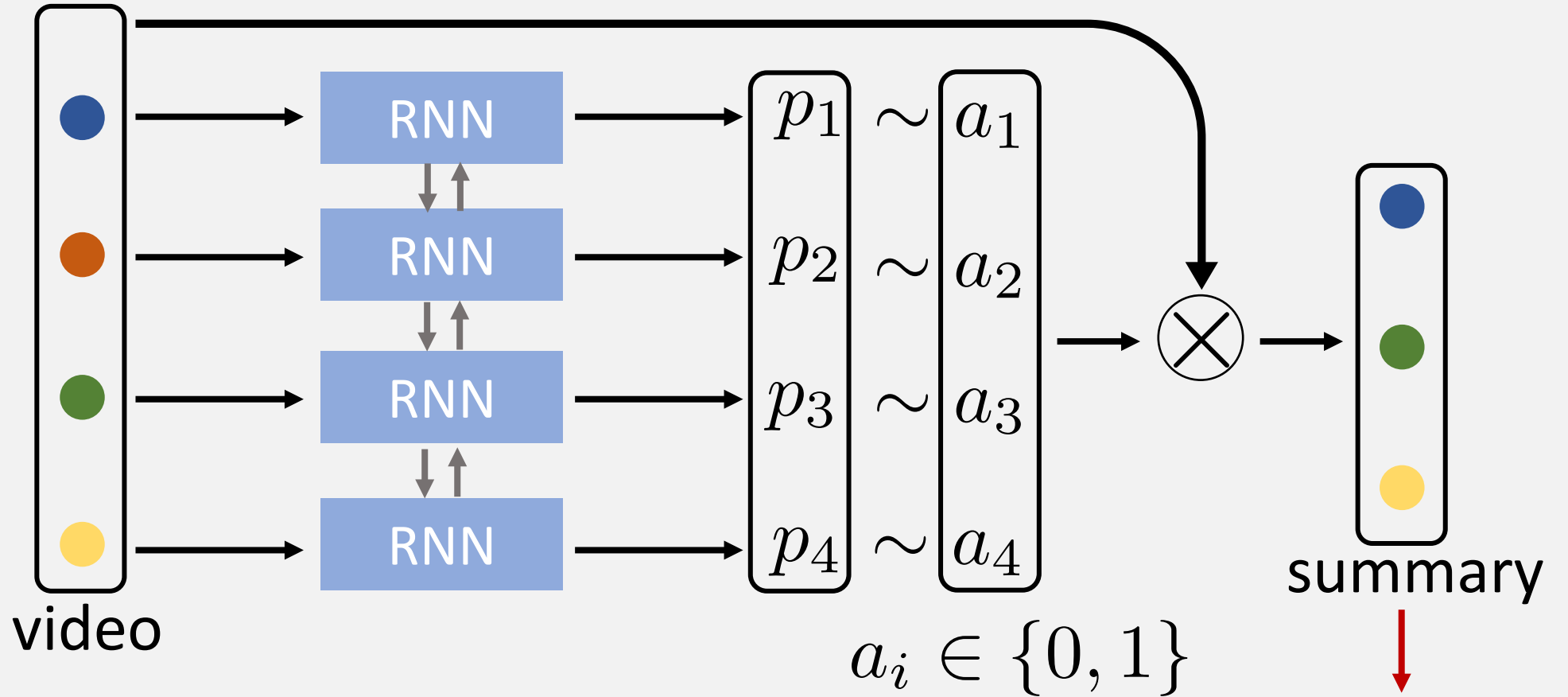
# Main idea

To mimic how humans summarize videos



Is summary diverse and representative?

# Model



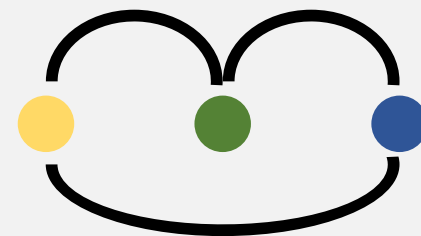
Diversity-representativeness reward



## Diversity reward

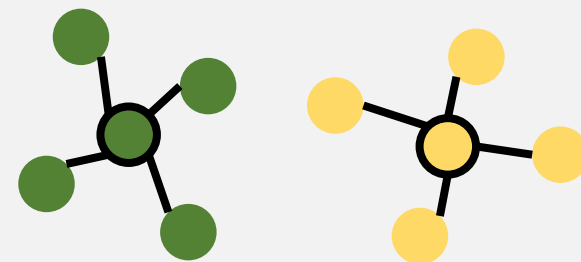
$$R_{\text{div}} = \frac{1}{|\mathcal{Y}|(|\mathcal{Y}|-1)} \sum_{t \in \mathcal{Y}} \sum_{\substack{t' \in \mathcal{Y} \\ t' \neq t}} d(x_t, x_{t'})$$

*Set of selected frames*



## Representativeness reward

$$R_{\text{rep}} = \exp\left(-\frac{1}{T} \sum_{t=1}^T \min_{t' \in \mathcal{Y}} \|x_t - x_{t'}\|_2\right)$$



# Optimization

Reward:

$$R = R_{\text{div}} + R_{\text{rep}}$$

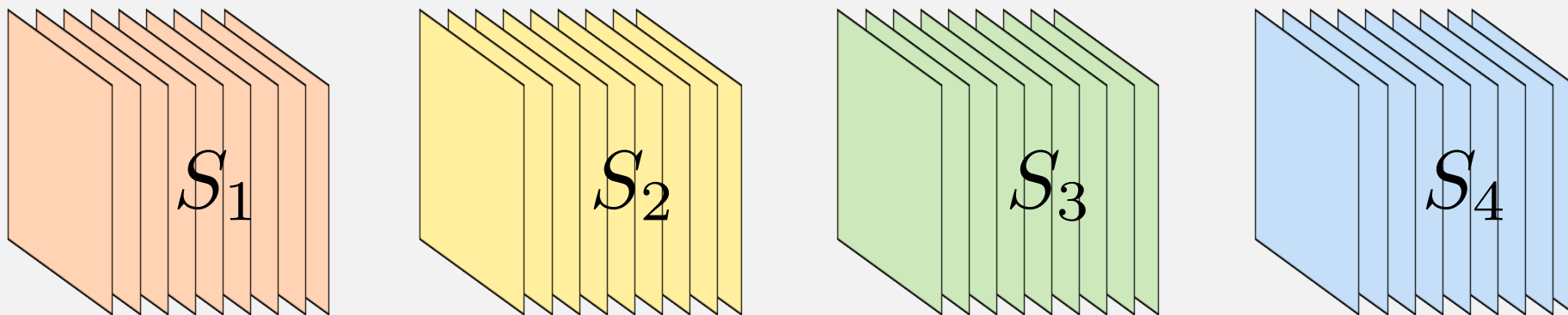
Objective function:

$$J(\theta) = \mathbb{E}[R]$$

Approximate gradients via REINFORCE:

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{n=1}^N \sum_{t=1}^T (R_n - b) \nabla_{\theta} \log \pi_{\theta}(a_t | h_t)$$

# Inference



Score prediction:

$$\{p_i\}_{i=1}^T = \text{RNN}(\{x_i\}_{i=1}^T)$$

Compute clip-level scores:

$$I(S_k) = \frac{1}{|S_k|} \sum_{i \in S_k} p_i$$

Select clips (0/1 Knapsack):

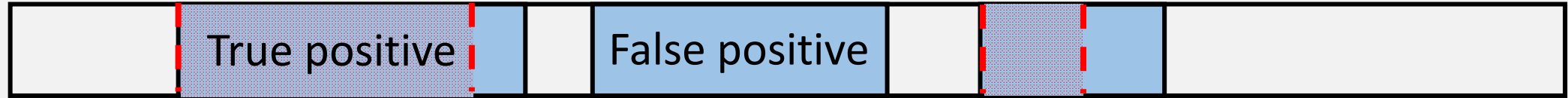
$$\arg \max_{\mu} \sum_k \mu_k I(S_k), \quad \sum_k \mu_k |S_k| \leq \gamma, \quad \mu_k \in \{0, 1\}$$

# Evaluation

Human summary



Machine summary



Metric:  $F\text{-score} = (2 \times \text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$

Dataset	# videos	Length (mins)	Description	# annotators per video
SumMe	25	1-6	User videos	15-18
TVSum	50	2-10	YouTube videos	20

# Quantitative Results

Table: Comparison with other unsupervised approaches.

Method	SumMe (%)	TVSum (%)
Video-MMR	26.6	-
Uniform sampling	29.3	15.5
K-medoids	33.4	28.8
Vsumm	33.7	-
Web image	-	36.0
Dictionary selection	37.8	42.0
Online sparse coding	-	46.0
Co-archetypal	-	50.0
GAN <sub>dpp</sub>	39.1	51.7
Ours	41.4	57.6

} ↑ 6%

} ↑ 11%

# Quantitative Results

Table: Comparison with other supervised approaches.

Method	SumMe (%)	TVSum (%)
Interestingness	39.4	-
Submodularity	39.7	-
Summary transfer	40.9	-
Bi-LSTM	37.6	54.2
DPP-LSTM	38.6	54.7
GAN <sub>sup</sub>	<b>41.7</b>	56.3
Ours	41.4	<b>57.6</b>

} ↑ 2%

# Quantitative Results

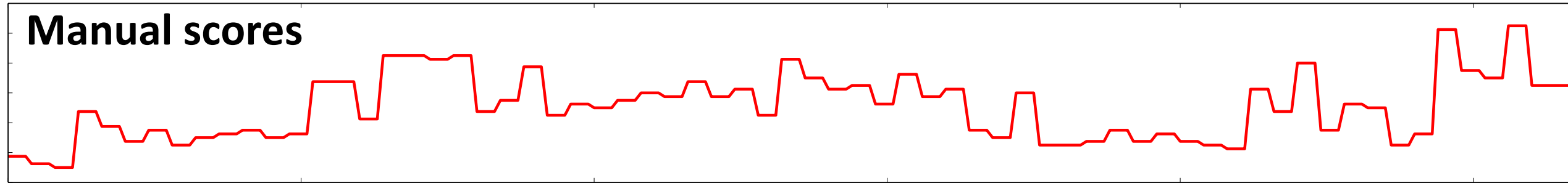
Table: Comparison with other supervised approaches.

Method	SumMe (%)	TVSum (%)
Interestingness	39.4	-
Submodularity	39.7	-
Summary transfer	40.9	-
Bi-LSTM	37.6	54.2
DPP-LSTM	38.6	54.7
GAN <sub>sup</sub>	41.7	56.3
Ours	41.4	57.6
Ours (supervised)	<b>42.1</b>	<b>58.1</b>

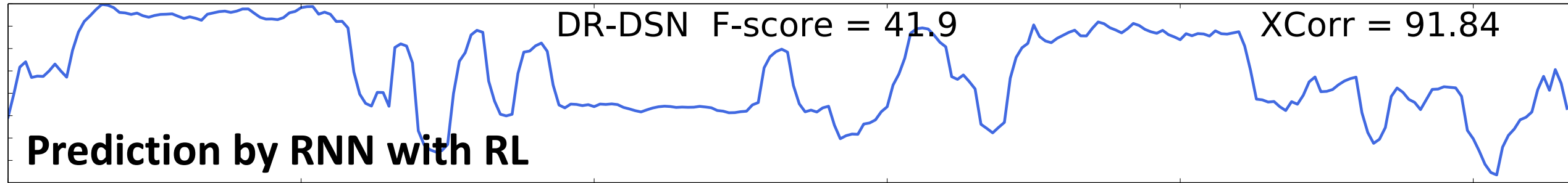
*For more experiments and details, please see our paper.*

# Qualitative Results

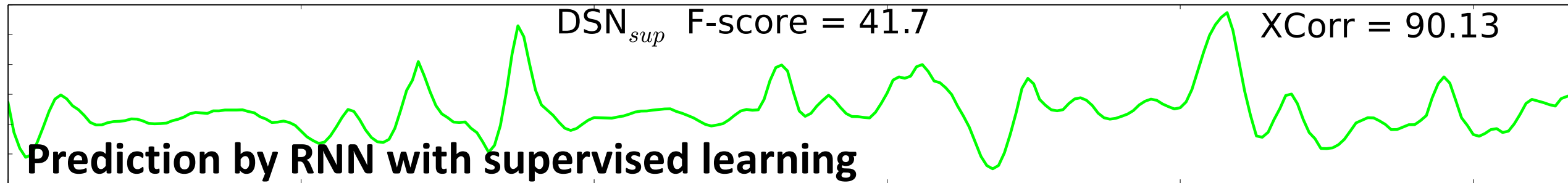
**Manual scores**



**Prediction by RNN with RL**



**Prediction by RNN with supervised learning**



Video #10 in TVSUM



# Summary

1. Proposed a label-free reward, i.e. diversity-representativeness reward.
2. Outperformed/competitive to other unsupervised/supervised ones.
3. Extended the unsupervised method to the supervised version.

Thanks!

Any questions?

please feel free to contact me at: [k.zhou@qmul.ac.uk](mailto:k.zhou@qmul.ac.uk)