Action recognition

in the spirit of object detection

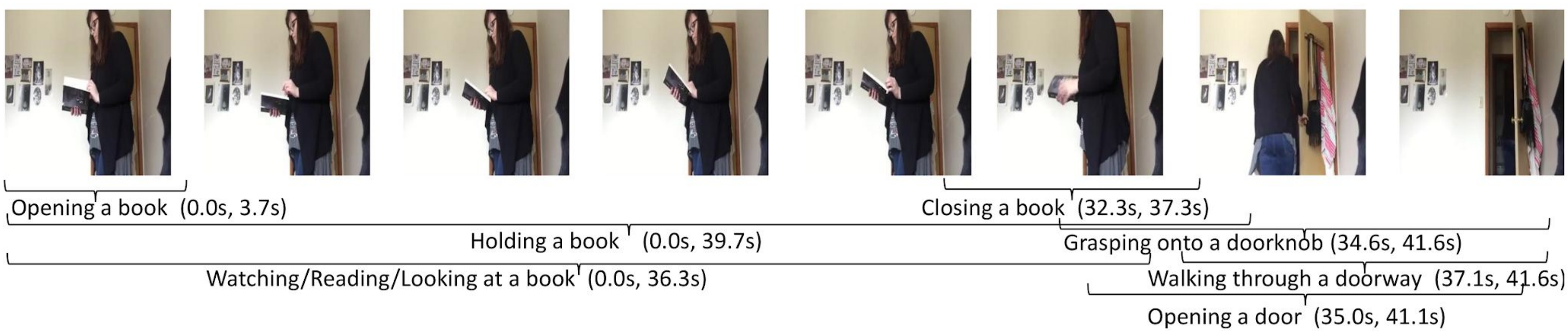
Nick Turner, Sven Dorkenwald

COS 598 - 04/23/18

Temporal activity detection

Classify

(1) Action
(2) Temporal window



Example from Charades

Fixed time contexts in prior approaches

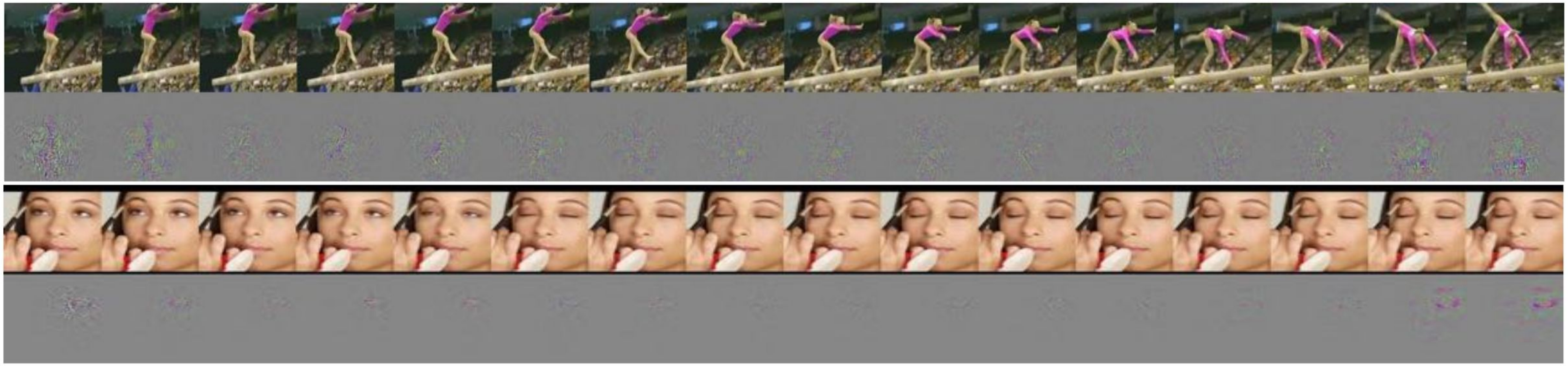
Prior two-step approaches:

(1) classify action → (2) agglomerate actions

Fixed time contexts in prior approaches

Prior two-step approaches:

(1) classify action → (2) agglomerate actions



16 frame input to C3D and extracted features in conv5b (last convolution)

“Advanced” temporal action localization

(1) **R-C3D** End-to-end model with combined activity proposal and classification stages

↓

(2) **CMS-RC3D** Contextual information is fused from multiple time scales

RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

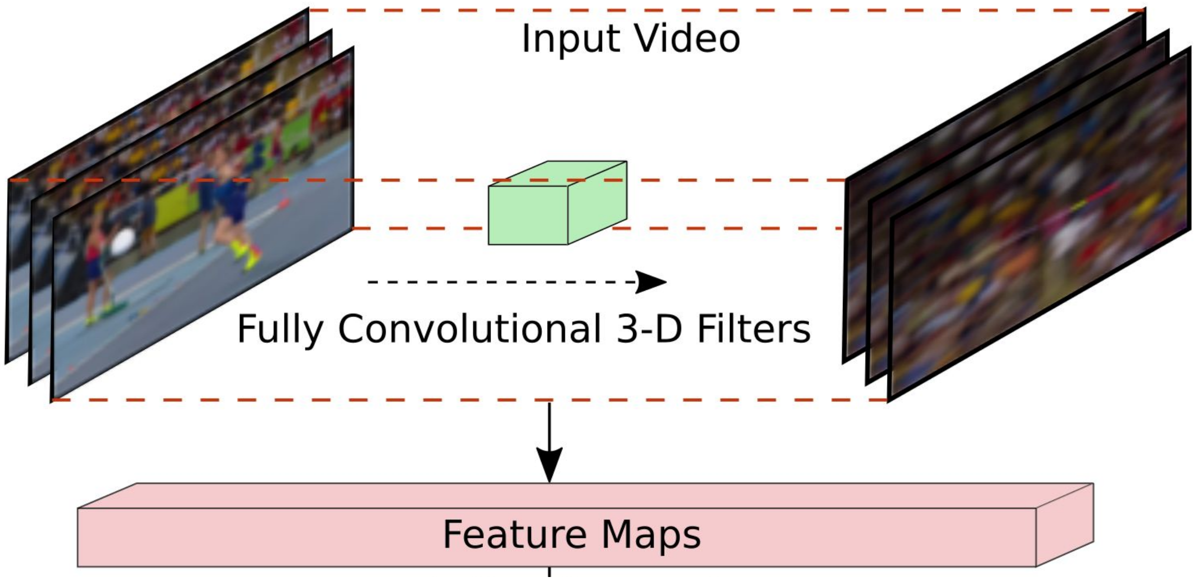
MOTIVATING PROBLEMS

NOVELTY

EXPERIMENTS

DISCUSSION II

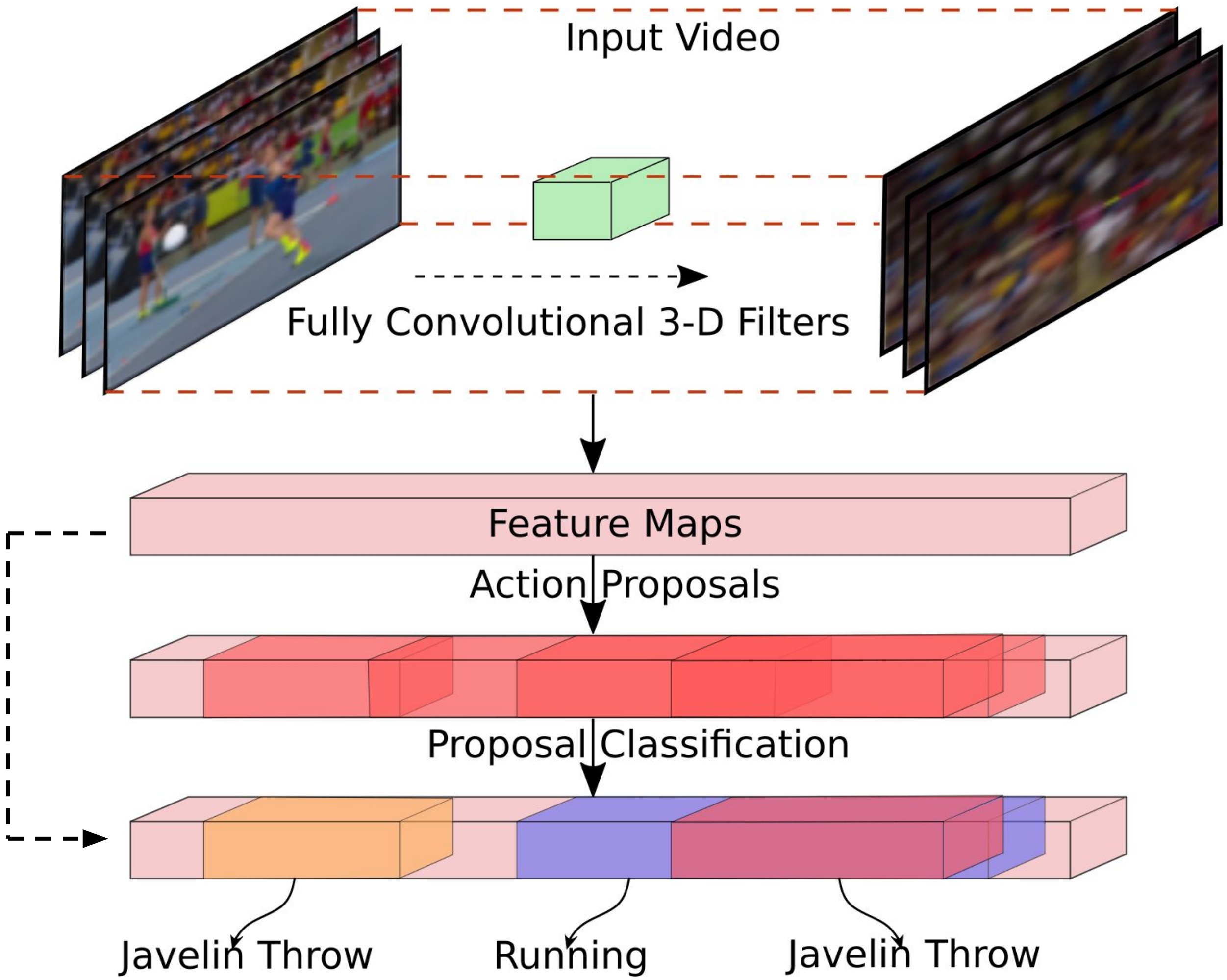
R-C3D uses features at **any** granularity



“Blown-up” C3D

___

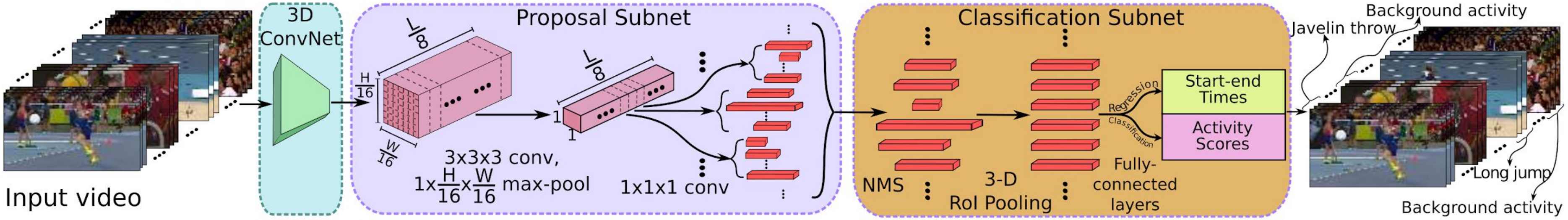
R-C3D uses features at **any** granularity



"Blown-up" C3D

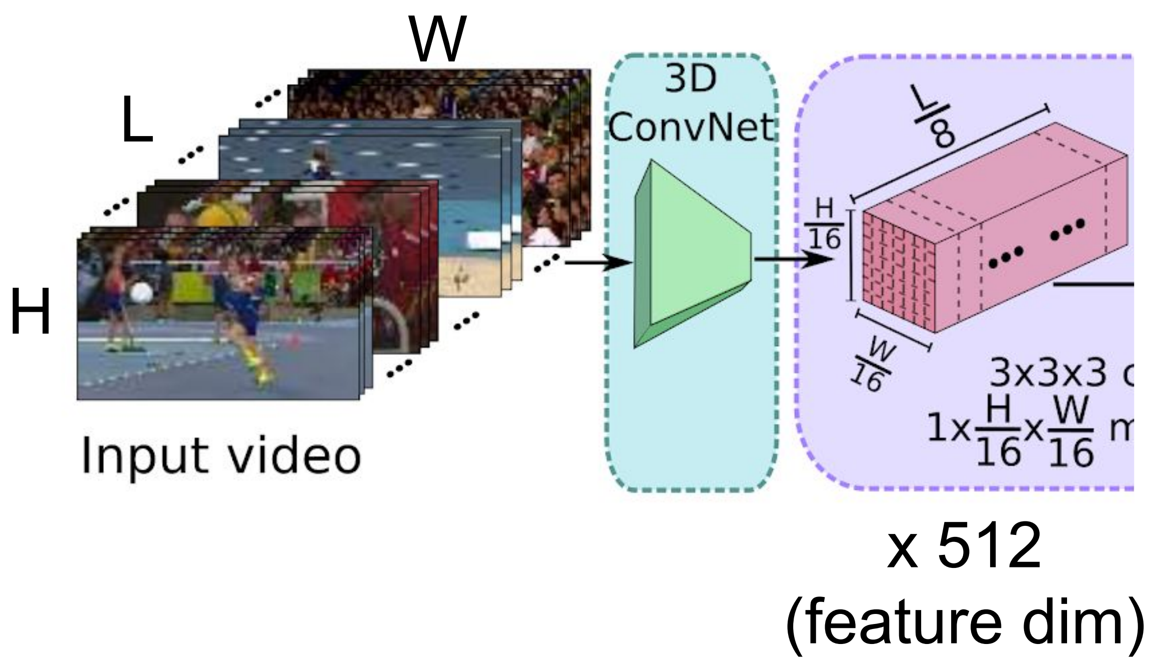
3D RoI Pooling

Model walkthrough



3D CNN feature extractor (C3D)

Goal: Extract spatio-temporal features

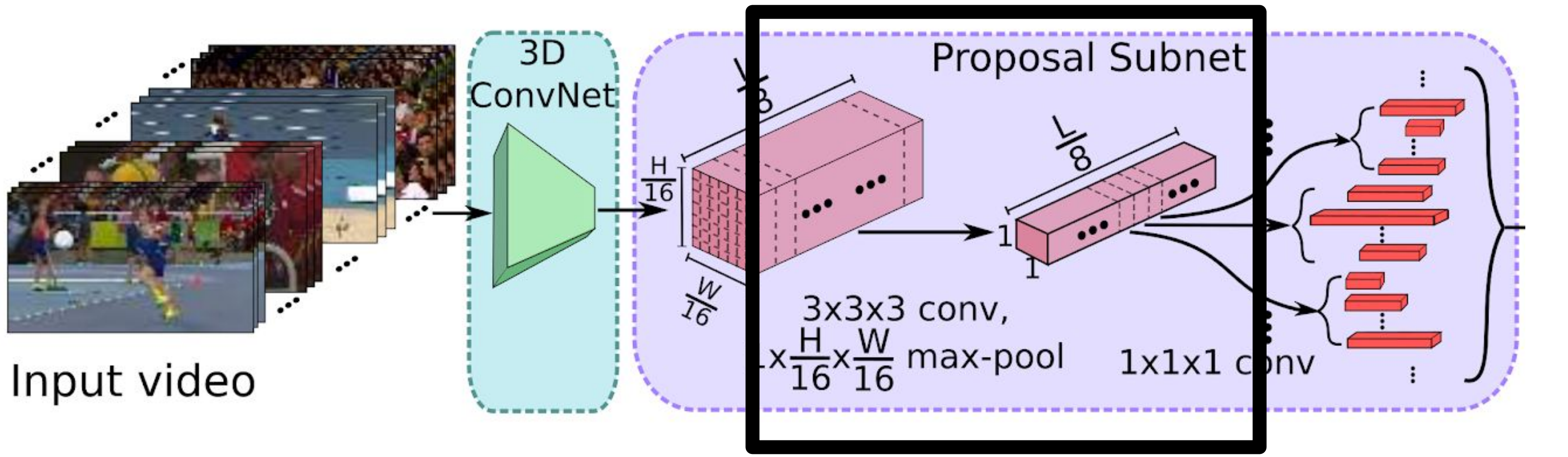


L: number of frames
(limited by memory)

H = W = 112

Proposal subnet

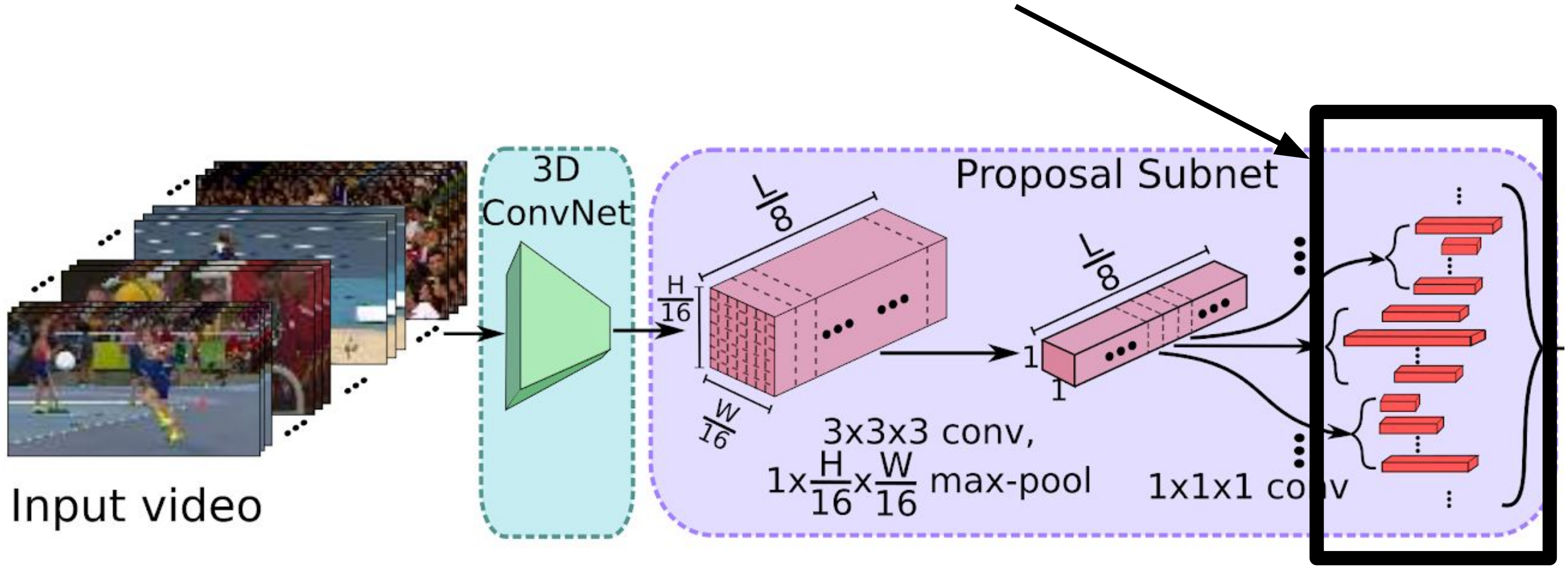
Goal: Predict which anchor segments contain actions



512 x L/8 x 1 x 1

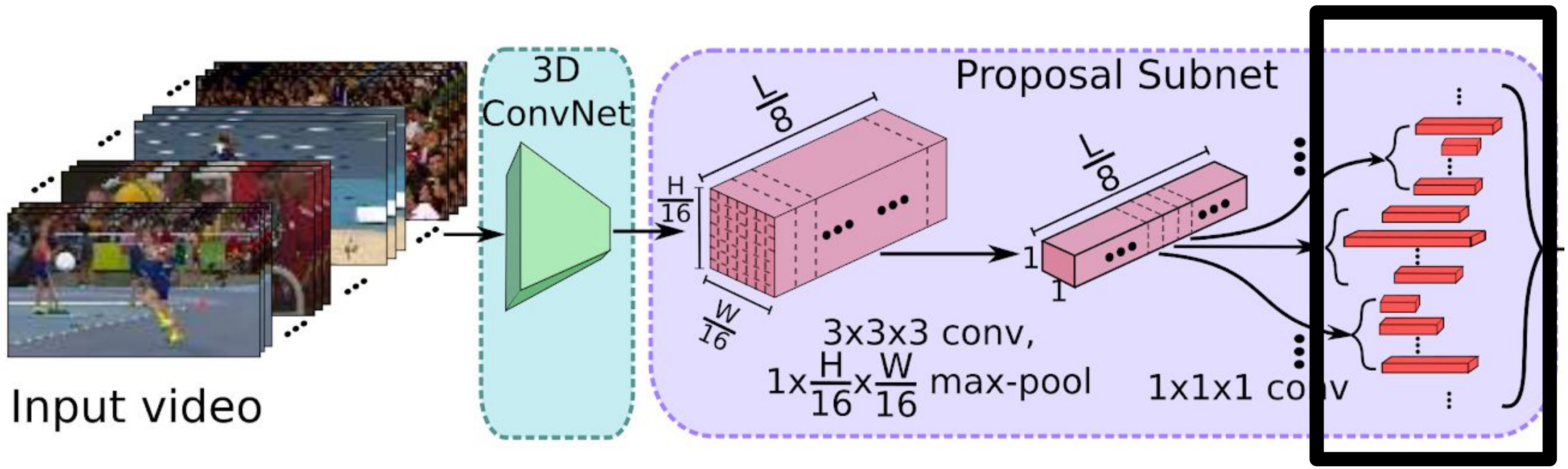
Proposal subnet

Goal: Predict which anchor segments contain actions



Proposal subnet

Goal: Predict which anchor segments contain actions

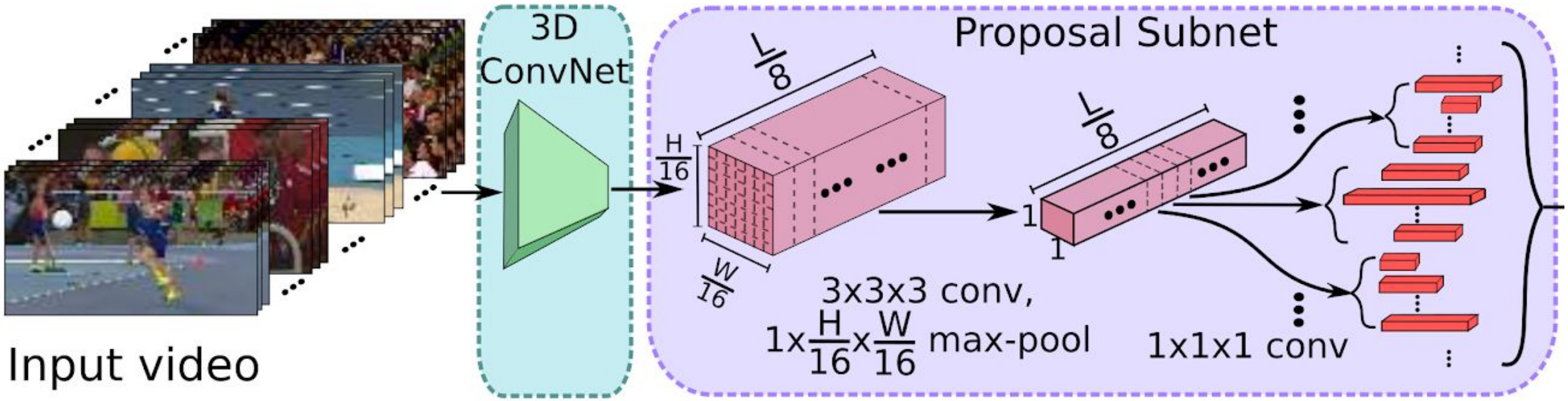


number of multiscale anchor segments = L / 8 * K

K: number of scales (“dataset dependent”)

Proposal subnet

Goal: Predict which anchor segments contain actions

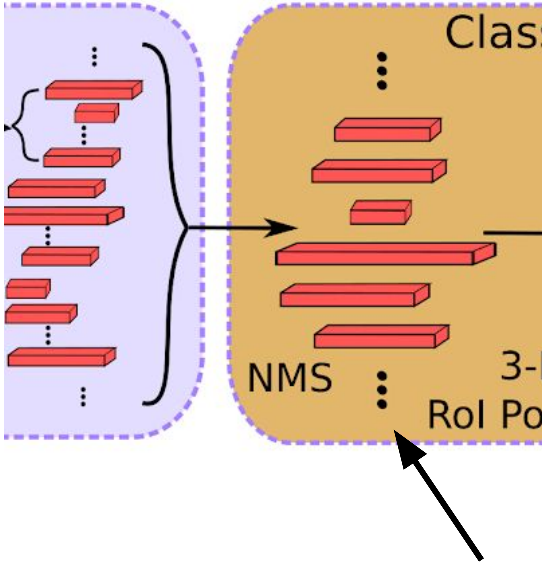


(1) Classify L / 8 * K segments as background vs action

(2) Infer (offset, length difference) from anchor segments

Classification subnet

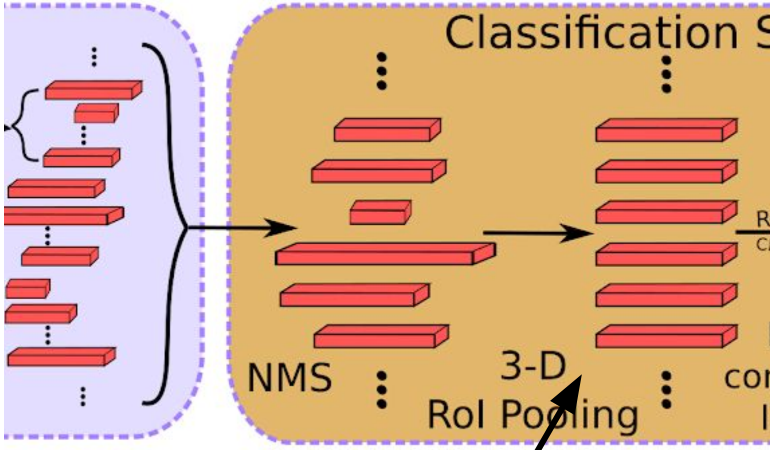
Goal: Select and classify proposals



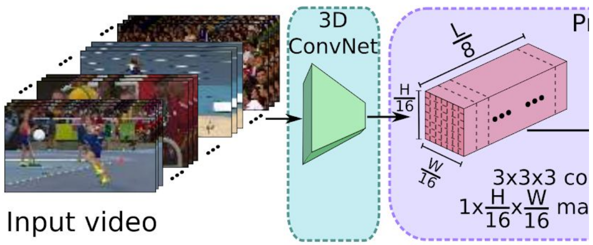
condensing the proposals

Classification subnet

Goal: Select and classify proposals

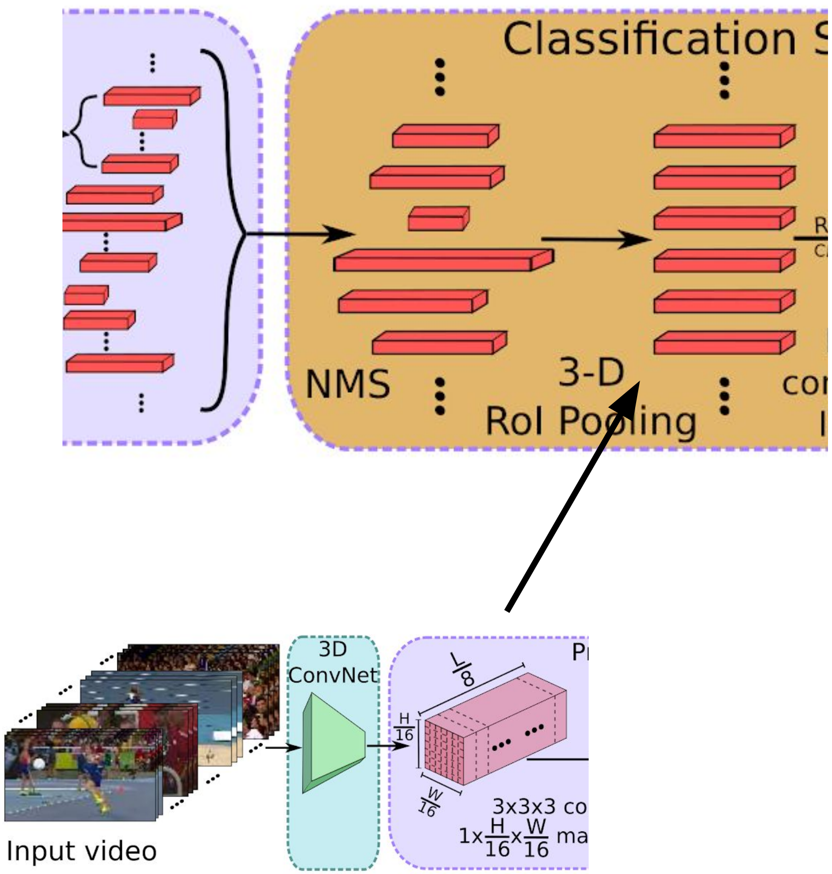


Problem: arbitrarily long regions

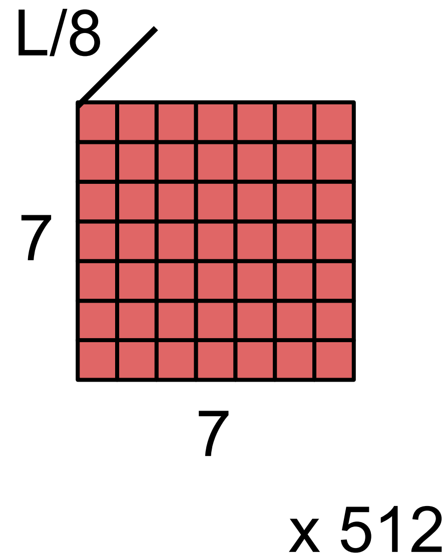


Classification subnet

Goal: Select and classify proposals



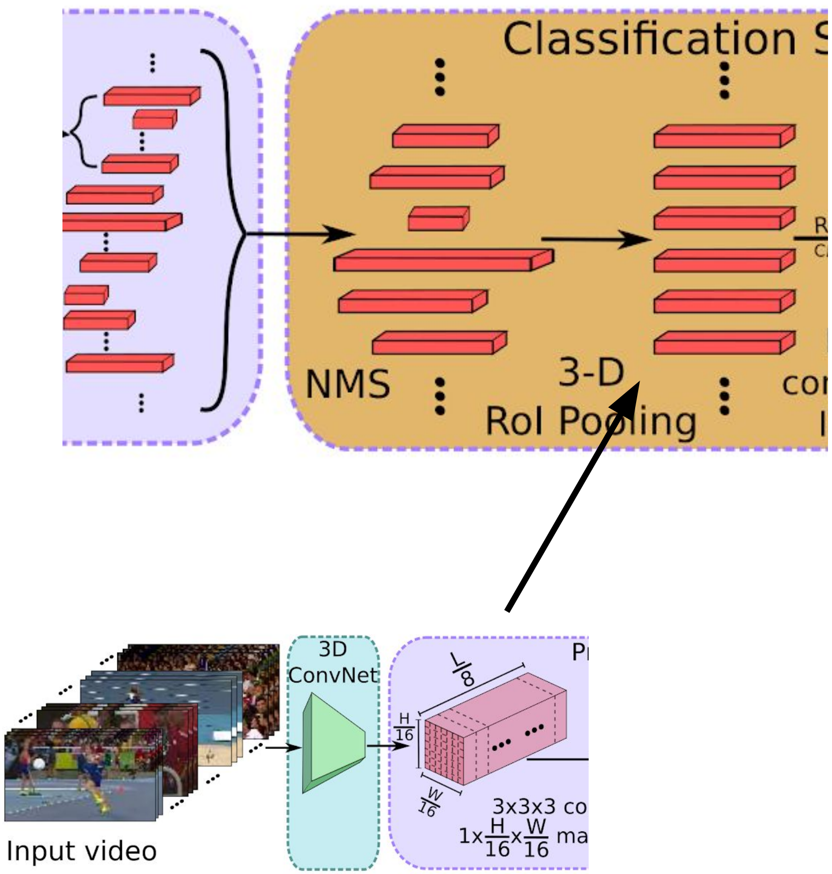
fixed sized 3D RoI pooling



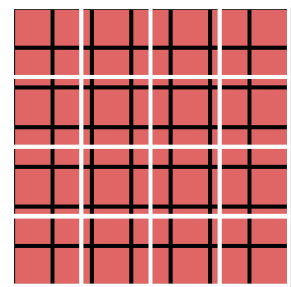
Classification subnet

Goal: Select and classify proposals

fixed sized 3D RoI pooling



4



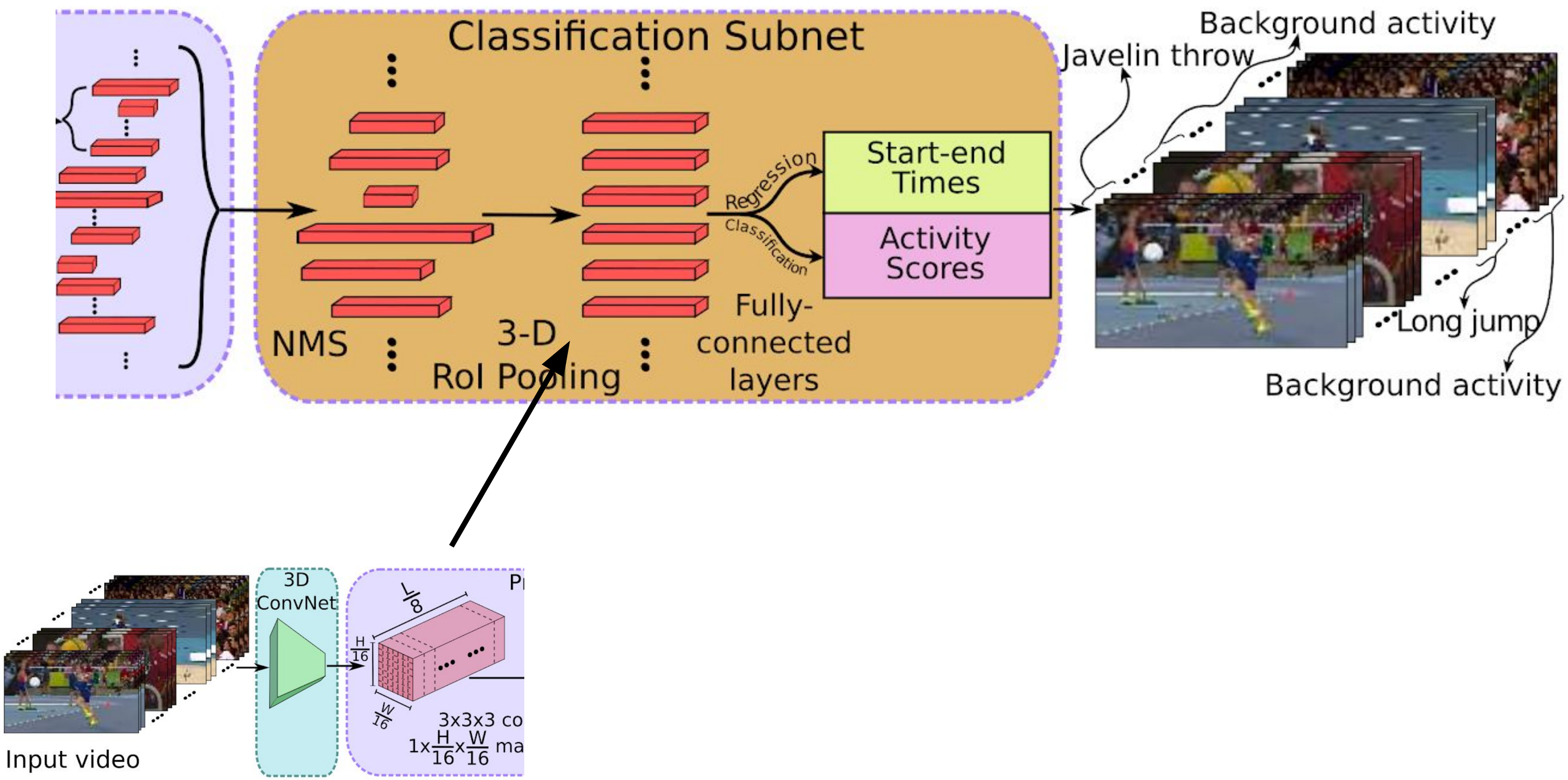
4

x 512

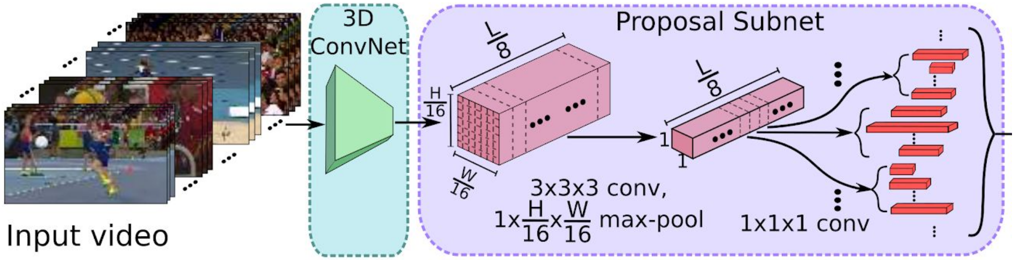
→ 8192 features

Classification subnet

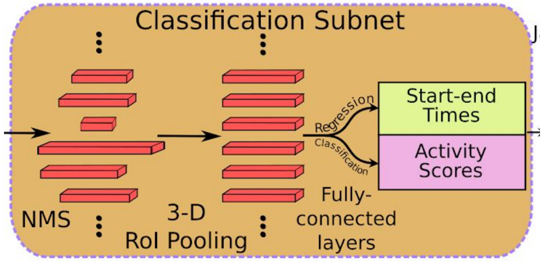
Goal: Select and classify proposals



Training the two subnets **jointly**



Regression on time-window
+
Classification on action / background



Regression on time-window
+
Classification on action

Loss function

Classification loss

Proposal net: single class
Classification net: multiclass

$$Loss = \overline{\frac{1}{N_{cls}} \sum_i L_{cls}(a_i, a_i^*)} + \underline{\lambda \frac{1}{N_{reg}} \sum_i a_i^* L_{reg}(t_i, t_i^*)}$$

Regression loss
on time window

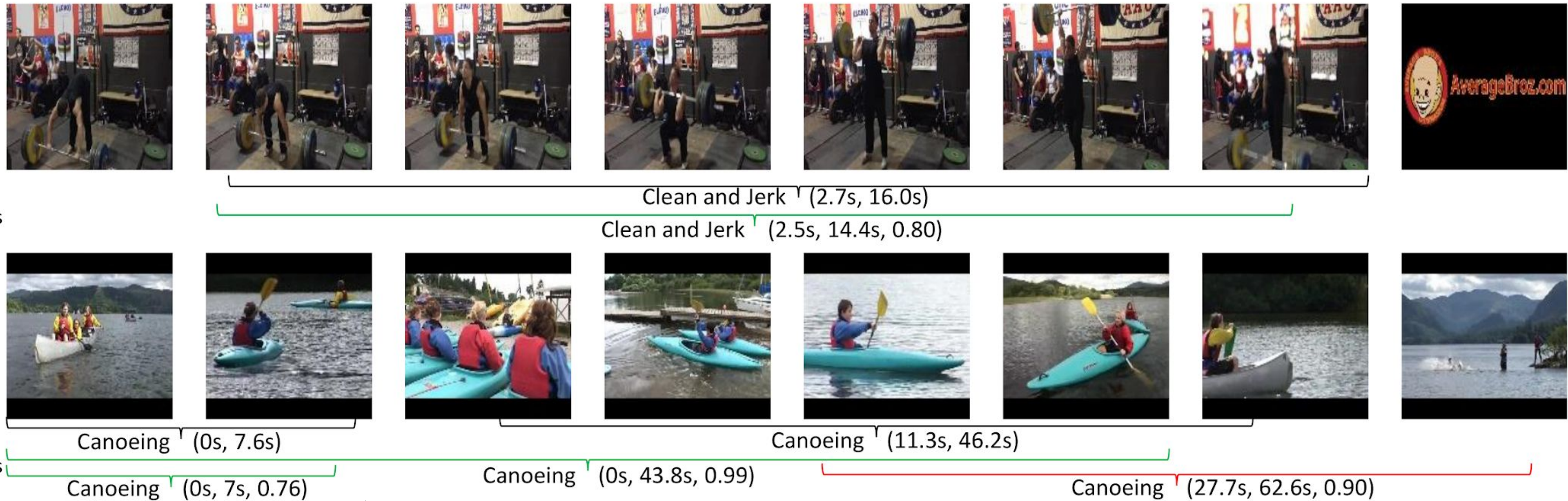
Time window:

$$t_i = \{\delta \hat{c}_i, \delta \hat{l}_i\} \qquad \begin{cases} \delta c_i = (c_i^* - c_i)/l_i \\ \delta l_i = log(l_i^*/l_i) \end{cases}$$

RC3D	**CMS-RC3D**

RC3D	**CMS-RC3D**
TASK REVIEW	MOTIVATING PROBLEMS
NOVELTY	NOVELTY
EXPERIMENTS	EXPERIMENTS
DISCUSSION I	DISCUSSION II

Qualitative evaluation on ActivityNet

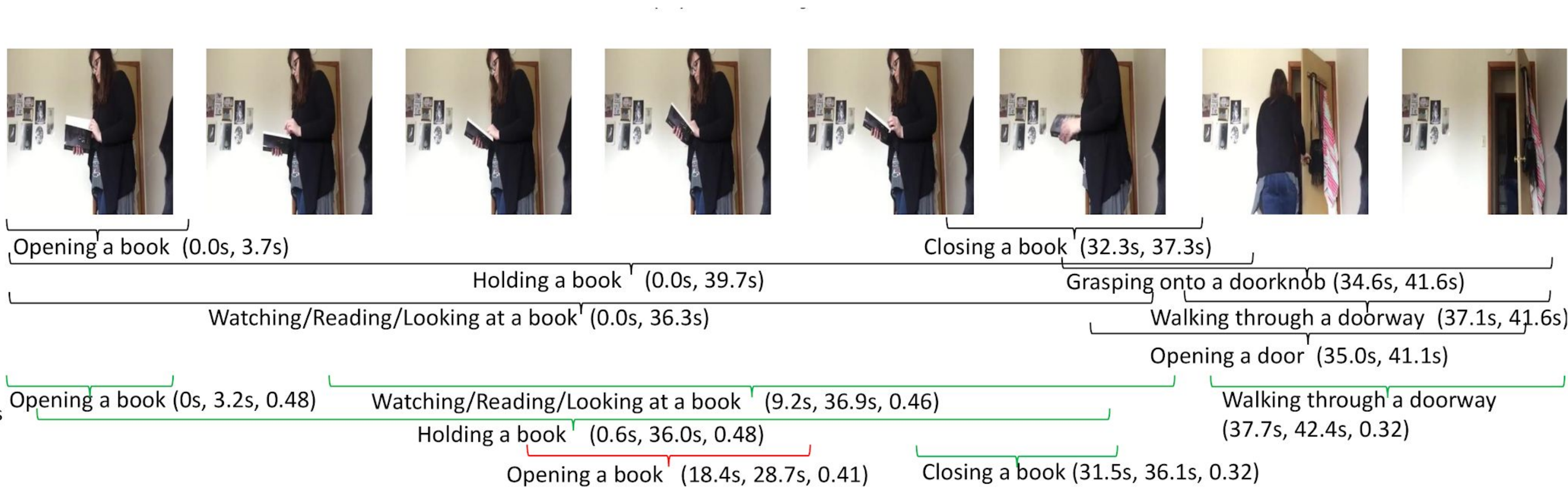


Overlapping actions

GT

R-C3D

Qualitative evaluation on Charades



GT

R-C3D

Results on THUMOS' 14

IoU

	α				
	0.1	0.2	0.3	0.4	0.5
Karaman et al. [13]	4.6	3.4	2.1	1.4	0.9
Wang et al. [37]	18.2	17.0	14.0	11.7	8.3
Oneata et al. [20]	36.6	33.6	27.0	20.8	14.4
Heilbron et al. [10]	-	-	-	-	13.5
Escorcia et al. [4]	-	-	-	-	13.9
Richard et al. [22]	39.7	35.7	30.0	23.2	15.2
Yeung et al. [39]	48.9	44.0	36.0	26.4	17.1
Yuan et al. [41]	51.4	42.6	33.6	26.1	18.8
Shou et al. [24]	47.7	43.5	36.3	28.7	19.0
Shou et al. [23]	-	-	40.1	29.4	23.3
R-C3D (our one-way buffer)	51.6	49.2	42.8	33.4	27.0
R-C3D (our two-way buffer)	**54.5**	**51.5**	**44.8**	**35.6**	**28.9**

mAP scores

proposal classification: 85% precision, 83% recall

Results on ActivityNet

mAP@0.5

	train data	validation	test
G. Singh *et. al*. [30]	train	34.5	36.4
B. Singh *et. al*. [29]	train+val	-	28.8
UPC [18]	train	22.5	22.3
R-C3D (ours)	train	**26.8**	**26.8**
R-C3D (ours)	train+val	-	**28.4**

RC3D is faster than existing methods

Inference speeds:

	FPS
S-CNN [24]	60
DAP [4]	134.1
R-C3D (ours on Titan X Maxwell)	**569**
R-C3D (ours on Titan X Pascal)	**1030**

R-C3D key takeaways

(1) An End-to-end solution allows for arbitrary time granularity
→ can handle overlapping activity
→ improvements in performance

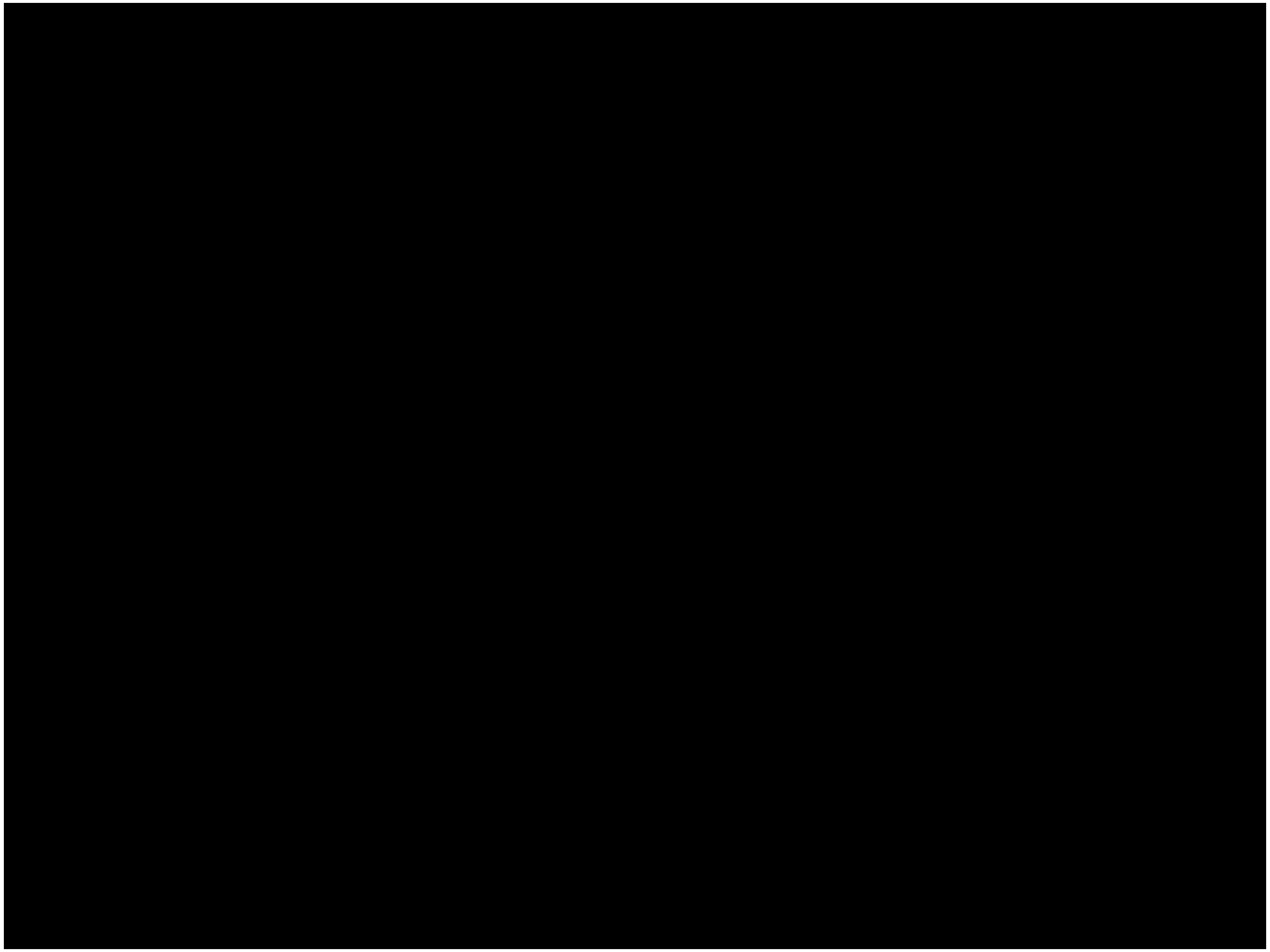
(2) Performance of the proposal net might / should allow for better activity prediction

(3) Newer graphics cards lead to large speed-ups

RC3D	**CMS-RC3D**
TASK REVIEW	MOTIVATING PROBLEMS
NOVELTY	NOVELTY
EXPERIMENTS	EXPERIMENTS
DISCUSSION I	DISCUSSION II

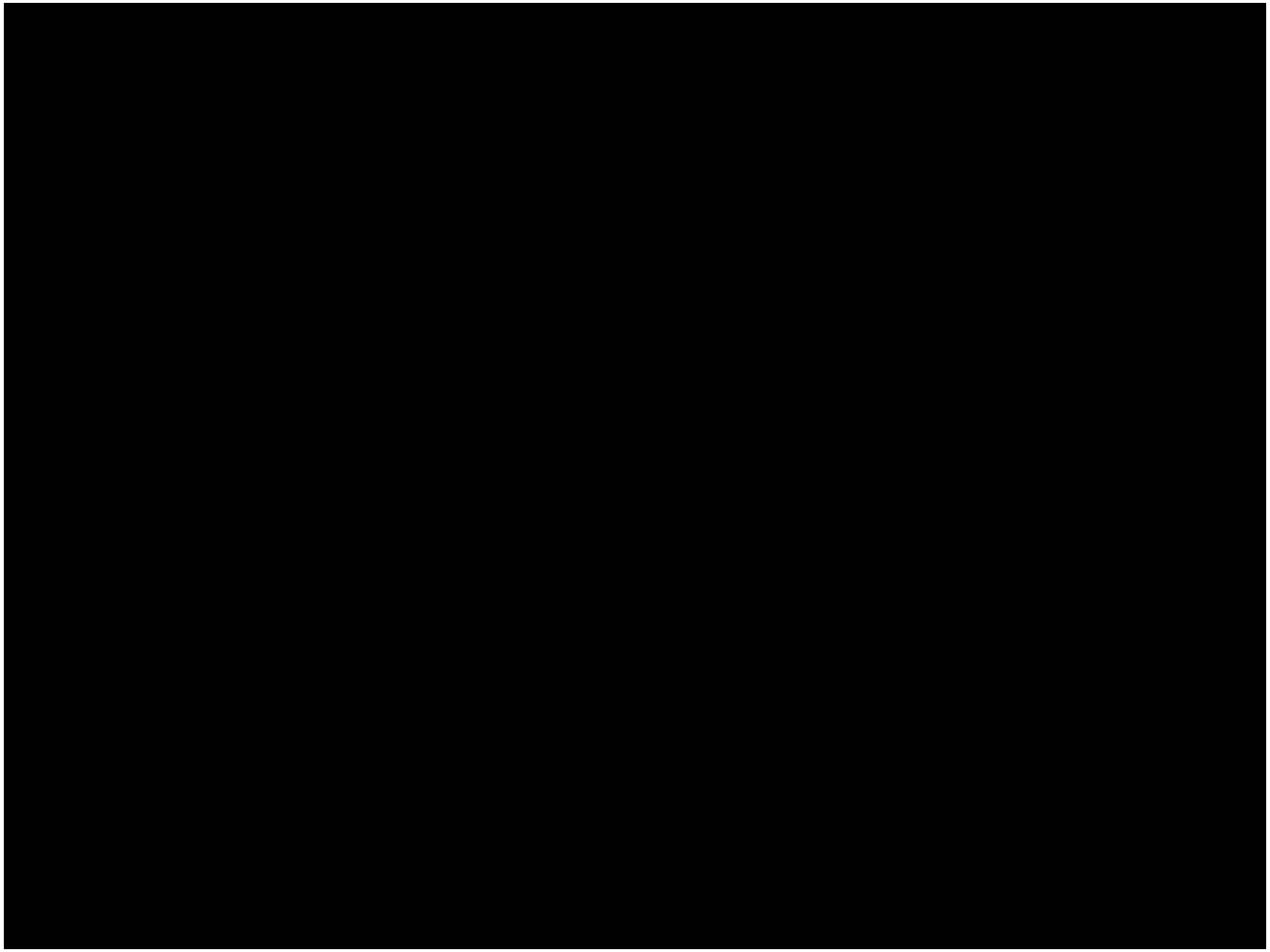
RC3D	CMS-RC3D
TASK REVIEW	**MOTIVATING PROBLEMS** Multiple Timescales Context
NOVELTY	NOVELTY
EXPERIMENTS	EXPERIMENTS
DISCUSSION I	DISCUSSION II

Activities take place over very different timescales



...perhaps representing multiple timescales will aid in activity detection

Context



Other approaches use context outside of the “activity window” itself to assist prediction

RC3D

CMS-RC3D

TASK REVIEW

MOTIVATING PROBLEMS

Multiple Timescales

Context

NOVELTY

NOVELTY

EXPERIMENTS

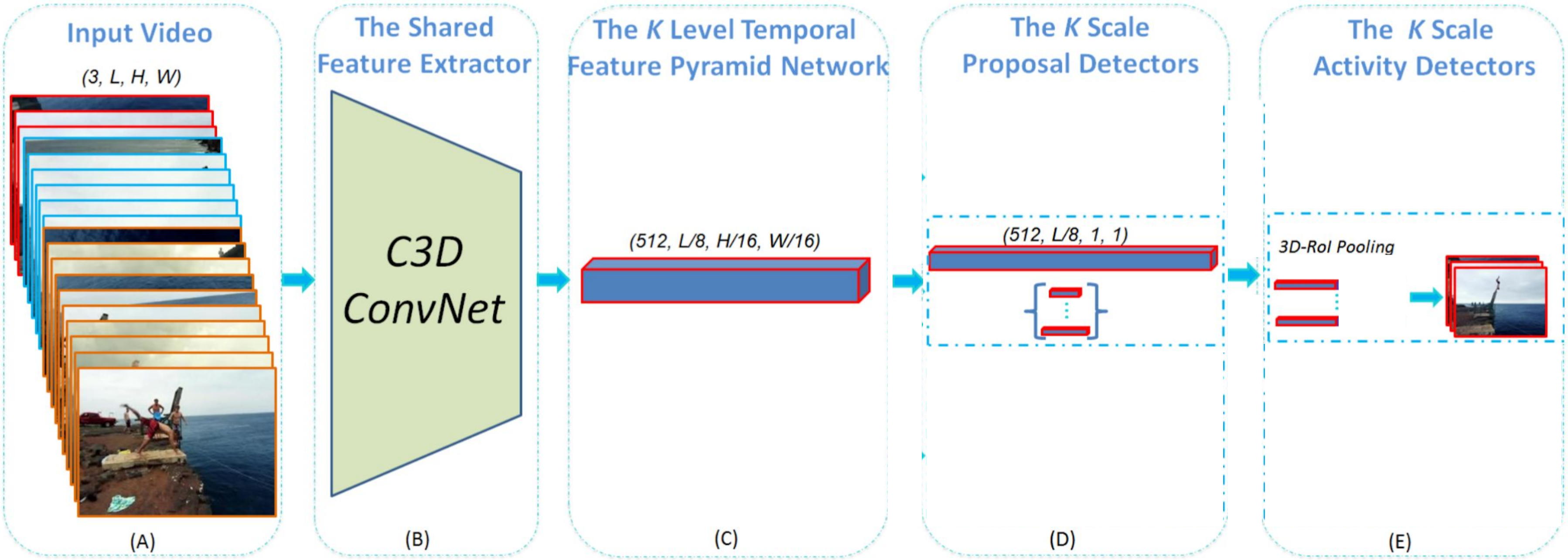
EXPERIMENTS

DISCUSSION I

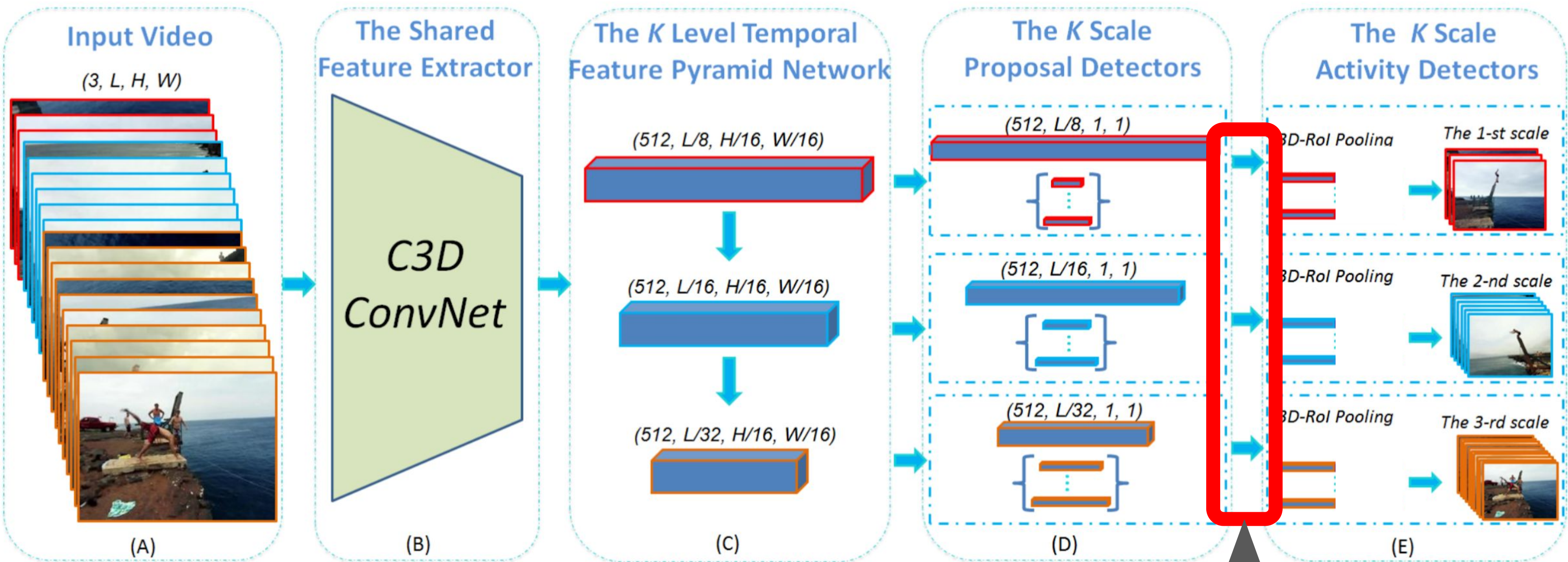
DISCUSSION II

RC3D	**CMS-RC3D**
TASK REVIEW	MOTIVATING PROBLEMS
NOVELTY	**NOVELTY** Multiple Timescales Context
EXPERIMENTS	EXPERIMENTS
DISCUSSION I	DISCUSSION II

Representing multiple time scales

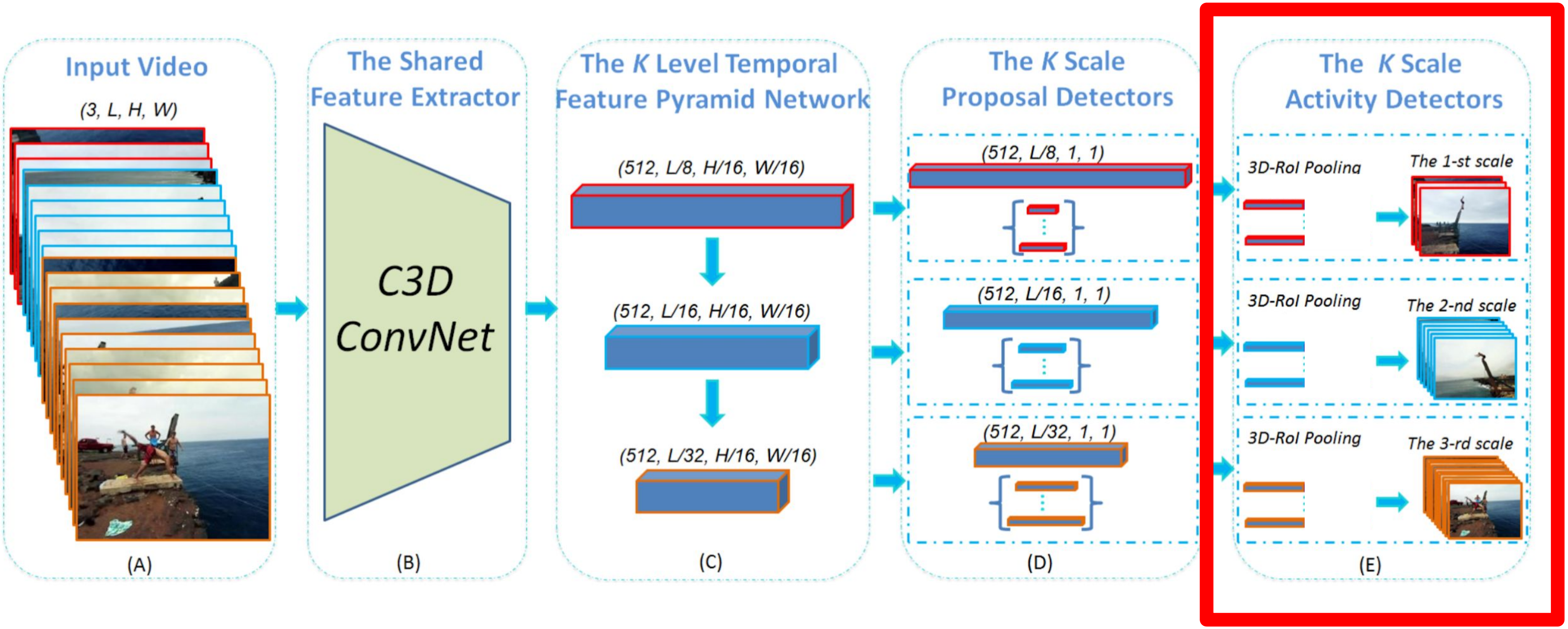


Representing multiple time scales

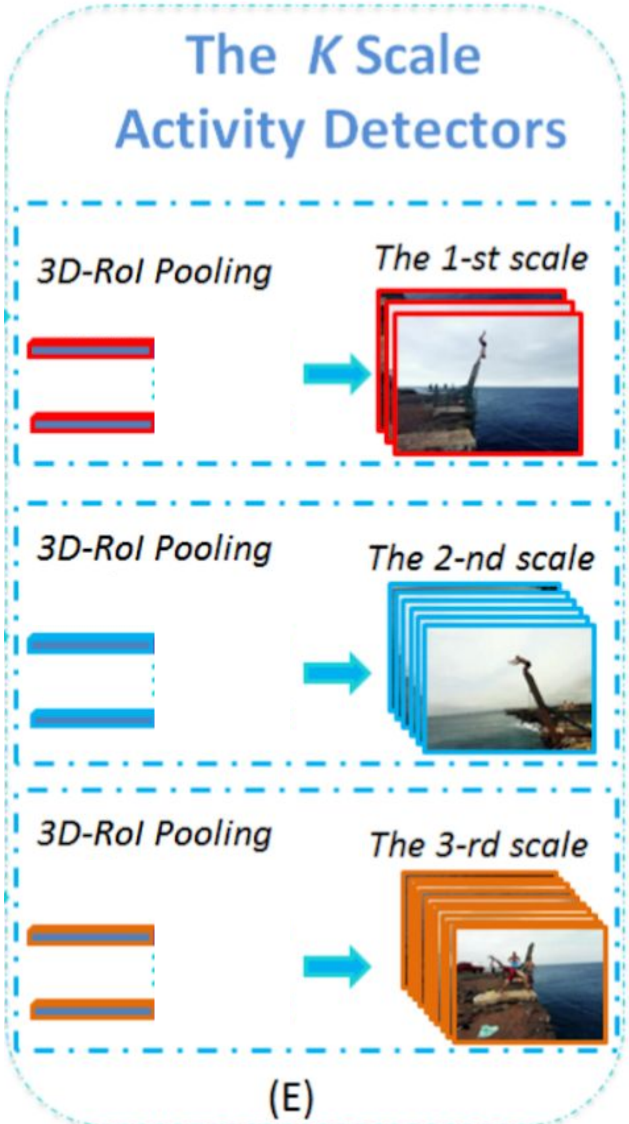


This is *slightly* misleading

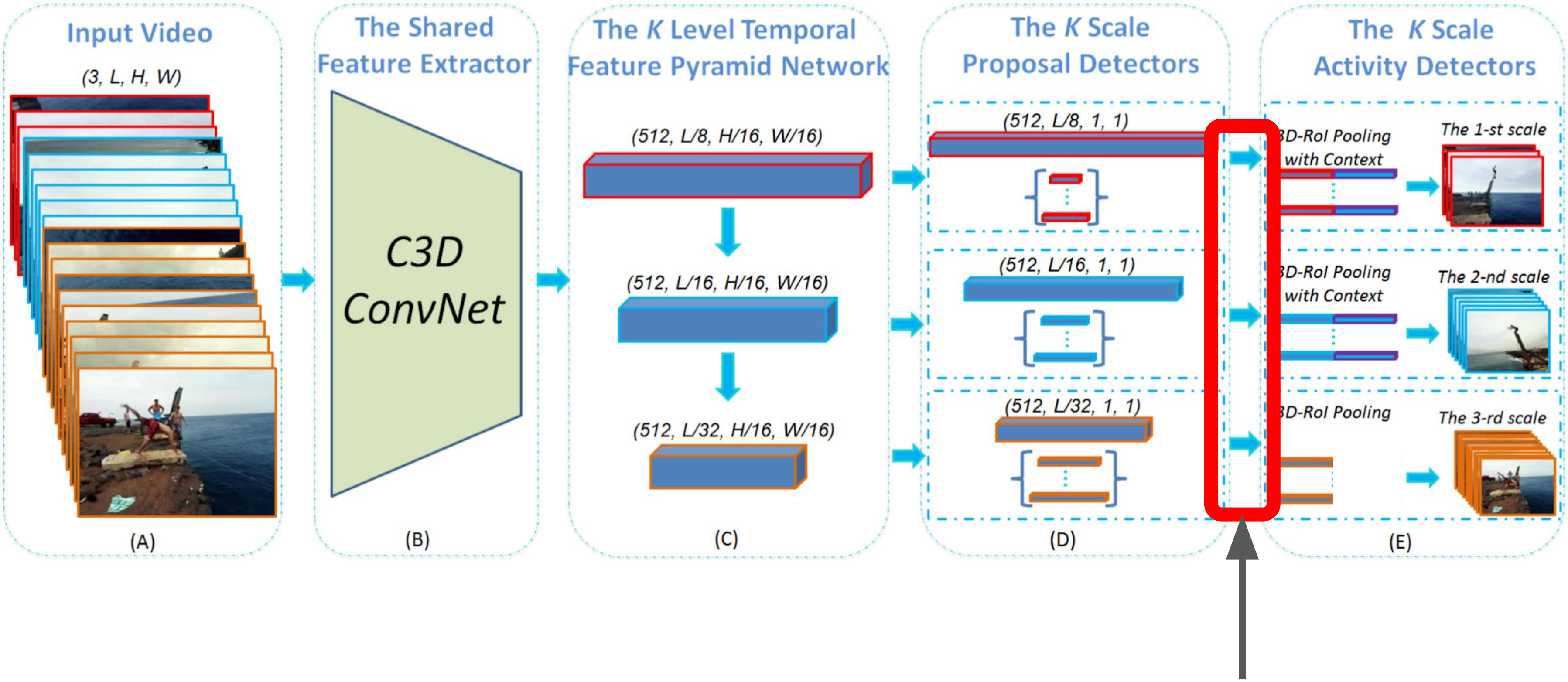
Adding context



Adding context

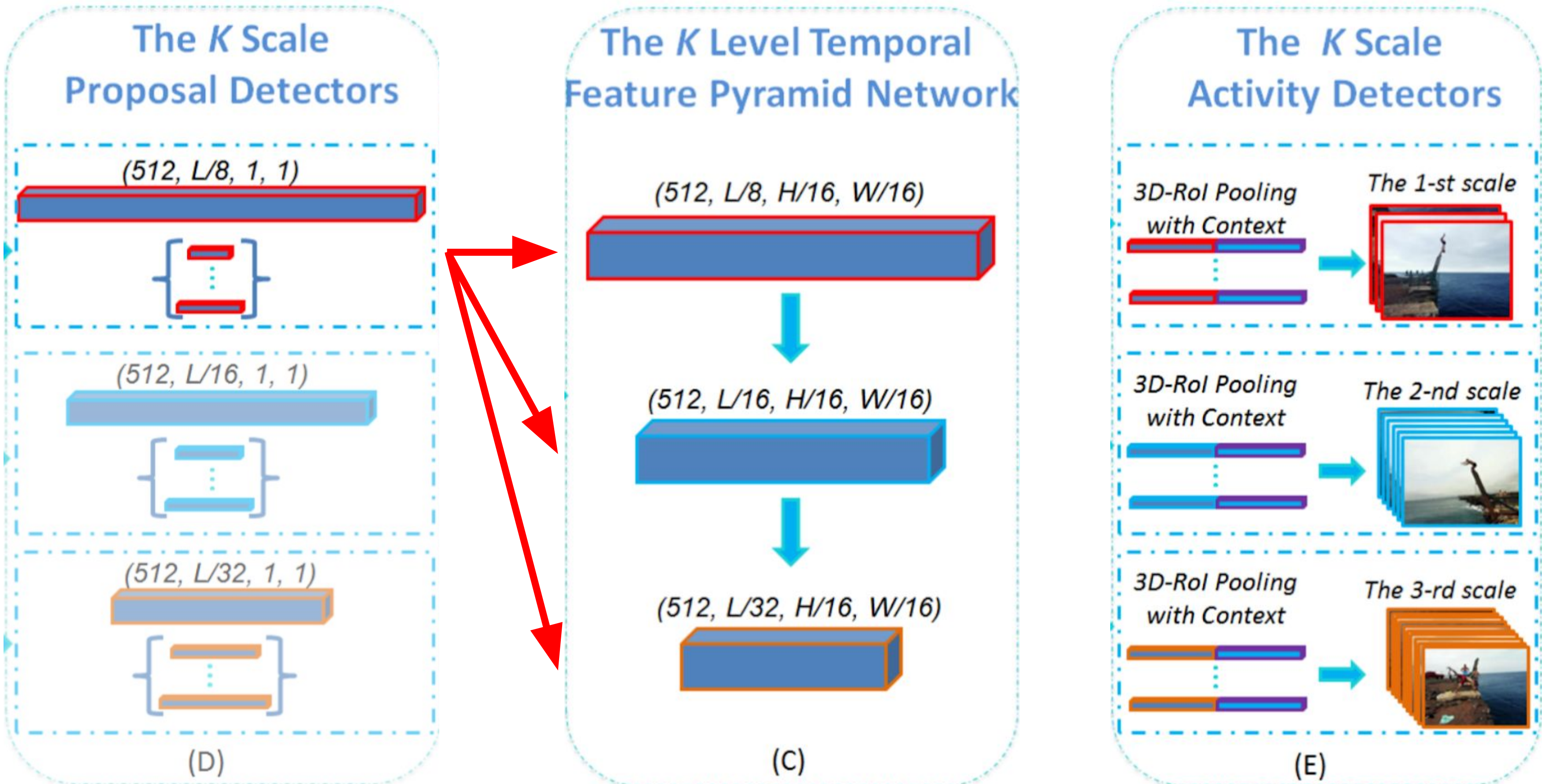


Do they add half the window to each side?
Or just double the length?



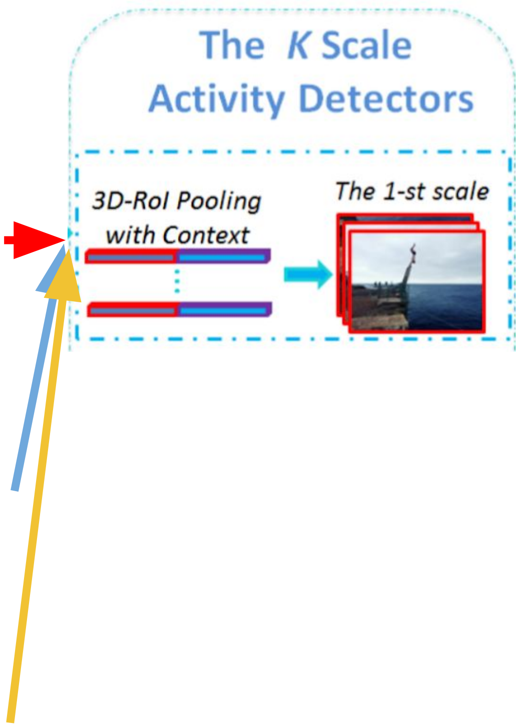
This is *slightly* misleading

How do we pick the scale at which to pool a given proposal?

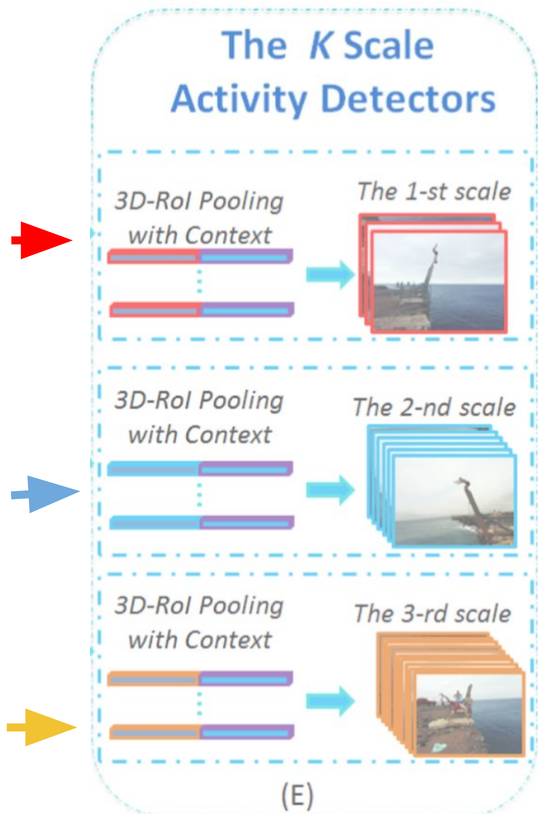


How do we pick the scale at which to pool a given proposal?

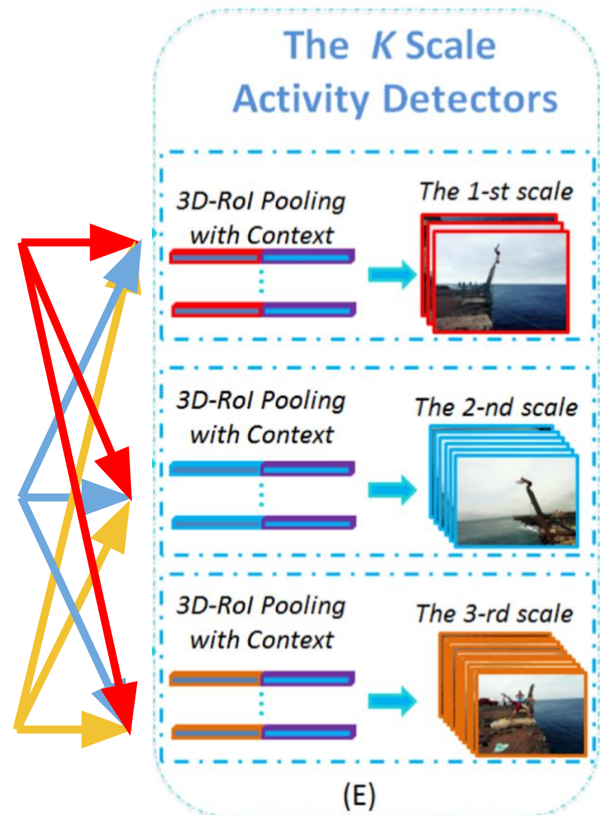
Strategy 1
“S1”



Strategy 2
“S2”



Strategy 3
“S3”



RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

MOTIVATING PROBLEMS

NOVELTY

Multiple Timescales

Context

EXPERIMENTS

DISCUSSION II

RC3D	**CMS-RC3D**
TASK REVIEW	MOTIVATING PROBLEMS
NOVELTY	NOVELTY
EXPERIMENTS	**EXPERIMENTS** Ablation Studies Evaluations
DISCUSSION I	DISCUSSION II

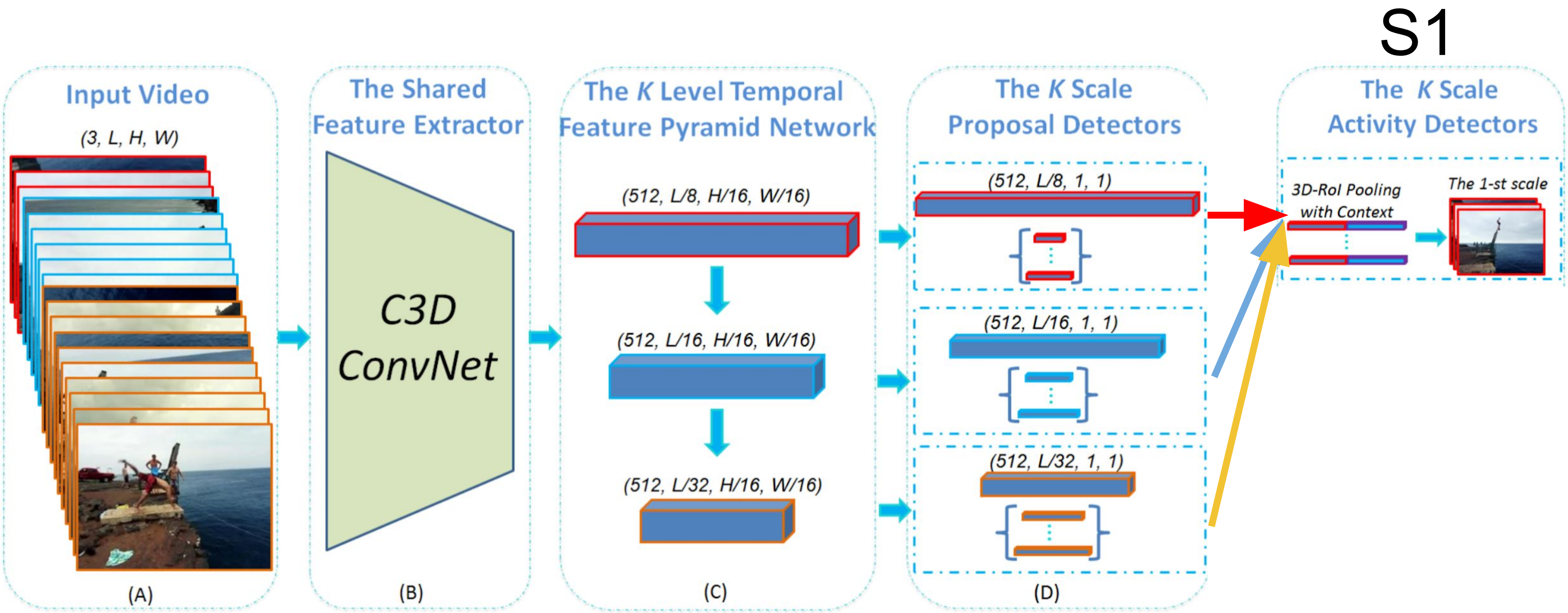
ActivityNet Evaluation

Method	0.5	0.75	0.95	Average
RC3D [33]	26.33	10.46	1.25	12.71

THUMOS '14 Evaluation

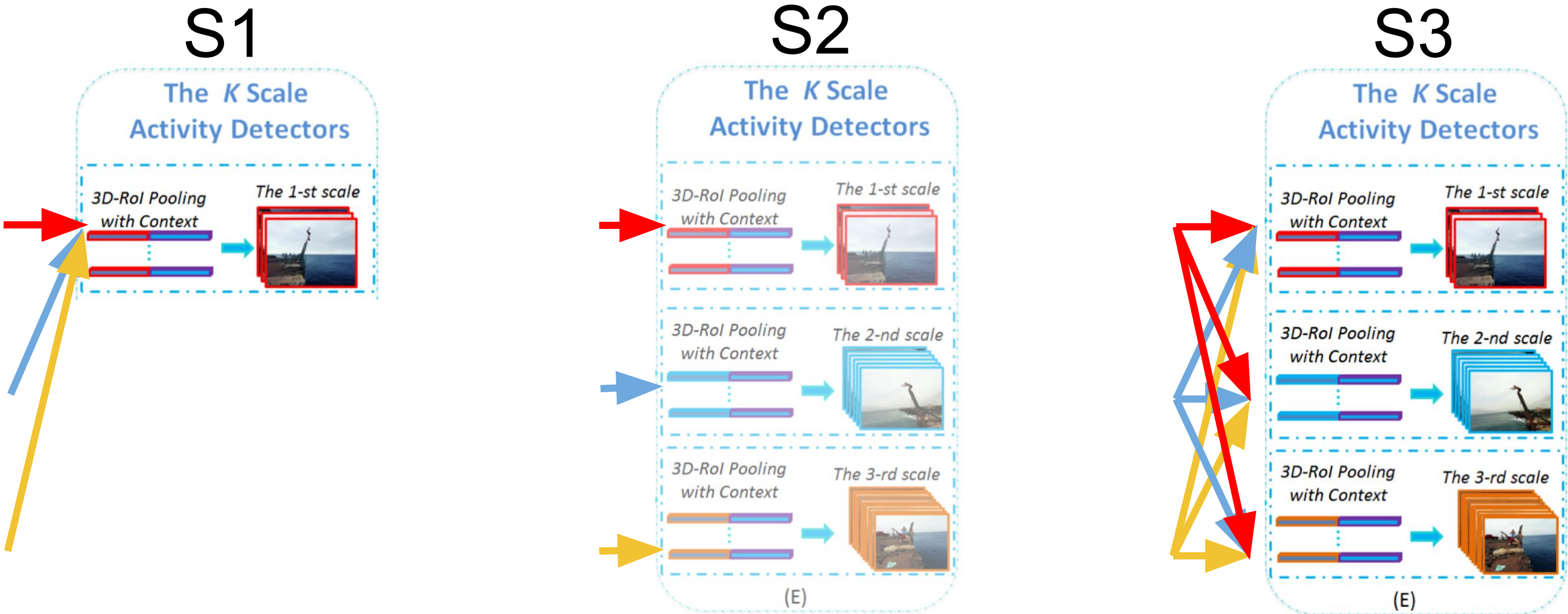
Method	0.1	0.2	0.3	0.4	0.5
RC3D [33]	54.5	51.5	44.8	35.6	28.9

Are multi-scale proposals useful?



Method	0.5	0.75	0.95	Average
RC3D [33]	26.33	10.46	1.25	12.71
MS(MAX)(S1)	27.65	13.93	1.12	14.91
MS(CONV)(S1)	28.01	13.80	1.20	15.12

How do we pick the scale at which to classify a given proposal?



Method	0.5	0.75	0.95	Average
MS(CONV)(S1)(CTX)	32.57	16.92	1.07	17.89
MS(CONV)(S2)(CTX)	31.89	17.23	1.16	17.72
MS(CONV)(S3)(CTX)	32.92	18.36	1.13	18.46

Both results together

Method	0.5	0.75	0.95	Average
RC3D [33]	26.33	10.46	1.25	12.71
MS(MAX)(S1)	27.65	13.93	1.12	14.91
MS(CONV)(S1)	28.01	13.80	1.20	15.12
MS(MAX)(S1)(CTX)	31.81	17.05	1.06	17.58
MS(CONV)(S1)(CTX)	32.57	16.92	1.07	17.89
MS(CONV)(S2)(CTX)	31.89	17.23	1.16	17.72
MS(CONV)(S3)(CTX)	32.92	18.36	1.13	18.46

Both results together

ABSOLUTE	No Multi-Scale	Multi-Scale
No Context	**12.71**	**15.01**
Context	**??**	**17.91**

RELATIVE	No Multi-Scale	Multi-Scale
No Context	**0.0**	**2.3**
Context	**??**	**5.2 (2.3+2.9?)**

RC3D | **CMS-RC3D**

TASK REVIEW | MOTIVATING PROBLEMS

NOVELTY | NOVELTY

EXPERIMENTS | **EXPERIMENTS**

Ablation Studies

Evaluations

DISCUSSION I | DISCUSSION II

RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

MOTIVATING PROBLEMS

NOVELTY

EXPERIMENTS

Ablation Studies

Evaluations

DISCUSSION II

THUMOS 2014

Method	0.1	0.2	0.3	0.4	0.5
Karaman *et al*. [16]	4.6	3.4	2.1	1.4	0.9
Wang *et al*. [31]	18.2	17.0	14.0	11.7	8.3
Oneata *et al*. [20]	36.6	33.6	27.0	20.8	14.4
SparseProp [4]	-	-	-	-	13.5
DAPs [9]	-	-	-	-	13.9
SLM [23]	39.7	35.7	30.0	23.2	15.2
FG [35]	48.9	44.0	36.0	26.4	17.1
PSDF [36]	51.4	42.6	33.6	26.1	18.8
S-CNN [25]	47.7	43.5	36.3	28.7	19.0
CDC [24]	-	-	40.1	29.4	23.3
TCN [8]	-	-	-	33.3	25.6
RC3D [33]	54.5	51.5	44.8	35.6	28.9
SS-TAD [1]	-	-	-	45.7	29.2
SSN [37]	66.0	59.4	51.9	41.0	29.8
Our RC3D	57.4	54.9	51.1	43.1	35.8
CMS-RC3D	61.6	59.3	54.7	48.2	40.0

THUMOS 2014

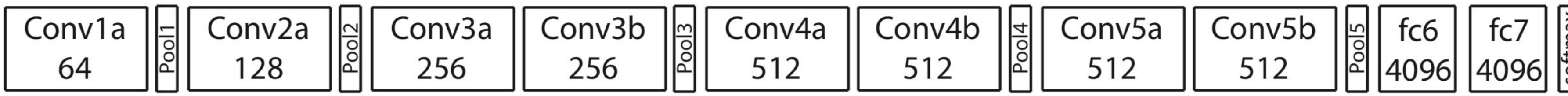
Method	0.1	0.2	0.3	0.4	0.5
PSDF [36]	51.4	42.6	33.6	26.1	18.8
TCN [8]	-	-	-	33.3	25.6
RC3D [33]	54.5	51.5	44.8	35.6	28.9
SSN [37]	66.0	59.4	51.9	41.0	29.8
Our RC3D	57.4	54.9	51.1	43.1	35.8
CMS-RC3D	61.6	59.3	54.7	48.2	40.0

Activity Net (version1.3)

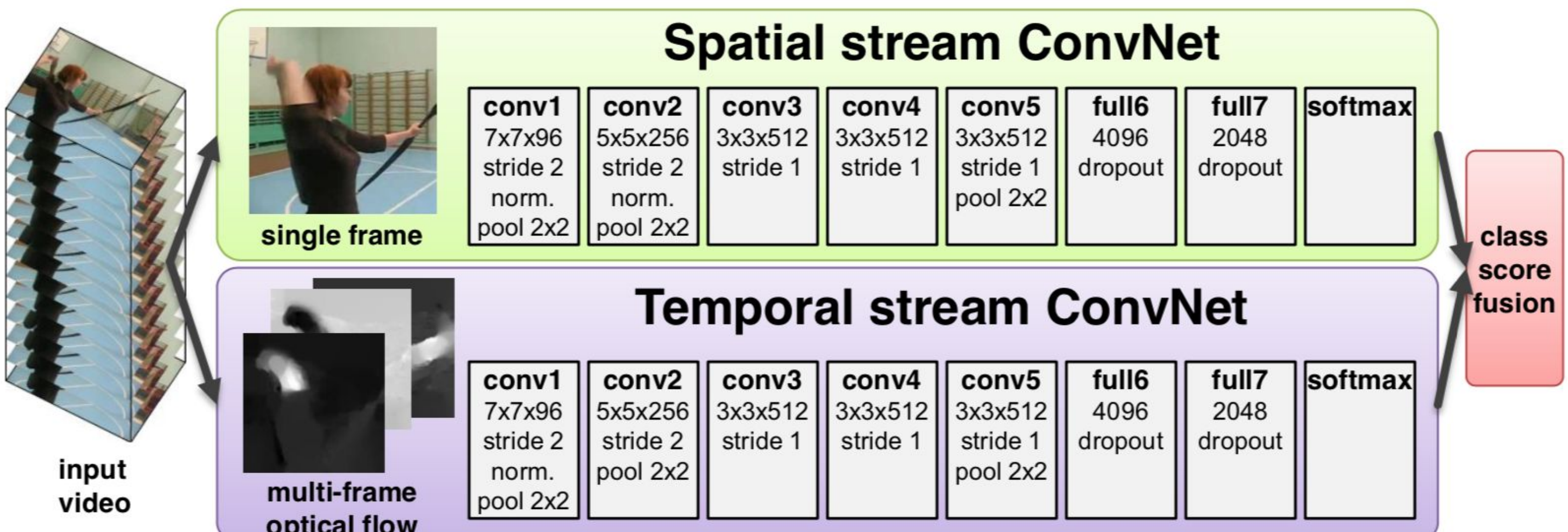
Method	0.5	0.75	0.95	Average
RC3D [33]	26.45	11.47	1.69	13.33
MSN [28]	28.67	17.78	2.88	17.68
TCN [8]	37.49	23.47	4.47	23.58
SSN [37]	43.26	28.70	5.63	28.28
CMS-RC3D	32.79	18.39	1.24	18.68

Shallower Feature Extractor?

C3D



Two-Stream Network



Shallower Feature Extractor?

From the **ORIGINAL** RC3D Paper

	mAP standard	mAP post-process
Random [25]	4.2	4.2
RGB [25]	7.7	8.8
Two-Stream [25]	7.7	10.0
Two-Stream+LSTM [25]	8.3	8.8
Sigurdsson et al. [25]	9.6	12.1
R-C3D (ours)	**12.4**	**12.7**

RC3D

TASK REVIEW

NOVELTY

EXPERIMENTS

DISCUSSION I

CMS-RC3D

MOTIVATING PROBLEMS

NOVELTY

EXPERIMENTS

Ablation Studies

<u>Evaluations</u>

DISCUSSION II

RC3D	**CMS-RC3D**
TASK REVIEW	MOTIVATING PROBLEMS
NOVELTY	NOVELTY
EXPERIMENTS	EXPERIMENTS
DISCUSSION I	**DISCUSSION II**

Lingering Thoughts

It doesn’t seem like the feature extractor is the core reason why TCN and SSN might outperform this system. Perhaps something dataset-specific is at work here?

Do windows with “context” include extra information both before and after? Or just after?

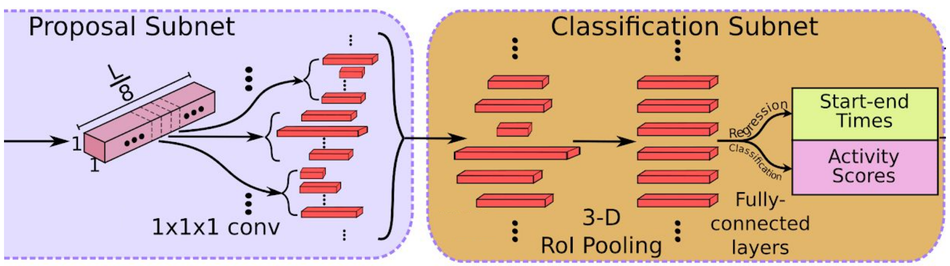
Summary

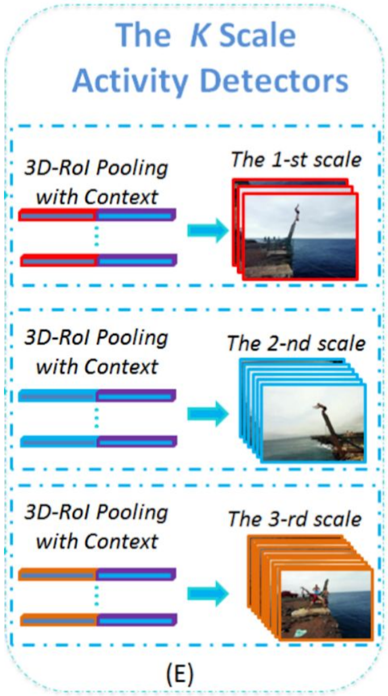
RC3D

CMS-RC3D

Time windows
similar to R-CNN

Same time windows
+ extra context
+ multiple resolutions





Thank You!

(this is in THUMOS2014)

Action recognition
in the spirit of object detection

Nick Turner, Sven Dorkenwald

COS 598 - 04/23/18

PROPOSED OUTLINE:

OVERVIEW

TASK DEFINITION / REVIEW (ask about this)

-What are we trying to do?

-What prior methods have we seen so far?

--C3D Architecture

NOVELTY:

Review R-CNN / Faster R-CNN

-Region proposals -> refined classifications

-R-C3D

EXPERIMENTS

-Training Procedure

-Representing ground truth activities

-Forming the loss function

-Performance Experiments

-Activity Detection Speed

DISCUSSION I (?)

-Lots of references to hand designed features. What's the true issue there?

-

CMS-RC3D

PROPOSED PROBLEMS WITH R-C3D

-Multiple time scales - show an example video

-Use of “contextual information”

NOVELTY

-Multiple time scales

-Contextual information

EXPERIMENTS

-Training Procedure

-Representing ground truth activities

-Forming the loss function

-Ablation Studies

(Do they analyze R-C3D with CTX but without MS anywhere?)

Which variables are most important? Reformat the results table?

DISCUSSION(?)

-Are there other experiment we wish they would do? What’s really most important?

END