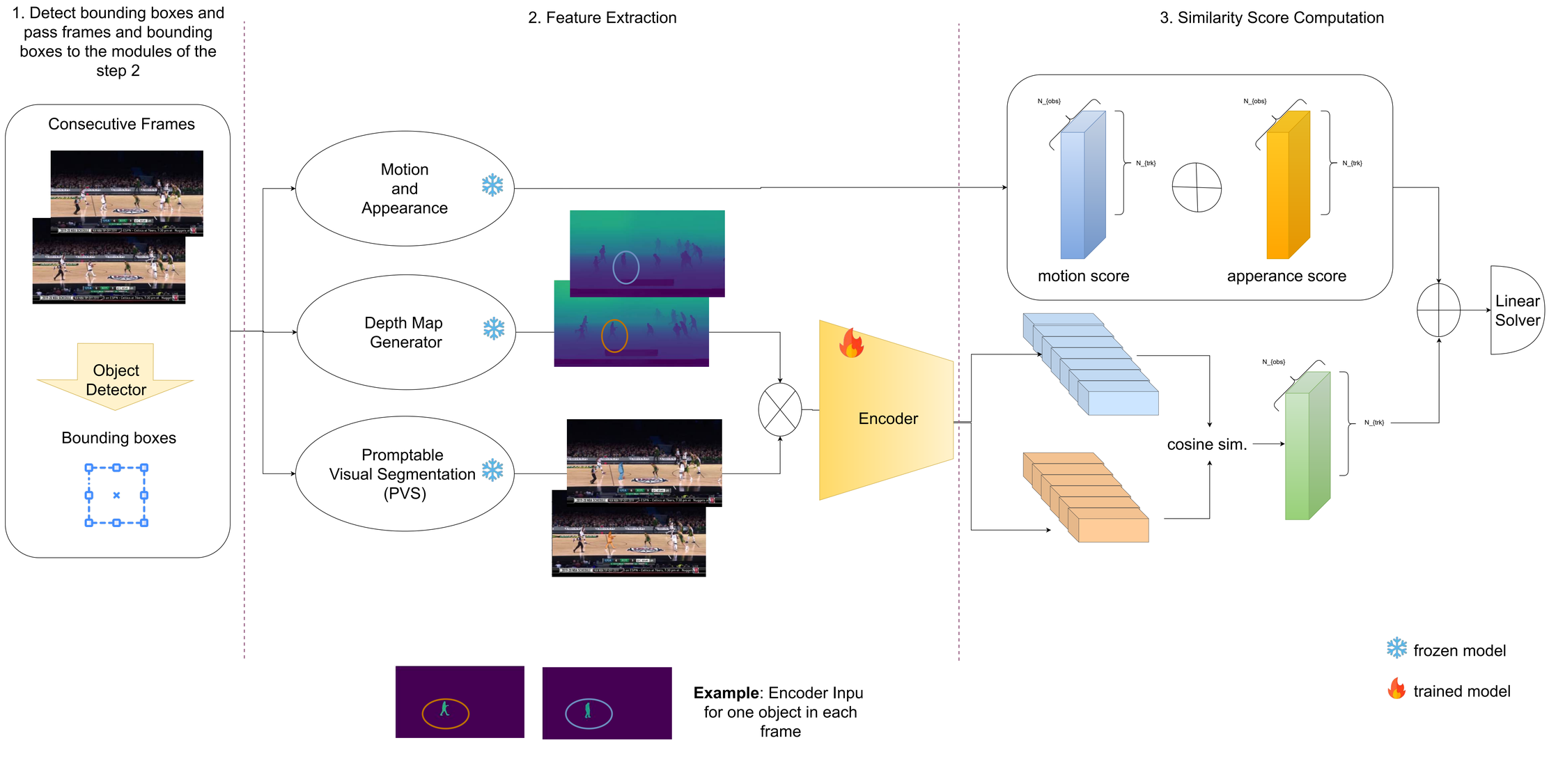
Fast Self-Supervised depth and mask aware Association for Multi-Object Tracking

Milad Khanchi, Maria Amer, Charalambos Poullis
Concordia University, Montreal, Quebec, Canada

BMVC 2025

Method



Results

Method	HOTA↑	IDF1↑	AssA↑	MOTA↑	FP(10^4)↓	FN(10^4)↓	ID_s↓	Frag↓	AssR↑
FairMOT [36]	59.3	72.3	58.0	73.7	2.75	11.7	3,303	8,073	63.6
CenterTrack [40]	52.2	64.7	-	67.8	1.8	1.6	3,039	-	-
TransTrack [29]	54.1	63.5	47.9	75.2	5.02	8.64	3,603	4,872	57.1
MOTRv2 [38]	**62.0**	**75.0**	**60.6**	**78.6**	-	-	-	-	-
ByteTrack [37]	63.1	77.3	62.0	80.3	2.55	8.37	2,196	2,277	68.2
MotionTrack [33]	65.1	80.1	-	**81.1**	2.38	**8.16**	1,140	-	-
OC-SORT [7]	63.2	77.5	63.2	78.0	**1.51**	10.8	1,950	2,040	67.5
StrongSORT++ [12]	64.4	79.5	64.4	79.6	2.79	8.62	1,194	1,866	**71.0**
GeneralTrack [24]	64.0	78.3	63.1	80.6	-	-	1,563	-	-
C-BIoU [35]	64.1	79.7	63.7	**81.1**	-	-	-	-	-
Deep OC-SORT [22]	64.9	80.6	65.9	79.4	1.66	9.88	1,023	2,196	70.1
CMTrack [28]	**65.5**	**81.5**	**66.1**	80.7	2.59	8.19	**912**	**1,653**	-
DiffMOT [21]	64.5	79.3	64.6	79.8	-	-	-	-	-
SelfTrEncMOT	63.48	78.12	63.25	79.16	1.9	9.6	1,008	-	68.8

Comparison on the MOT17 test set. The JDR methods are in the upper part, and the TBD methods are in the lower part. Methods using YOLOX are highlighted in blue.

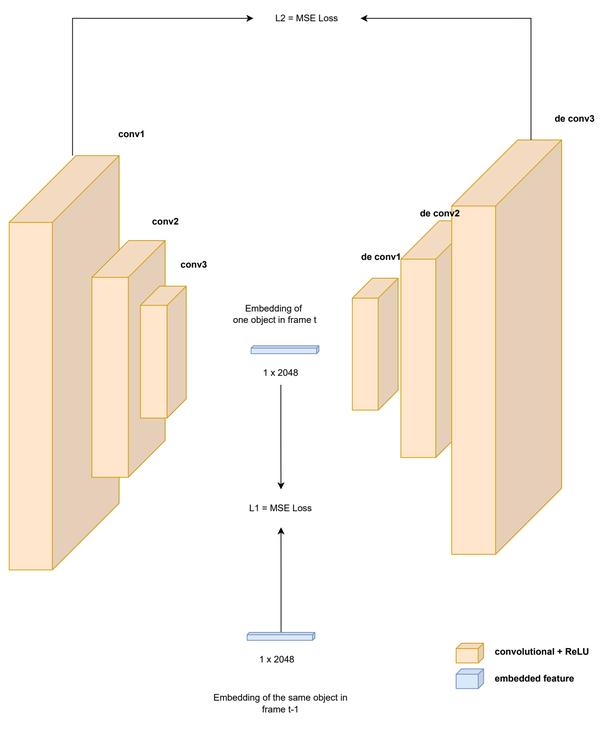
Method	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
FairMOT [36]	49.3	53.5	34.7	86.4	70.2
CenterTrack [40]	62.7	60.0	48.0	90.8	81.7
TransTrack [29]	**68.9**	**71.5**	**57.5**	**92.6**	**82.7**
ByteTrack [37]	62.8	69.8	51.2	94.1	77.1
BoT-SORT [1]	68.7	70.0	55.9	94.5	84.4
OC-SORT [7]	71.9	72.2	59.8	94.5	86.4
DiffMOT [21]	72.1	72.8	60.5	94.5	86.0
*ByteTrack [37]	64.1	71.4	52.3	95.9	78.5
*MixSort-Byte [10]	65.7	74.4	58.4	96.2	78.8
*OC-SORT [7]	73.7	74.0	61.5	96.5	88.5
*MixSort-OC [10]	74.1	74.4	62.0	96.5	88.5
*GeneralTrack [24]	74.1	76.4	61.7	96.8	89.0
*DiffMOT [21]	76.2	76.1	65.1	**97.1**	**89.3**
*SelfTrEncMOT	**76.4**	**77.1**	**66.0**	95.84	88.4

Comparison with MOT trackers on the SportsMOT test set. The lower part shows TBD methods, which are relevant to ours. Methods marked with * indicate the detector was trained on SportsMOT's train and validation sets. YOLOX-based methods are highlighted in blue.

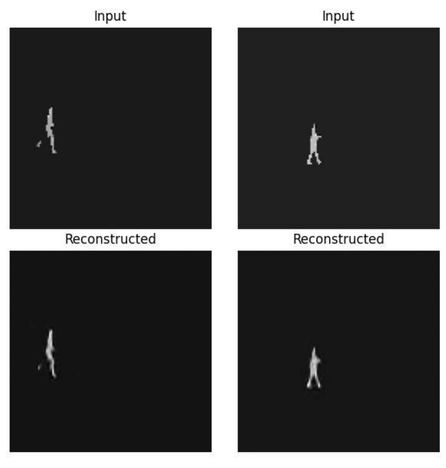
Method	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
FairMOT [36]	39.7	40.8	23.8	82.2	66.7
CenterTrack [40]	41.8	35.7	22.6	86.8	78.1
TransTrack [29]	45.5	45.2	27.5	85.4	75.9
MOTRv2 [38]	**69.9**	**71.7**	**59.0**	**91.9**	**83.0**
ByteTrack [37]	47.3	52.5	31.4	89.5	71.6
MotionTrack [33]	52.9	53.8	34.7	91.3	80.9
OC-SORT [7]	55.1	54.2	38.0	89.4	80.3
StrongSORT++ [12]	55.6	55.2	38.6	91.1	80.7
GeneralTrack [24]	59.2	59.7	42.8	91.8	82.0
C-BIoU [35]	60.6	61.6	45.4	91.8	81.3
Deep OC-SORT [22]	61.3	61.5	45.8	92.3	82.2
CMTrack [28]	61.8	63.3	46.4	92.5	-
DiffMOT [21]	62.3	63.0	47.2	**92.8**	**82.5**
SelfTrEncMOT	**64.14**	**66.47**	**50.85**	90.08	81.06

Comparison on the DanceTrack test set. Methods are grouped into JDR in the upper part and TBD in the lower part. Methods using YOLOX are highlighted in blue.

Autoencoder



Qualitative Results



Qualitative Results of the Depth-Segmentation Autoencoder: The encoder preserves key spatial details and object boundaries, supporting robust association, despite variations in scale and structure.

Contributions

Self-supervised encoder design: We introduce a self-supervised encoder that enhances the *temporal stability* and *discriminability* of depth-segmentation features.

Novel MOT framework: Our method is the **first** to leverage the self-supervised encoder to refine segmentation masks and incorporate them directly into the matching score, **without computing mask IoU**.

Strong performance: We achieve **competitive results** across diverse tracking scenarios, showing particular robustness in **occluded scenes**.