



# Learning Transferable Self-attentive Representations for Action Recognition in Untrimmed Videos with Weak Supervision

Xiao-Yu Zhang<sup>1</sup> Haichao Shi<sup>1,2,4</sup> Changsheng Li<sup>3,4</sup>

Kai Zheng<sup>3,4</sup> Xiaobin Zhu<sup>5</sup> Lixin Duan<sup>3,4</sup>

<sup>1</sup>*Institute of Information Engineering, Chinese Academy of Sciences*

<sup>2</sup>*School of Cyber Security, University of Chinese Academy of Sciences*

<sup>3</sup>*University of Electronic Science and Technology of China*

<sup>4</sup>*Youedata Co., Ltd., Beijing*

<sup>5</sup>*Beijing Technology and Business University*



Presenter: Haichao Shi

# Outline

---

1. Introduction
2. Our Method (TSRNet)
  - Two-Stream Feature Extraction
  - Self-attentive Action Classification
  - Knowledge Transfer
  - Temporal Action Detection
3. Evaluation
4. Conclusion

# Action Recognition in Videos

---

## ■ Videos

- **Trimmed**: fully semantic annotations (*UCF101, HMDB51, etc.*)
- **Untrimmed**: typically long, may contain multiple activities, difficult for temporal annotation (*THUMOS, ActivityNet, etc.*)

## ■ Opportunities

- **Videos** provide huge and rich data for visual learning
- **Action** is important in motion perception and has many applications

## ■ Challenges

- Temporal models and representations
- High computational and memory cost
- Noisy and weakly labels

# Motivation

---

## ■ Existing Methods on **Weakly Supervised Action Detection**

- **UntrimmedNets [1]** utilizes a soft selection module for untrimmed video classification along with activity localization.
- **STPN [2]** utilizes a sparsity constraint to detect the activities.
- **W-TALC [3]** improve the localization results by optimizing two complimentary loss functions.

## ■ Limitations

- Limited training videos.
- Difficult to learn the specific high-level features for untrimmed videos.
- External background information affect the model performance greatly.

[1] Limin Wang et al. UntrimmedNets for Weakly Supervised Action Recognition and Detection, in CVPR 2017

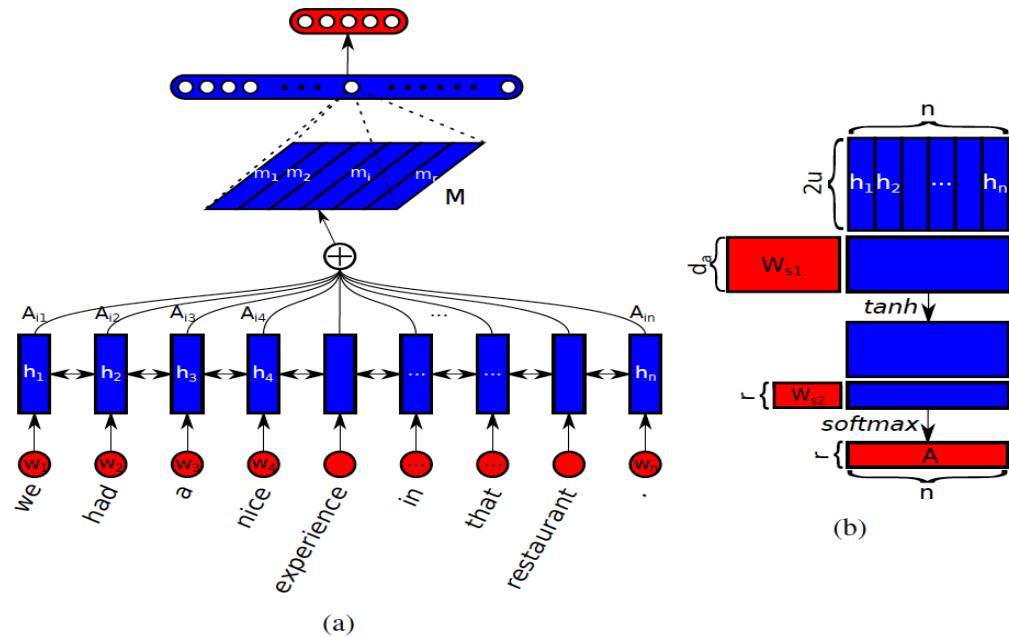
[2] Phuc Nguyen et al. Weakly Supervised Action Localization by Sparse Temporal Pooling Network, in CVPR 2018

[3] Sujoy Paul et al. W-TALC: Weakly-supervised Temporal Activity Localization and Classification, in ECCV 2018

# Inspiration

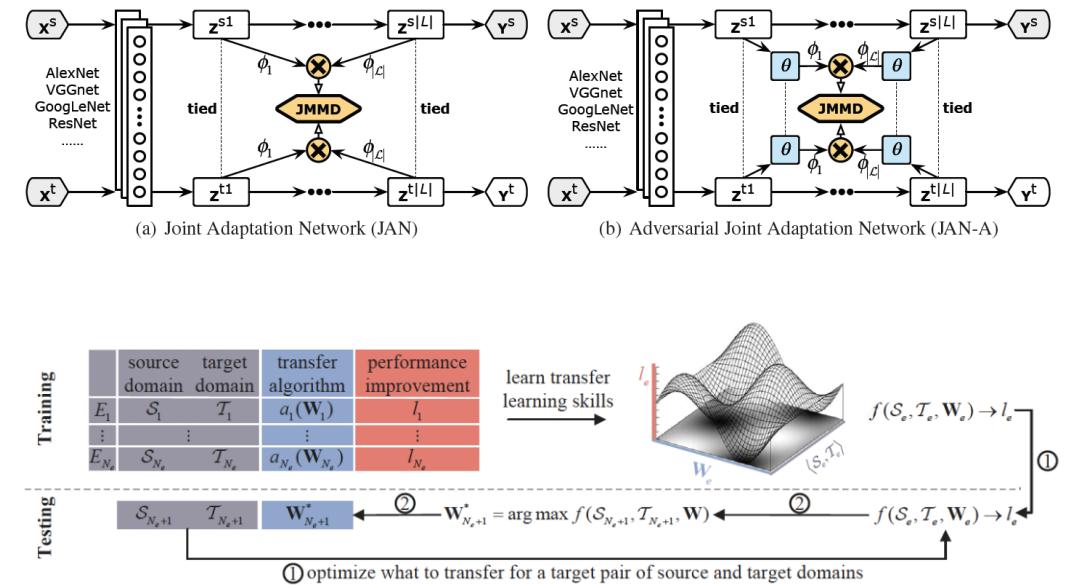
## ■ Self-Attention Mechanism

- Intra-domain exploring



## ■ Transfer Learning

- Inter-domain exploring



[1] Ashish Vaswani et al. Attention Is All You Need, in NIPS 2017

[2] Zhouhan Lin et al. A Structured Self-attentive Sentence Embedding, in ICLR 2017

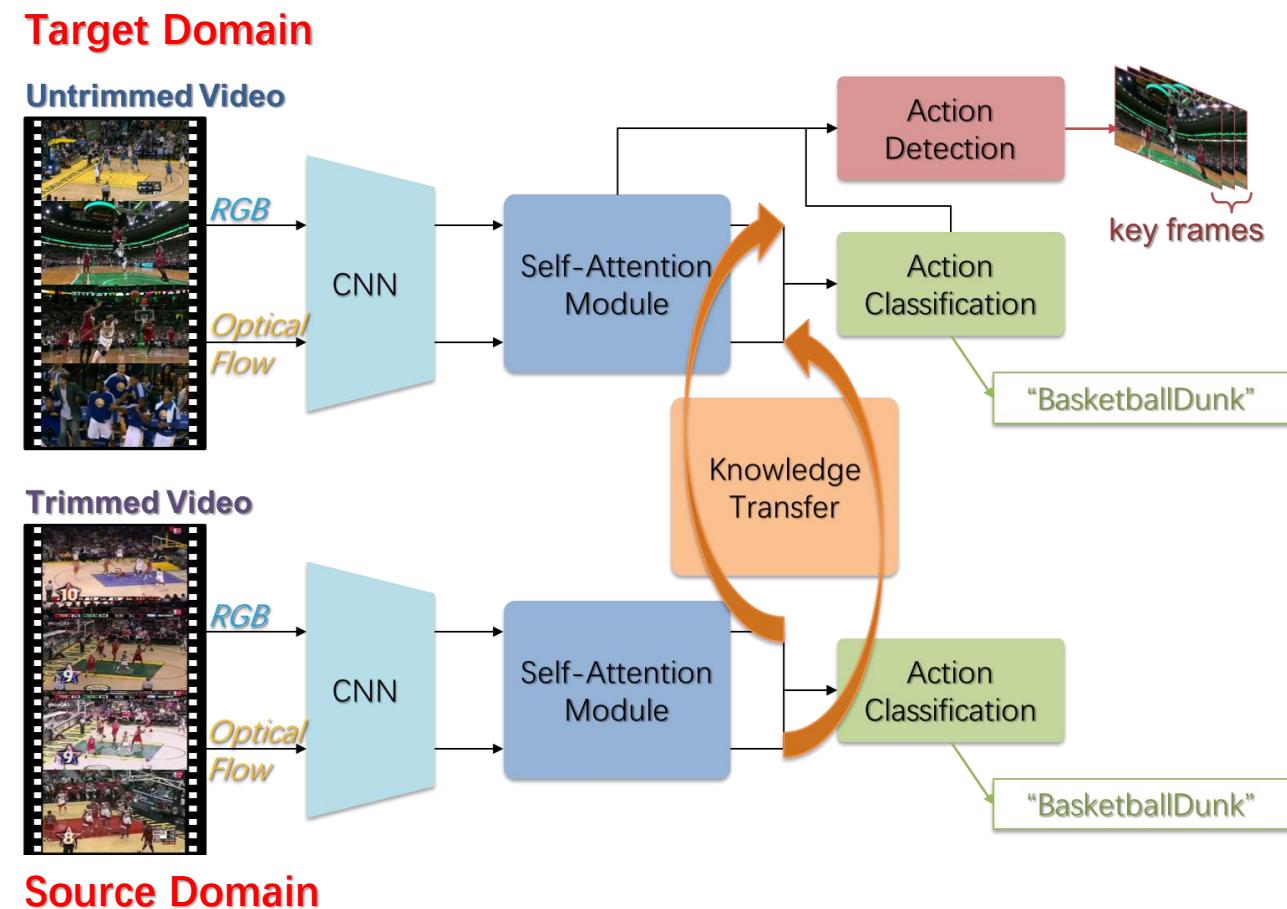
[3] Mingsheng Long et al. Deep Transfer Learning with Joint Adaptation Networks, in ICML 2017

[4] Ying Wei et al. Transfer Learning via Learning to Transfer, in ICML 2018

# Our Method (TSRNet)

- TSRNet: Transferable Self-attentive Representation learning based deep neural Network

- **Self-Attention Module:** capture domain-specific properties
- **Transfer Module:** capture general properties shared by domains



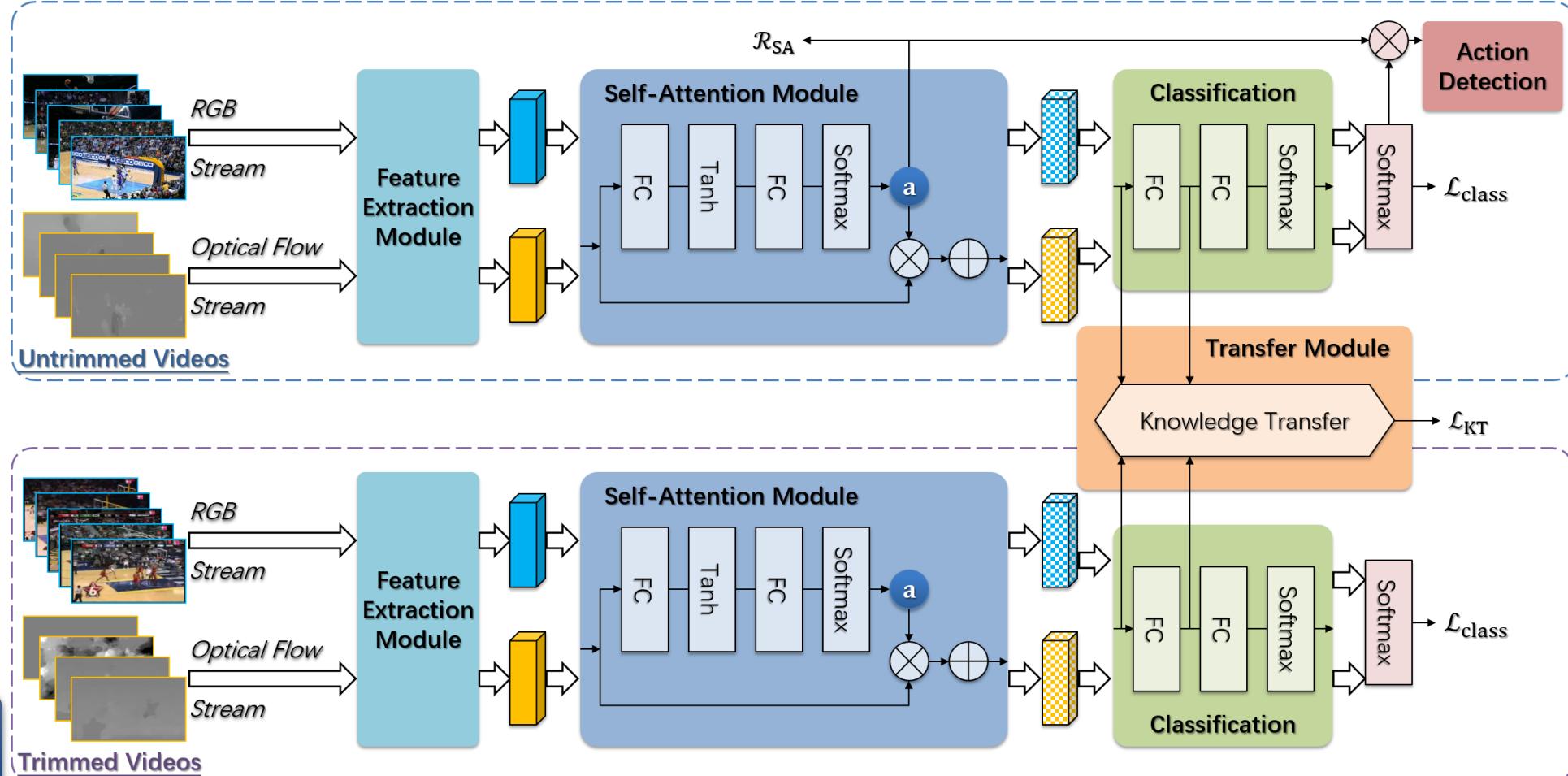
# Framework

Two-stream feature extraction

Self-attentive action classification

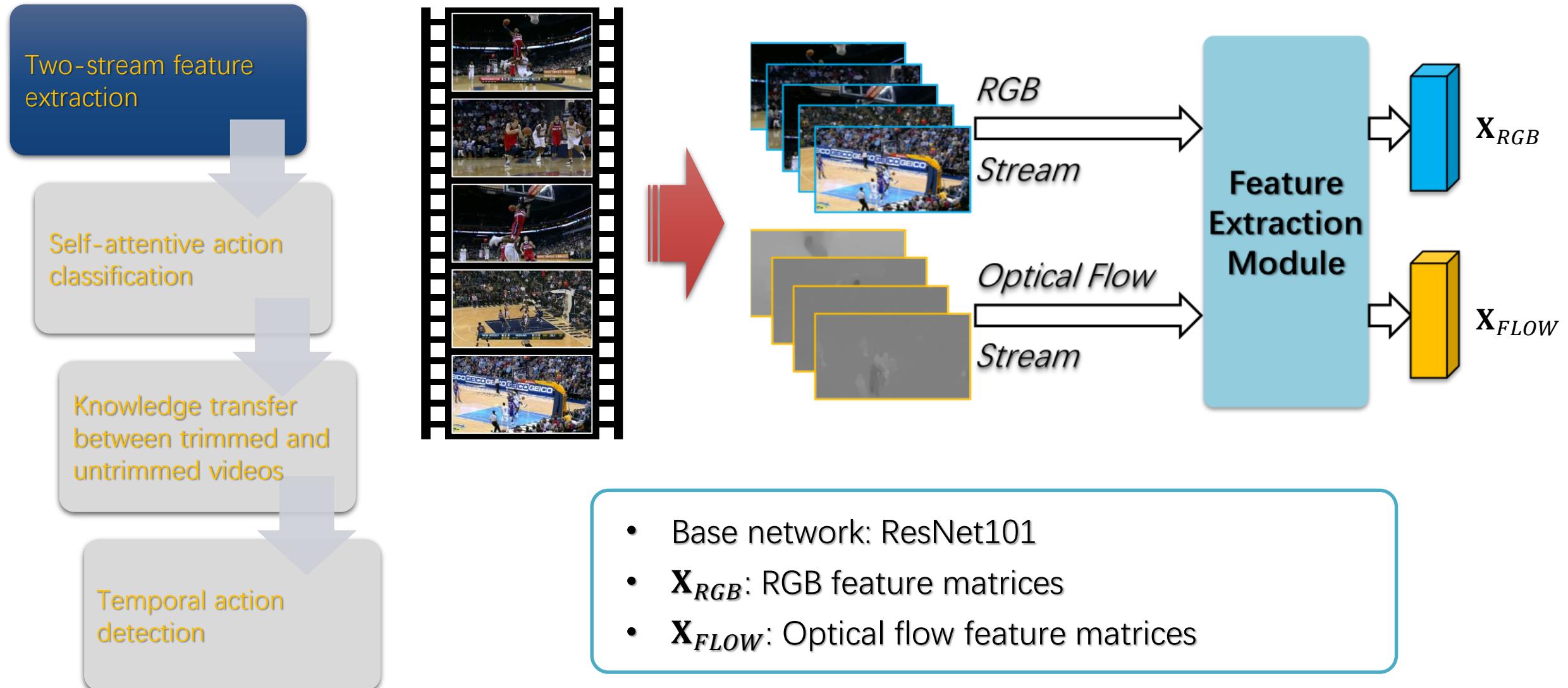
Knowledge transfer between trimmed and untrimmed videos

Temporal action detection



Overall loss:  $\mathcal{L} = \mathcal{L}_{SA} + \mathcal{L}_{KT}$

# Two-Stream Feature Extraction



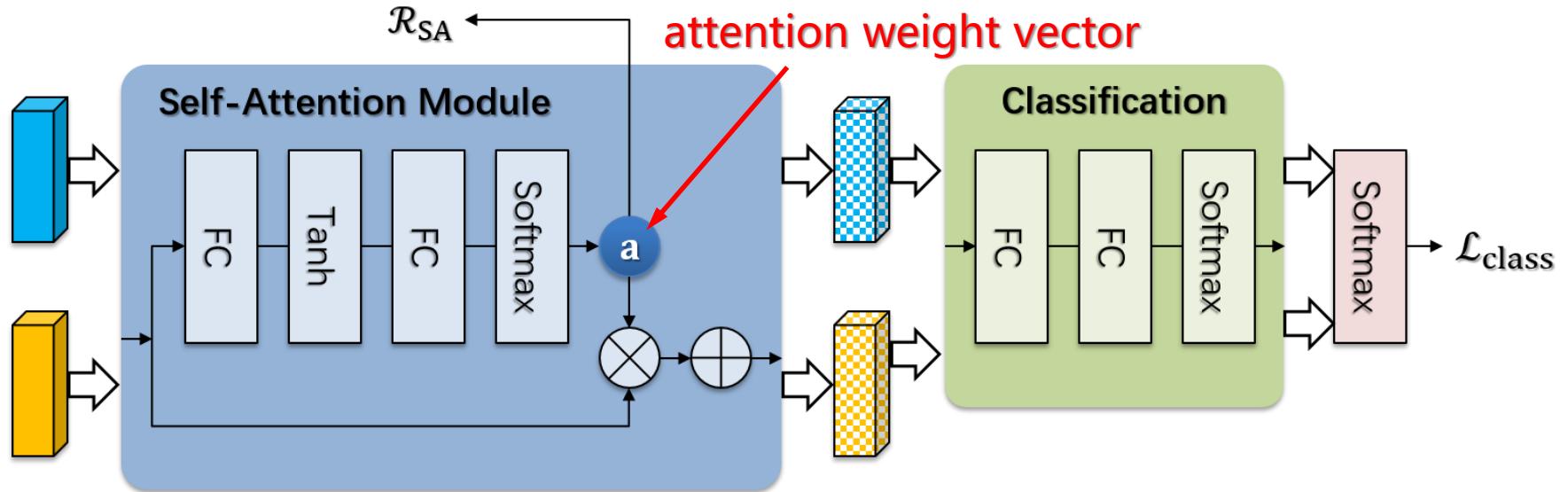
# Self-Attentive Action Classification

Two-stream feature extraction

Self-attentive action classification

Knowledge transfer between trimmed and untrimmed videos

Temporal action detection



- $\mathbf{m} = \mathbf{X}\mathbf{a} = \mathbf{X} \left( \text{softmax}(\mathbf{w}_2 \cdot \tanh(\mathbf{W}_1 \mathbf{X})) \right)^T$
- $\mathcal{L}_{SA} = \mathcal{L}_{class} + \mathcal{R}_{SA}$ 
  - $\mathcal{L}_{class}$ : the standard multi-label cross-entropy loss
  - $\mathcal{R}_{SA} = \alpha \mathcal{R}_{smooth} + \beta \mathcal{R}_{sparsity}$ 
    - $\mathcal{R}_{smooth} = \sum_{i=1}^{n-1} (a_i - a_{i+1})^2$
    - $\mathcal{R}_{sparsity} = \|\mathbf{a}\|_1$

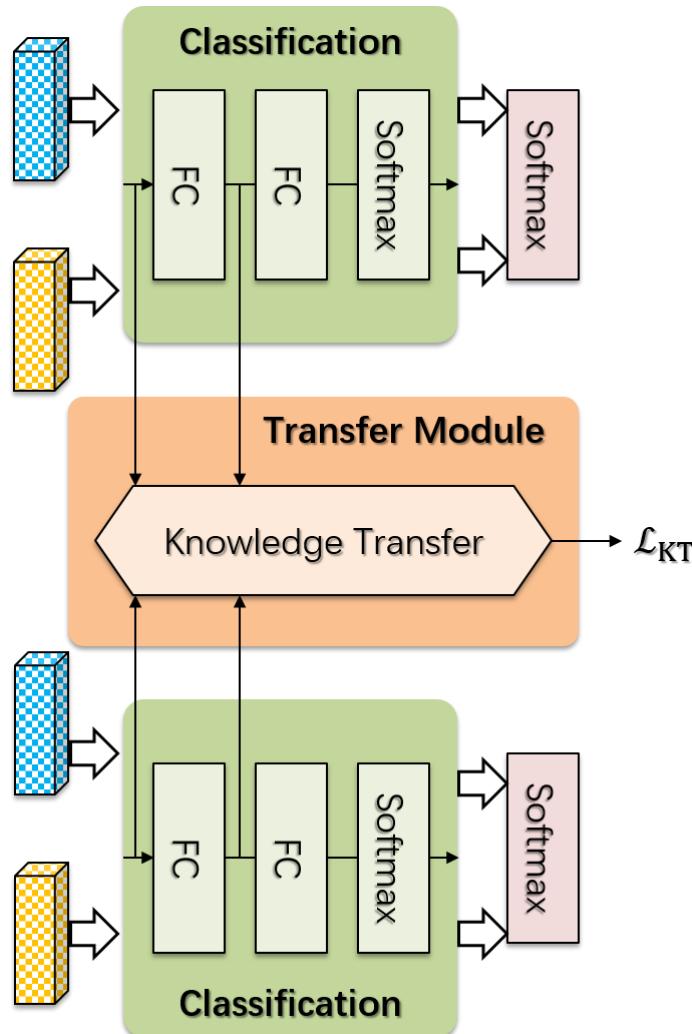
# Knowledge Transfer

Two-stream feature extraction

Self-attentive action classification

Knowledge transfer between trimmed and untrimmed videos

Temporal action detection



- $\mathcal{L}_{KT} = \mathcal{L}_{FC1} + \mathcal{L}_{FC2}$ 
  - $\mathcal{L}_{FC1} = \text{MMD}^2(\mathcal{T}, \mathcal{U}) = \frac{1}{n_T^2} \sum_{i=1}^{n_T} \sum_{j=1}^{n_T} k(\mathbf{t}_i, \mathbf{t}_j) + \frac{1}{n_U^2} \sum_{i=1}^{n_U} \sum_{j=1}^{n_U} k(\mathbf{u}_i, \mathbf{u}_j) - \frac{2}{n_T \cdot n_U} \sum_{i=1}^{n_T} \sum_{j=1}^{n_U} k(\mathbf{t}_i, \mathbf{u}_j)$
  - $\mathcal{L}_{FC2} = \text{MMD}^2(FC1(\mathcal{T}), FC1(\mathcal{U}))$

$\mathcal{T} = \{\mathbf{t}_i | i=1 \dots n_T\}$ : the set of features of trimmed videos  
 $\mathcal{U} = \{\mathbf{u}_i | i=1 \dots n_U\}$ : the set of features of untrimmed videos  
 $k(\cdot, \cdot)$ : the Gaussian kernel function

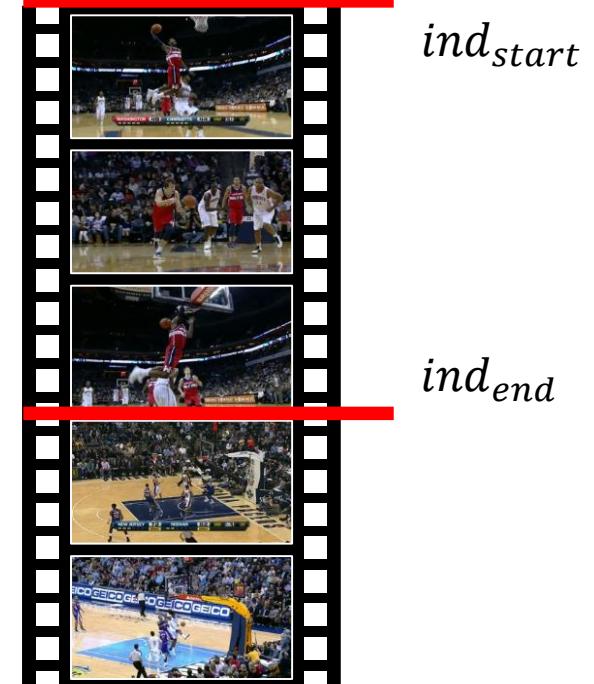
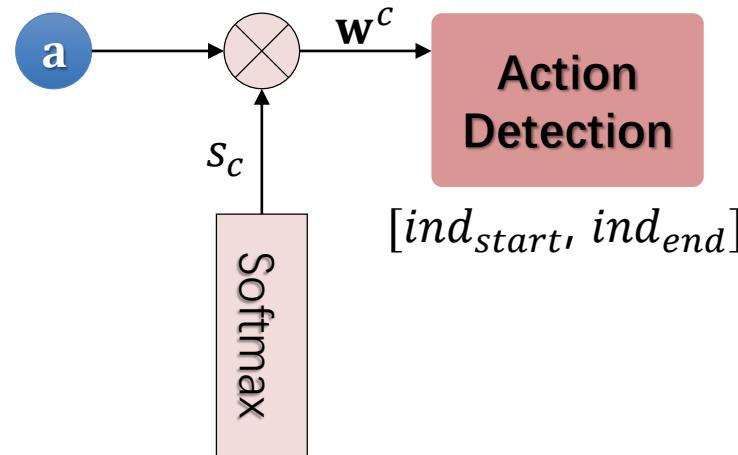
# Temporal Action Detection

Two-stream feature extraction

Self-attentive action classification

Knowledge transfer between trimmed and untrimmed videos

Temporal action detection



- $w_i^c = a_i s_c$
- $\bar{w}_i^c = \theta \cdot w_{i,RGB}^c + (1 - \theta) \cdot w_{i,Flow}^c$
- $t_{start} = \frac{ind_{start}}{F}, t_{end} = \frac{ind_{end}}{F}$

$w_i^c$ : the weighted score of frame  $i$  for class  $c$ .

$s_c = [s_1, s_2, \dots, s_m]^T \in \mathbb{R}^{m \times 1}$ : the output of softmax layer.

$[ind_{start}, ind_{end}]$ : the frame indices of starting and ending positions.

$F$ : the fps (frames per second) of videos.

# Experiments – Settings

---

## ■ Evaluation Datasets

	# Training Data	# Testing Data
THUMOS14	1,010	1,574
ActivityNet1.3	10,024	4,926

## ■ Transfer Dataset

# Overlapping Classes	THUMOS14	ActivityNet1.3
UCF101	20	200

# Experiments – Settings

---

## ■ Implementations Details

Hyper-parameters	Settings
Batch Size	16
Momentum	0.9
Dropout	0.8
Learning Rate	0.0001(RGB) / 0.0005(Optical Flow)
Sampling Rate	30 fps (frames per second)
Decay Rate	Decrease every 5,000 iterations by 10

# Results – Action Classification

## ■ Action recognition on **THUMOS14**

Table 1: Classification accuracy (%) of all the methods on the THUMOS14 dataset for action recognition. Note that SRNet is a simpler version of TSRNet, which excludes the knowledge transfer module.

	RGB	Optical Flow	Fusion
(Wang and Schmid 2013)	-	-	63.1
(Wang et al. 2016)(3 seg)	-	-	78.5
(Wang et al. 2017)	-	-	82.2
Two-Stream	68.2	71.6	73
SRNet	72.3	76.2	79.4
<b>TSRNet</b>	<b>74.4</b>	<b>79.6</b>	<b>87.1</b>

**Two-Stream**: TSRNet w/o (Self-Attention & Knowledge Transfer module)

**SRNet**: TSRNet w/o Knowledge Transfer module

# Results – Action Classification

---

## ■ Action recognition on **ActivityNet1.3**

Table 2: Classification accuracy (%) of all the methods on the ActivityNet1.3 dataset for action recognition. Note that SRNet is a simpler version of TSRNet, which excludes the knowledge transfer module.

	RGB	Optical Flow	Fusion
Two-Stream	71.4	73.5	79.2
SRNet	74.3	80.1	86.9
TSRNet	<b>79.7</b>	<b>84.3</b>	<b>91.2</b>

**Two-Stream**: TSRNet w/o (Self-Attention & Knowledge Transfer module)

**SRNet**: TSRNet w/o Knowledge Transfer module

# Results – Action Detection

## ■ Action detection on **THUMOS14**

Table 3: Comparisons on the THUMOS14 dataset for action detection.

	Method	mAP@IoU (%)								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Full supervision	(Richard and Gall 2016)	39.7	35.7	30.0	23.2	15.2	-	-	-	-
	(Shou, Wang, and Chang 2016)	47.7	43.5	36.3	28.7	19.0	10.3	5.3	-	-
	(Yeung et al. 2016)	48.9	44.0	36.0	26.4	17.1	-	-	-	-
	(Alwassel, Heilbron, and Ghanem 2017)	49.6	44.3	38.1	28.4	19.8	-	-	-	-
	(Lin, Zhao, and Shou 2017)	50.1	47.8	43.0	35.0	24.6	-	-	-	-
	(Yuan et al. 2016)	51.4	42.6	33.6	26.1	18.8	-	-	-	-
	(Shou et al. 2017)	-	-	40.1	29.4	23.3	13.1	7.9	-	-
	(Xu, Das, and Saenko 2017)	54.5	51.5	44.8	35.6	28.9	-	-	-	-
Weak supervision	(Zhao et al. 2017)	<b>66.0</b>	<b>59.4</b>	<b>51.9</b>	<b>41.0</b>	<b>29.8</b>	-	-	-	-
	(Wang et al. 2017)	44.4	37.7	28.2	21.1	13.7	-	-	-	-
	(Singh and Lee 2017)	36.4	27.8	19.5	12.7	6.8	-	-	-	-
	(Nguyen et al. 2017)	52.0	44.7	35.5	25.8	16.9	9.9	4.3	1.2	0.1
	(Nguyen et al. 2017)	45.3	38.8	31.1	23.5	16.2	9.8	5.1	2.0	0.3
	TSRNet (w/o $\mathcal{L}_{FC2}$ )	53.5	45.3	35.9	26.5	17.2	10.4	5.31	1.93	0.21
TSRNet		<b>55.9</b>	<b>46.9</b>	<b>38.3</b>	<b>28.1</b>	<b>18.6</b>	<b>11.0</b>	<b>5.59</b>	<b>2.19</b>	<b>0.29</b>

**TSRNet (w/o  $\mathcal{L}_{FC2}$ )**: TSRNet w/o the 2<sup>nd</sup> Knowledge Transfer

# Results – Action Detection

## ■ Action detection on **ActivityNet1.3**

Table 4: Comparisons on the ActivityNet1.3 dataset for action detection.

	Methods	mAP@IoU (%)			
		0.5	0.75	0.95	Average
Full supervision	(Singh and Cuzzolin 2016)	34.5	-	-	11.3
	(Xu, Das, and Saenko 2017)	26.8	-	-	-
	(Xiong et al. 2017)	29.1	23.5	5.5	-
	(Heilbron et al. 2017)	40.0	17.9	4.7	21.7
	(Shou et al. 2017)	45.3	26.0	0.2	23.8
	(Zhao et al. 2017)	39.12	23.48	5.49	23.98
	(Lin et al. 2018)	<b>52.50</b>	<b>33.53</b>	<b>8.85</b>	<b>33.72</b>
Weak supervision	(Nguyen et al. 2017)	29.3	16.9	2.6	-
	TSRNet (pretrained:[ResNet101@ImageNet])	29.9	17.2	2.71	19.56
	TSRNet (pretrained:[TSRNet@overlap30])	<b>33.1</b>	<b>18.7</b>	<b>3.32</b>	<b>21.78</b>

**TSRNet (pretrained: [ResNet101@ImageNet])**: using ResNet101 pretrained on ImageNet to initialize the feature extraction module of TSRNet

**TSRNet ( pretrain: [TSRNet@overlap30])**: using the overlapping 30 classes between UCF101 and ActivityNet1.3 to initialize the entire TSRNet

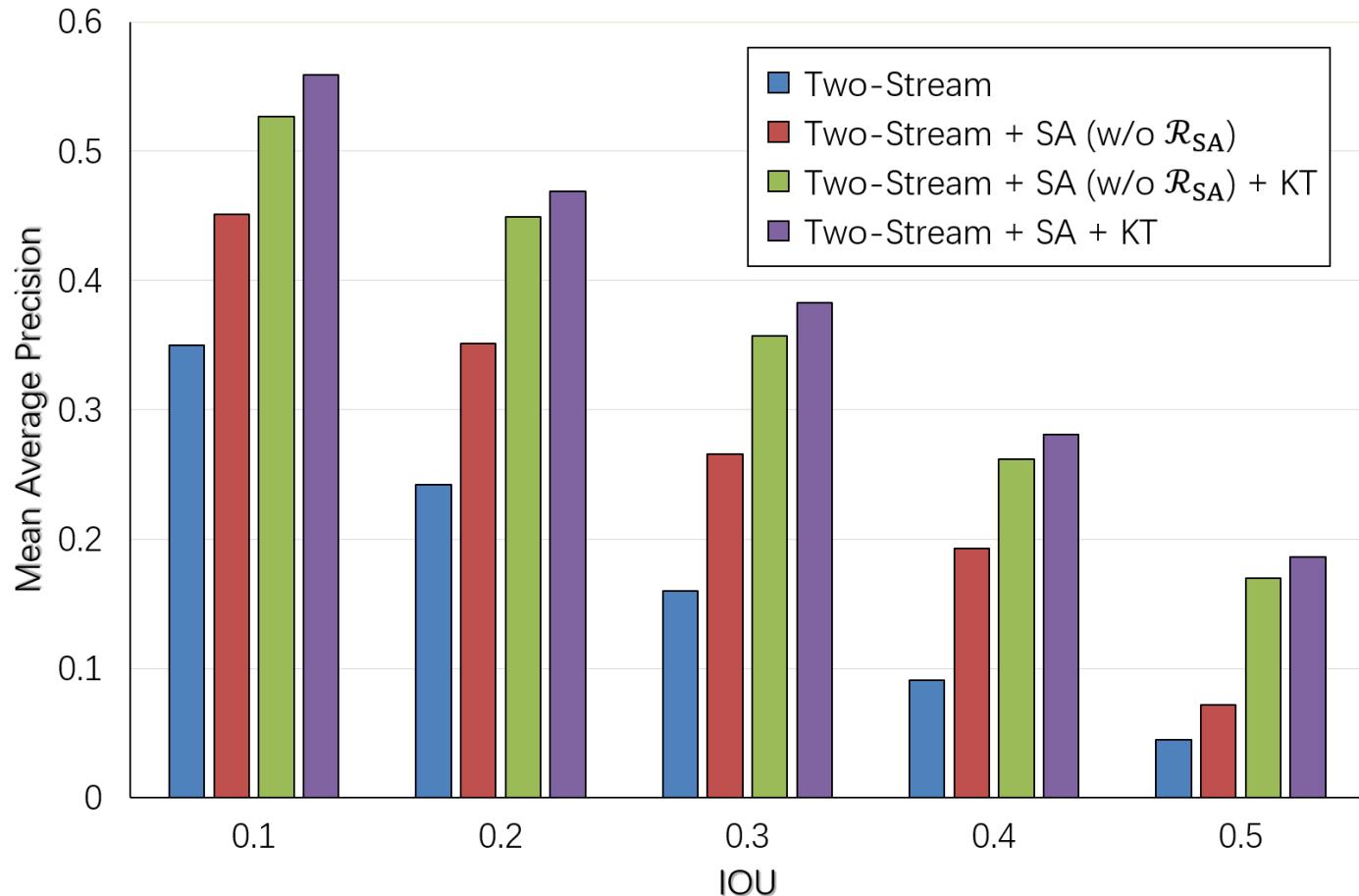
# Results – Ablation Study

## ■ Ablation study on **THUMOS14**

**SA**: the self-attention module

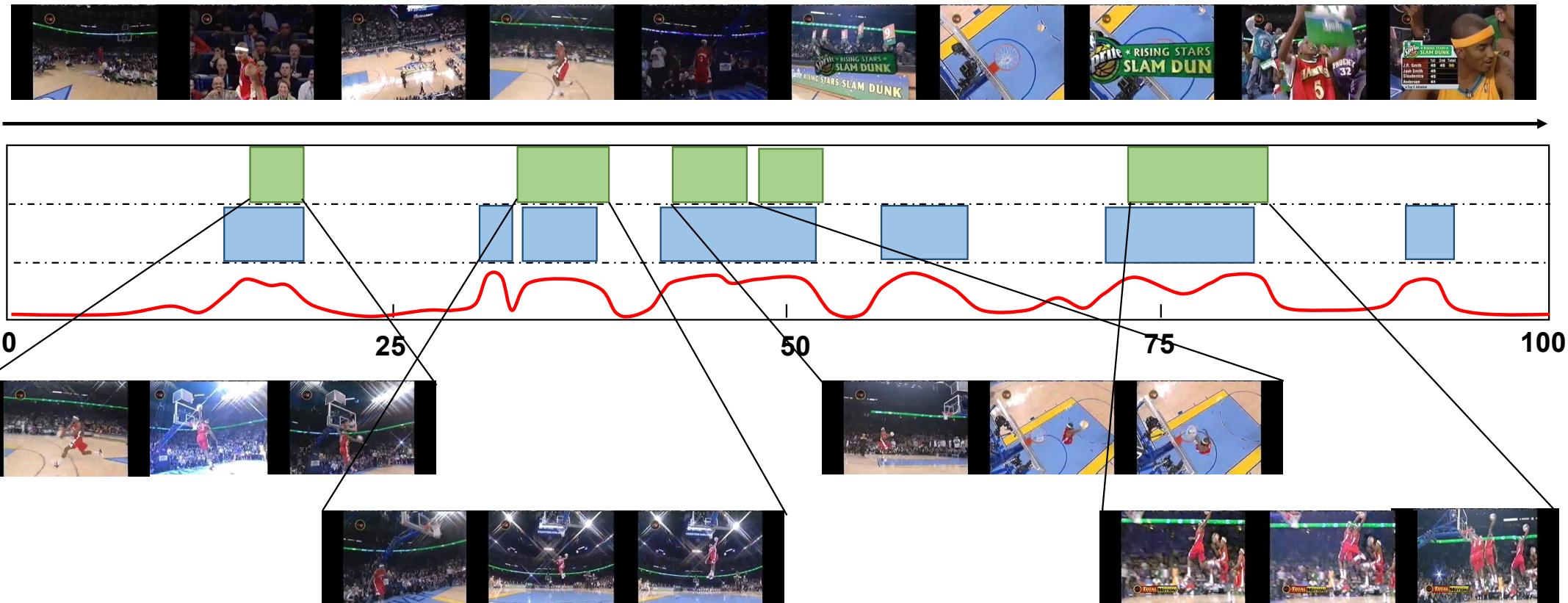
**KT**: the knowledge transfer module

**Two-Stream + SA + KT**: the full implementation of TSRNet



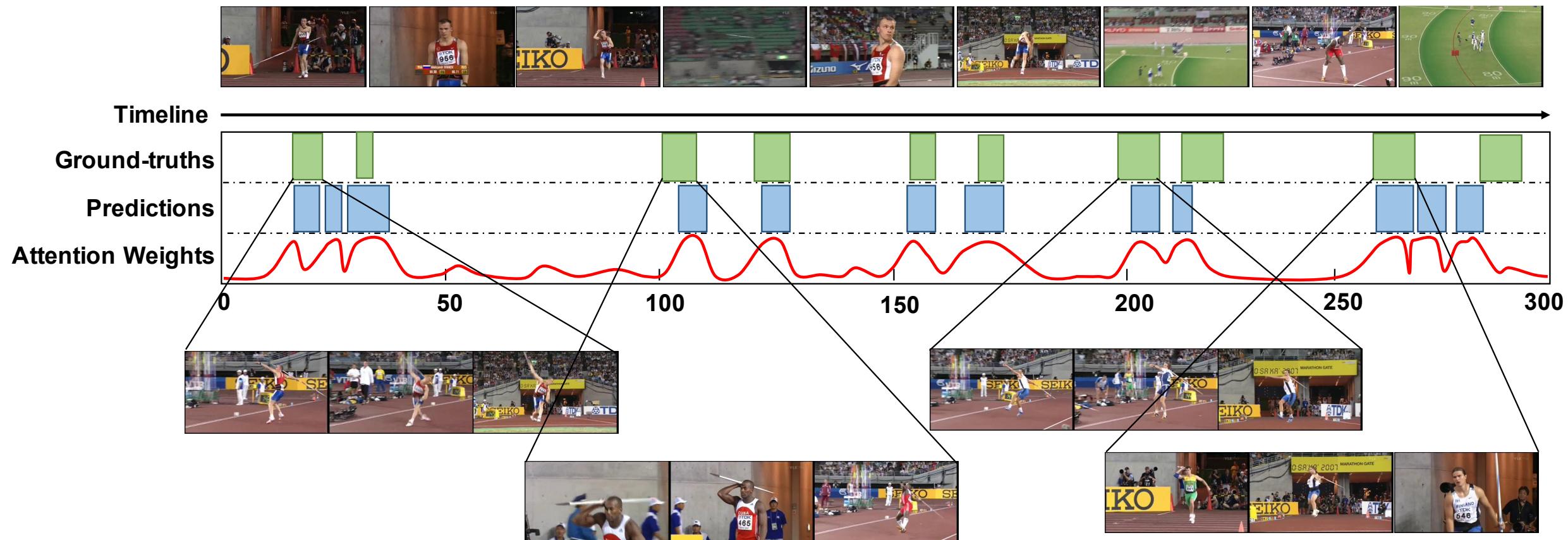
# Results – Qualitative Evaluation

## ■ Qualitative evaluation on **THUMOS14**



# Results – Qualitative Evaluation

## ■ Qualitative evaluation on **ActivityNet1.3**



# Conclusion

---

- TSRNet is the first to introduce **Knowledge Transfer** for action recognition in untrimmed videos with weak supervision
  - Knowledge of additional trimmed videos is effectively leveraged and transferred to improve the classification performance for untrimmed ones.
- TSRNet adopts **Self-Attention** mechanism to obtain frame-levels analysis
  - Frames with higher self-attention weights can be selected out for the purpose of temporal action localization/detection in videos.
- TSRNet outperforms the existing state-of-the-art competitors
  - Extensive experiments on two challenging untrimmed video datasets (i.e., **THUMOS14** and **ActivityNet1.3**) show promising results



Thank you!  
Questions & Answers

