Learning a Neural Association Network for Self-supervised Multi-object Tracking

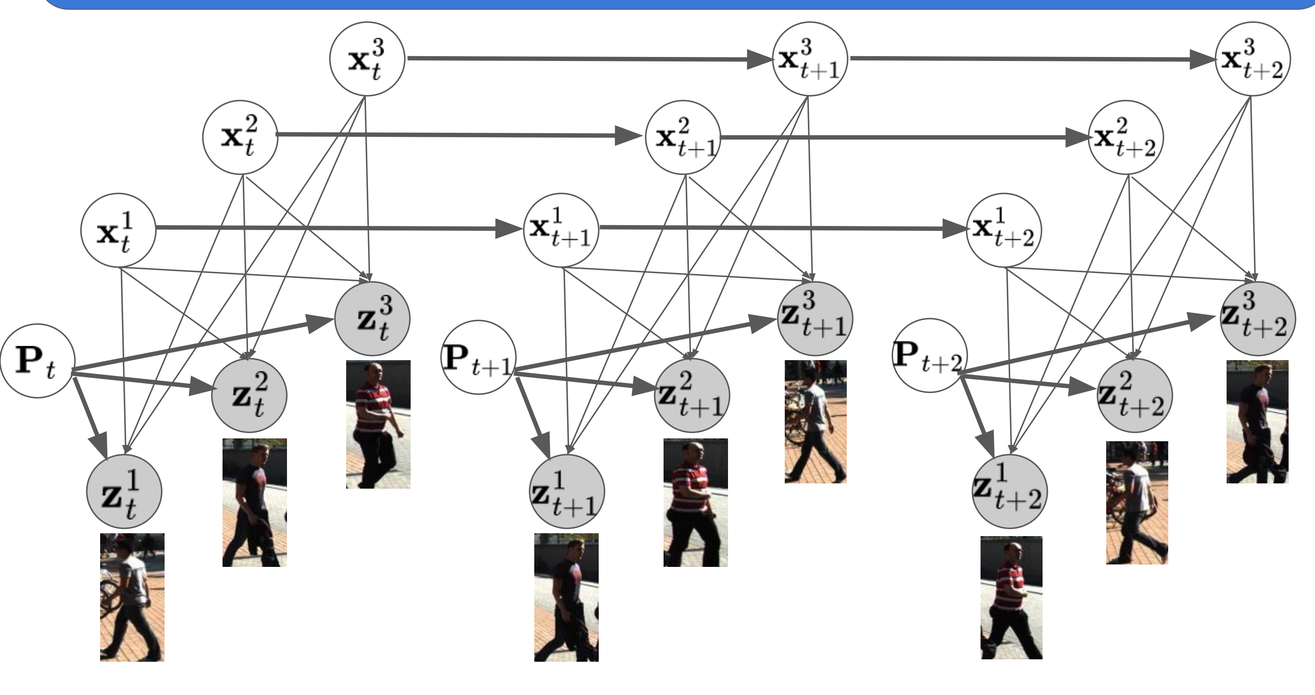
Shuai Li[1], Michael Burke[2], Subramanian Ramamoorthy[3], Juergen Gall[1]
[1]University of Bonn [2]Monash University [3]University of Edinburgh

BMVC 2025

Introduction

1. Train a NN to associate objects for tracking WITHOUT manual labels
2. Principled learning objective
3. Network is trained in just 10 minutes

Technical Approach



KF Prediction:

$$p(\mathbf{x}_t|\mathbf{z}_{1:t-1},\mathbf{P}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1},\mathbf{P}_{1:t-1})d\mathbf{x}_{t-1}$$
$$= \mathcal{N}(\mathbf{x}_t;\mathbf{F}\mu_{t-1},\mathbf{F}\Sigma_{t-1}\mathbf{F}^T+\mathbf{Q}_t)$$
$$= \mathcal{N}(\mathbf{x}_t;\hat{\mu}_t,\hat{\Sigma}_t).$$

KF Update:

$$p(\mathbf{x}_t|\mathbf{z}_{1:t},\mathbf{P}_{1:t}) = \frac{p(\mathbf{z}_t|\mathbf{x}_t,\mathbf{P}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1},\mathbf{P}_{1:t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1}\mathbf{P}_{1:t})}$$
$$= \mathcal{N}(\mathbf{x}_t;\mu_t,\Sigma_t)$$

Marginal of Observations:

$$p(\mathbf{z}_t|\mathbf{z}_{1:t-1},\mathbf{P}_{1:t}) = \int p(\mathbf{z}_t|\mathbf{x}_t,\mathbf{P}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1},\mathbf{P}_{1:t-1})d\mathbf{x}_t$$
$$= \mathcal{N}(\mathbf{z}_t;\mathbf{H}_t\mathbf{P}_t\hat{\mu}_t,(\mathbf{H}_t\mathbf{P}_t)\hat{\Sigma}_t(\mathbf{H}_t\mathbf{P}_t)^T+\mathbf{R}_t)$$

Learning objective:

$$\arg\min_\theta -\sum_{t=1}^T \log\mathcal{N}(\mathbf{z}_t;\mathbf{H}_t\mathbf{P}_t\tilde{\mu}_t,(\mathbf{H}_t\mathbf{P}_t)\tilde{\Sigma}_t(\mathbf{H}_t\mathbf{P}_t)^T+\mathbf{R}_t)$$

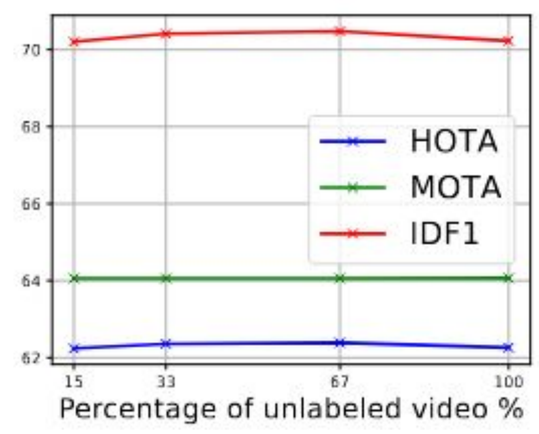
Sinkhorn layer enables end-to-end training
Online inference with a Kalman Filter

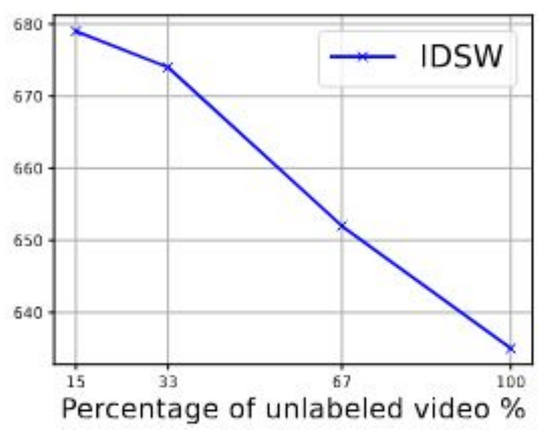
Quantitative Results

Method	Sup.	HOTA ↑	MOTA ↑	IDF1 ↑	IDSW ↓
MHT_BiLSTM [17]	✓	41.0	47.5	51.9	2069
Tracktor++ [2]	✓	44.8	56.3	55.1	1987
TrackFormer [28]	✓	-	57.6	62.3	4018
SUSHI [6]	✓	54.6	62.0	71.5	1041
SORT [4]	✗	-	43.1	39.8	4852
UNS20regress [1]	✗	46.4	56.8	58.3	1320
UnsupTrack [15]	✗	46.9	**61.7**	58.1	1864
Lu *et al.* [26]	✗	<u>49.0</u>	58.8	<u>61.2</u>	**1219**
Ours	✗	**50.3**	<u>60.3</u>	**63.4**	<u>1266</u>

Method	Sup.	HOTA ↑	MOTA ↑	IDF1 ↑	IDSW ↓
Tracktor++V2 [2]	✓	42.1	52.6	52.7	1648
ArTist [36]	✓	-	53.6	51.0	1531
ApLift [14]	✓	46.6	58.9	56.5	2241
SUSHI [6]	✓	55.4	61.6	71.6	1053
SORT20 [4]	✗	36.1	42.7	45.1	4470
Ho *et al.* [12]	✗	-	41.8	-	5918
UnsupTrack [15]	✗	<u>41.7</u>	<u>53.6</u>	<u>50.6</u>	<u>2178</u>
Ours	✗	**47.4**	**59.5**	**58.5**	**1656**

Method	Sup.	mHOTA ↑	mIDF1 ↑	IDF1 ↑	IDSW ↓
MOTR *et al.* [42]	✓	-	43.5	-	-
Yu *et al.* [41]	✓	-	44.5	66.8	8315
QDTrack [33]	✓	41.7	50.8	71.5	6262
ByteTrack [43]	✓	-	54.8	70.4	9140
SORT [4]	✗	27.9	33.8	56.4	**9647**
Ours	✗	**34.5**	**42.2**	**62.3**	23143





Qualitative Results

Conclusions

1. Principled learning objective
2. Fast training (minutes vs hours)
3. SOTA among self-supervised methods (given publice detections) on MOT17/20 and BDD100k