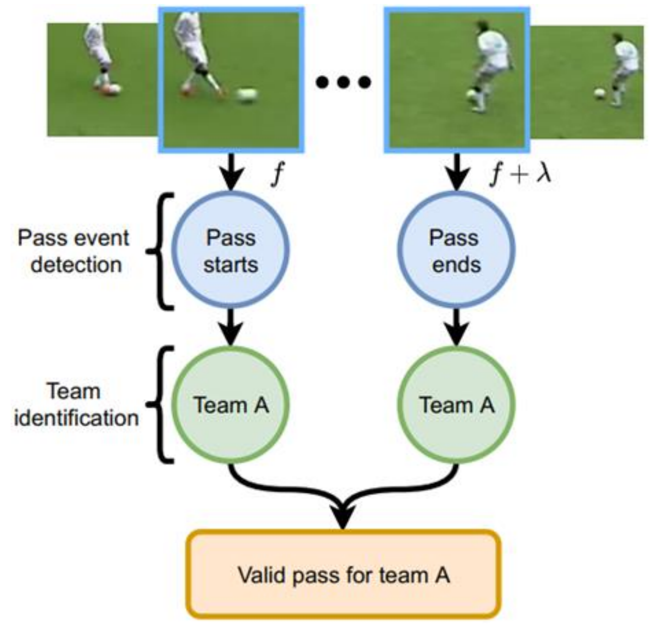
Watch and Act: Dual Interacting Agents for Automatic Generation of Possession Statistics in Soccer

Saikat Sarkar[1], Dipti Prasad Mukherjee[2], Amlan Chakrabarti[1]

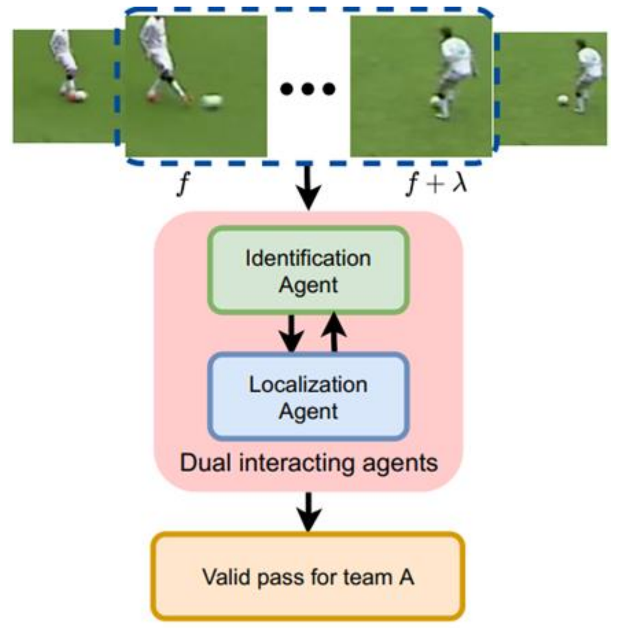
[1]University of Calcutta, [2]Indian Statistical Institute

1. Overview

- Existing pass detection methods [1] follow two steps, pass event detection followed by team identification
- This two-step process is complex and irrecoverable to errors
- We propose a dual interacting agent based model for single-step pass detection
- $Possession\ stat\ of\ team\ i = \frac{\#Valid\ passes\ by\ team\ i}{\#Valid\ passes\ by\ both\ teams}$



Existing approach



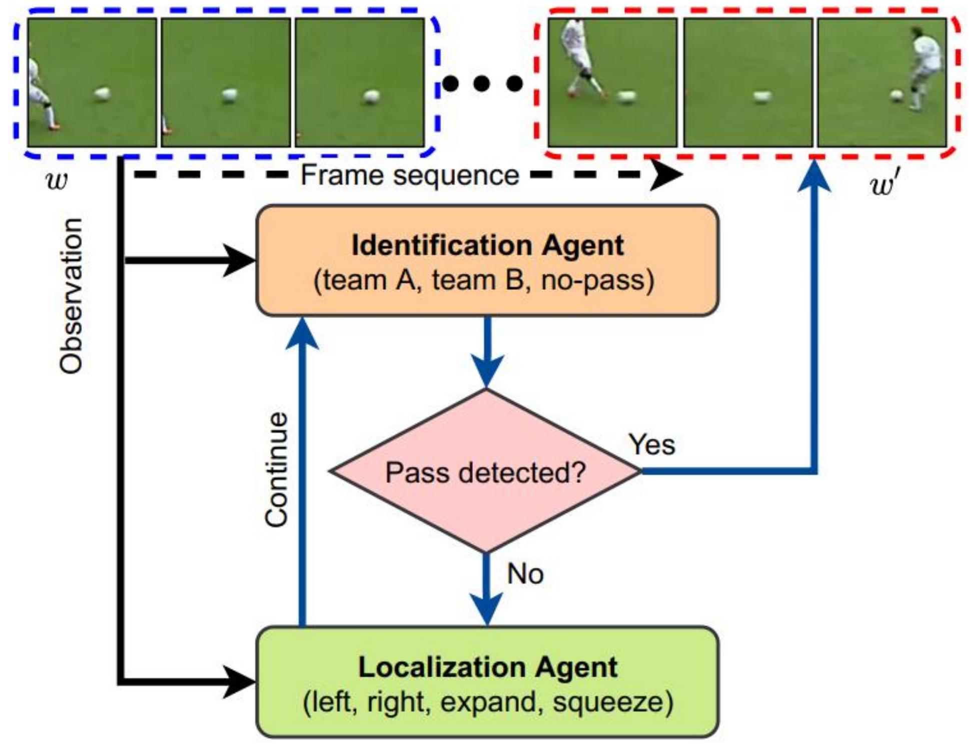
Proposed approach

2. Proposed method

Soccer video → Temporal window initialization → Pass detection using dual interacting agents → Possession statistics generation

3. Flow chart

1. Identification agent decides if the temporal window w contains a pass
2. If no-pass, the localization agent moves and/or rescales w to w'
3. If a pass is detected, w is repositioned



4. Localization agent

- **Task**: To localize a pass
- **Actions**: $a_L = \{left, right, expand, squeeze\}$
- **Reward**:

$$R_L(s, a_L) = \begin{cases} +1 & \text{if IoU}(w', w_g) \geq \tau, \\ +0.1 & \text{else if} \\ & (D(w, w_g) - D(w', w_g)) \geq 0 \\ -1 & \text{otherwise.} \end{cases}$$

Ground truth window

Threshold

State

Boundary distance

5. Identification agent

- **Task**: To identify a valid pass
- **Actions**: $a_I = \{team-A, team-B, no-pass\}$
- **Reward**:

$$R_I(s, a_I) = \begin{cases} +1 & \text{if IoU}(w, w_g) \geq \tau \text{ AND} \\ & a_I == team(w_g), \\ +0 & \text{if IoU}(w, w_g) < \tau \text{ AND} \\ & a_I == no\text{-}pass, \\ -1 & \text{otherwise.} \end{cases}$$

Team label

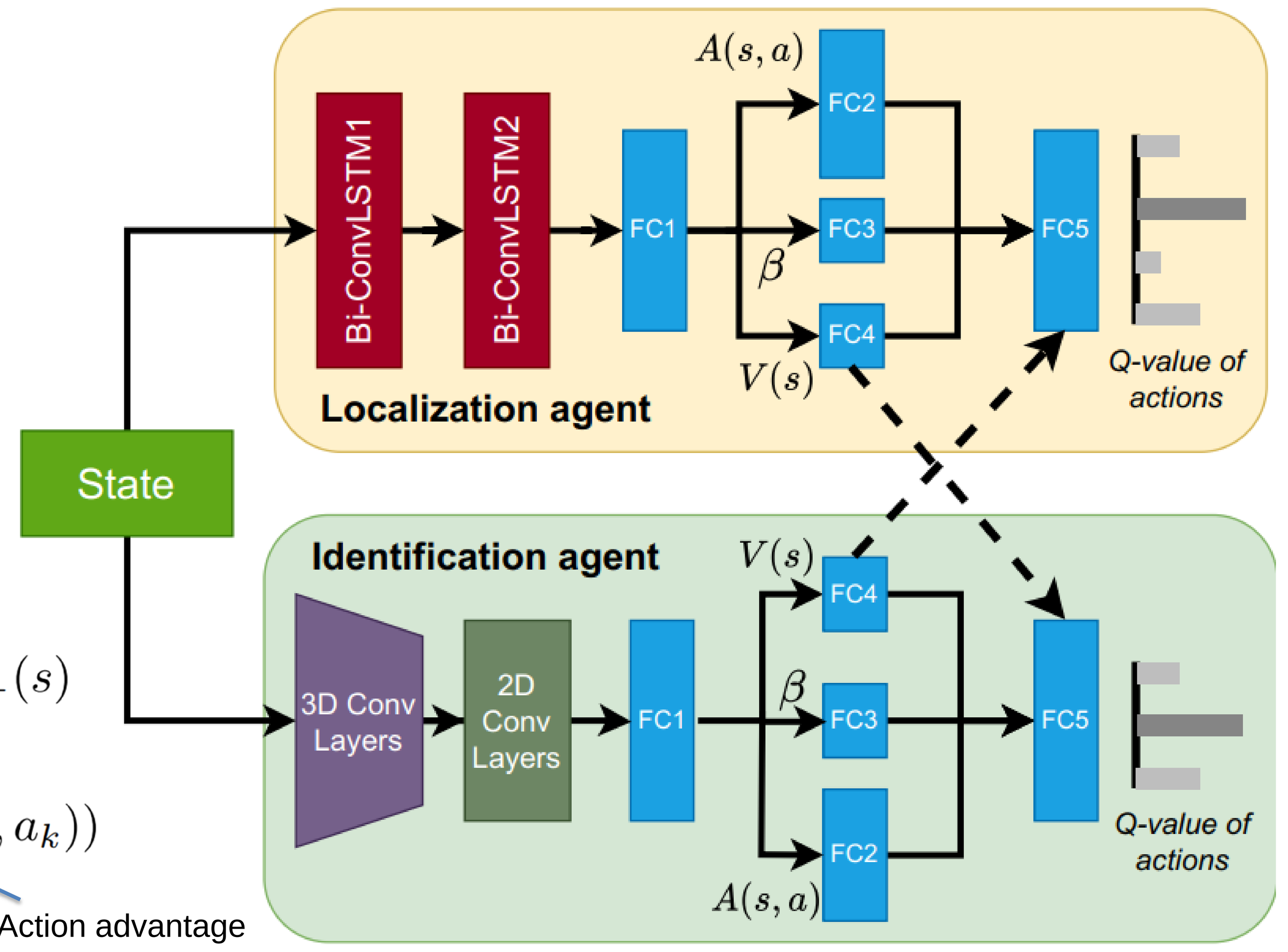
6. Communication between agents

State value of the other agent

State value of the current agent

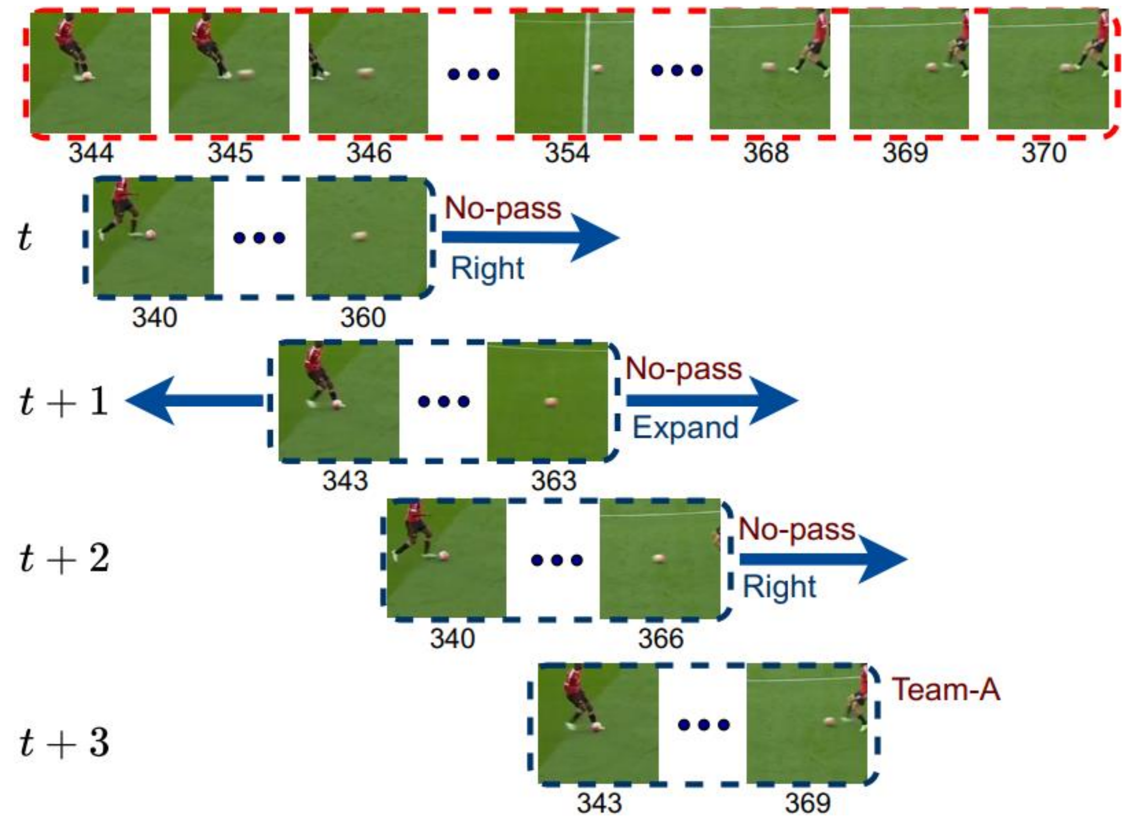
$$Q_i(s, a_i) = \beta V_i(s) + (1-\beta) V_{i-}(s) + (A_i(s, a_i) - \frac{1}{k}\sum_{a_k} A_i(s, a_k))$$

Action advantage



7. Experimental results

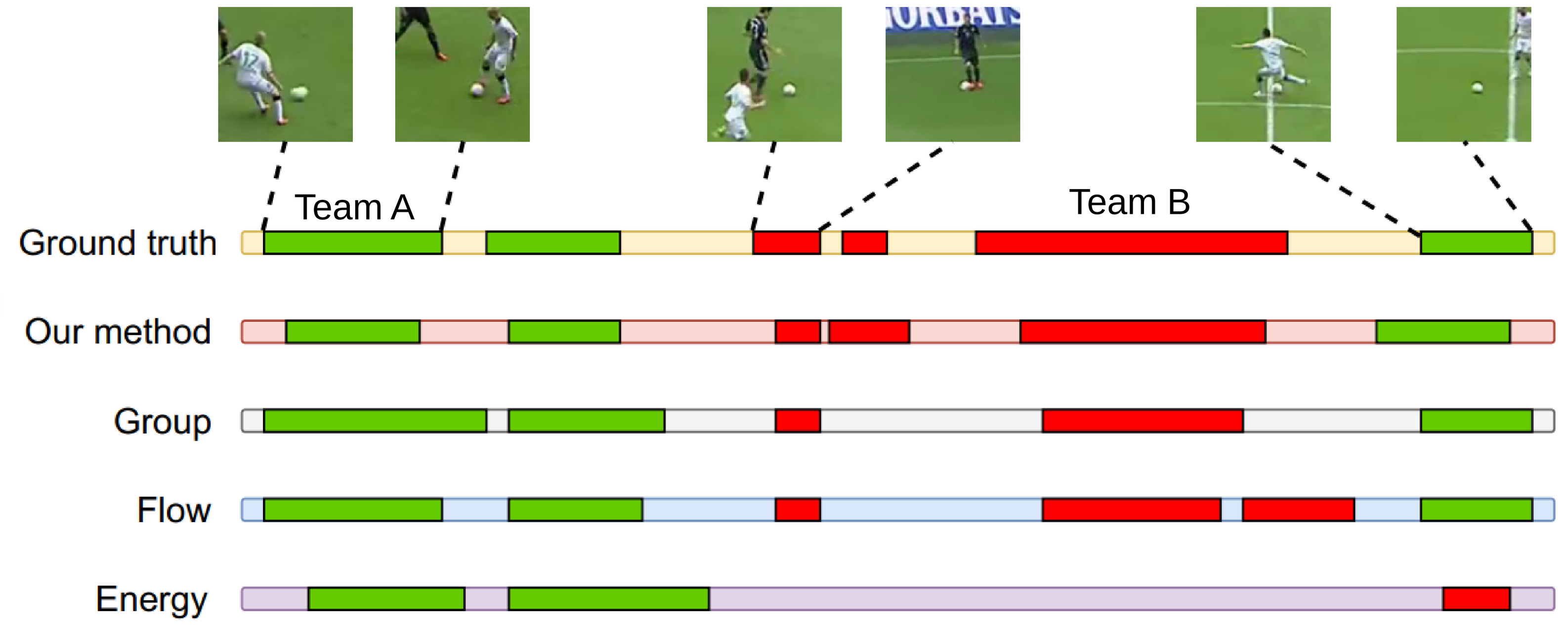
Typical steps of a pass detection



Method	Pass detection error (%)		Possession stat error (%)		Processing
	team-A	team-B	team-A	team-B	time (sec)
Ours	20.5	16.4	13.3	13.4	0.05 (GPU)
Group	11.8	24.0	11.7	12.5	21.8
Flow	26.7	25.9	15.3	15.4	6.86
Energy	33.0	35.4	18.8	18.9	0.08

Comparison of error

Comparison of pass detection



8. References

[1] Saikat Sarkar, Dipti Prasad Mukherjee, and Amlan Chakrabarti. From soccer video to ball possession statistics. *Pattern Recognition*, page 108338, 2022.

[2] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003. PMLR, 2016.

Ball 3D Localization from a Single Calibrated Image

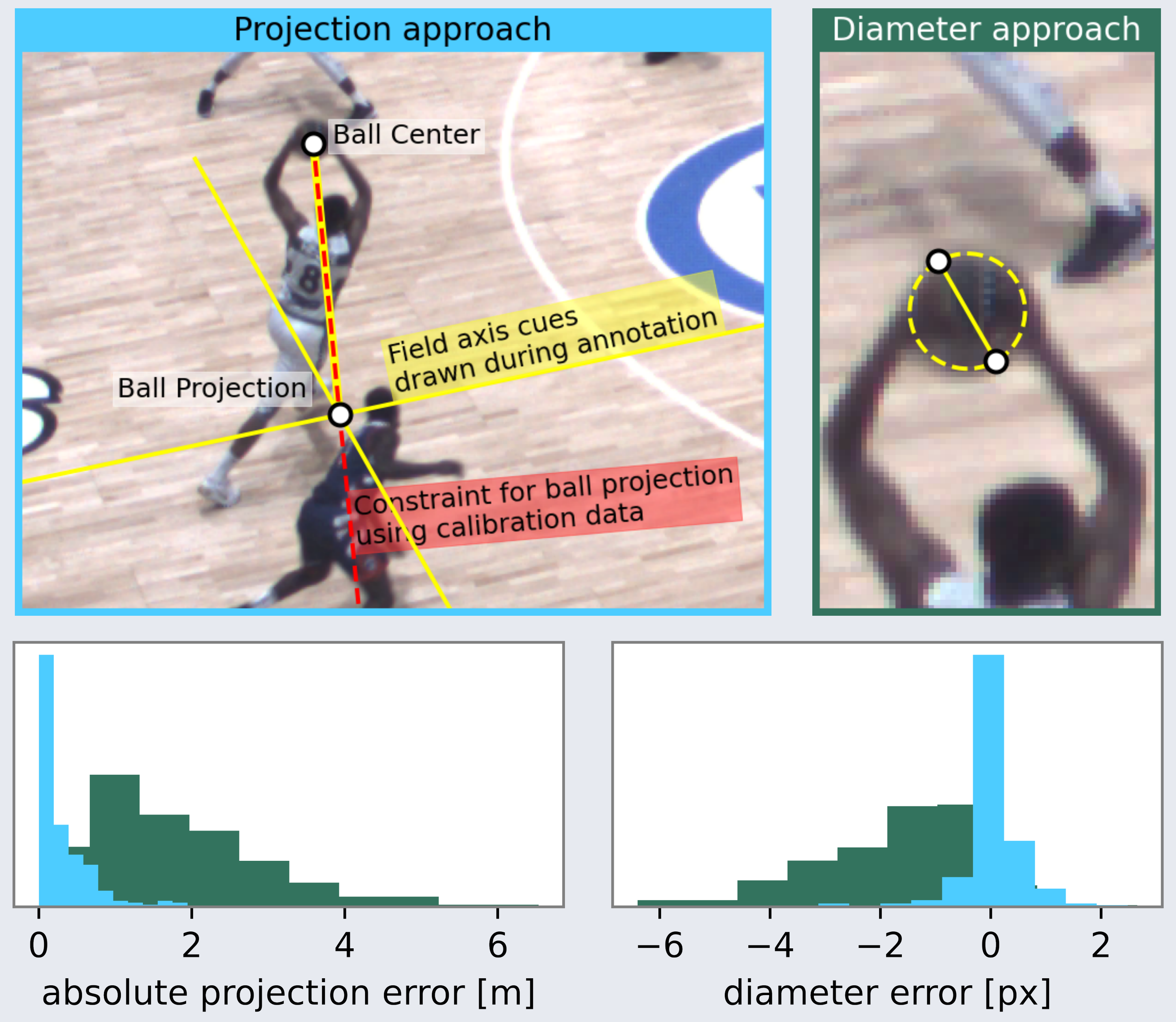
Gabriel Van Zandycke and Christophe De Vleeschouwer

UCLouvain

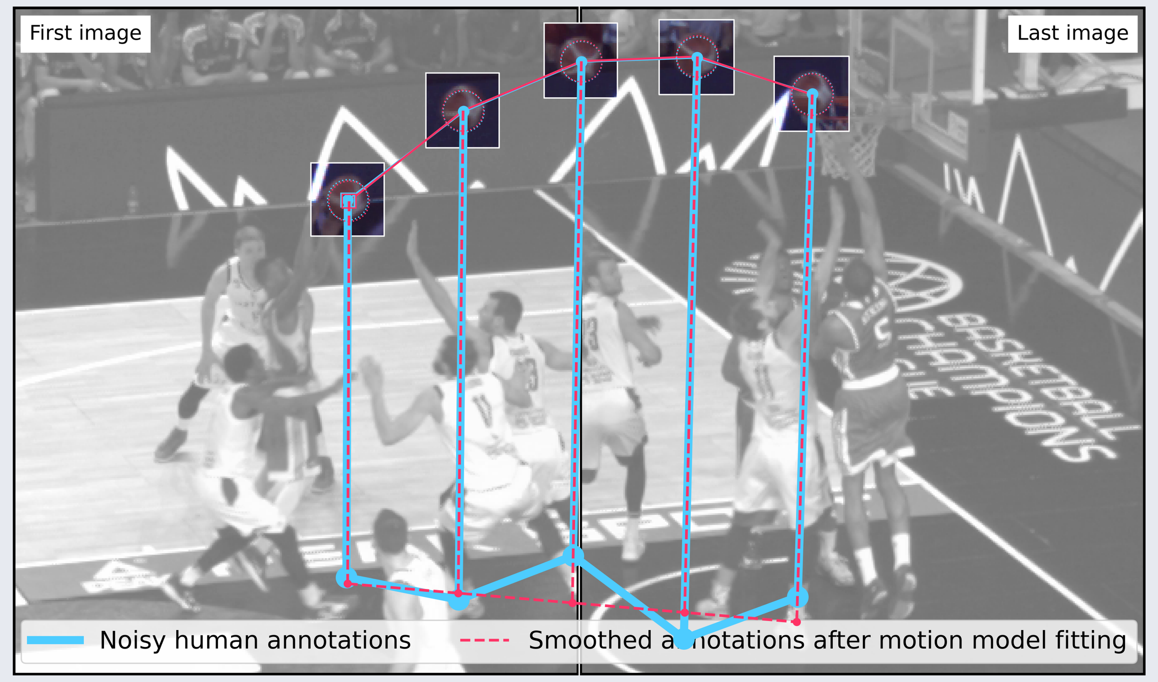
sportradar

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

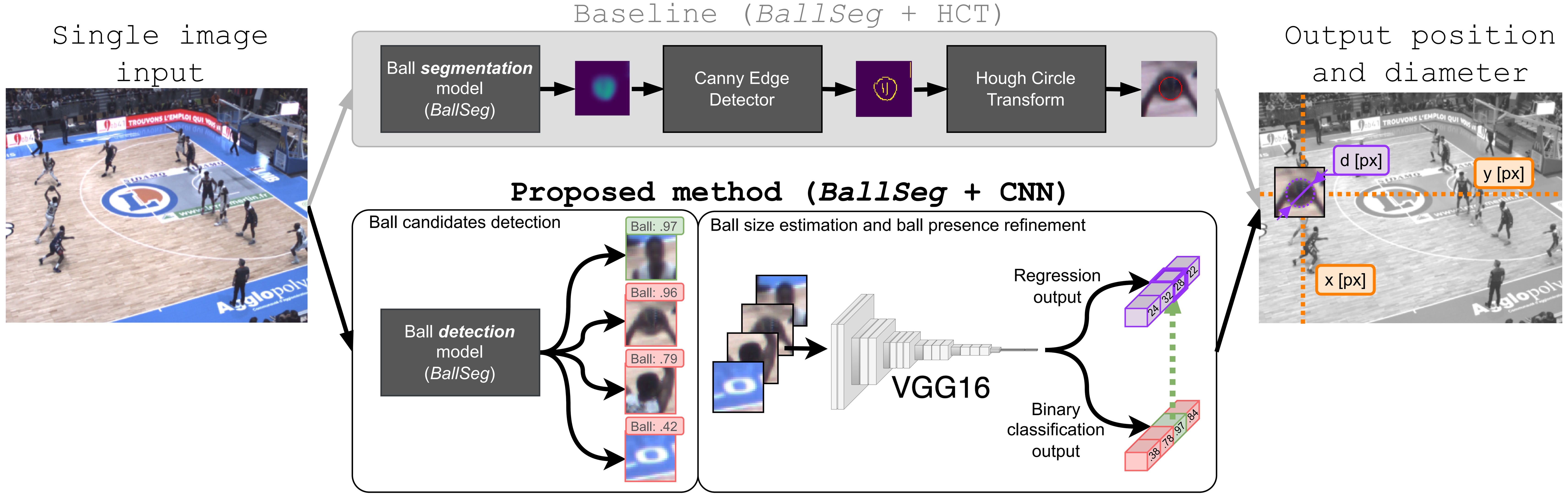
Single Image Ball 3D Annotation



High Quality Ball 3D Evaluation Set

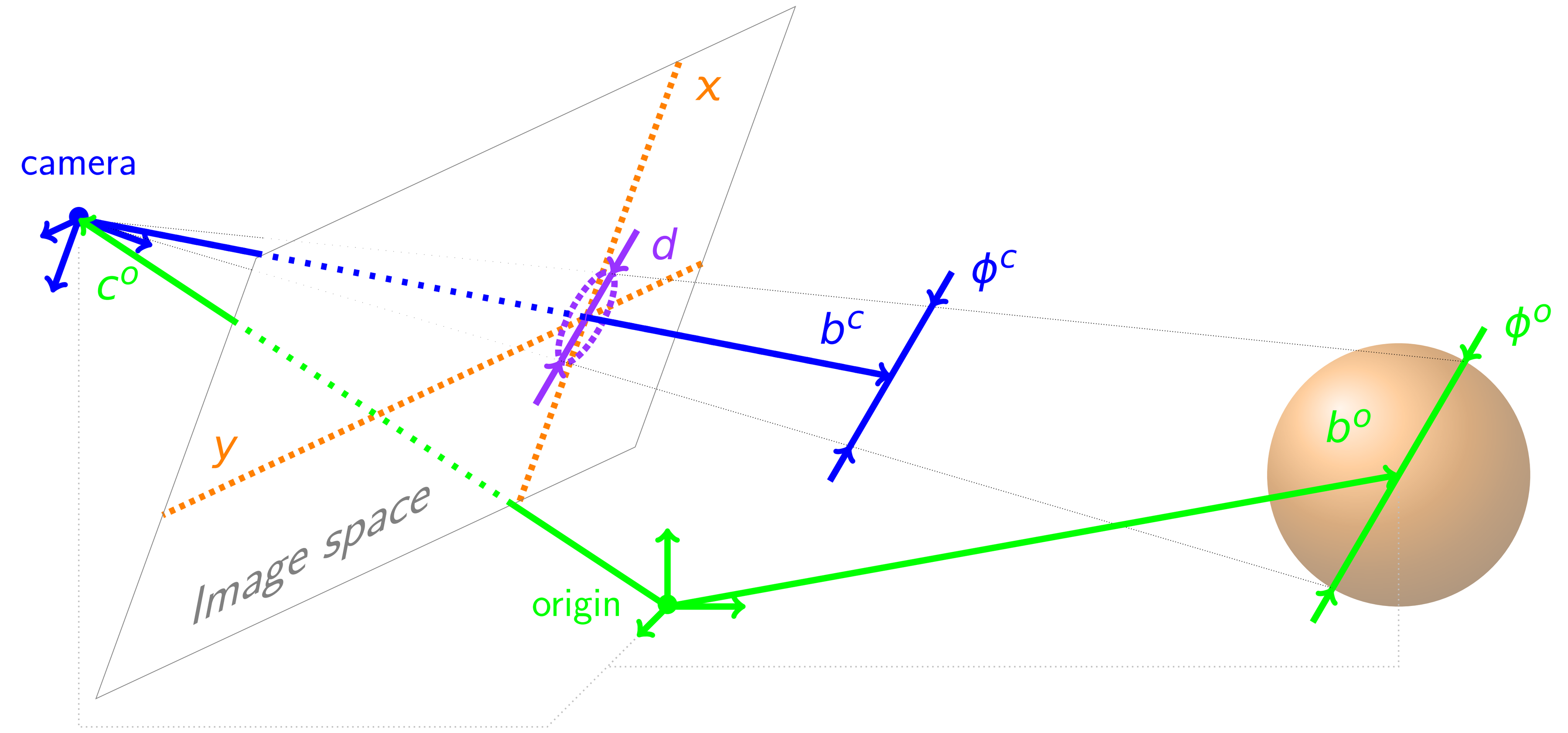


35 ballistic trajectories between 4 and 17 images.
Made publicly available here:
`https://www.kaggle.com/datasets/gabrielvanzandycke/ballistic-raw-sequences`



Given:

- K: camera intrinsic matrix
- R: camera rotation matrix
- $\mathcal{R}$: lens distortion rectification
- ϕ^o: ball diameter in world coordinates system



$$\mathbf{b}^c = K^{-1} \cdot \mathcal{R}\left(\begin{bmatrix} x \\ y \\ 1 \end{bmatrix}\right)$$

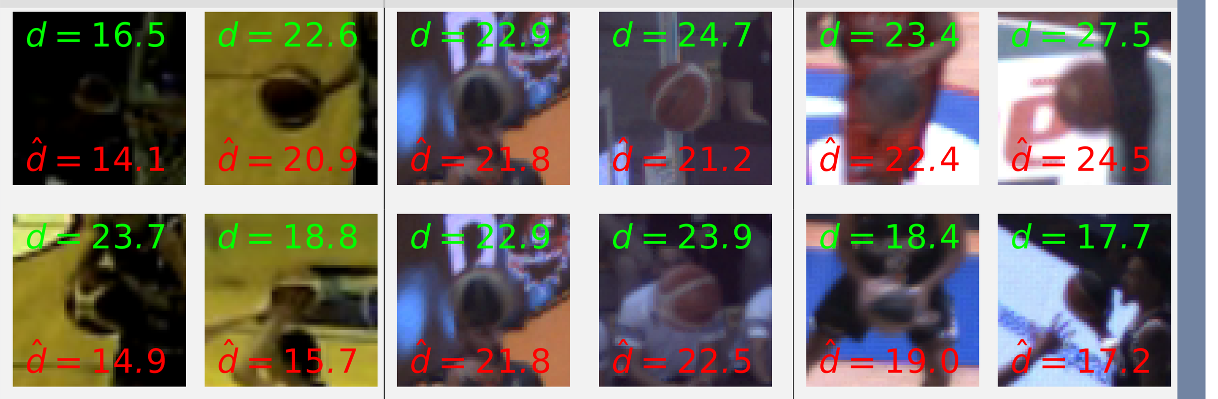
$$\mathbf{b}^o = R^T \cdot \frac{\phi^o \cdot \mathbf{b}^c}{\phi^c} + \mathbf{c}^o$$

Localization results

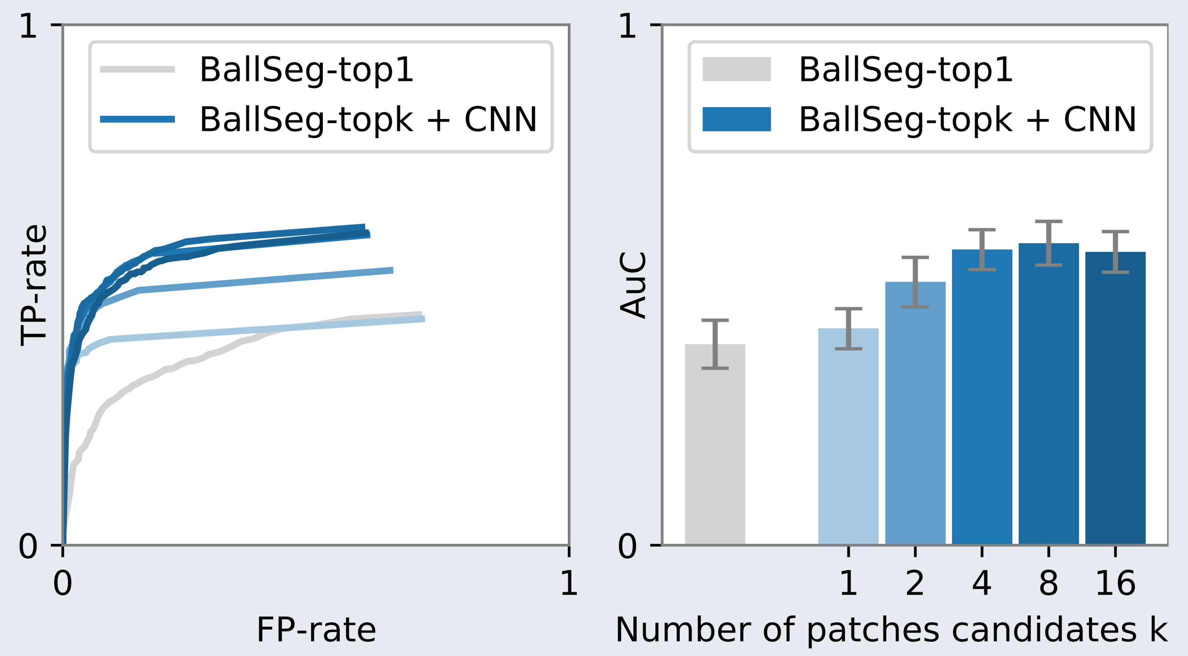
	$MAE_{[px]}$	$MAE_{[m]}$	$MAE_{[\%]}$
Baseline: *BallSeg*[1] + HCT	$4.6 \pm .5$	$5.1 \pm .5$	24 ± 4
Proposed method: ***BallSeg*[1] +CNN**	$1.6 \pm .2$	$1.8 \pm .2$	$10 \pm .7$
Oracle + CNN	$1.5 \pm .1$	$1.7 \pm .1$	$10 \pm .5$

Our high-quality evaluation set

APIDIS dataset[2]	High Quality Ball 3D Evaluation Set	DeepSport dataset[3]



Detection quality improved



References

[1] G. Van Zandycke and C. De Vleeschouwer, "Real-time cnn-based segmentation architecture for ball detection in a single view setup," *MMSports 2019*.

[2] P. Parisot and C. De Vleeschouwer, "Consensus-based trajectory estimation for ball detection in calibrated cameras systems," *Journal of Real-Time Image Processing*, 2019.

[3] G. Van Zandycke, "DeepSport dataset: `https://www.kaggle.com/gabrielvanzandycke/deepsport-dataset`," 2021.

Efficient tracking of team sport players with few game-specific annotations

Adrien Maglo, Astrid Orcesi and Quoc-Cuong Pham

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France, {firstname.lastname}@cea.fr

OBJECTIVE

Track team sport players from one team during a **full game** thanks to **few** human **annotations**

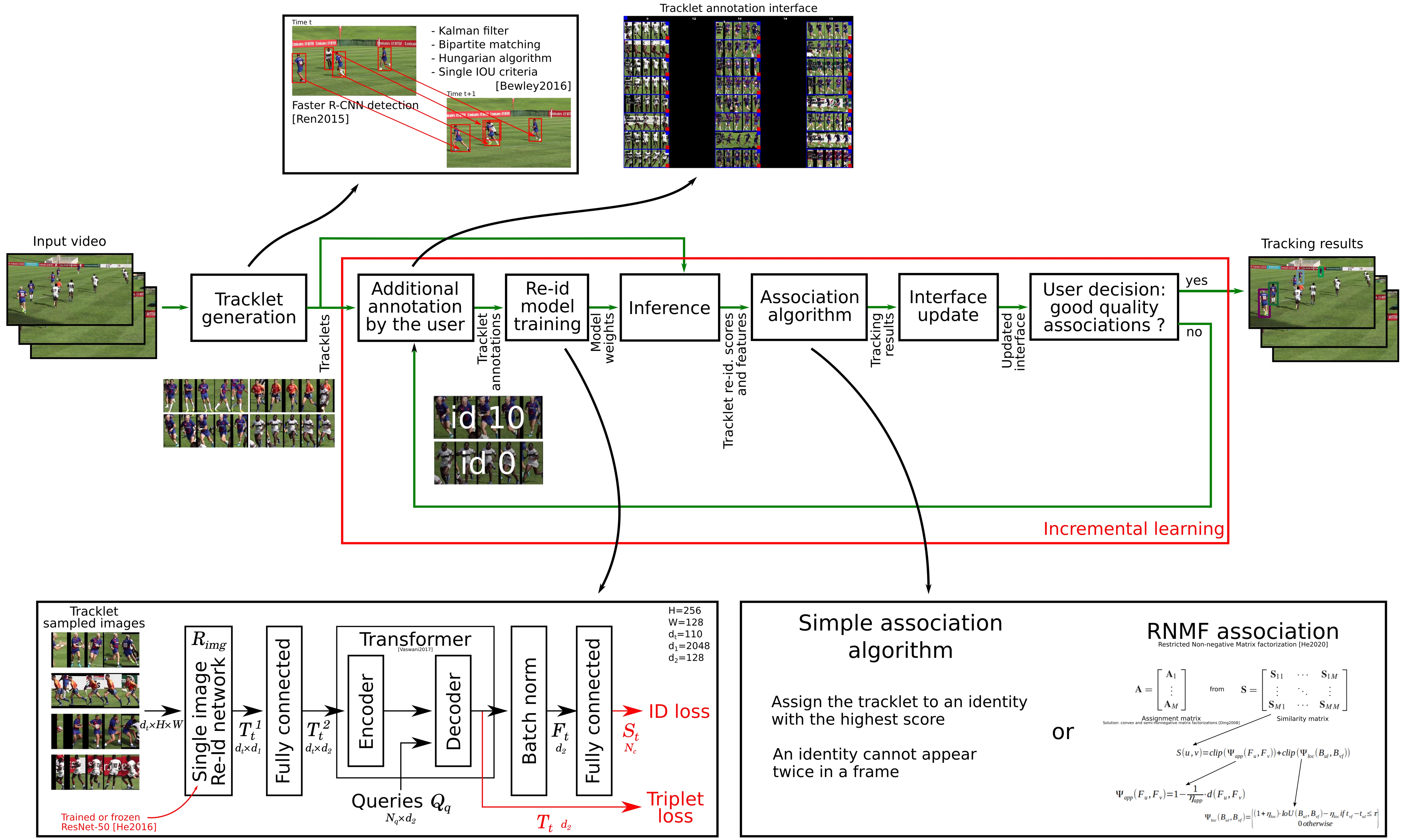
CHALLENGES

- Fast movements
- Similar player appearances
- Various poses
- Occlusions
- Multiple entries and exits of the field of view

NEW TRACKING RUGBY DATASET

- 7 rugby Dubaï 2021 Tournament
- 3 sequences of 40 s. at 1080p 50 FPS
- Publicly released at https://kalisteo.cea.fr/index.php/free-resources/

PROPOSED METHOD



TRACKING RESULTS

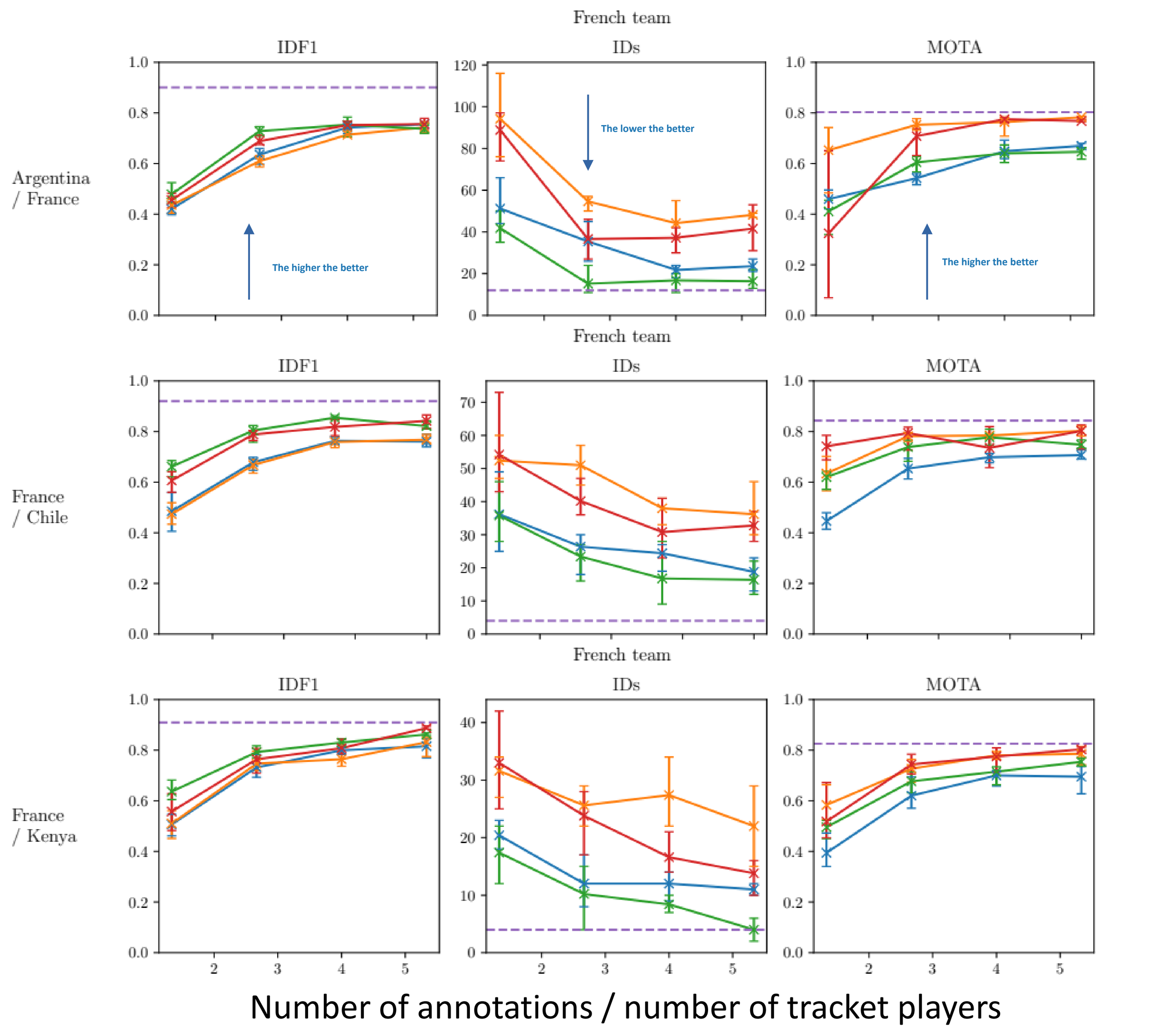
Tracking performances **increases with the number of annotations**

R_{img} frozen with the iterative association

R_{img} frozen with the RNMF association

R_{img} trained with the iterative association

R_{img} trained with the RNMF association



DETECTION AND IDENTIFICATION RESULTS

Detection and identification performances **better for big player** bounding boxes

R_{img}	assoc.	Det. recall	Team class. recall	Id. class. recall	Total recall
All detected bounding boxes					
frozen	iter.	75.8	58.4±2.1	73.8±4.5	32.7±2.4
frozen	RNMF		74.6±2.5	60.9±6.5	34.5±4.6
trained	iter.		75.9±3.9	**84.0±3.4**	48.3±3.0
trained	RNMF		**89.1±2.0**	79.4±2.6	**53.6±1.8**
Big detected bounding boxes (area superior to 25214 pixels)					
frozen	iter.	89.7	60.8±2.2	77.3±6.8	42.1±3.1
frozen	RNMF		72.3±2.2	66.4±5.0	43.1±4.4
trained	iter.		76.2±3.5	**87.4±5.2**	59.7±4.2
trained	RNMF		**90.8±0.9**	83.5±3.4	**67.9±2.6**

France Kenya – French team – 32 frames – 6 annotations / player

CONCLUSION

- **New** semi-automatic team sport player tracking method
- **New** rugby tracking dataset

[Bewley2016] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In IEEE International Conference on Image Processing, pages 3464–3468. IEEE, 2016.
[Ding2008] Chris HQ Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(1):45–55, 2008.
[He2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.
[He2020] Yuhang He, Xing Wei, Xiaopeng Hong, Weiwei Shi, and Yihong Gong. Multi-target multi-camera tracking by tracklet-to-target assignment. IEEE Transactions on Image Processing, 29:5191–5205, 2020.
[Ren2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 28:91–99, 2015.
[Vaswani2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems, 30, 2017.

End-to-End High-Risk Tackle Detection System for Rugby

Naoki Nonaka[1], Ryo Fujihira[1], Monami Nishio[1], Hidetaka Murakami[2], Takuya Tajima[3], Mutsuo Yamada[4], Akira Maeda[5,6] and Jun Seita[1]

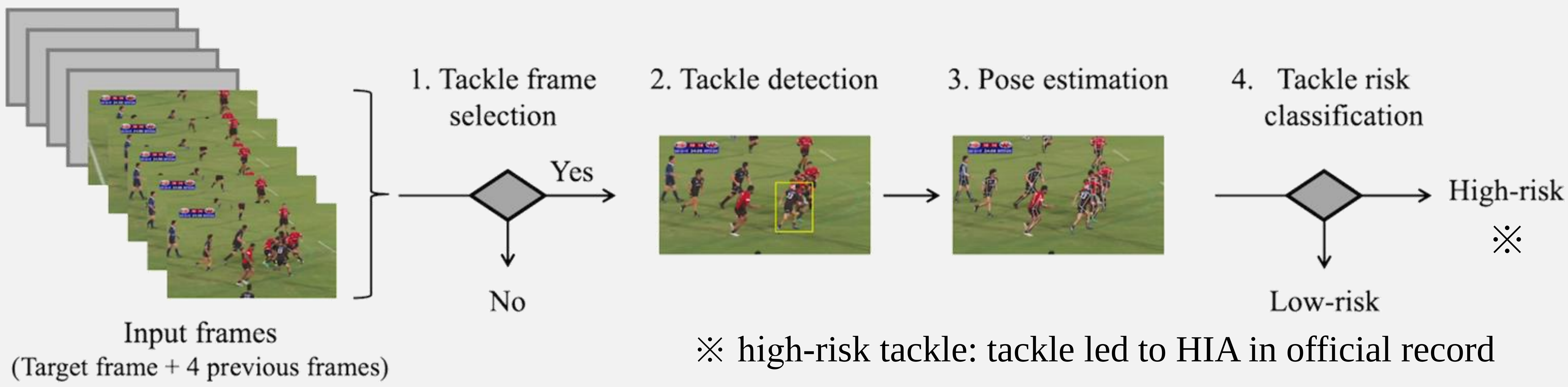
[1]Advanced Data Science Project, RIKEN Information R&D and Strategy Headquarters [2] Murakami Surgical Hospital
[3] Faculty of Medicine, University of Miyazaki [4] Faculty of Health and Sport Sciences, Ryutsu Keizai University
[5] Hakata Knee & Sports Clinic [6] Faculty of Human Health, Kurume University

Background

- Concussion raises the risk of harmful aftereffect and is the most common injury in Rugby Union [1].
- World Rugby introduced Head Injury Assessment (HIA) protocol to identify suspected concussion.
- HIA is conducted by human professional, thus affordable only for elite league.

Develop a high-risk tackle detection system without human intervention

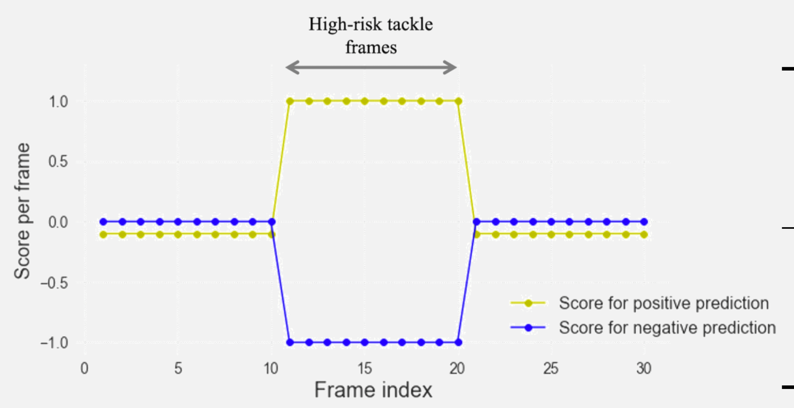
System



※ high-risk tackle: tackle led to HIA in official record

- Consists of 4 models (frame selection, tackle detection, pose estimation, tackle risk classification).
- Takes 5 sequential frames and return risk of tackle, when tackle is in given frame.

Result



Ground Truth	Prediction True	Prediction False
True	+1	-1
False	-0.1	0

Evaluation metric for high-risk tackle detection system.
For each frame in video, we give score shown on right table and subsequently, sum up per frame score and normalize obtained scores.

Frame selection model	Tackle detection model	Pose estimation model	Score	Recall
Human labels	RetinaNet	HRNet	0.3449	0.583
		CenterTrack	0.4905	0.833
	DETR	HRNet	0.2249	0.417
		CenterTrack	0.5397	0.917
No selection	RetinaNet	HRNet	0.2312	0.583
		CenterTrack	0.2759	1.000
	DETR	HRNet	0.2204	0.583
		CenterTrack	0.2224	1.000
ResNet Mixed Convolution	RetinaNet	HRNet	0.1837	0.333
		CenterTrack	0.0793	0.167
	DETR	HRNet	0.1825	0.333
		CenterTrack	0.1680	0.333
ResNet 2+1D	RetinaNet	HRNet	0.0840	0.167
		CenterTrack	0.2807	0.500
	DETR	HRNet	0.000	0.000
		CenterTrack	0.2719	0.500
ResNet 3D	RetinaNet	HRNet	0.0867	0.167
		CenterTrack	0.0400	0.083
	DETR	HRNet	0.0866	0.167
		CenterTrack	0.0820	0.167

- Trained and tested with TV broadcasted match video of Japanese elite league.
- Combination of 3 frame selection, 2 tackle detection and 2 pose estimation models were tested.
- Combination of ResNet2+1D, RetinaNet and CenterTrack performed best.

Discussion/Conclusion

- Developed end-to-end high-risk tackle detection system.
- System could detect 50% of high-risk tackles.
- Further room for improvement, especially in tackle frame selection.

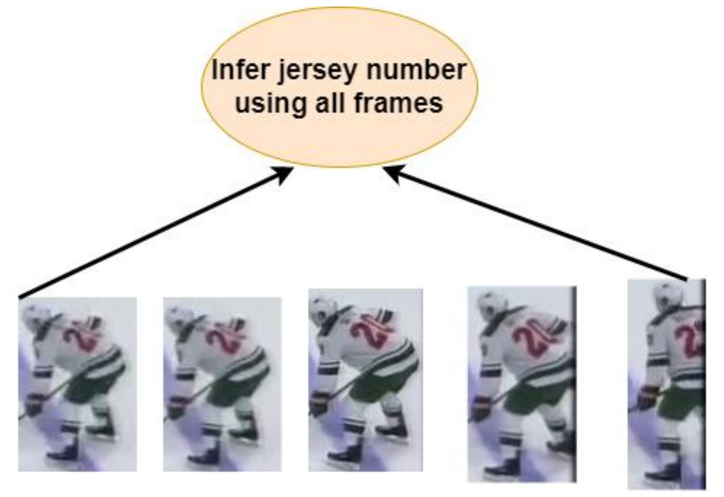
[1]: CW Fuller, Aileen Taylor, Marc Douglas, and Martin Raftery. Rugby world cup 2019 injury surveillance study. *South African Journal of Sports Medicine*, 32(1), 2020.

Ice hockey player identification via transformers and weakly supervised learning

Kanav Vats, William McNally, Pascale Walters, David A. Clausi and John S. Zelek

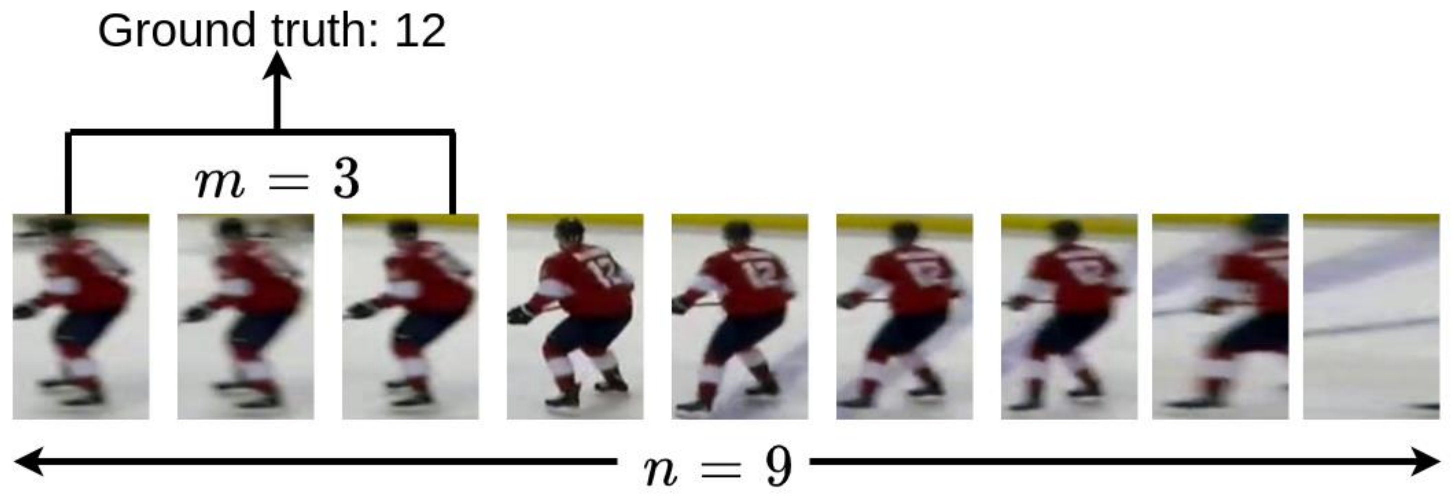
Problem statement

Identify jersey number (JN) from player tracklet.



Motivation

1. Previous works [1,2] sample a fixed number of frames from anywhere in a tracklet, with no knowledge of JN presence.



Sampled frames have no JN visible, but GT=12

2. Only a subset of roster player on the rink; use of play-by-play data can help boost identification accuracy

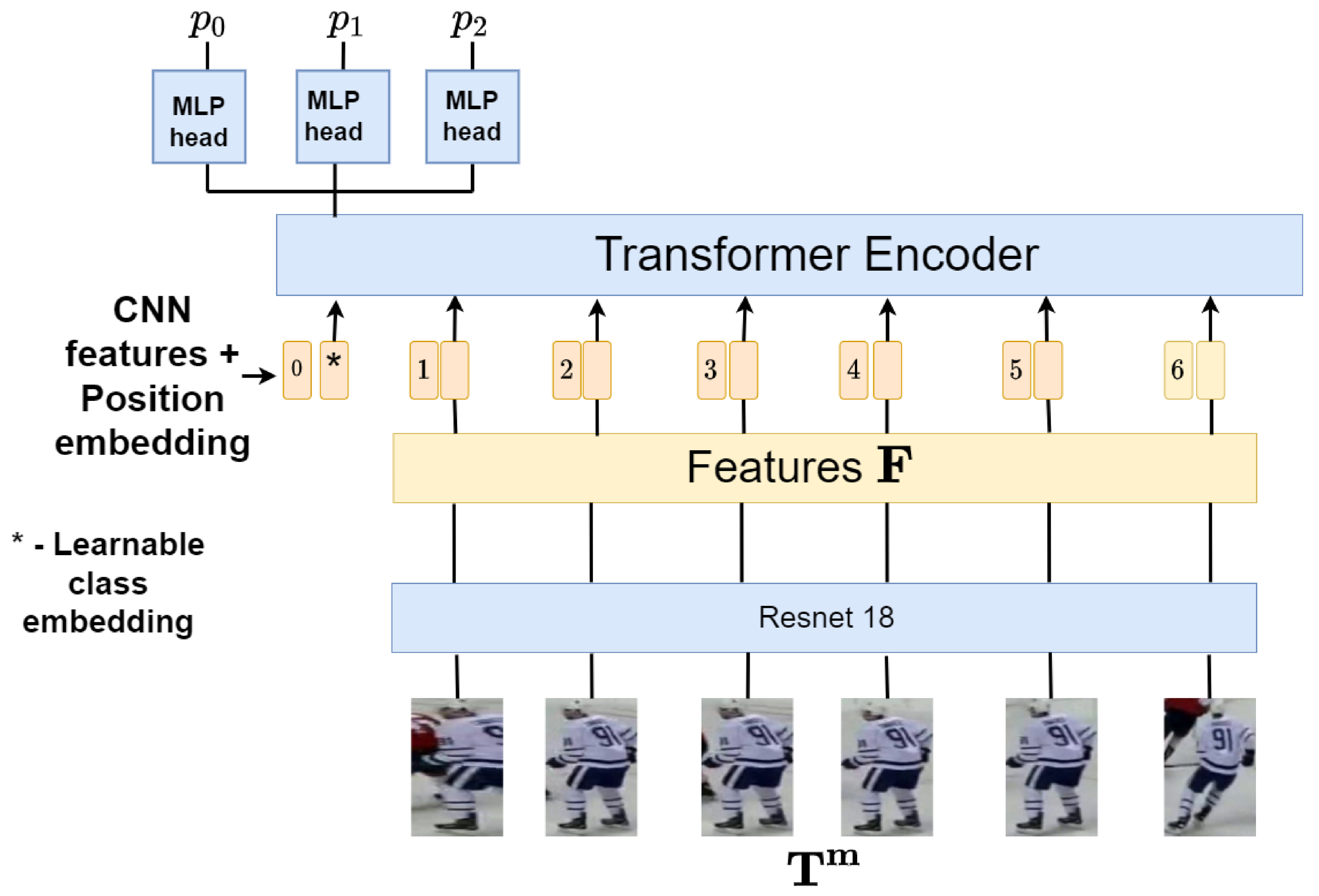
3. Use recent vision based transformer networks for player identification

Contributions

1. Incorporate player shift times into the inference using OCR, increasing accuracy by 6%.
2. Transformer based network outperforming the previous benchmark on the dataset[1]
3. Weakly-supervised training strategy achieving faster network convergence

References:
[1] Vats *et al. Arxiv preprint* 2110.03090
[2] Chan *et al. Expert Systems with Application, 2021*
[3] Vats *et al.* ACM MMSports 2021
[4] Kendall *et al.* CVPR 2018

Network



Input: *m* frames sampled from a tracklet.
Output: Jersey number probability of first, second digits and overall holistic number.
Loss: Multi-task cross-entropy loss[3] with learned weights[4].

Weakly supervised training

1. Generate weak/approximate labels for jersey number presence using a network trained to infer jersey number from static images.
2. Train the transformer network by sampling tracklet frames where jersey number is visible.

Incorporating player shifts

Given a player shifts database,

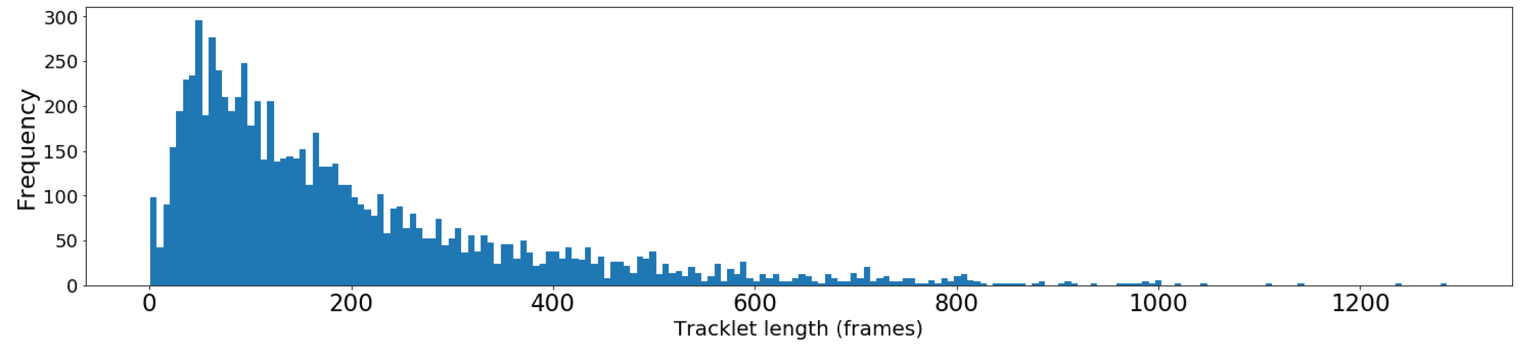
STEP1: Use an OCR to read game time

STEP2: Use game times to extract players present on ice during a game clip from the database

STEP3: Create *shift vectors* encoding shift information from *STEP2* and multiply with final logits.

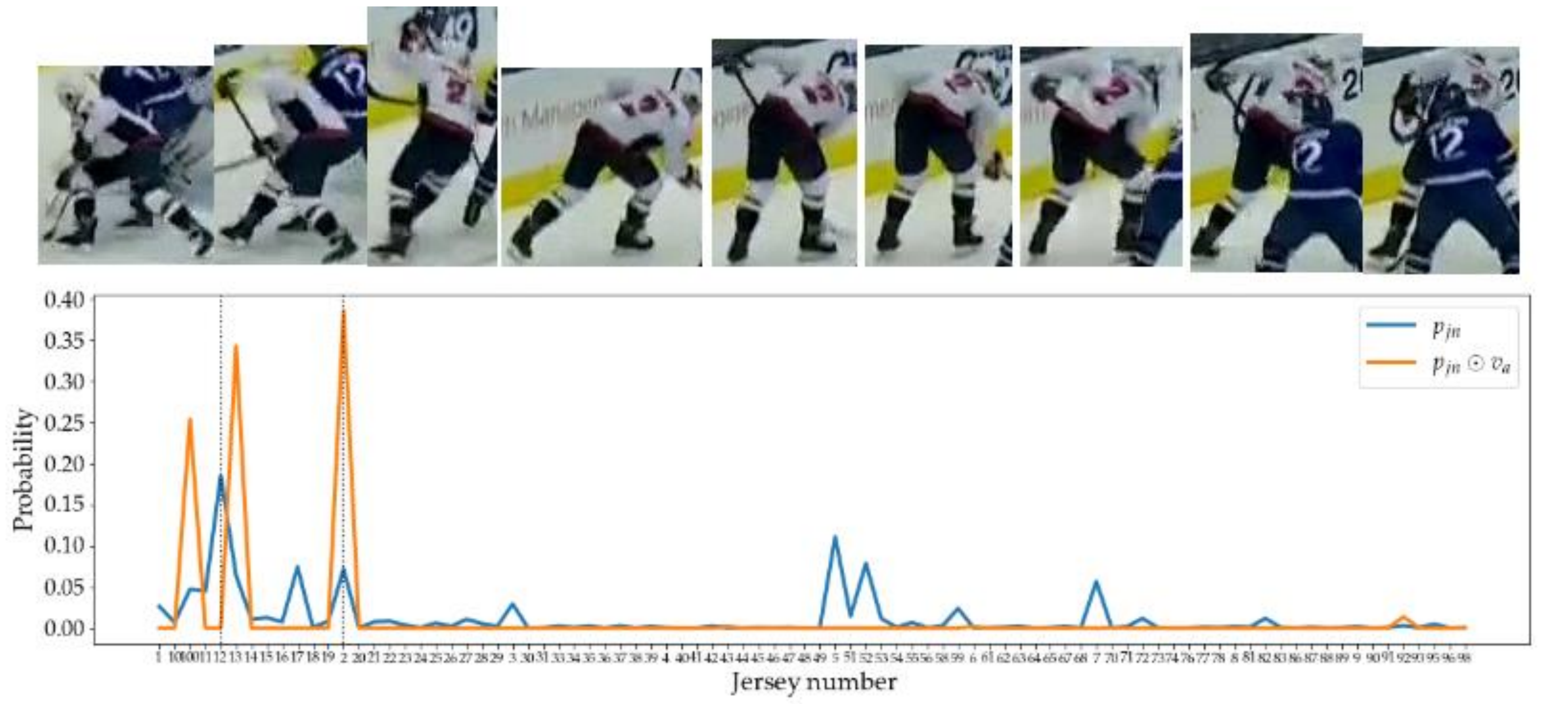
Dataset

84 clips/sequences
Trainval:71, Test:13
Game wise split made
Tracklets:
Training:2829 Validation:176 Test:505



Tracklet length distribution in dataset

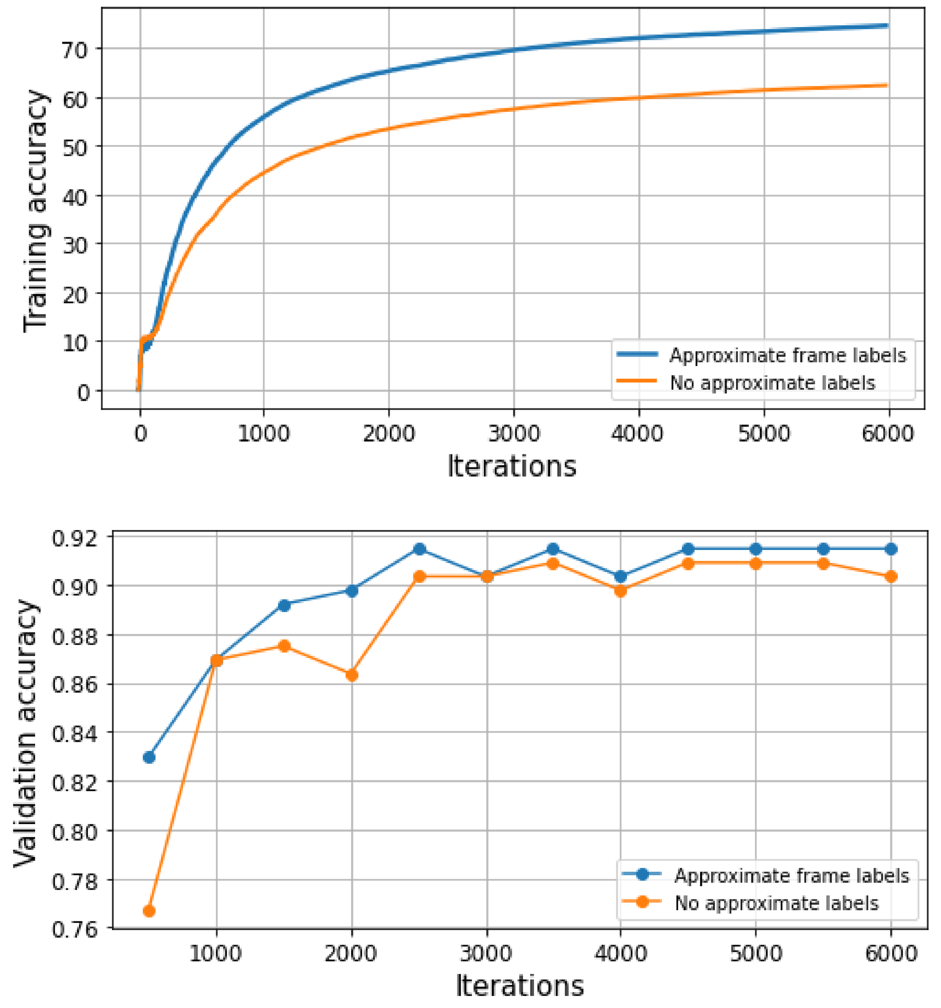
Results



Example where incorporating player shifts resulted in correct prediction
Ground truth 2, using shifts (red) prediction:2, without shifts prediction: 12

Video number	Ours w/ shift data	Ours w/ roster data	Ours w/o shift/roster data
1	90.70%	95.35%	90.60%
2	**91.43%**	85.71%	74.29%
3	**87.72%**	87.72%	84.2%
4	**80.00%**	76.0%	72.00%
5	**83.33%**	83.33%	81.48%
6	**90.00%**	90.0%	90.00%
7	**85.07%**	80.60%	73.13%
8	**93.75%**	93.75%	91.6%
9	**94.45%**	93.18%	88.6%
10	**93.02%**	88.37%	83.72%
11	**82.22%**	80.00%	71.11%
12	**84.85%**	84.85%	84.85%
13	**86.11%**	83.33%	80.56%
Mean	**87.97%**	86.32%	82.02%

Incorporating shifts leads of a performance increase of almost 6%



Faster convergence and better accuracy using weakly supervised training

Network	Accuracy
Proposed	83.37%
Temporal 1d CNN[1]	83.17%
Chan *et al.[2]*	73.1%

Interaction Classification with Key Actor Detection in Multi-Person Sports Videos

Farzaneh Askari[1], Rohit Ramaprasad[2], James J. Clark[1], Martin D. Levine[1]

[1]McGill University, Montreal, QC, Canada, [2]Birla Institute of Technology and Science, Pilani, Rajasthan, India

Motivation:

- Human actions and their interactions with each other and their environment plays a significant role in video understanding, especially sports analysis.
- Sports broadcast scenes are often crowded. Some of the actors participate in the main event (i.e., **key actors**), and the rest are present in the scene without being part of the actual event.
- Ice hockey broadcast videos include complex scenes due to frequent occlusions, camera viewpoints, camera motion
- Penalties are complicated human interactions during a sports game that significantly affect the dynamics and directions of the game.

Contribution:

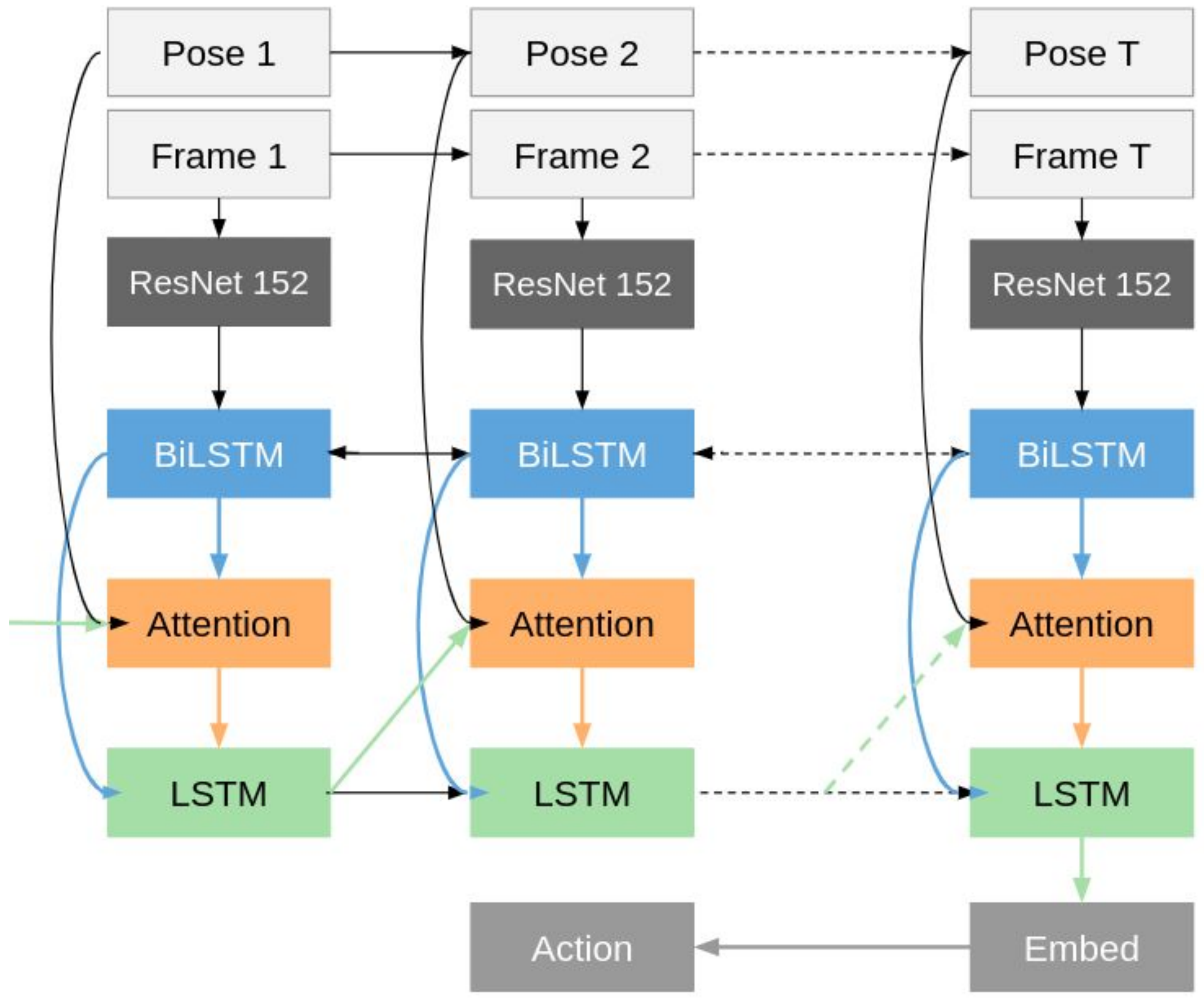
- **We propose a CNN-RNN based model equipped with an attention mechanism that recognizes penalties from ice hockey broadcast videos while isolating the players involved in the event.**

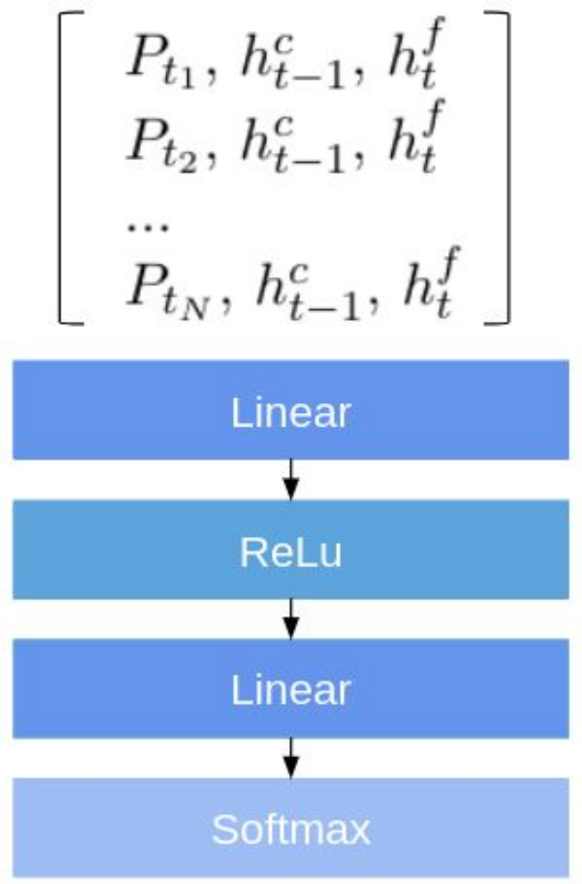
Dataset:

- Multi person penalty videos with pose and hockey stick annotations
- Classes: Tripping (80), Slashing(76), No penalty (98)

An example of tripping class with pose and stick annotation

Method:





Attention mechanism

$$h^f_t = \mathbf{BiLSTM}_f(h^f_{t-1}, h^f_{t+1}, f_t) \tag{1}$$

$$h^c_t = \mathbf{LSTM}_c(h^c_{t-1}, h^f_t, pa_t) \tag{2}$$

$$pa_t = \sum_{i=1}^{N_t} \mathbf{Softmax}\left(MLP([p_{ti}, h^f_t, h^c_{t-1}])\right) p_{ti} \tag{3}$$

$$loss = -\sum_{k=1}^{K} y^c_{T_i} \log y^c_{T_i} \tag{4}$$

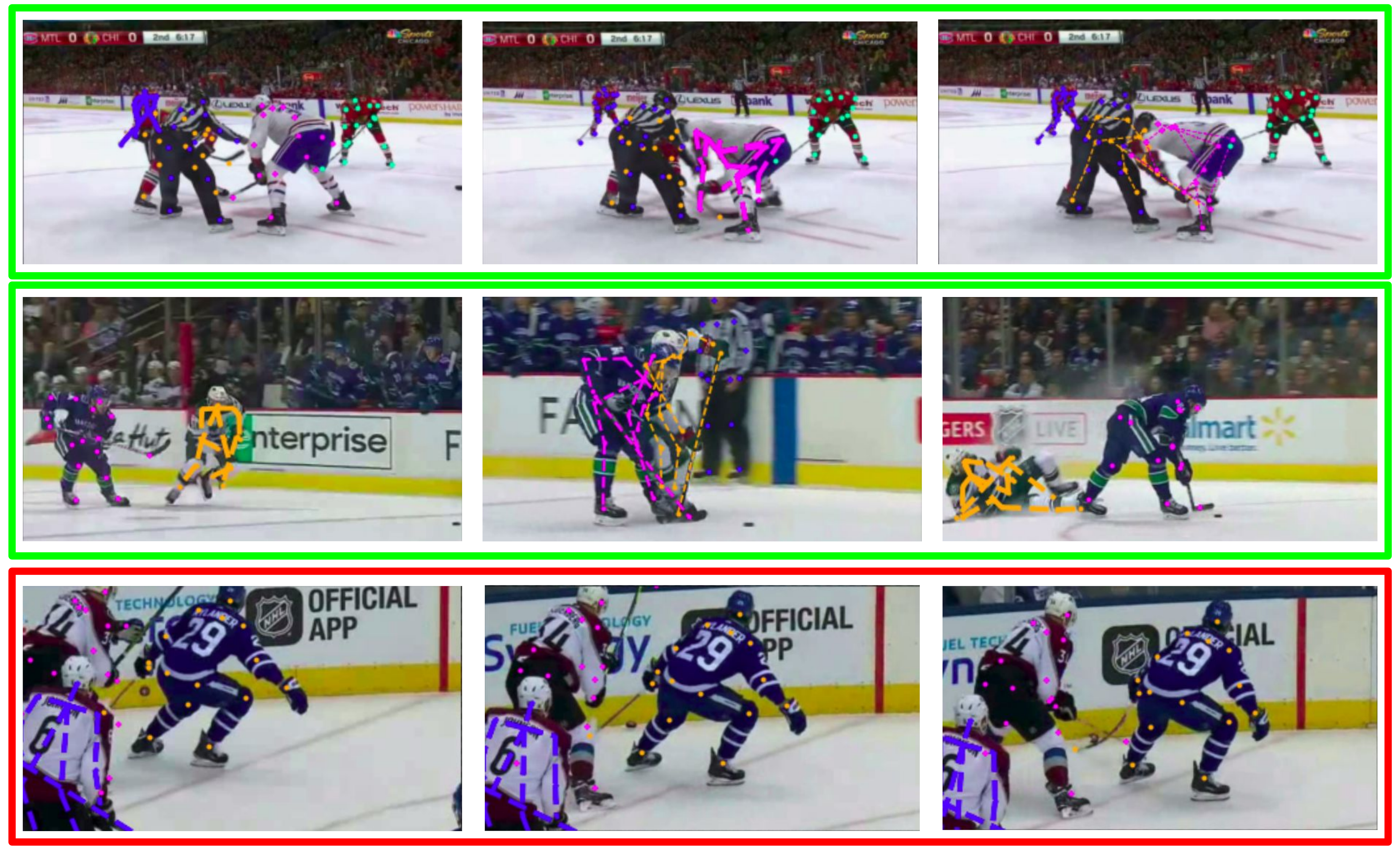
Results:

Model	Accuracy (%)
Model1: only frames (no Att)	87.43
Model2: only pose (Att)	80.66
Model3: frames and pose fusion (Att)	**93.93**

Table 1. Penalty classification accuracy

Model	Accuracy (%)
Model2: only pose (Att)	**80.66**
Model2 wo stick: only pose (Att)	74.86
Model3: frames and pose fusion (Att)	**93.93**
Model3 wo stick: frames and pose fusion (Att)	90.46

Table 2. The effect of stick keypoints on penalty classification



THE UNIVERSITY OF TOKYO
Tokyo Institute of Technology
Data Stadium

Pass Receiver Prediction in Soccer using Video and Players' Trajectories

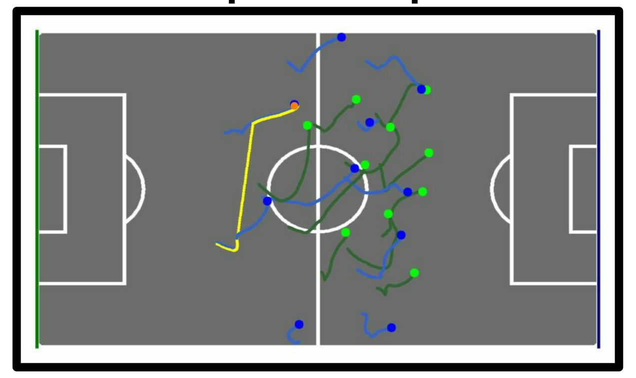
Yutaro Honda 1 Rei Kawakami 2 Ryota Yoshihashi 2 Kenta Kato 3 Takeshi Naemura 1
1 The University of Tokyo 2 Tokyo Institute of Technology 3 Data Stadium Inc.

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Introduction

Our goal: to create a prediction model that can be used for some analyses or applications.

For that goal, we combine geometric (left) and visual (right) information to improve prediction accuracy.



Alignment Process

Key technologies.

1. *Using detected points by YOLO as well as trajectories.*

Trajectories

: Tracked positions of 20 players cannot be projected precisely.

Detection points

: Accurate positions of each player but include errors such as,

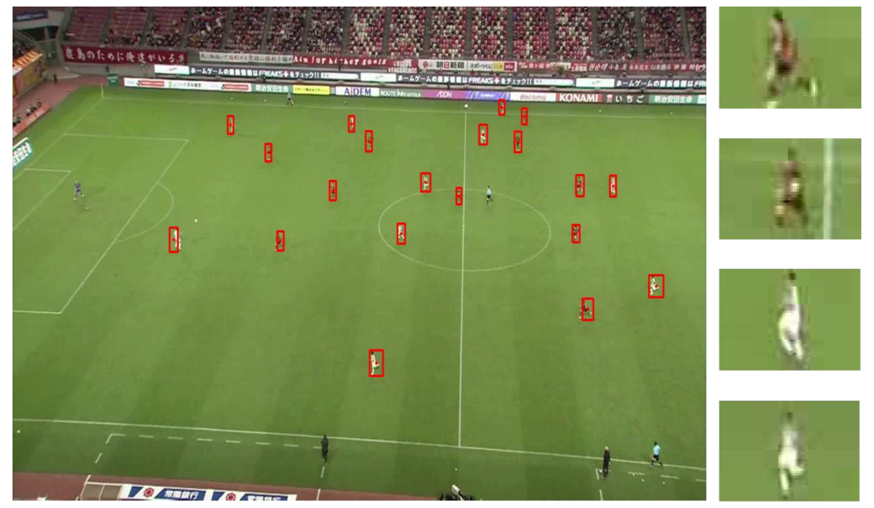
1. Miss detections. 2. Unnecessary detections.

2. *Correction of YOLO errors using ICP + Hungarian Alg.*

   1. **ICP between trajectory and detections.**
      Miss detected players may be identified.
      We call them pseudo detection points.

   2. **Hungarian matching to filter points.**
      Red points w/o circle
      : Pseudo detection points
      Blue points w/o circle
      : Unnecessary detection points

3. *CPD was used for final alignment.*
We obtained only 20 players tracked positions in image coordinates system.



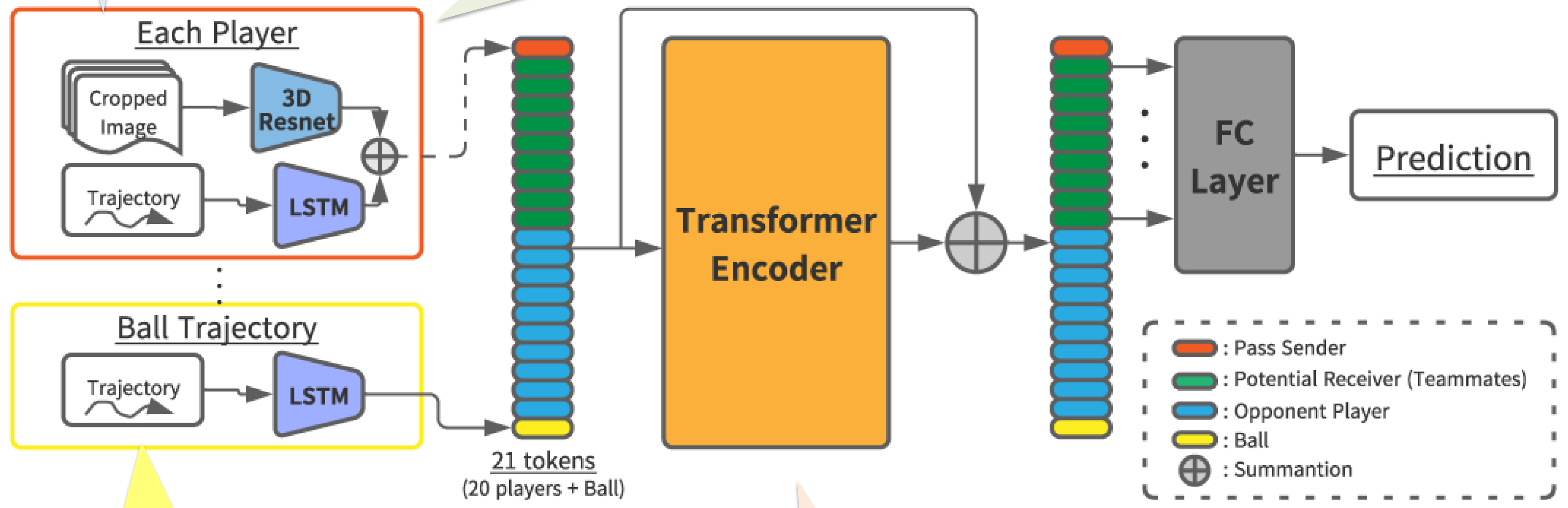
Method

Combining three basic methods: 3DCNN, LSTM, Transformer

- ***Input***: Players' video frames & trajectories, ball trajectory (as context info)
- ***Output***: Possibilities for receiving a next pass of each player.

The cropped player video is used for two reasons:
1. To use the trajectory and the video simultaneously.
2. To prevent loss of visual information.

The cropped images. → The movement of the body.
The trajectories. → The spatial movement on the field.



Ball movement is considered as a unique context information.

The transformer takes into account the interaction between players through an attention mechanism

Experiment

Predicting a receiver out of 9 teammates (excluded goad keeper).

- Rule based: Treating the closest teammate as the receiver.
- CNN: Simple CNN that considers the players' position just before the pass.

	Top-1	Top-3	Top-5
Rule based	30.84	68.13	82.82
CNN	38.84	77.78	91.31
Our (trajectory)	48.48	84.27	94.80
Ours (trajectory + RGB)	61.10	91.52	97.47

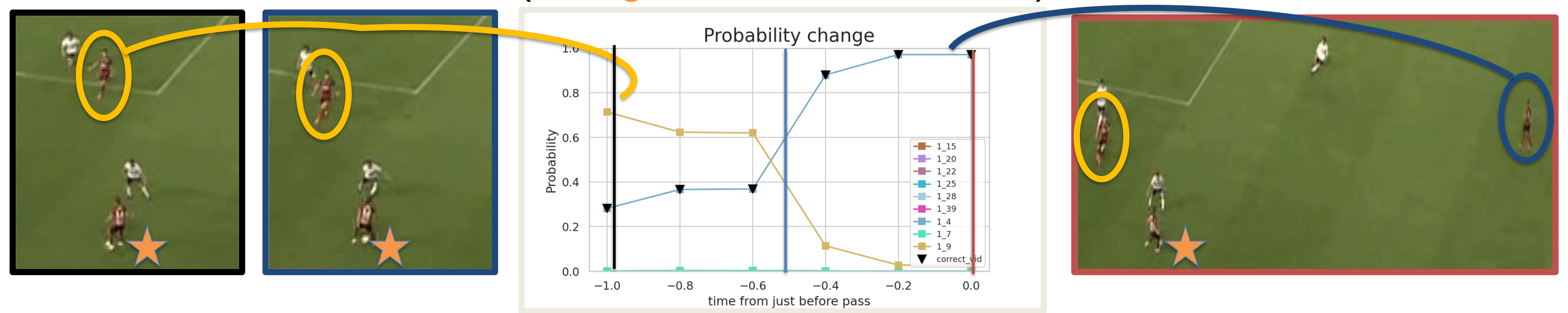
Possible Application

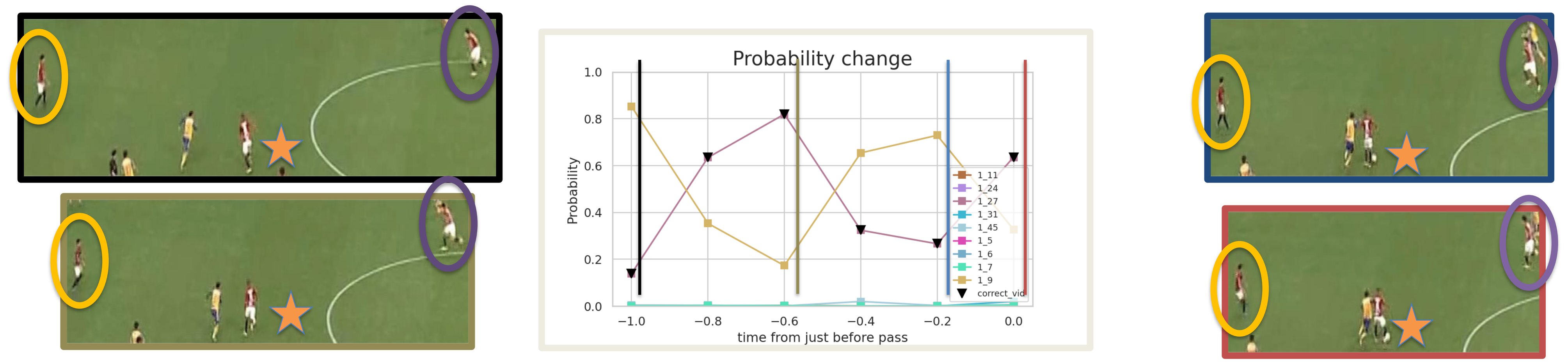
Detecting key timing from probability change.

The change of players' decision appeared as probability change.

What happens in these scenes?

(Orange stars indicate sender)

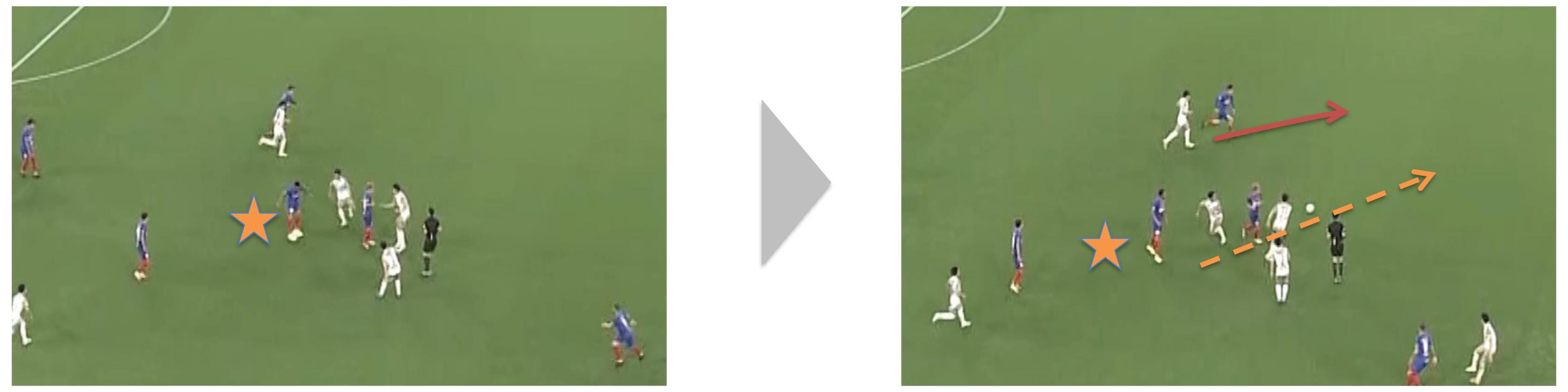




Searching high-level pass scenes as prediction errors.

Many of prediction errors are as follows: *headings, side changes, long passes*, and plays where even the kicker does not have full control of the ball.

Some errors are high-level pass scenes: it deceives our prediction model.



High-skilled through pass.

Conclusion

- Considering visual information directly improved the prediction accuracy.
- We developed a pipeline that aligns player trajectories and video frames.
- We presented possible applications using our prediction model.

KATJA LUDWIG, DANIEL KIENZLE, RAINER LIENHART

RECOGNITION OF FREELY SELECTED KEYPOINTS ON HUMAN LIMBS

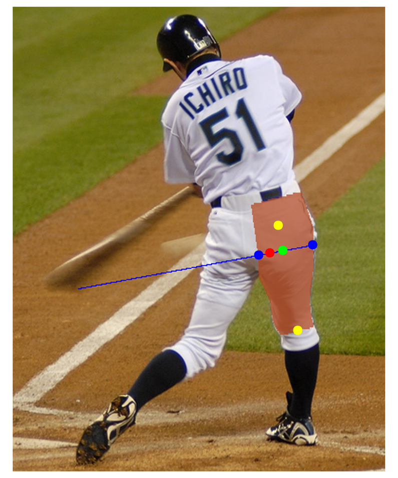
Universität Augsburg University

INTRODUCTION

- Video analysis is popular for performance evaluation and improvement of athletes' capabilities based on the results of the analysis
- For individual sports, the location of keypoints is of main interest - their detection can be automated by human pose estimation models
- Annotations are expensive, only necessary keypoints for the specific task are annotated
- More keypoints open possibilities for new and/or extended types of analyses, but are too time consuming to annotate
- Our approach introduces a method to estimate arbitrary keypoints on human limbs without any additional keypoint annotations

KEYPOINT GENERATION

Random keypoints are generated using segmentation masks, the masks can be created using detectron2 [1] if the dataset does not contain any:



1. A point b_p (green) is randomly sampled on the line (= projection line) between two fixed keypoints b_i, b_j (yellow) enclosing a body part.
2. A line orthogonal to the projection line is created and the boundary points c_1, c_2 (blue) of the body part are determined as the intersection of that line and the boundary of the segmentation mask.
3. The random point b_t (red) is generated by randomly sampling a point on the line segment between the boundary points, while points on both sides are equally probable.

KEYPOINT REPRESENTATION

1. Representation as Keypoint and Thickness Vectors

- Let n be the number of keypoints in a dataset and p_b denote the distance from b_p to b_i divided by the distance from b_i to b_j (= percentage of the projection line). Then the keypoint vector $v^k \in \mathbb{R}^n$ is designed as

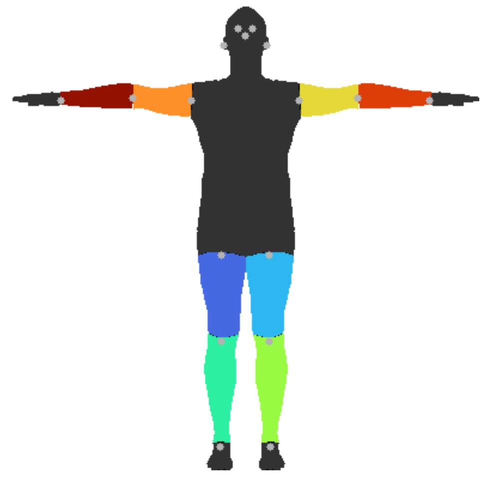
$$v_l^k = \begin{cases} 1 - p_b, & l = i \\ p_b, & l = j \\ 0, & l \neq i \wedge l \neq j \end{cases} \qquad l = 1, \dots, n$$

- Let c denote the intersection point that is closer to b_t. Let p_t be the distance from b_t to c divided by the distance from b_p to c (= percentage of thickness). Then, the thickness vector $v^t \in \mathbb{R}^3$ is designed as

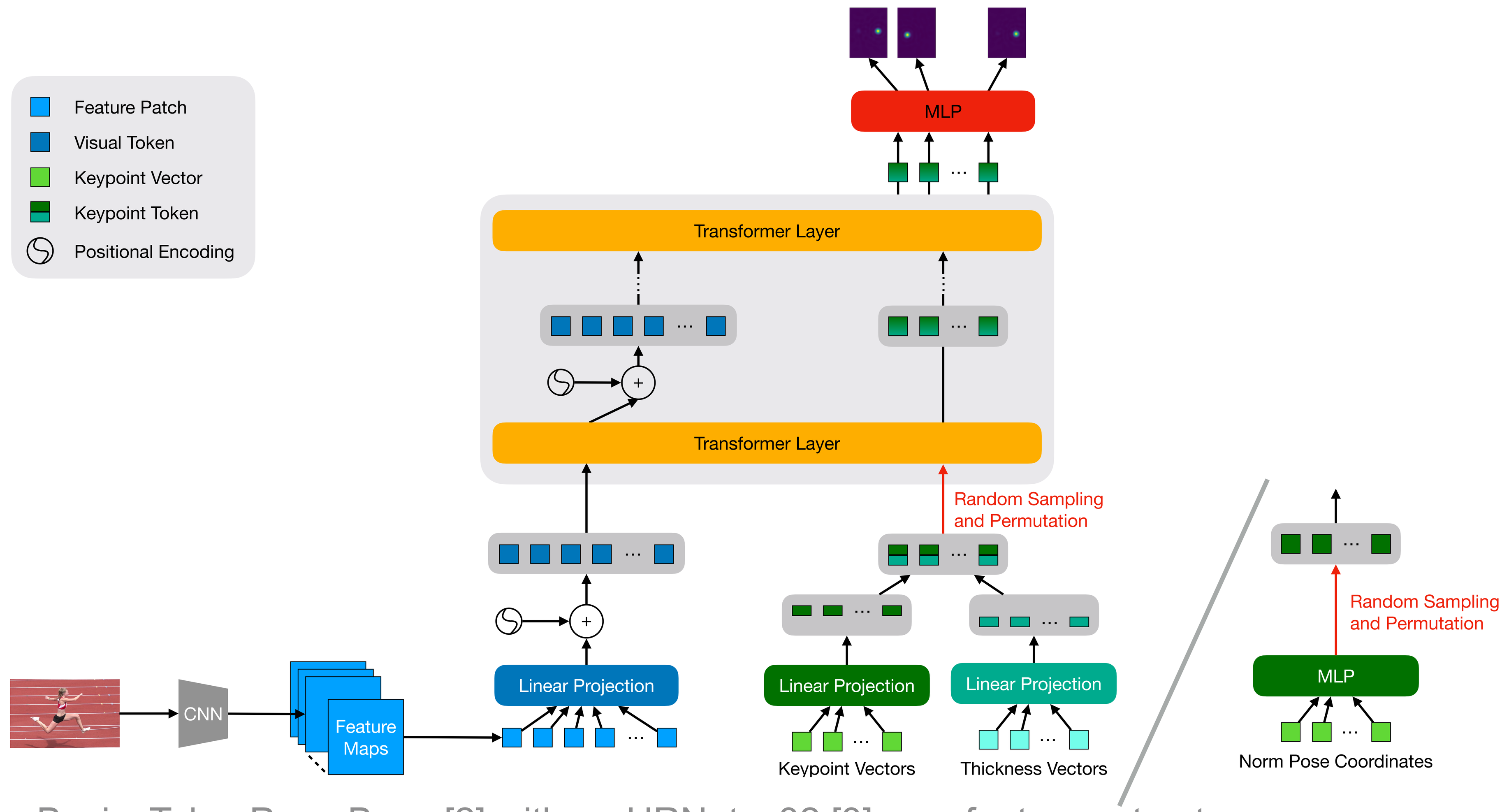
$$v^t = \begin{cases} (p_t, 1 - p_t, 0)^T, & b_t \text{ closer to } c_1 \\ (0, 1 - p_t, p_t)^T, & b_t \text{ closer to } c_2 \end{cases}$$

2. Representation as Norm Pose Point

- All keypoints are represented in normalized x- and y-coordinates of the following norm pose:

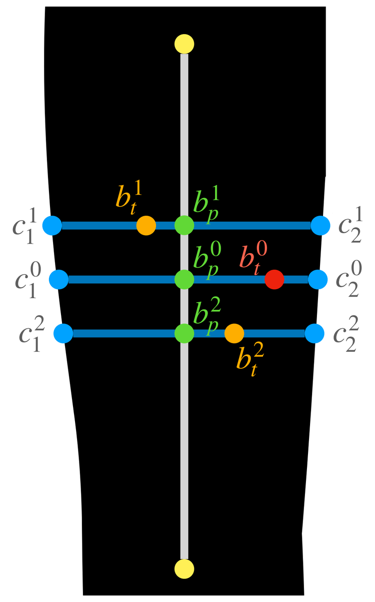


MODEL ARCHITECTURE



- Basis: TokenPose-Base [2] with an HRNet-w32 [3] as a feature extractor
- Keypoint and thickness vectors are transformed to keypoint tokens via a linear transformation just like feature patches and visual tokens
- Norm pose coordinates are transformed to keypoint tokens via a multi layer perceptron

THICKNESS METRICS



- Problem: Models predicting only the projection points b_p instead of b_t achieve high PCK and OKS scores, although the model does not learn the semantic of the body part shapes $\longrightarrow$ new metrics are necessary
- Let b_t^0 be the desired ground truth keypoint, b_p^0 the corresponding projection point, c_1^0 the intersection point on the other side of b_p^0 and c_2^0 the intersection point on the same side, w.l.o.g. (see figure)
- The ground truth thickness is $t_0 = \dfrac{||b_t^0 - b_p^0||_2}{||c_2^0 - b_p^0||_2}$
- If the model predicts a point b_t^2 on the same side of the projection line as b_t^0, the predicted thickness is $t_2 = \dfrac{||b_t^2 - b_p^2||_2}{||c_2^2 - b_p^2||_2}$ and the **thickness error** $e_2 = |t_0 - t_2|$
- For points b_t^1 on the opposite side, the thickness error is $e_1 = \dfrac{||b_t^1 - b_p^1||_2}{||c_1^1 - b_p^1||_2} + t_0$
- Used metrics: ***Mean Thickness Error (MTE)*** and ***Percentage of Correct Thickness (PCT)***, defined analogous to PCK

EXPERIMENTS

- DensePose [1] split of COCO [4] dataset:
  - 39,210 persons for training, 2,243 for validation and 7,297 for testing - 17 keypoints
  - Correction of ~3,500 segmentation masks (left-right errors, published on our website)
- Triple and long jump dataset:
  - 4,101 images for training, 464 for validation and 1,461 for testing - 20 keypoints
  - Segmentation masks from detectron2 [1], no need for manual annotations

Model	DensePose - COCO					Triple & Long Jump			
	AP	Avg PCK	Full PCK	MTE	PCT	Avg PCK	Full PCK	MTE	PCT
TokenPose	**84.6**	84.1				**91.3**			
Keypoint & Thickness Vectors	84.0	**84.2**	**87.2**	**25.5**	**68.1**	90.9	**93.6**	**16.2**	**81.4**
Norm Pose Linear	78.5	80.5	83.1	33.0	56.4	90.3	93.5	17.0	79.0
Norm Pose 4-Layer MLP	83.1	83.7	87.1	25.7	66.9	90.9	**93.6**	16.8	79.8

References

[1] Iasonas Kokkinos, Rıza Alp Güler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. 2018.

[2] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. arXiv preprint arXiv:2104.03516, 2021.

[3] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 2020.

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dolla ́r, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, pages 740–755. Springer, 2014.

Semi-Supervised Training to Improve Player and Ball Detection in Soccer

Renaud Vandeghen*, Anthony Cioppa*, Marc Van Droogenbroeck

* these authors contributed equally.

Paper Code

https://github.com/rvandeghen/SST

Abstract

- We propose a novel **semi-supervised learning** method for leveraging unlabeled data by generating pseudo labels with a teacher-student approach.
- We introduce **three loss parametrizations** to introduce **doubt** in the pseudo labels based on their **confidence scores**.

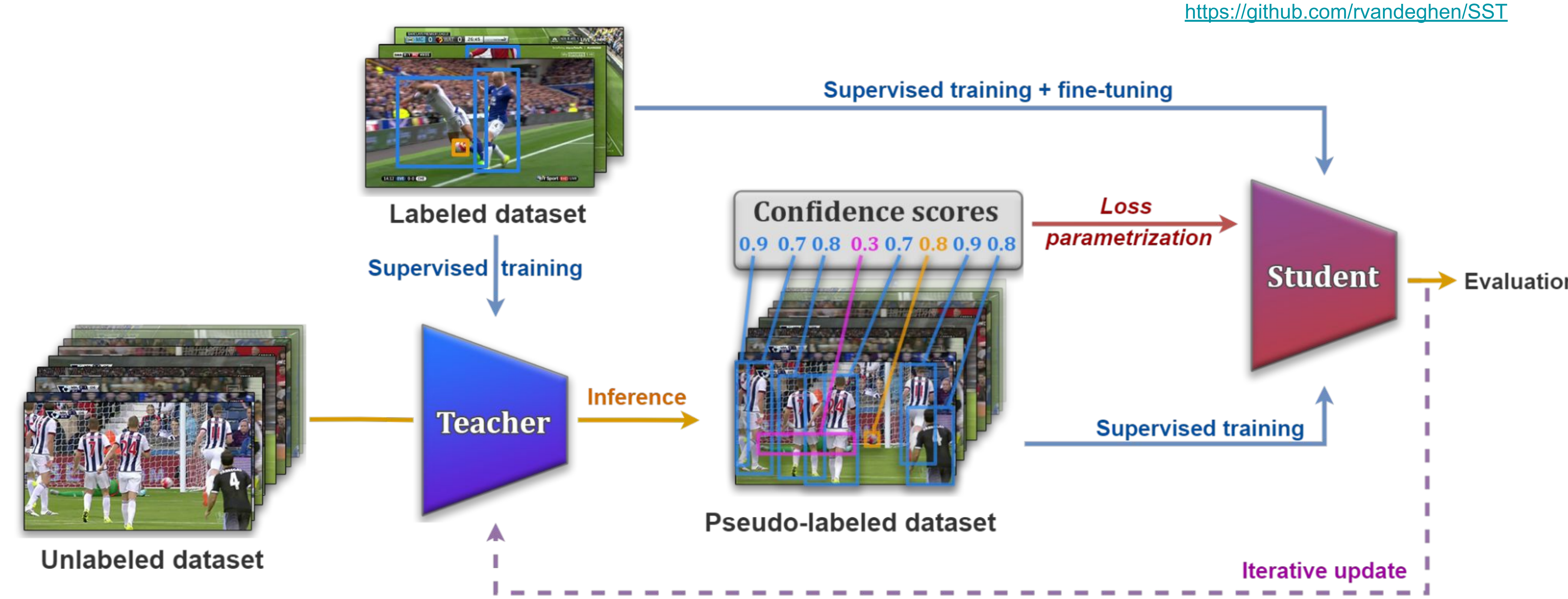
Motivations

- It is **expensive** in time or money to **annotate** large amounts of data.
- Unlabeled data are collected but often left unused.

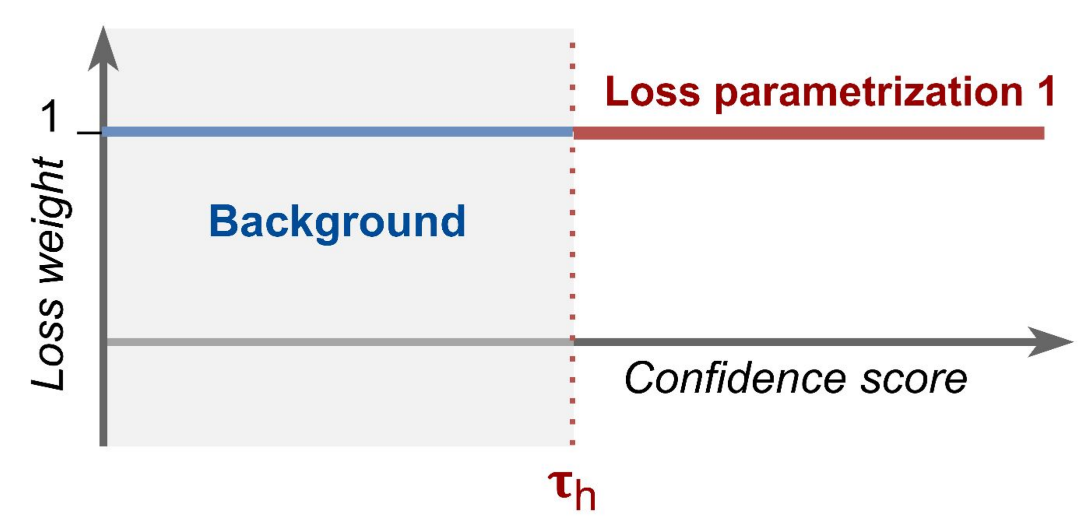
➢ Let's use them to **improve our models**!

Methodology

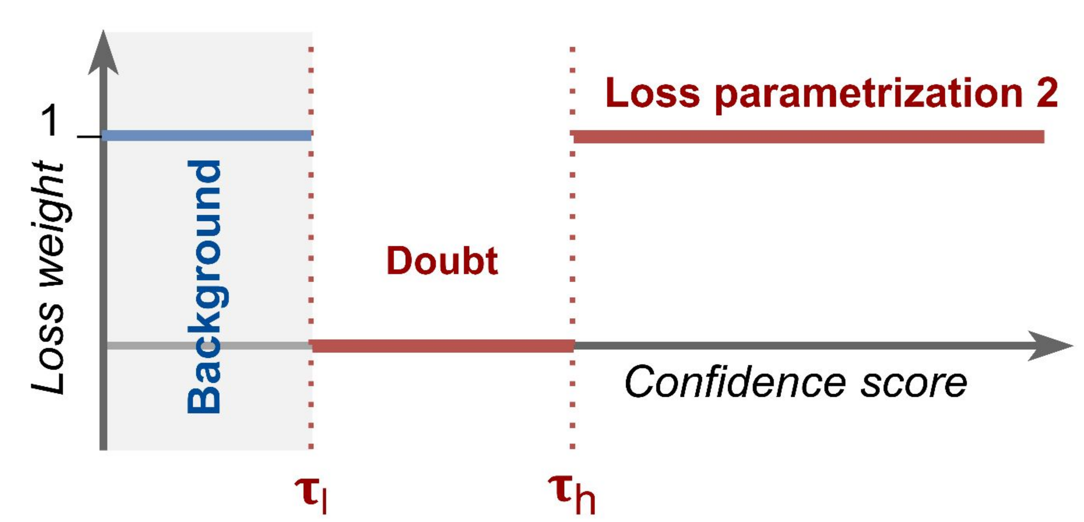
- **Step 1: Training the teacher:** We train a teacher model with the **labeled data** in a supervised way.
- **Step 2: Generating pseudo labels:** We use the trained teacher to **generate pseudo labels** on the unlabeled data.
- **Step 3: Training the student:** We train a student model with **the labeled and pseudo-labeled data**. We **introduce doubt for unsure predictions** of the teacher by **parametrizing the loss** and we **fine-tune** the student model with the labeled data.
- **Step 4: Iterating with a new teacher:** The fine-tuned **student becomes the new teacher** and it is used to generate new pseudo labels.



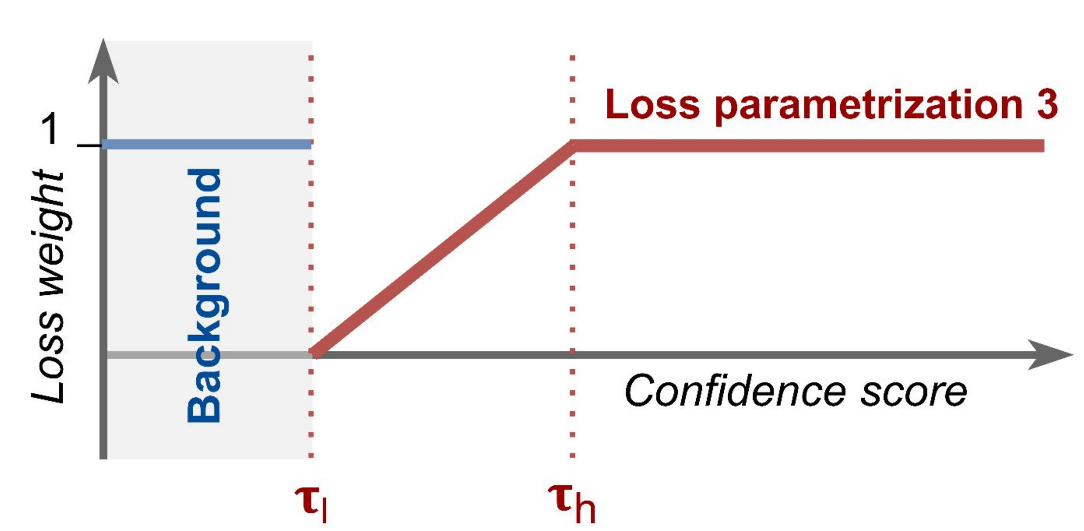
Loss parametrizations



Single separation between the objects and the background



Doubt for unsure predictions of the teacher



Progressive doubt for unsure predictions of the teacher

Experimental Results

Quantitative results

Metric: mAP

Method	l	h	Validation set 1%	5%	10%	100%	Test set 100%
Teacher	-	-	18.1	31.9	39.5	52.7	51.0
Param. 1	-	0.99	25.8†	38.6	**44.3**	53.7	–
Param. 2	0.9	0.99	26.0	38.7	**44.3**	**53.8**	–
Param. 3	0.9	1	**26.2**	**38.9**	43.7	**53.8**	**52.3**

- All parametrizations improve the performance.
- Param. 3 leads to the best performance.

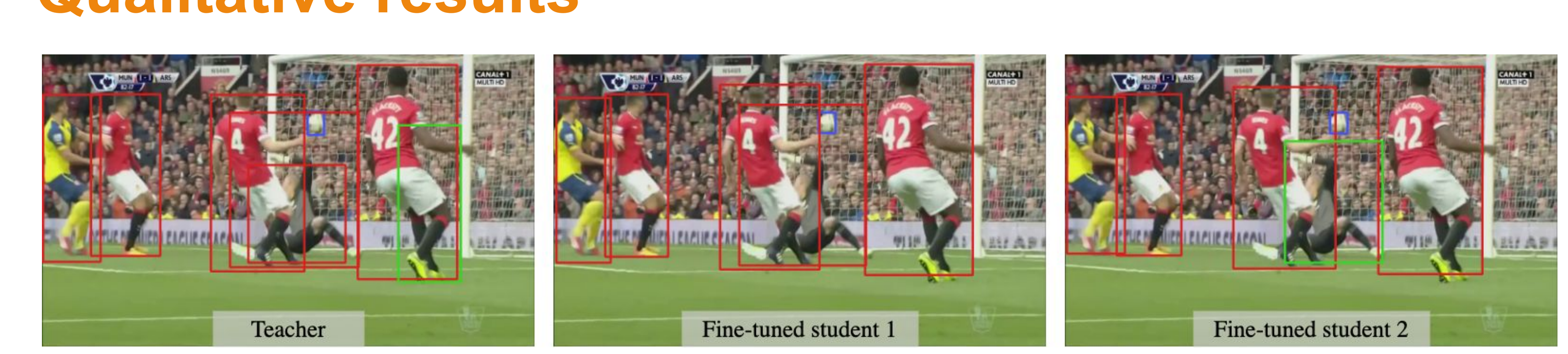
Effect of fine-tuning

Metric: mAP

Method	1%	5%	10%	100%
Teacher	18.1	31.9	39.5	52.7
Param. 1	22.6† → 25.8	36.0 → 38.6	42.3 → 44.3	52.6 → 53.7
Param. 2	23.1 → 26.1	36.6 → 38.7	43.0 → 44.3	52.6 → 53.8
Param. 3	23.0 → 26.2	36.1 → 38.9	41.9 → 43.7	52.7 → 53.8

- Fine-tuning on the labeled data further improves the performances for each parametrization.

Qualitative results



SoccerTrack

A Dataset and Tracking Algorithm for Soccer with Fish-eye and Drone Videos

Atom Scott[*1,2], I. Uchida[*1,2], M. Onishi[2], Y. Kameda[1], K. Fukui[1] and K. Fujii[3]

[1]University of Tsukuba, [2]National Institute of Advanced Industrial Science and Technology, [3]Nagoya University *Indicates equal contribution

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Introduction

Motivation

✗ Broadcast videos do not show the entire pitch
✗ No large public dataset for tracking with a full pitch view

Contributions

- A new soccer tracking dataset called *SoccerTrack (50,000+ frames)!*
  - ✓ **Wide-view** (fish-eye camera in 8K)
  - ✓ **Top-view** (drone camera in 8K)
  - ✓ **GNSS** location data
  - ✓ **Bounding boxes** w/ ID
- The codebase for camera calibration, tracking (players and ball) and other pre/post-processing tools.

Dataset Construction

Collection Method

The device used for GNSS data collection

Participants

- ✓ College-level athletes
- ✓ University of Tsukuba, Japan
- ✓ Ethics committee approved

Semi-Automatic

1. Collect video and GNSS data
2. Perform object detection on video
3. Project bounding boxes and GNSS points to pitch coordinates via homography transform.
4. Assign IDs to bounding boxes w/ bipartite matching

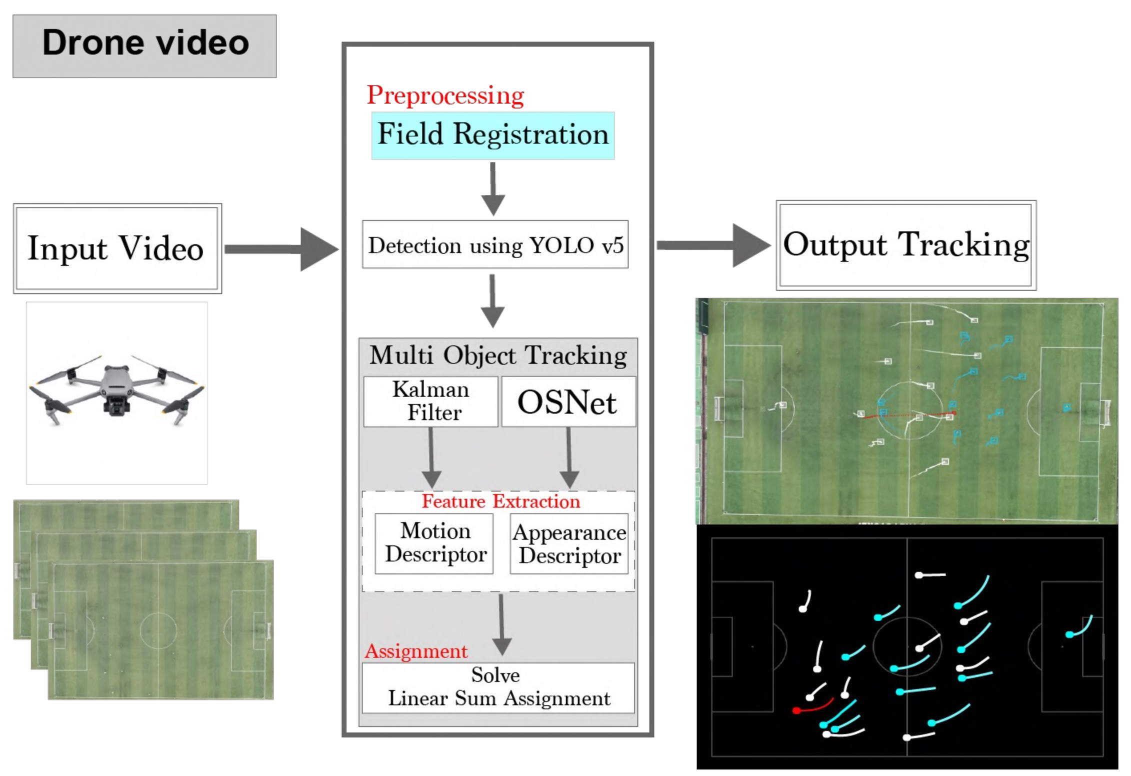
Dataset Overview

	Wide-view camera	Top-view camera	GNSS
Device	Z CAM E2-F8	DJI Mavic 3	STATSPORTS APEX 10 Hz
Resolution	8 K (7,680 × 4,320 pixels)	4 K (3,840 × 2,160 pixels)	Abs. err. in 20-m run: 0.22 ± 0.20 m [4]
FPS	30	30	10
Player tracking	✓	✓	✓
Ball tracking	✓	✓	✗
Bounding box	✓	✓	—
Location data	✓	✓	✓
Player ID	✓	✓	✓

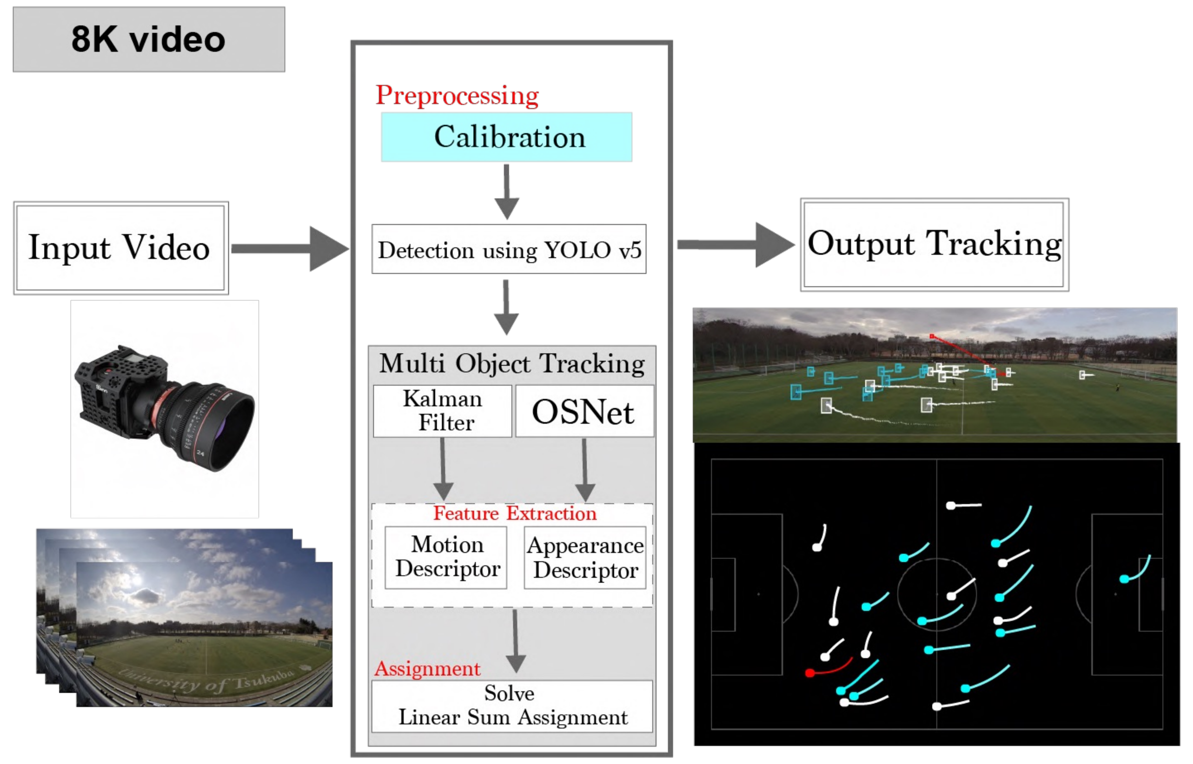
SoccerTrack Dataset

SoccerTrack Algorithm

Top-view Camera Tracking Pipeline



Wide-view Camera Tracking Pipeline



Results

Annotation Accuracy [mean ± std]

Top-view Camera

KP Projection Error	0.06 ± 0.03 m
Discrepancy w/ GNSS	2.76 ± 2.86 m

Wide-view Camera

KP Projection Error	0.56 ± 0.42 m
Discrepancy w/ Top-view	2.77 ± 4.47 m

Tracking Performance (Ave.)

Top-view Camera

MOTA Score	50.5%
ID Switches	5

Wide-view Camera

MOTA Score	14.2%
ID Switches	19

Conclusion

- Both top and wide camera views can be used for tracking
- Annotation evaluations showed reasonable accuracy
- The tracking algorithm can be improved

Extra Details

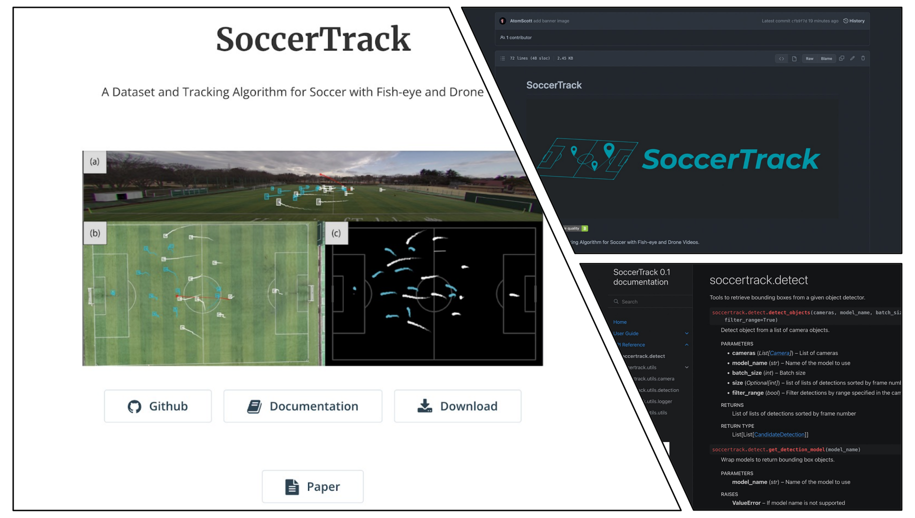
Comparison with other Tracking Datasets

Dataset	Camera	Wide-view	Top-view	GNSS/LPS	Location data	Bounding box	Tracking code
D'Orazio et al. [11]	✓	✗	✗	✗	✓	✗	✗
Pettersen et al. [32]	✓	Panorama	✗	LPS	✓	✗	✗
Pappalardo et al. [31]	✗	✗	✗	✗	✓	✗	✗
GFootball [23]	✓	—	—	—	✓	✗	✗
SoccerNet v1 [14]	✓	✗	✗	✗	✗	✗	✗
SoccerNet v2 [9]	✓	✗	✗	✗	✓	✓	✓
SoccerTrack (ours)	✓	Fish-eye	Drone	GNSS	✓	✓	✓

Public Release Schedule

Date	Content
06/20	**10 minutes** of top/wide view (30 secs x 20 clips)
08/01	**20 minutes** of top/wide view (30 secs x 40 clips)
09/01	**30 minutes** of top/wide view (30 secs x 60 clips)

Documentation / Data Downloads



Find our webpage at https://github.com/AtomScott/SoccerTrack

…Or just google "SoccerTrack"!

Sports Field Registration via Keypoints-aware Label Condition

Yen-Jui Chu[1] Jheng-Wei Su[1] Kai-Wen Hsiao[1] Chi-Yu Lien[1] Shu-Ho Fan[1] Min-Chun Hu[1] Ruen-Rone Lee[2] Chih-Yuan Yao[3] Hung-Kuo Chu[1]
[1]National Tsing Hua University [2]Delta Research Center [3]National Taiwan University of Science and Technology

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Webpage

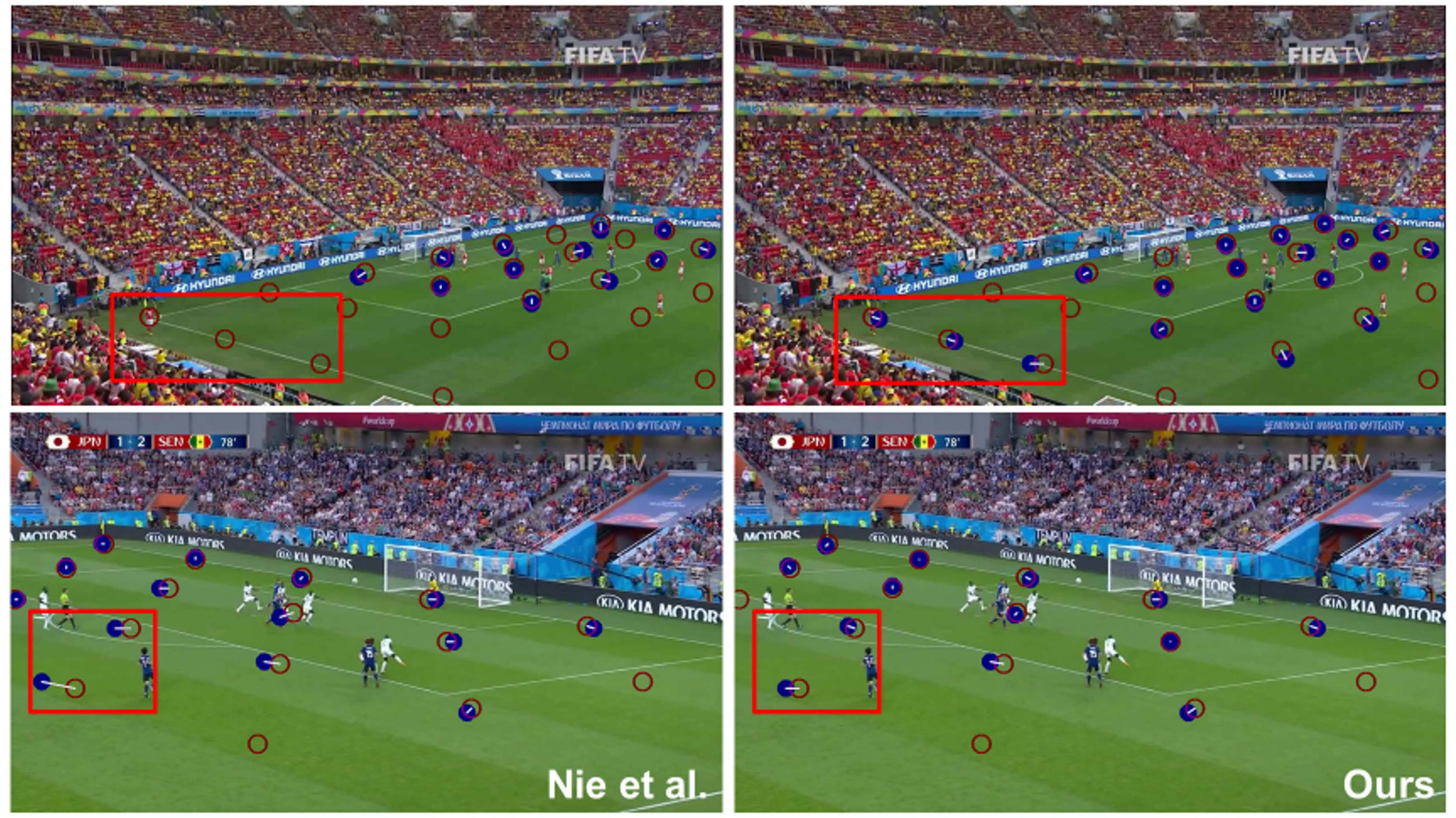
➢ Introduction

Background

Sports Field Registration is to estimate homography transformation using field-features between 2D field model and image. A wide variety of sports applications requires a robust sports field registration such as virtual advertising and true-view replay.

Motivation

Real-world field images usually present a uniform and textureless appearance, extracting sparse field-features due to camera zoom-in or occlusions caused by the players. Those cases make the homography estimation a non-trivial and challenging task. Inspired by keypoints detection method, which may suffer the missing and misalignment problems due to uniform appearance, we use similar idea to tackle this problem differently. Below is missing and misalignment case between the state-of-the-art method and ours.



Contributions

- We combine instance segmentation with dynamic filter learning to detect a grid of uniformly distributed keypoints over the entire field image.
- We introduce a new soccer dataset, called TS-WorldCup, with detailed field markings on 3812 **time-sequence** field images.

➢ Model Architecture

Standard Encoder-decoder

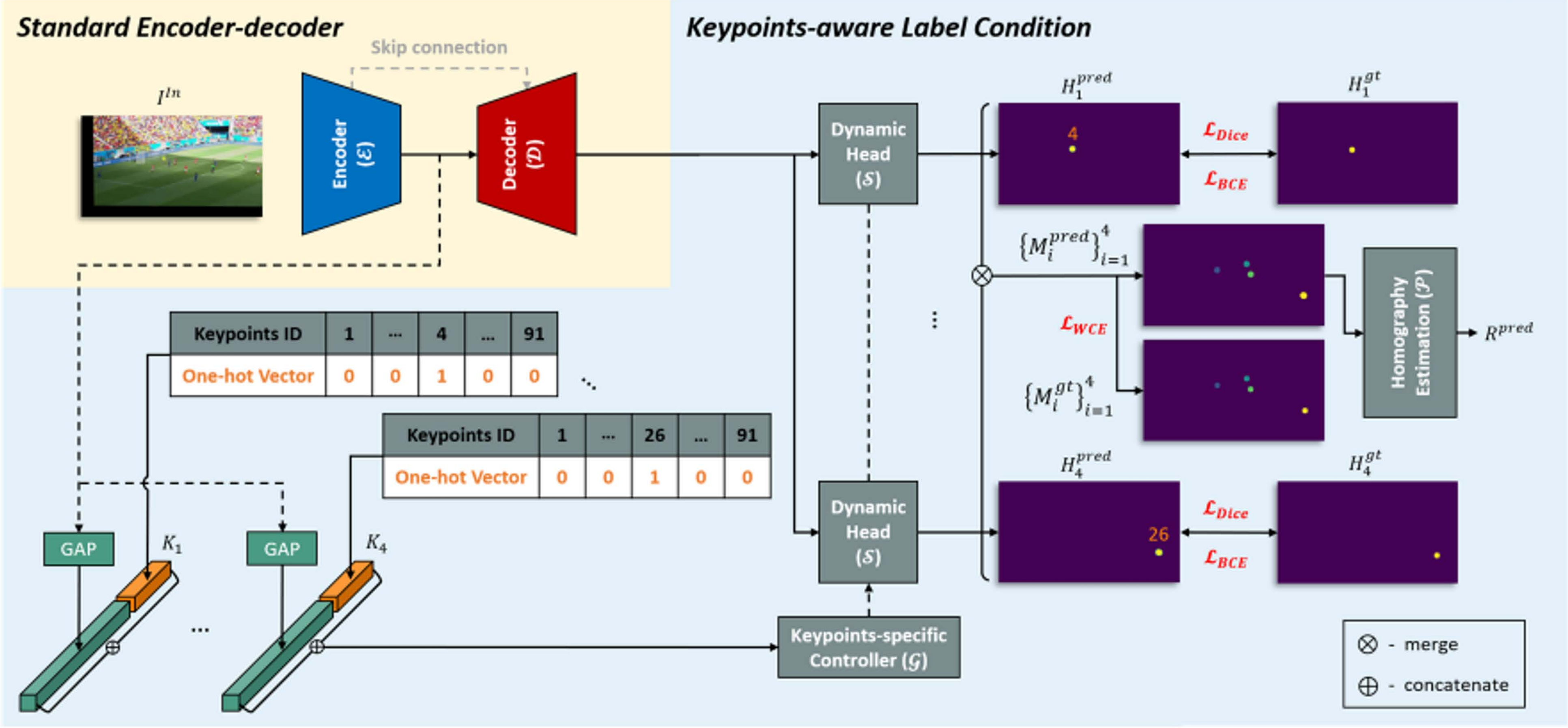
We adopt a encoder-decoder structure to extract the feature map of input field image.

Keypoints-aware Label Condition

- Dynamic Filter Generation

  The parameters of convolution filters are generated dynamically by keypoints-specific controller and conditioned on both the field image and the assigned keypoints.
- Dynamic Head

  Leverage three convolution layers based on learned convolution filters and employ soft aggregation for merging heatmaps to get the final result.



Loss functions

We adopt three loss functions to train our model.

- Binary Dice Loss

  It is proven helpful for addressing the data imbalance problem between foreground and background.
- Binary Cross Entropy Loss

  It is commonly used in the binary classification problems.
- Weighted Cross Entropy Loss

  It tackles the data imbalance problem by assigning weight to each class.

➢ Evaluation

TS-WorldCup Dataset

We create a new soccer dataset with detailed field markings on 3812 field images from 43 videos of Soccer World Cup 2014 and 2018. It is beneficial for temporal evaluation due to contains time-sequence frames.

Quantitative Results

Our method outperforms state-of-the-arts on our collected TS-WorldCup dataset. The symbol $*$ denotes the methods that are finetuned on the TS-WorldCup training set.

Method	$\mathbf{IOU_{whole}}$ (%) ↑		$\mathbf{IOU_{part}}$ (%) ↑		**Proj. (meter)** ↓		**Re − Proj.** ↓	
	mean	median	mean	median	mean	median	mean	median
Chen et al.	89.0	92.2	96.8	97.6	0.65	0.47	0.020	0.017
Nie et al.	90.1	92.8	96.6	97.4	0.57	0.51	0.015	0.012
Ours	**93.2**	**94.3**	**97.6**	**97.7**	**0.45**	**0.41**	**0.012**	**0.011**
Chen et al. *	90.7	94.1	96.8	97.4	0.54	0.38	0.016	0.013
Nie et al. *	92.5	94.2	97.4	97.8	0.43	0.38	0.011	0.010
Ours *	**94.8**	**95.4**	**98.1**	**98.2**	**0.36**	**0.33**	**0.009**	**0.008**

Qualitative Results

➢ Conclusion

- We estimate a robust homography based on a grid of uniformly distributed keypoints and instance segmentation with dynamic filter learning.
- We compile a new soccer dataset, called TS-WorldCup, by annotating time-sequence field-frames.