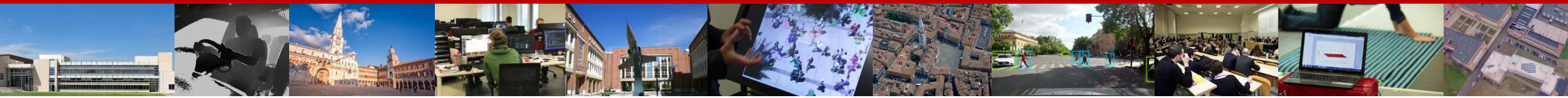
$\mathcal{RMS}$-Net:
Regression and Masking for Soccer Event Spotting



Matteo Tomei[1], Lorenzo Baraldi[1], Simone Calderara[1], Simone Bronzin[2], Rita Cucchiara[1]

[1]University of Modena and Reggio Emilia (name.surname@unimore.it)

[2]Metaliquid (name.surname@meta-liquid.com)

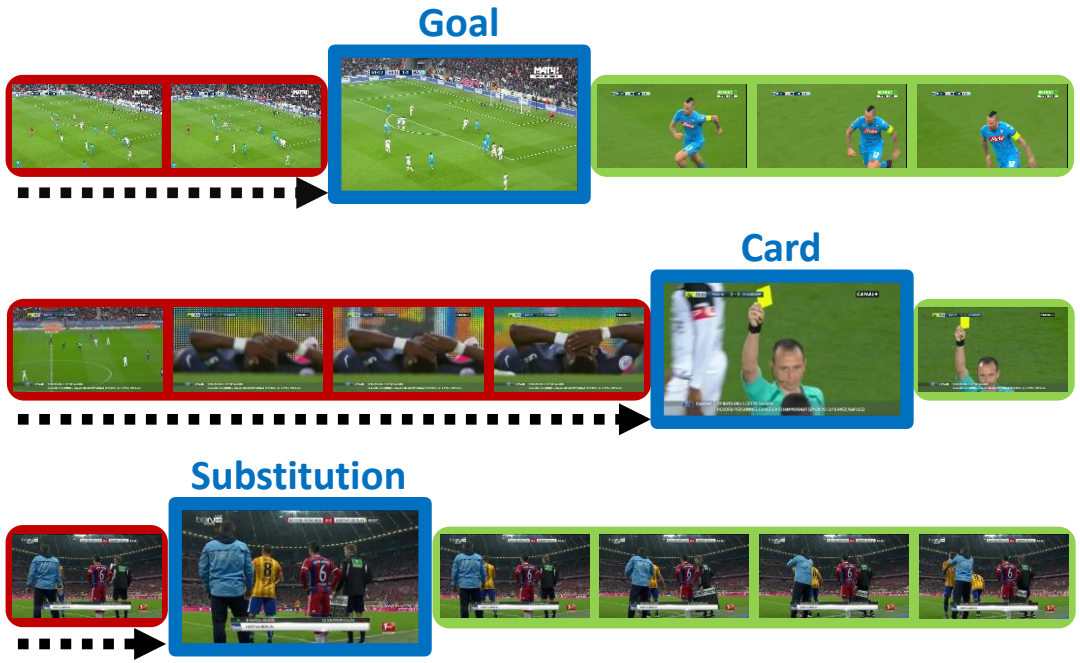
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

Centro Interdipartimentale di Ricerca
Softech: ICT per le Imprese

RMS-Net: Regression and Masking for Soccer Event Spotting

Action Spotting:
Action Spotting consists in finding the anchor time (or spot) that identifies an event [1].
Since perfectly spotting a target is intrinsically arduous, a temporal tolerance δ is introduced to consider an event spotted by a candidate.



SoccerNet [1] dataset:

- 500 full broadcast soccer matches
- 300 games for train, 100 for val, 100 for test
- 3 classes (Goal, Card, Substitution)

Related Works:

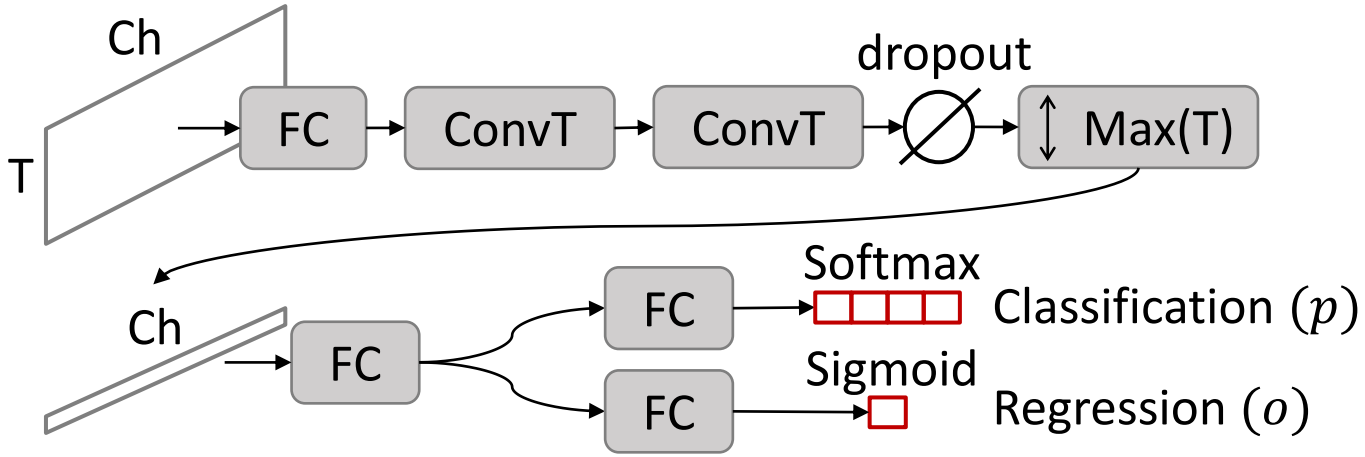
- SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos [1]
- Improved soccer action spotting using both audio and video streams [2]
- Event detection in coarsely annotated sports videos via parallel multi-receptive field 1d convolutions [3]
- A context-aware loss function for action spotting in soccer videos [4]

[1] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "Soccernet: A scalable dataset for action spotting in soccer videos", CVPR Workshops, 2018.
[2] B. Vanderplaetse and S. Dupont, "Improved soccer action spotting using both audio and video streams", CVPR Workshops,2020.
[3] K. Vats, M. Fani, P. Walters, D. A. Clausi, and J. Zelek, "Event detection in coarsely annotated sports videos via parallel multi-receptive field 1d convolutions" CVPR Workshops, 2020.
[4] A. Cioppa, A. Deliege, S. Giancola, B. Ghanem, M. V. Droogenbroeck,R. Gade, and T. B. Moeslund, "A context-aware loss function for action spotting in soccer videos", CVPR, 2020.

RMS-Net: Regression and Masking for Soccer Event Spotting

Action Spotting:

1. Consider a video clip $X = (x_1, x_2, \dots, x_T)$ from a match.
2. Minimize the cross-entropy loss between predicted event class p_e and ground truth event e.
3. Minimize the squared-error loss between predicted relative offset o and ground truth relative offset r.
4. Convert relative timestamp to absolute timestamp.

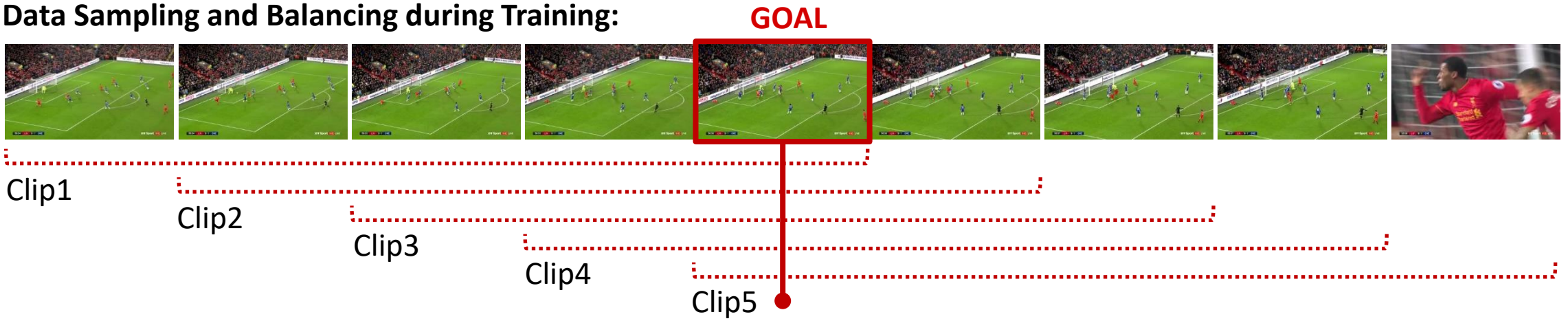


$$\mathcal{L}_{cls} = -\sum_{c=0}^{C} 1_{c=e} \log(p_c)$$

$$\mathcal{L}_{regr} = (\sigma(o) - r)^2$$

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{regr}$$

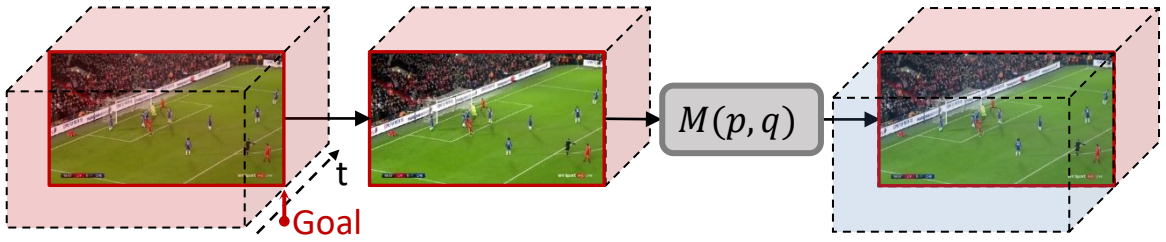
Data Sampling and Balancing during Training:



RMS-Net: Regression and Masking for Soccer Event Spotting

Masking Strategy:
Since the majority of visual cues that contribute to the recognition of an event occur just after the event [1], we propose a masking function which encourages the network to learn robust features after the event. During training, our function randomly replaces the frames before an event with a background clip, as follows:

$$M(p,q)(X) = \begin{cases} (z_1, \ldots, z_{t-s-1}, x_{t-s}, \ldots, x_T) \\ \quad \text{if } r \leq q,\ u < p \\ \\ (x_1, \ldots, x_{t-s-1}, x_{t-s}, \ldots, x_T) \\ \quad \text{otherwise,} \end{cases}$$



Where:

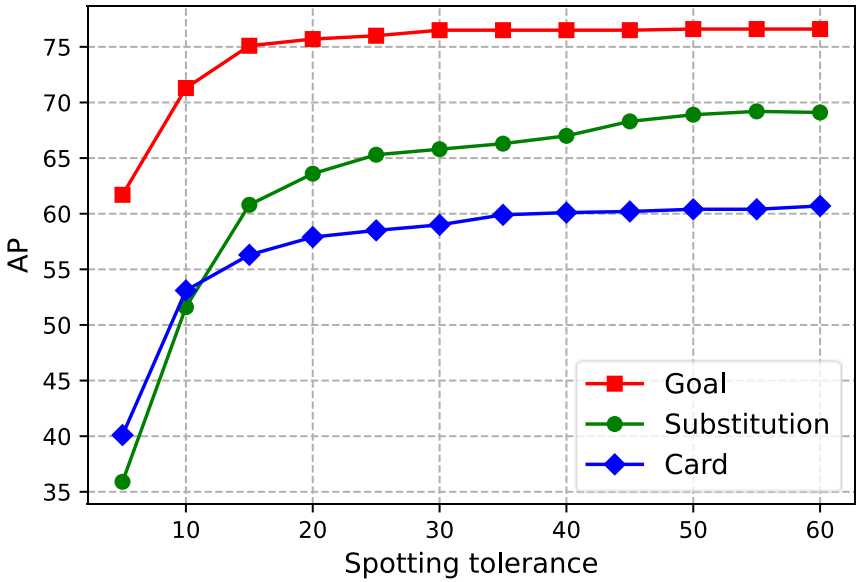
- p is a fixed masking probability
- q is the maximum relative temporal offset in the clip to allow masking
- s is the starting absolute timestamp of the video clip
- r is the relative timestamp of the event in the clip
- $(z_i)_{i=1}^{t-s-1}$ is a sequence of frames selected from a random background clip
- u is a random value sampled from the uniform distribution $U[0,1]$

RMS-Net: Regression and Masking for Soccer Event Spotting

Evaluation Metric:

Given a temporal tolerance δ, the AP for a class is computed by considering a prediction as positive if the distance from its closest ground truth spot is less than δ. The mAP is the average of the AP of each class. The **Average-mAP** is the area under the mAP curve obtained by varying δ from 5 to 60 seconds.

Per-class Average Precision,
as a function of spotting tolerance.



Comparison with baselines and state-of-the-art approaches using ResNet-152 features released with SoccerNet [1].

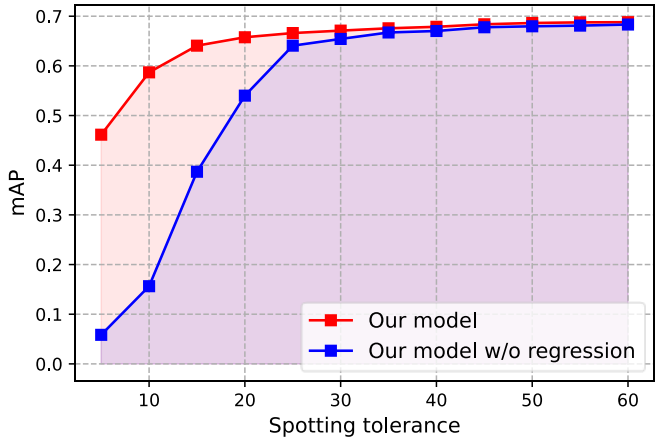
Model	Clip length (s)	Features	Val Avg-mAP	Test Avg-mAP
SoccerNet baseline [1]	5	ResNet-152 (PCA)	-	34.5
SoccerNet baseline [1]	60	ResNet-152 (PCA)	-	40.6
SoccerNet baseline [1]	20	ResNet-152 (PCA)	-	49.7
Vanderplaetse *et al.* [2]	20	ResNet-152 (PCA) + Audio	-	56.0
Vats *et al.* [3]	15	ResNet-152 (PCA)	-	60.1
Cioppa *et al.* [4]	120	ResNet-152 (PCA)	-	62.5
Ours	20	ResNet-152 (PCA)	**67.8**	**65.5**

[1] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "Soccernet: A scalable dataset for action spotting in soccer videos", CVPR Workshops, 2018.
[2] B. Vanderplaetse and S. Dupont, "Improved soccer action spotting using both audio and video streams", CVPR Workshops,2020.
[3] K. Vats, M. Fani, P. Walters, D. A. Clausi, and J. Zelek, "Event detection in coarsely annotated sports videos via parallel multi-receptive field 1d convolutions" CVPR Workshops, 2020.
[4] A. Cioppa, A. Deliege, S. Giancola, B. Ghanem, M. V. Droogenbroeck,R. Gade, and T. B. Moeslund, "A context-aware loss function for action spotting in soccer videos", CVPR, 2020.

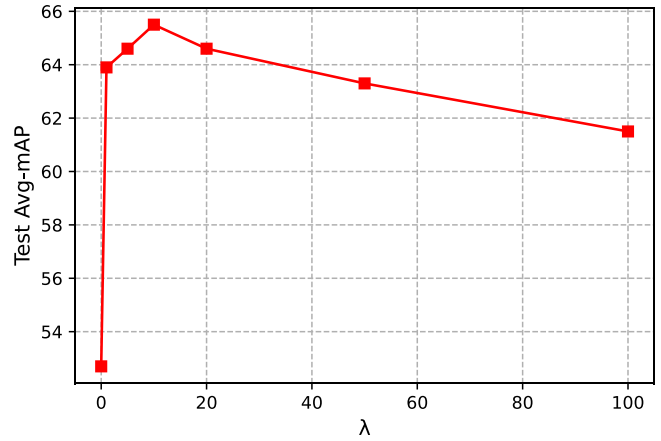
RMS-Net: Regression and Masking for Soccer Event Spotting

AImage Lab

mAP when $\delta \in [5, 60]$ seconds, for our model and when removing the regression branch.



Performance when varying the regression weight λ.



Performance when varying the masking probability p and the maximum relative temporal offset q for masking.

p	q	Val Avg-mAP	Test Avg-mAP
1/5	0.5	66.8	64.6
1/4	0.5	67.0	64.4
1/3	0.5	**67.8**	**65.5**
1/2	0.5	67.1	64.4
1	0.5	64.7	60.7
1/3	0.1	65.5	63.4
1/3	0.25	67.4	64.7
1/3	0.5	**67.8**	**65.5**
1/3	0.75	66.5	64.0
1/3	1	64.7	62.6

Performance when varying p, keeping $q = 0.5$ fixed and masking frames after the event.

p	q	Val Avg-mAP	Test Avg-mAP
1/5	0.5	65.2	62.5
1/4	0.5	64.6	62.9
1/3	0.5	63.8	61.8
1/2	0.5	61.4	60.7
1	0.5	54.1	54.1

RMS-Net: Regression and Masking for Soccer Event Spotting

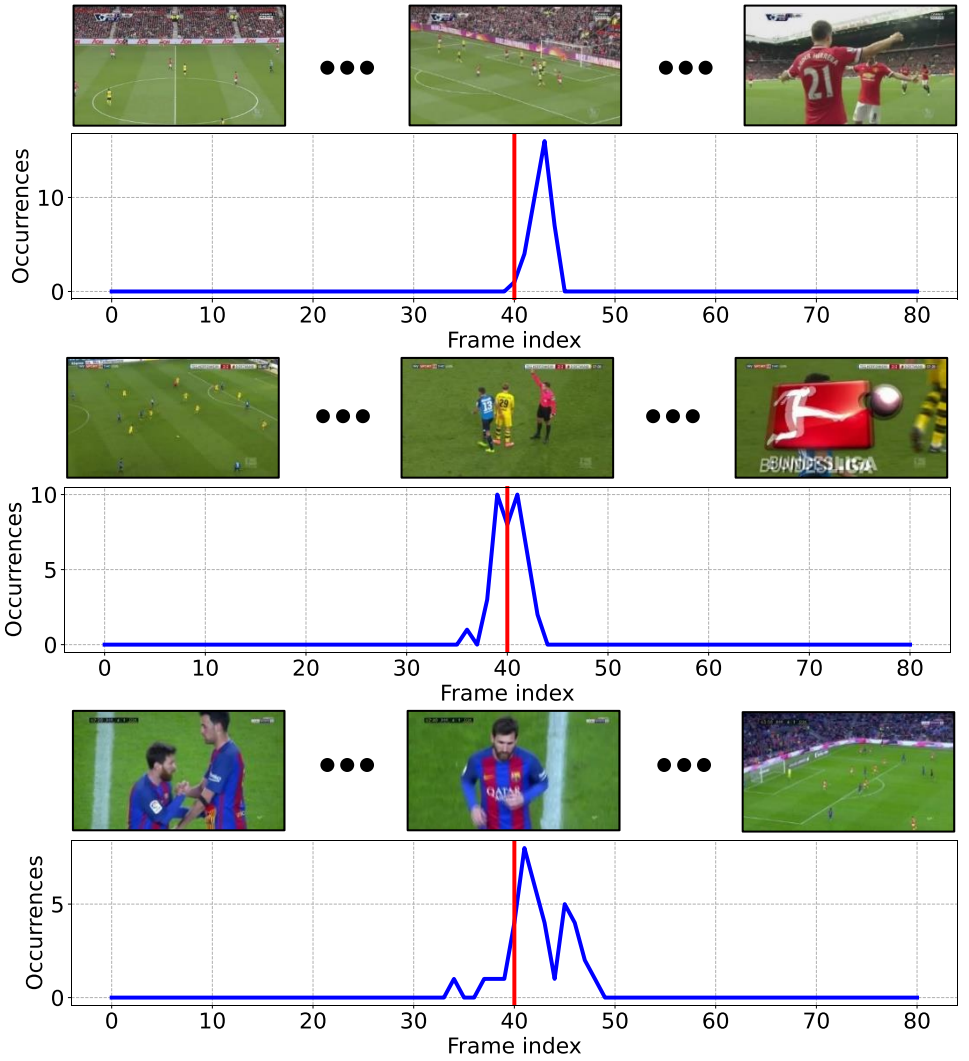
AImage Lab

Performance of the proposed model when removing key components.

Model	Val Avg-mAP	Test Avg-mAP
Ours	**67.8**	**65.5**
Ours w/o uniformly distributed offsets	48.7	46.2
Ours w/o offset regression branch	58.5	55.7
Ours w/o masking	66.5	64.0

Performance analysis when finetuning different variants of ResNet.

Model	Pre-train	Val Avg-mAP	Test Avg-mAP
ResNet-18 + Our	ImageNet	73.8	70.9
ResNet-50 + Our	ImageNet	76.6	74.9
ResNet-152 + Our	ImageNet	**77.5**	**75.1**



Thank you for your attention

$\mathcal{RMS}$-Net: Regression and Masking for Soccer Event Spotting

Matteo Tomei[1], Lorenzo Baraldi[1], Simone Calderara[1], Simone Bronzin[2], Rita Cucchiara[1]

[1]University of Modena and Reggio Emilia (name.surname@unimore.it)

[2]Metaliquid (name.surname@meta-liquid.com)