SportMamba: Adaptive Non-Linear Multi-Object Tracking with State Space Models for Team Sports

Dheeraj Khanna, Jerrin Bright, Yuhao Chen, John Zelek
University of Waterloo, Waterloo, Ontario, Canada

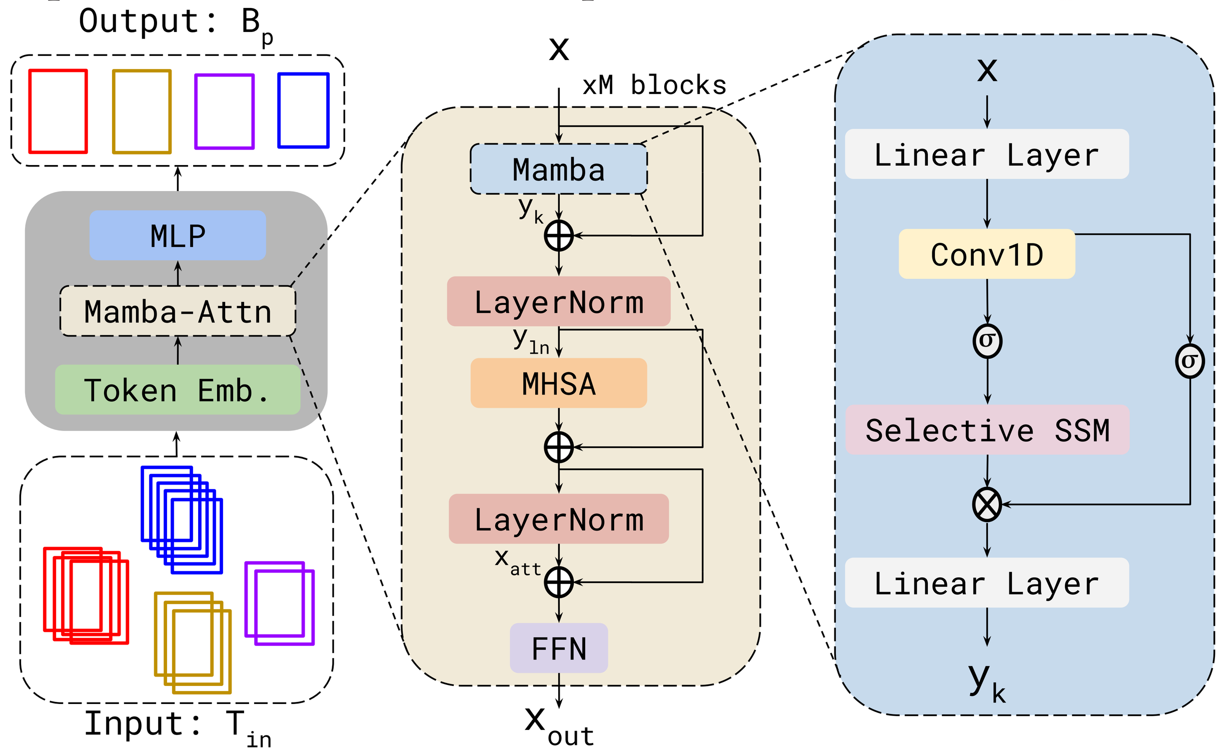
CVPR Nashville, June 11-15, 2025

Key Contributions

- **SportMamba:** A hybrid online tracker designed for fast, non-linear motion in sports.
- **Motion Model:** Integrates Mamba state-space and self-attention for accurate motion prediction.
- **Spatial Matching:** Uses height-adaptive IoU and buffered association for robust tracking.
- **Performance:** Achieves SOTA on SportsMOT and strong zero-shot generalization to VIP-HTD.

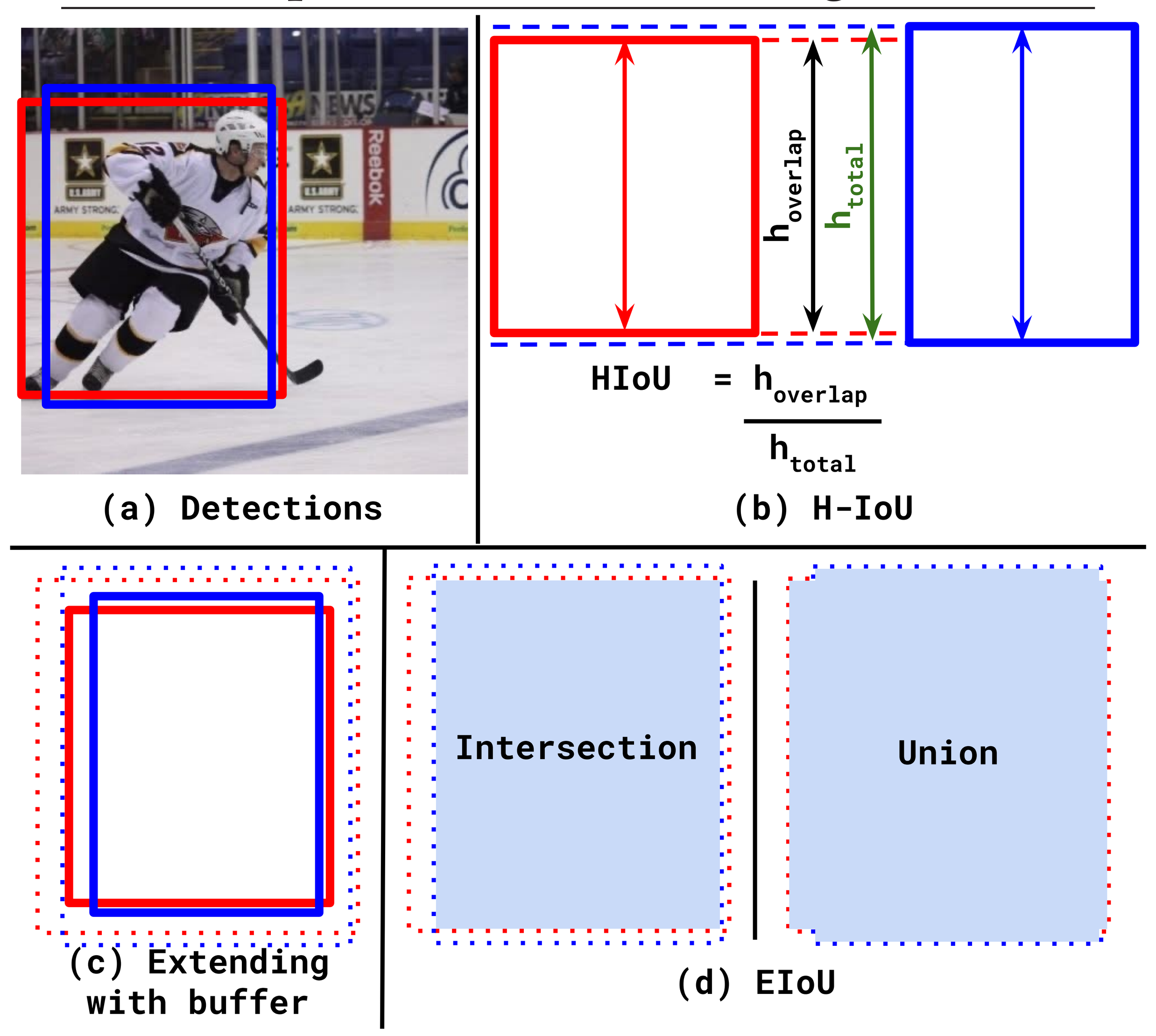
Motion Model

Token embeddings encode past trajectories, which are processed through Mamba-Attention blocks and MHSA to capture long-range motion dependencies. This representation is passed to an MLP head to predict future bounding boxes.



Spatial Matching

Visual Representation of HIoU against EIoU

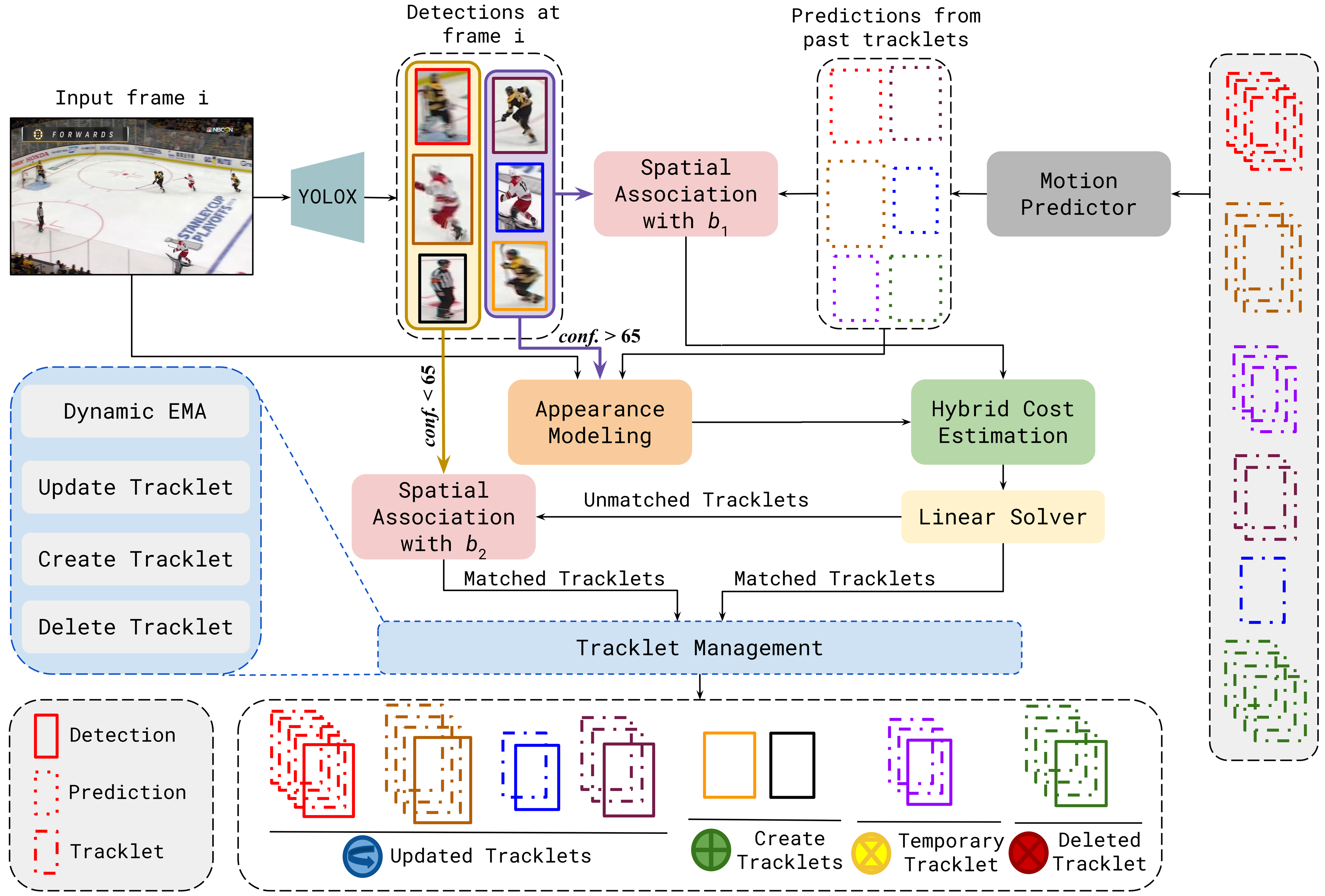


Spatial association is done using Height-Adaptive EIoU (HA-EIoU) defined as:

$$\text{HA-EIoU} = \text{HIoU}.\text{EIoU} \tag{1}$$

SportMamba

Motivation:
Tracking players in team sports is extremely challenging due to fast, non-linear motion, frequent occlusions, and similar appearances (e.g., jerseys). Traditional Kalman Filter-based or appearance-only trackers **struggle in these settings**. While transformer-based models offer better modeling, they are often too heavy for real-time use.



Method:

1. **Fine-tuned Detector:** Identifies players in each frame.
2. **Motion Predictor:** Forecasts future positions using past trajectories.
3. **High-Confidence Matching:** Uses appearance features and height-adaptive IoU for robust associations.
4. **Fallback Matching:** Applies relaxed IoU when appearance cues are unreliable.
5. **Tracklet Management:** Creates, updates, deletes tracklets and updates features via dynamic EMA.

Qualitative Results

Hybrid Cost Estimation

The hybrid cost matrix is estimated as a weighted sum of appearance and spatial similarity, defined as:

$$\begin{aligned}\mathcal{J}_f &= \lambda_{reid}\mathcal{J}_{reid} + \lambda_{ssim}\mathcal{J}_{ssim}\\ \mathcal{J}_f &= \lambda_{reid}(1 - S_{reid}) + \lambda_{ssim}(1 - \text{HA-EIoU})\end{aligned} \tag{2}$$

where,

$$S_{reid}(i,j) = \frac{e_i^T \cdot e_j^D}{\|e_i^T\|\|e_j^D\|} \tag{3}$$

Loss Functions

The overall objective function for the motion predictor:

$$\mathcal{L} = \lambda_{l1}^s \mathcal{L}_{L1}^s + \lambda_{ciou}\mathcal{L}_{ciou} \tag{4}$$

where,

$$\mathcal{L}_{L1}^s = \begin{cases}\frac{1}{2}(P_t - G_t)^2, & \text{if } |P_t - G_t| < 1,\\ |P_t - G_t| - \frac{1}{2}, & \text{otherwise.}\end{cases} \tag{5}$$

Quantitative Results

Tracking results for SportsMOT test set

Method	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
Filter-based Methods					
ByteTrack	62.8	69.8	51.2	94.1	77.1
OC-SORT	71.9	72.2	59.8	94.5	86.4
*OC-SORT	73.7	74.0	61.5	96.5	88.5
Learning-based Methods					
DiffMOT	72.1	72.8	60.5	94.5	86.0
*ByteSSM	74.4	74.5	62.4	96.8	88.8
Ours	77.3	77.7	66.8	96.9	89.5

Zero-shot Tracking results for VIP-HTD test set

Method	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
Filter-based Methods					
ByteTrack	64.4	81.1	64.8	73.9	64.2
OC-SORT	61.0	75.4	58.9	74.6	63.4
Deep OC-SORT	59.4	73.4	56.1	74.5	56.1
Learning-based Methods					
DiffMOT	64.1	79.4	63.6	76.1	65.0
ByteSSM	63.4	77.7	61.8	76.2	65.4
Ours	65.1	80.1	64.6	76.2	65.9

Acknowledgement

VIPlab Vision and Image Processing Research Group vip.uwaterloo.ca
NSERC CRSNG
Mitacs
Stathletes
SCAN ME