Deeper Action

ICCV 2021 Workshop

Person-Context Cross Attention for Spatio-Temporal Action Detection

1st Place Solution to MultiSports Track of DeeperAction Challenge 2021

Zhiqing Ning[1*] Qiaokang Xie[2†*] Wengang Zhou[2] Liangwei Wang[1] Houqiang Li[2]

[1]Huawei Noah's Ark Lab [2]University of Science and Technology of China

NOAH'S ARK LAB

中国科学技术大学 University of Science and Technology of China

*Equal contribution. †Intern at Huawei Noah's Ark Lab.

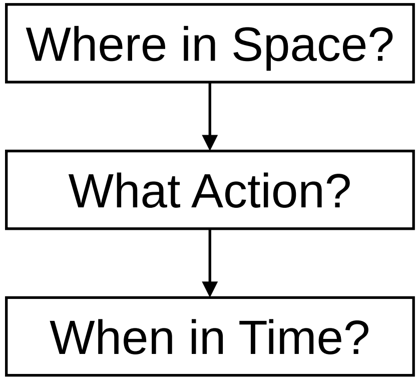
Outline

- □ 1. Overview
- □ 2. Pipeline
- □ 3. Details & Analysis
  - ■ 3.1 Person Detection
  - ■ 3.2 Video Feature Extraction
  - ■ 3.3 Relation Modeling
  - ■ 3.4 Action Prediction
  - ■ 3.5 Training & Inference
- □ 4. Conclusion

1. Overview

□ Spatio-Temporal Action Detection

- Localize actions in both space and time
- Evaluation: Frame mAP and Video mAP



Where in Space?

What Action?

When in Time?

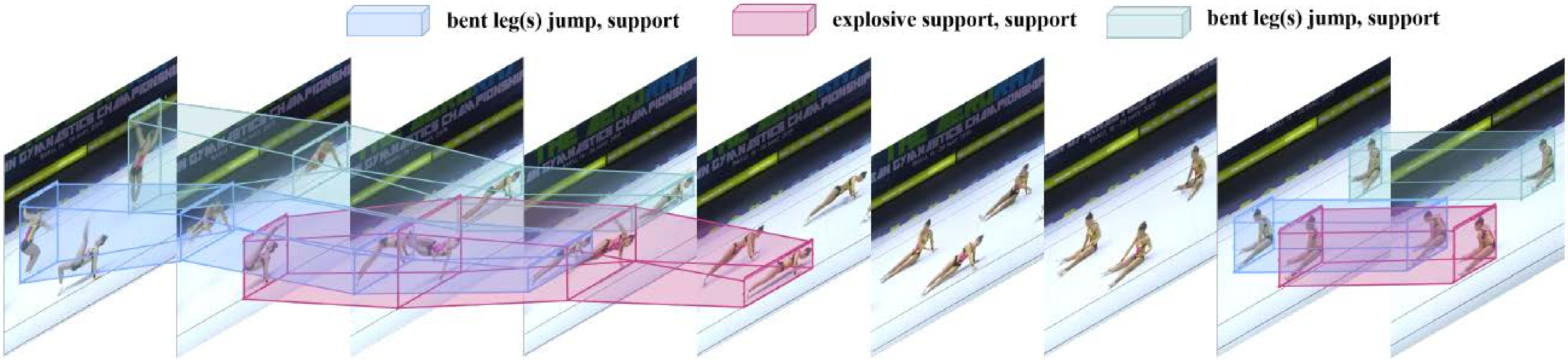
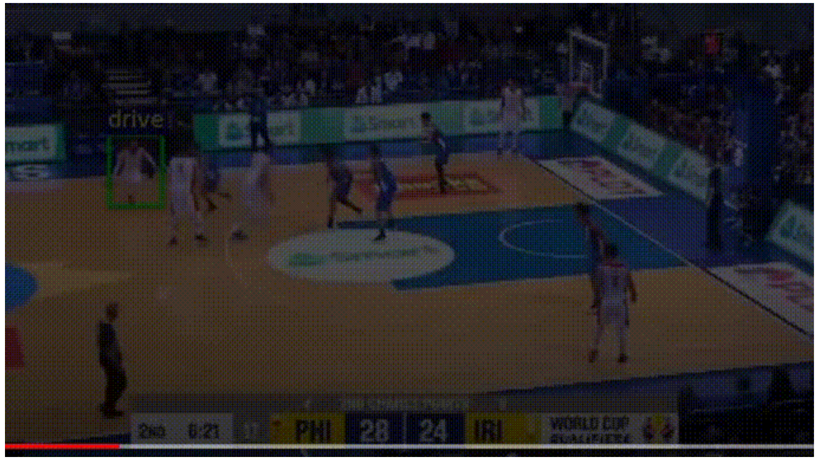

bent leg(s) jump, support
explosive support, support
bent leg(s) jump, support

Figure: Li, Yixuan, et al. "MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions." arXiv preprint arXiv:2105.07404 (2021).

1. Overview

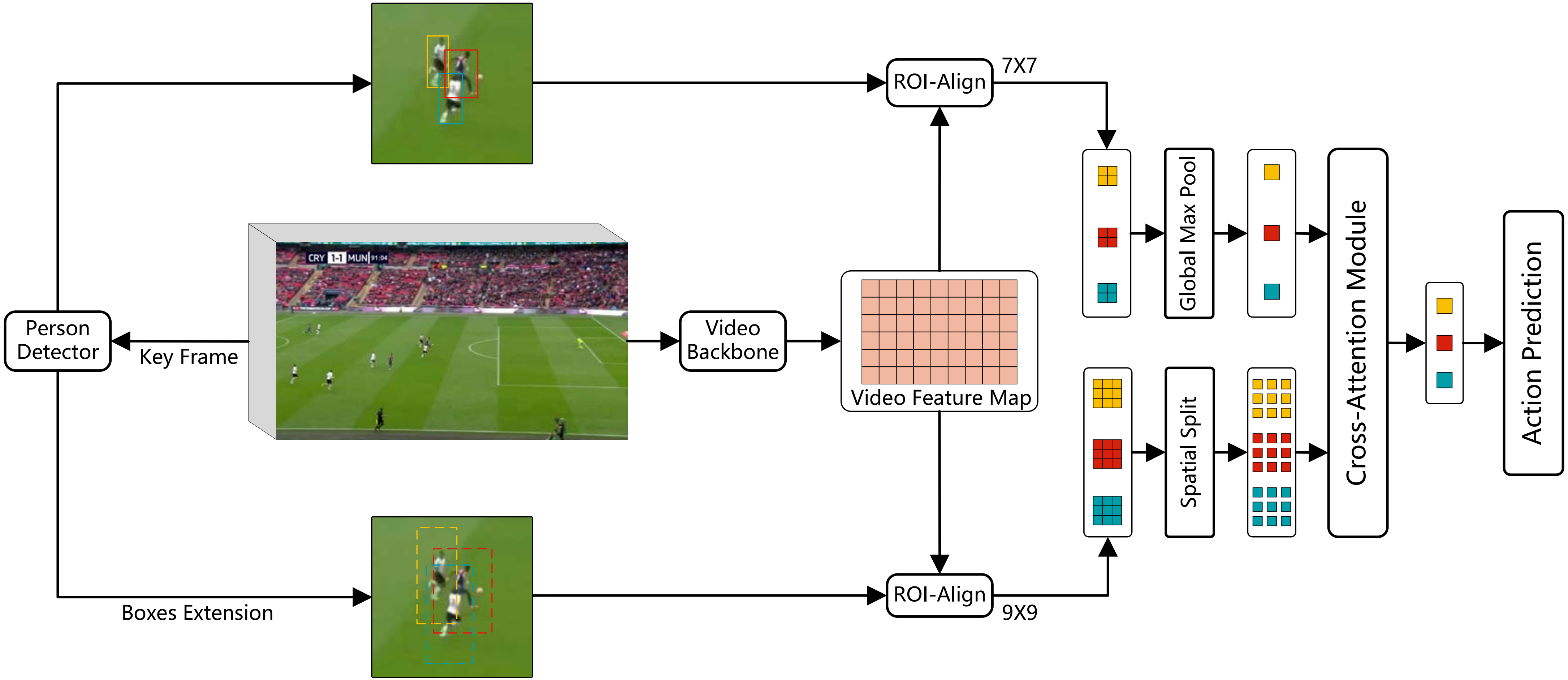
□ MultiSports Dataset

- ■ 66 fine-grained action categories selected from 4 sports
- ■ ~3.2k video clips, ~37.8k action instances
- ■ Action instances labeled at 25 FPS, resulting in ~907k bounding boxes



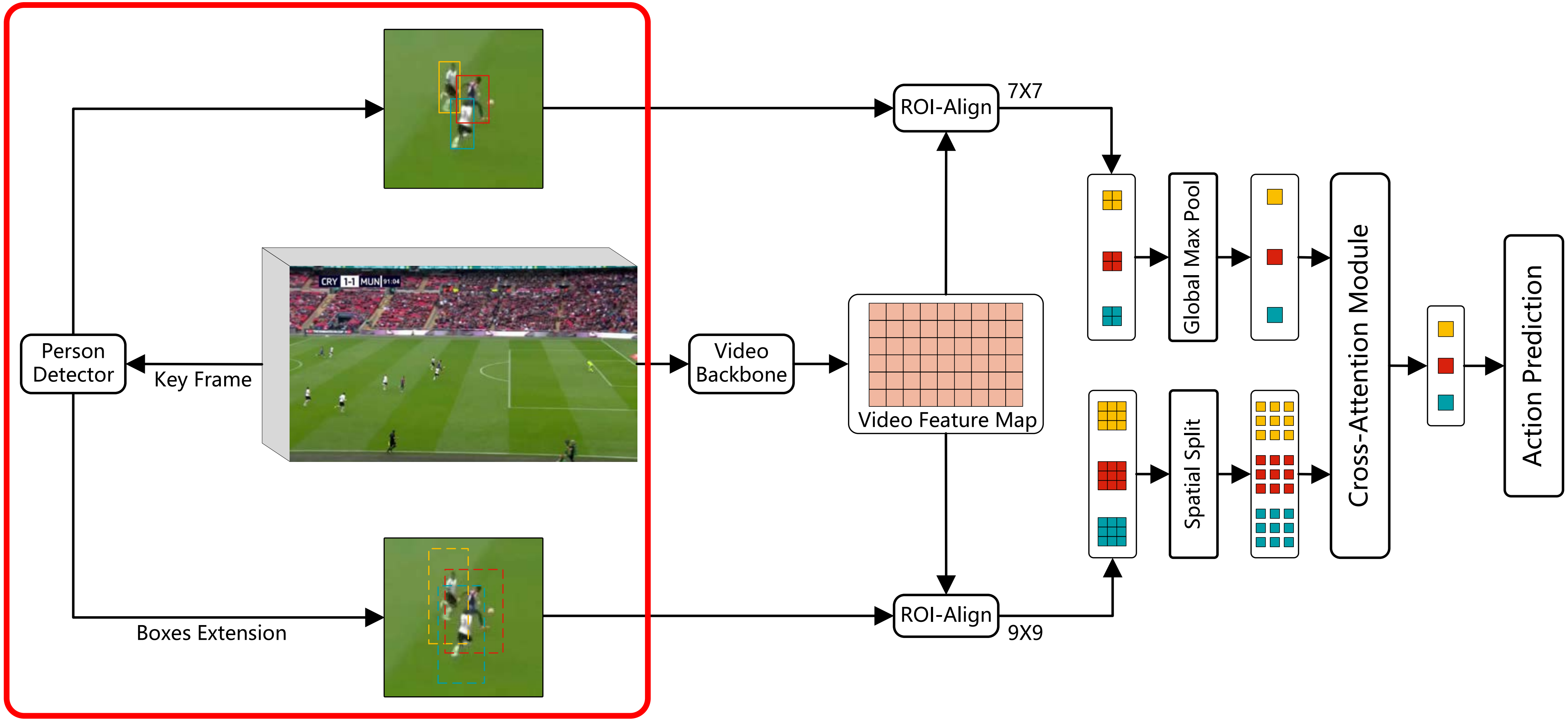
2. Pipeline

Person Detection + Video Feature Extraction + Relation Modeling + Action Prediction



2. Pipeline

Person Detection + Video Feature Extraction + Relation Modeling + Action Prediction

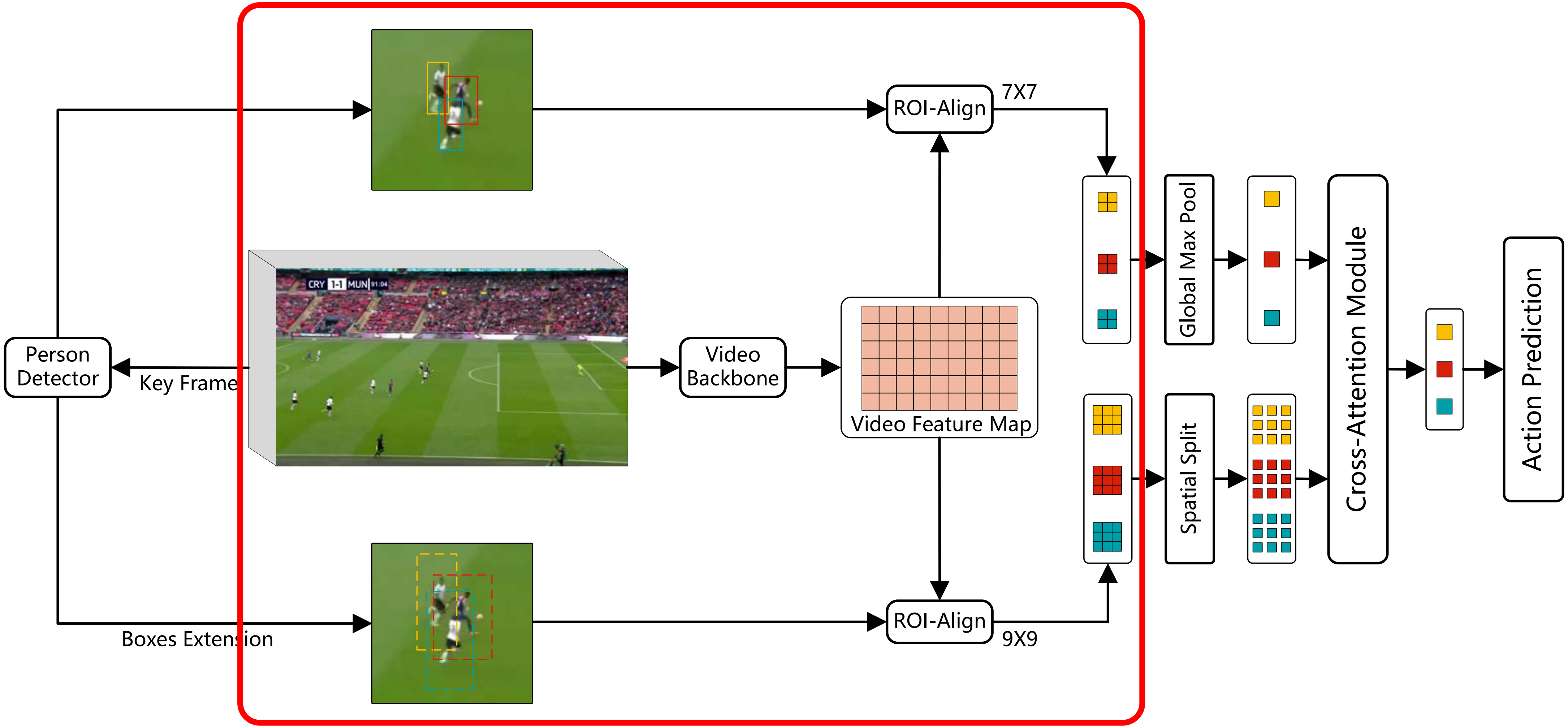


Deeper Action

2. Pipeline

ICCV 2021 Workshop

Person Detection + Video Feature Extraction + Relation Modeling + Action Prediction

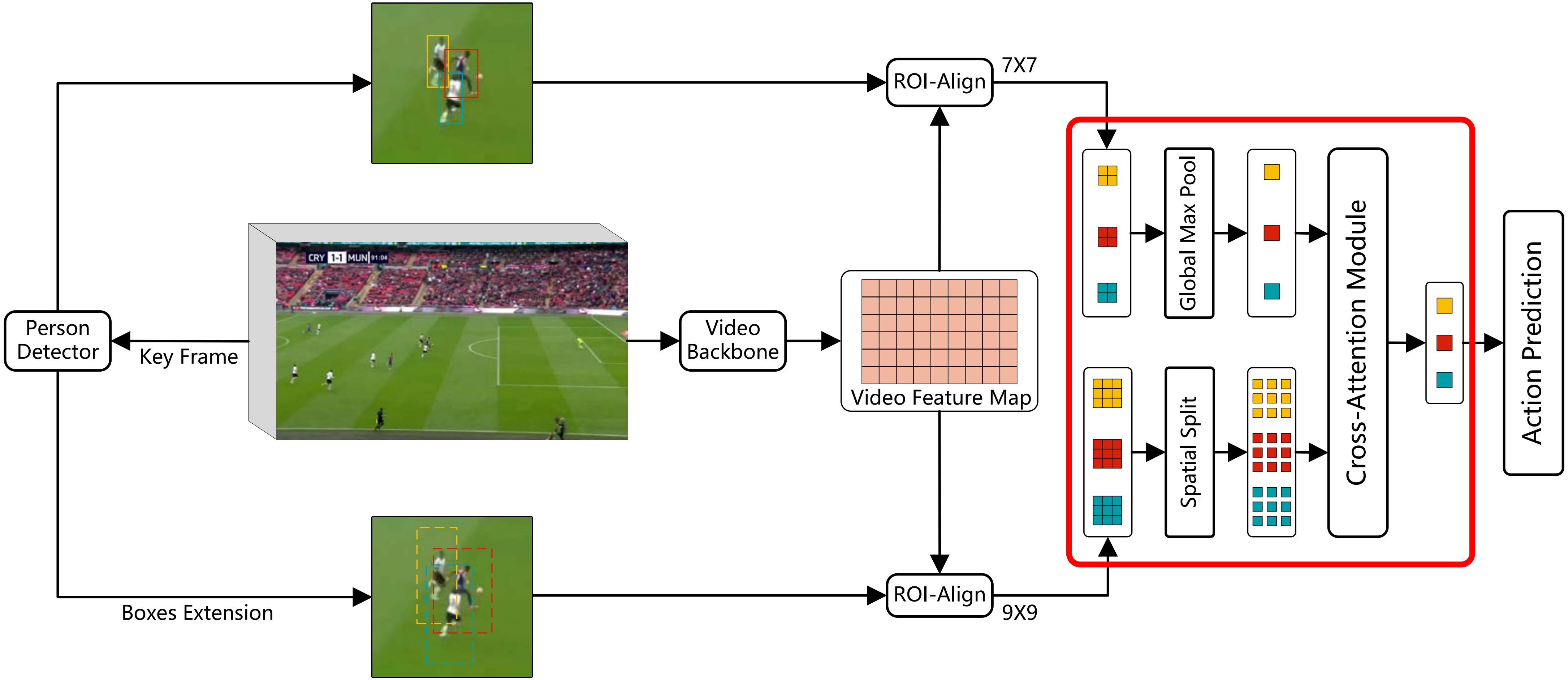


Deeper Action

2. Pipeline

ICCV 2021 Workshop

Person Detection + Video Feature Extraction + Relation Modeling + Action Prediction

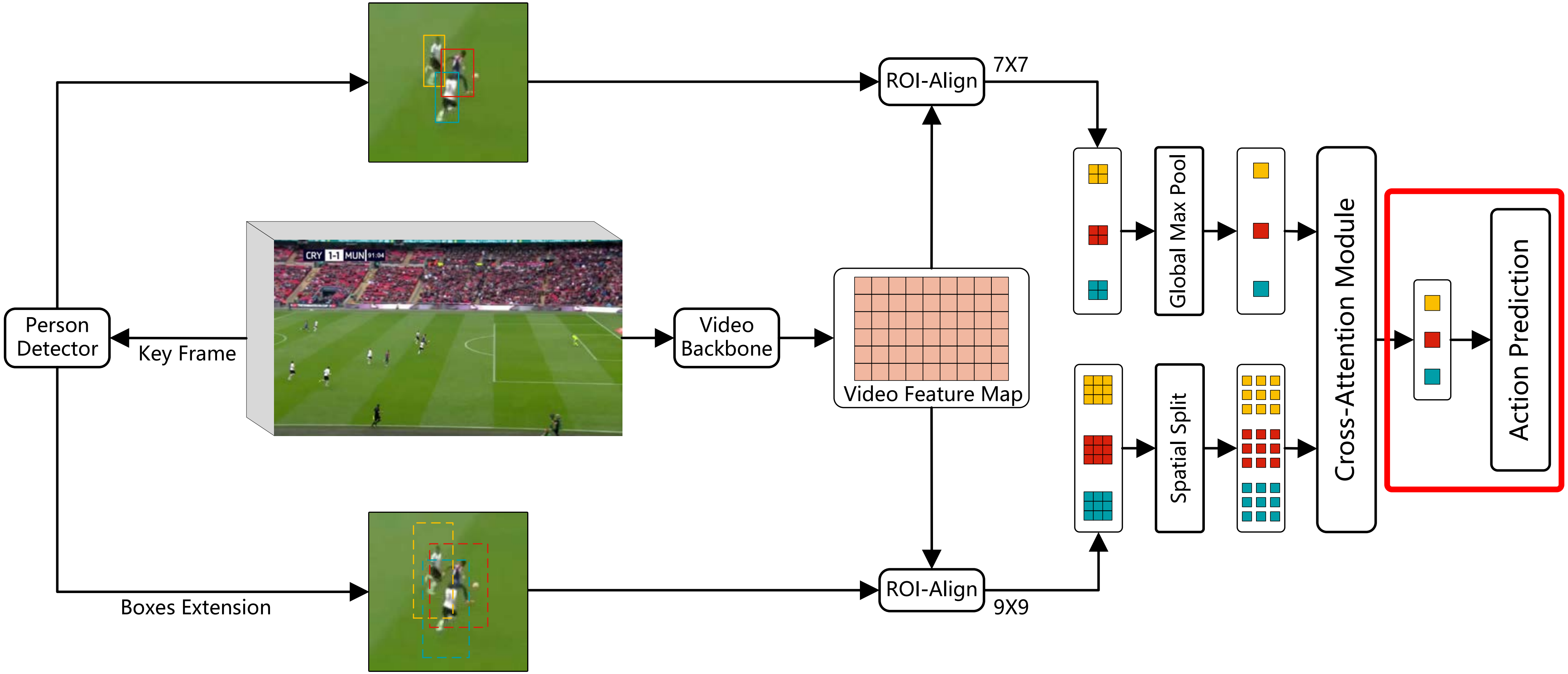


Deeper Action

2. Pipeline

ICCV 2021 Workshop

Person Detection + Video Feature Extraction + Relation Modeling + Action Prediction



Deeper Action

Outline

ICCV 2021 Workshop

□ 1. Overview

□ 2. Pipeline

□ 3. Details & Analysis

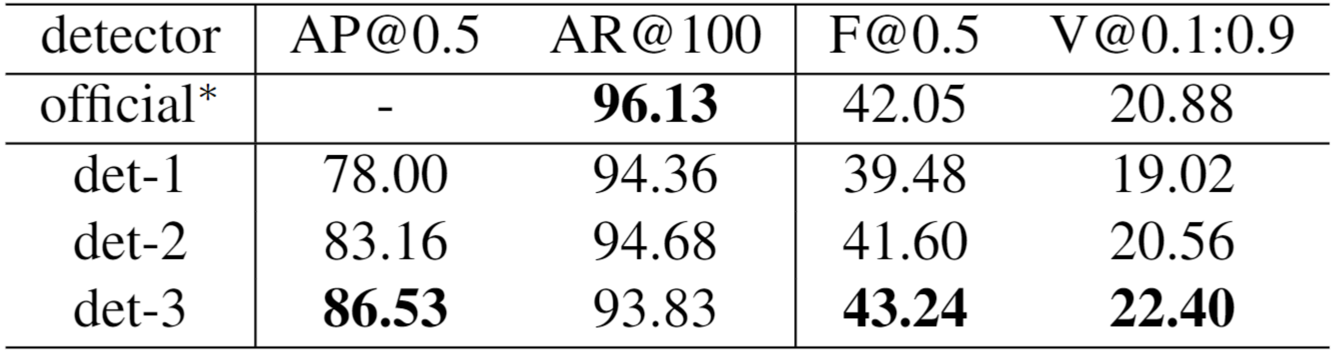
- ■ 3.1 Person Detection
- ■ 3.2 Video Feature Extraction
- ■ 3.3 Relation Modeling
- ■ 3.4 Action Prediction
- ■ 3.5 Training & Inference

□ 4. Conclusion

3. Details & Analysis

□ 3.1 Person Detection

- Faster R-CNN with ResNeXt-101-FPN backbone
  - Pre-trained on ImageNet and COCO person keypoint images
  - Fine-tuned on the training set of MultiSports for higher detection precision



detector	AP@0.5	AR@100	F@0.5	V@0.1:0.9
official*	-	**96.13**	42.05	20.88
det-1	78.00	94.36	39.48	19.02
det-2	83.16	94.68	41.60	20.56
det-3	**86.53**	93.83	**43.24**	**22.40**

Results on val set. AP and AR are only evaluated on frames with annotations.
AP@0.5: average precision of person detections with IoU > 0.5; AR@100: average recall with top 100 detections each frame.

=> Higher AP gives better performance !

* : Official Person Boxes: https://github.com/MCG-NJU/MultiSports

3. Details & Analysis

□ 3.2 Video Feature Extraction

- Backbone: SlowFast*
  - Two pathways with different FPS are used to capture spatial semantics and motion information.
  - Depth: R101
  - $T \times \tau = 8 \times 8$
  - $\alpha = 4$
  - Pretrained on Kinetics-600 dataset.
- The video backbone is used to extract 3D features maps

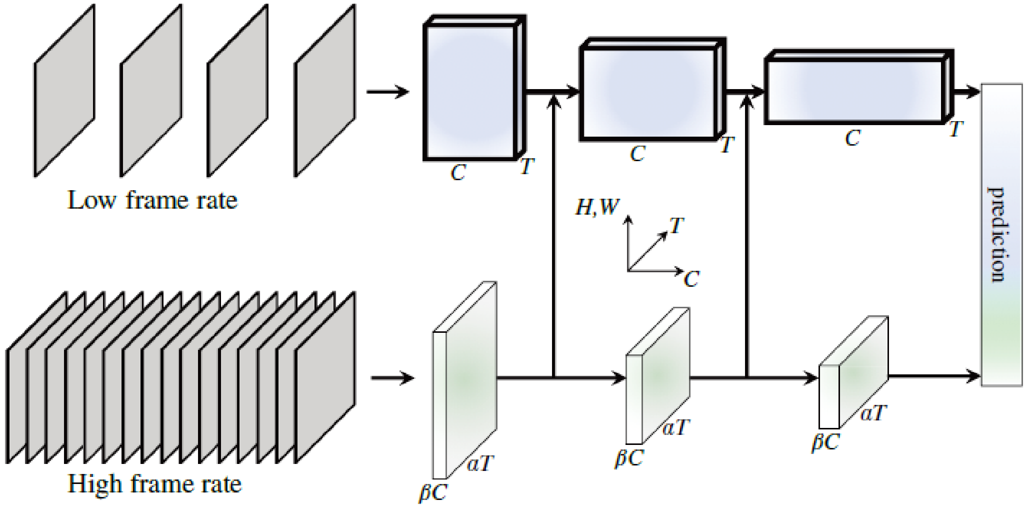


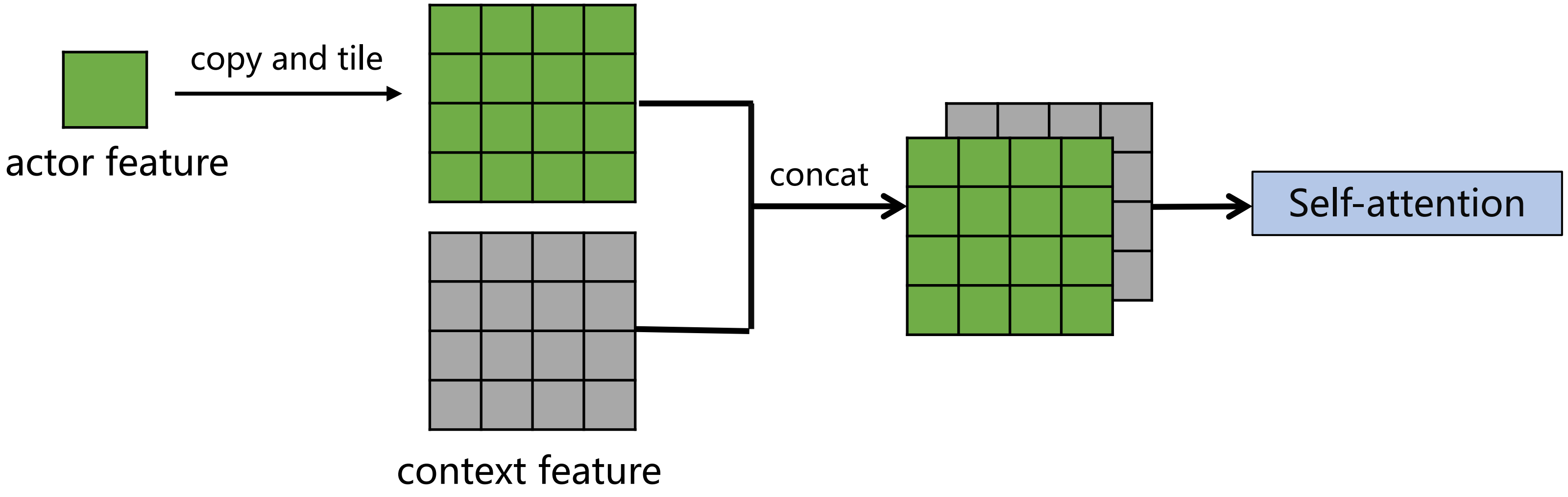
Figure 1. **A SlowFast network** has a low frame rate, low temporal resolution *Slow* pathway and a high frame rate, $\alpha\times$ higher temporal resolution *Fast* pathway. The Fast pathway is lightweight by using a fraction (β, *e.g.*, 1/8) of channels. Lateral connections fuse them.

* and figure : Feichtenhofer, Christoph, et al. "Slowfast networks for video recognition." ICCV. 2019.

3. Details & Analysis

□ 3.3 Relation Modeling

- How to utilize spatio-temporal context for relation modeling.
  - Alphaction[1]: person-person & person-object
  - ACAR[2]: person-context



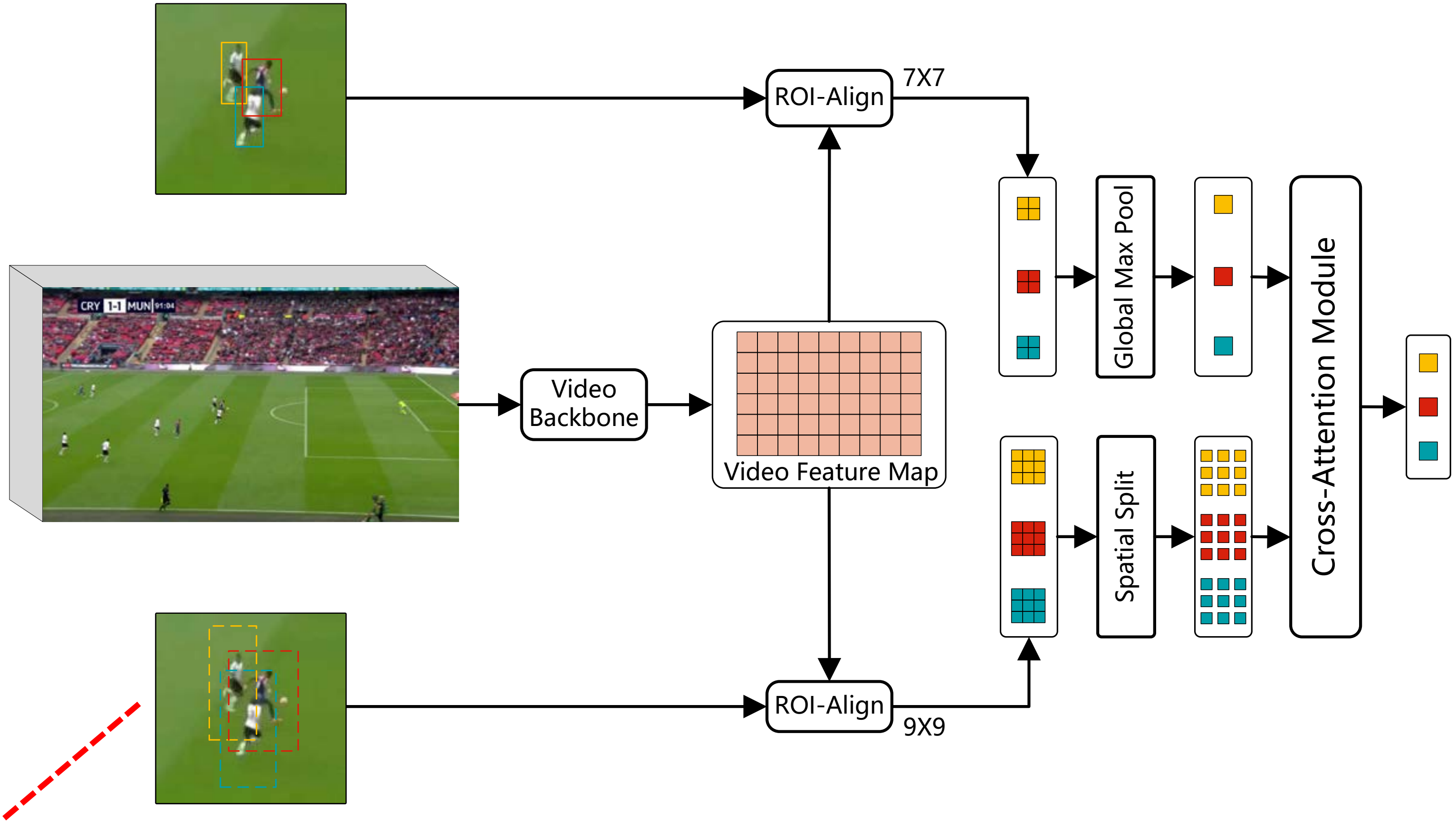
ACAR Head

[1] : Tang, Jiajun, et al. "Asynchronous interaction aggregation for action detection." ECCV, 2020.
[2] : Pan, Junting, et al. "Actor-context-actor relation network for spatio-temporal action localization.“ CVPR, 2021.

3. Details & Analysis

□ 3.3 Relation Modeling

- Action is usually related to the surroundings near the person in MultiSports.
- For computational efficiency consideration.

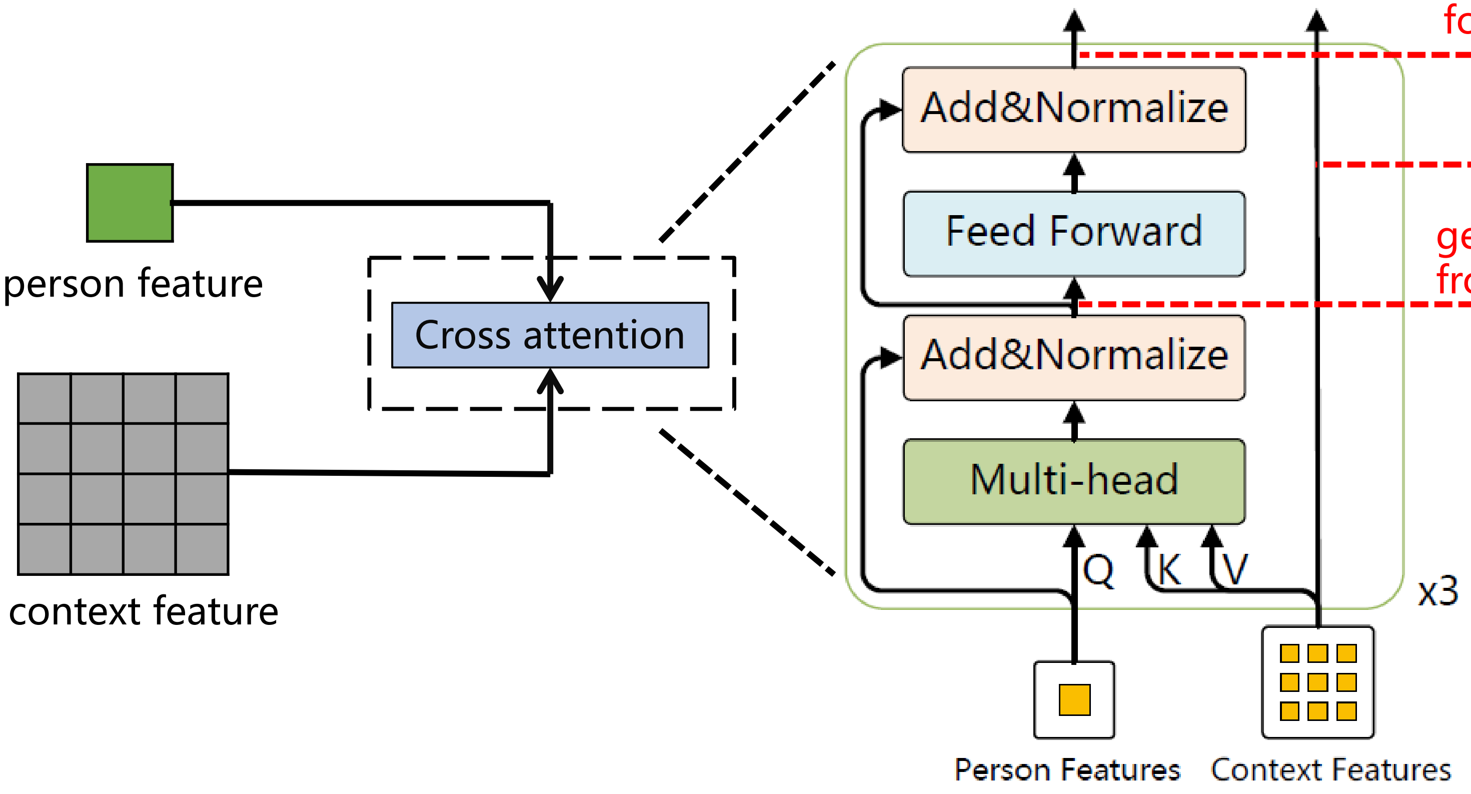


Expand the box scale to twice the previous size

3. Details & Analysis

□ 3.3 Relation Modeling

■ Person-Context Cross Attention



forward and keep attending to the context

preserve rich representations

get relations from context

$$Q^0 = A_i, \qquad (1)$$

$$\boxed{X^j} = LN(MultiHead(Q^{j-1}, \boxed{C_i, C_i}) + Q^{j-1}), \qquad (2)$$

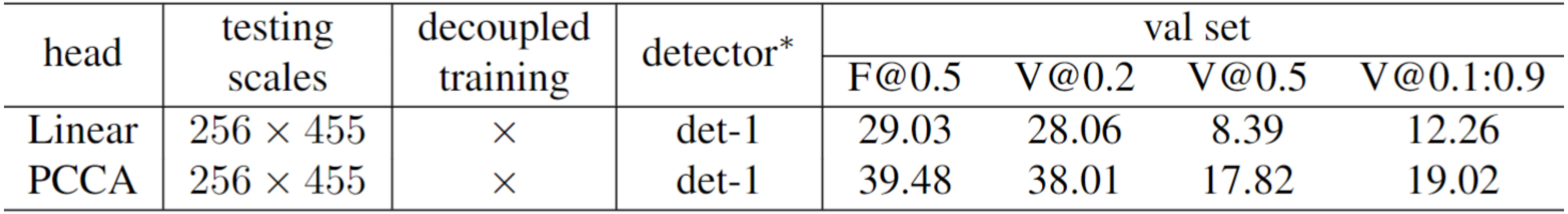
$$\boxed{Q^j} = LN(FFN(X^j) + X^j), \qquad (3)$$

3. Details & Analysis

□ 3.3 Relation Modeling

■ Influence of Person-Context Cross Attention

- ➢ Frame AP@0.5: +10.45
- ➢ Video AP@0.1:0.9: +6.76



head	testing scales	decoupled training	detector*	val set			
				F@0.5	V@0.2	V@0.5	V@0.1:0.9
Linear	256 × 455	×	det-1	29.03	28.06	8.39	12.26
PCCA	256 × 455	×	det-1	39.48	38.01	17.82	19.02

Results on val set. Backbone SlowFast R101 8x8, scale 256x455.

Deeper Action

3. Details & Analysis

ICCV 2021 Workshop

□ 3.4 Action Prediction

- Classification: Sigmoid + BCE
- Long-tailed distribution in MultiSports：Decoupled learning

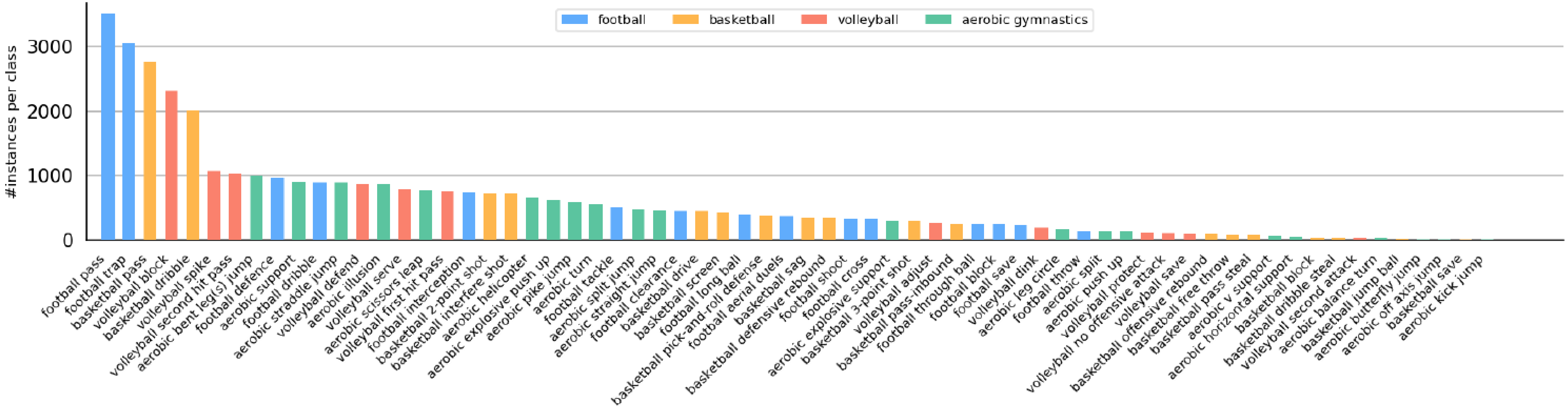


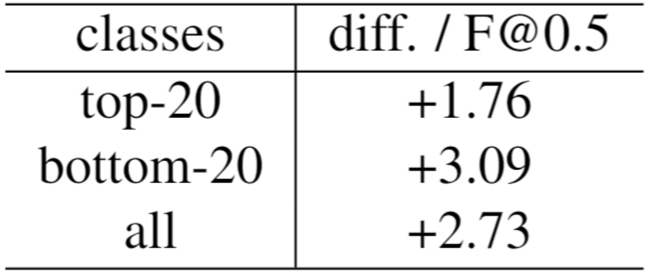
Figure: Li, Yixuan, et al. "MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions." arXiv preprint arXiv:2105.07404 (2021).

3. Details & Analysis

□ 3.4 Action Prediction

- Classification: Sigmoid + BCE
- Long-tailed distribution in MultiSports : Decoupled learning
  - Phase 1: Standard random data sampling for normal representation learning.
  - Phase 2: Class-balanced data sampling for classifier learning.

(freezing the parameters of the model except the final classifier)

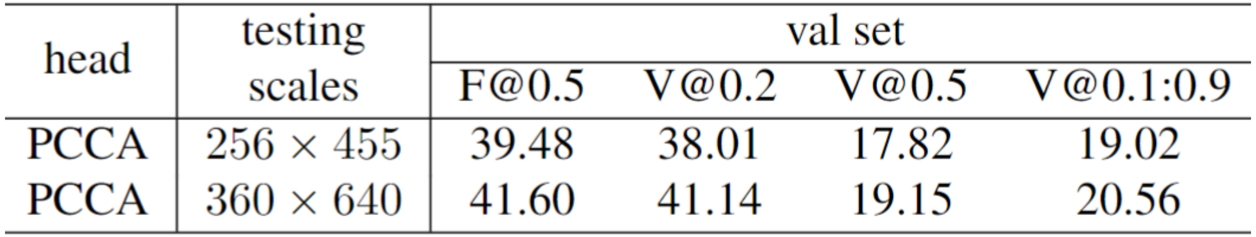


classes	diff. / F@0.5
top-20	+1.76
bottom-20	+3.09
all	+2.73

Influence of decoupled learning on val set.
Classes are ranked by their numbers of labeled samples

3. Details & Analysis

□ 3.5 Training & Inference



head	testing scales	val set			
		F@0.5	V@0.2	V@0.5	V@0.1:0.9
PCCA	256 × 455	39.48	38.01	17.82	19.02
PCCA	360 × 640	41.60	41.14	19.15	20.56

■ Training

- ➢ Spatial scales: {256x455, 360x640}
- ➢ SGD, with a batch size {32 for 256x455, 24 for 360x640}
- ➢ Base lr 0.1, linear warm-up (3 epochs), weight decay 1e-4 and Nesterov momentum of 0.9
- ➢ Stepwise learning rate at epoch [5, 8, 10] by a factor of 0.1
- ➢ Max epochs: 12 for training on train set only, and 15 for train+val set

■ Inference

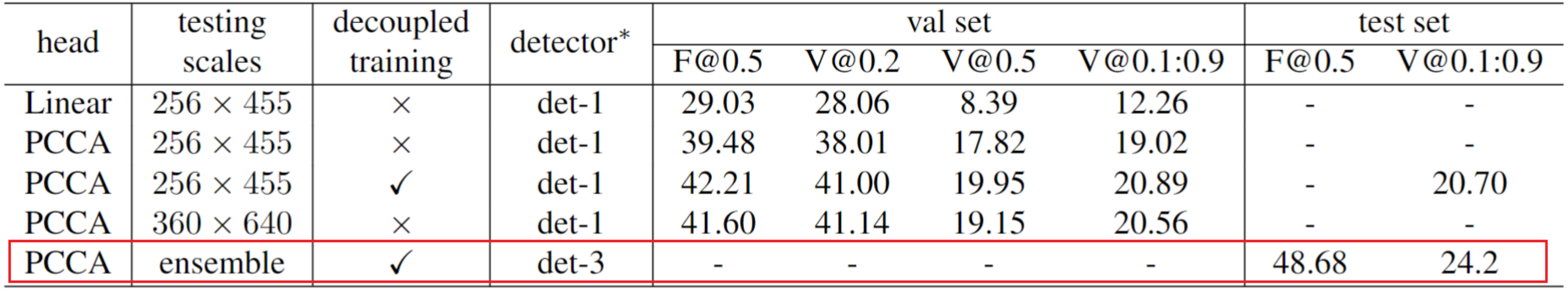
- ➢ On person detections with confidence ≥ 0.6
- ➢ Tube linking: the same link algorithm as MOC* with minimal modifications adapted for frame-level predictions.

* : Li, Yixuan, et al. "Actions as moving points." ECCV, 2020.

4. Conclusion

□ Final results

- Combine train set and val set for training
- Ensemble: Two spatial scales {256x455, 360x640} results with horizontal flips



head	testing scales	decoupled training	detector*	val set				test set	
				F@0.5	V@0.2	V@0.5	V@0.1:0.9	F@0.5	V@0.1:0.9
Linear	256 × 455	×	det-1	29.03	28.06	8.39	12.26	-	-
PCCA	256 × 455	×	det-1	39.48	38.01	17.82	19.02	-	-
PCCA	256 × 455	✓	det-1	42.21	41.00	19.95	20.89	-	20.70
PCCA	360 × 640	×	det-1	41.60	41.14	19.15	20.56	-	-
PCCA	ensemble	✓	det-3	-	-	-	-	48.68	24.2

□ Future work

- How to utilize the clear temporal boundaries in Multisports?

Q&A

Thanks for watching!