Online Multi-Object Tracking with Dual Matching Attention Networks

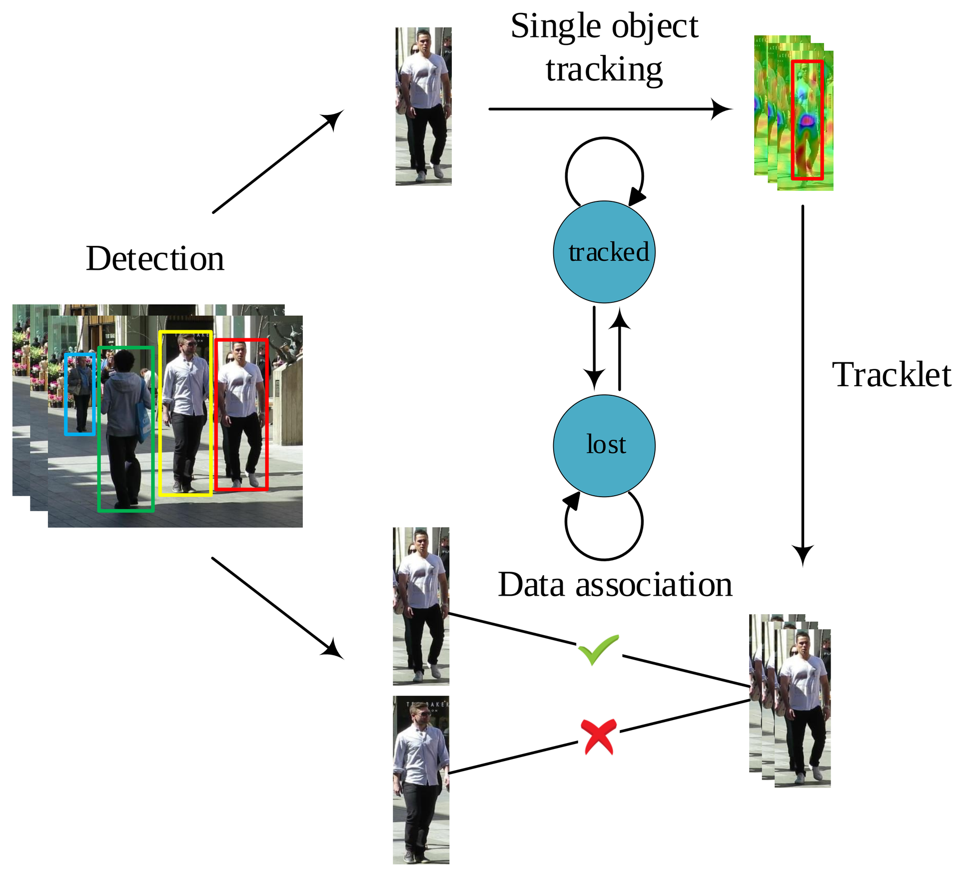
Ji Zhu[1,2] Hua Yang[1] Nian Liu[3] Minyoung Kim[4] Wenjun Zhang[1] Ming-Hsuan Yang[5,6]

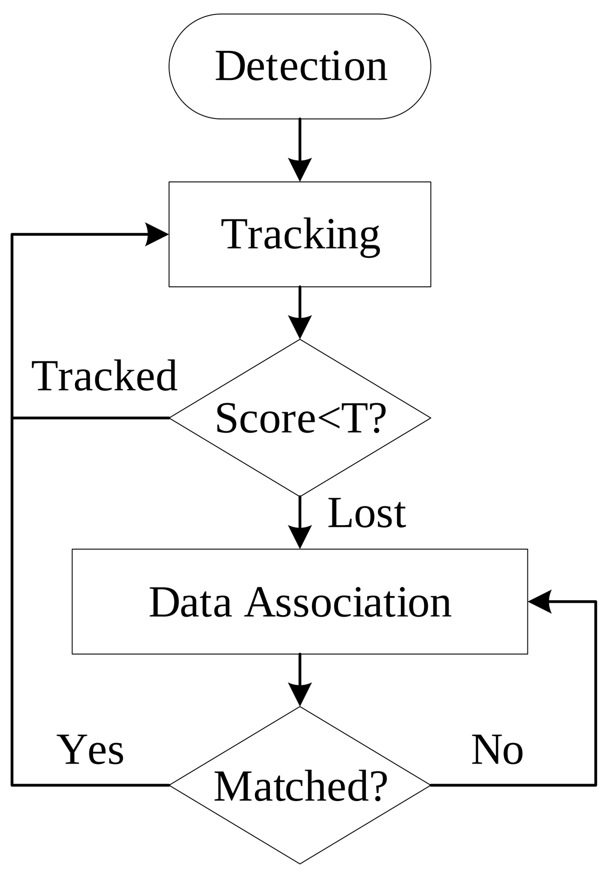
[1]Shanghai Jiao Tong University [2]Visbody Inc [3]Northwestern Polytechnical University

[4]Massachusetts Institute of Technology [5]University of California at Merced [6]Google Cloud AI

https://github.com/jizhu1023/DMAN_MOT

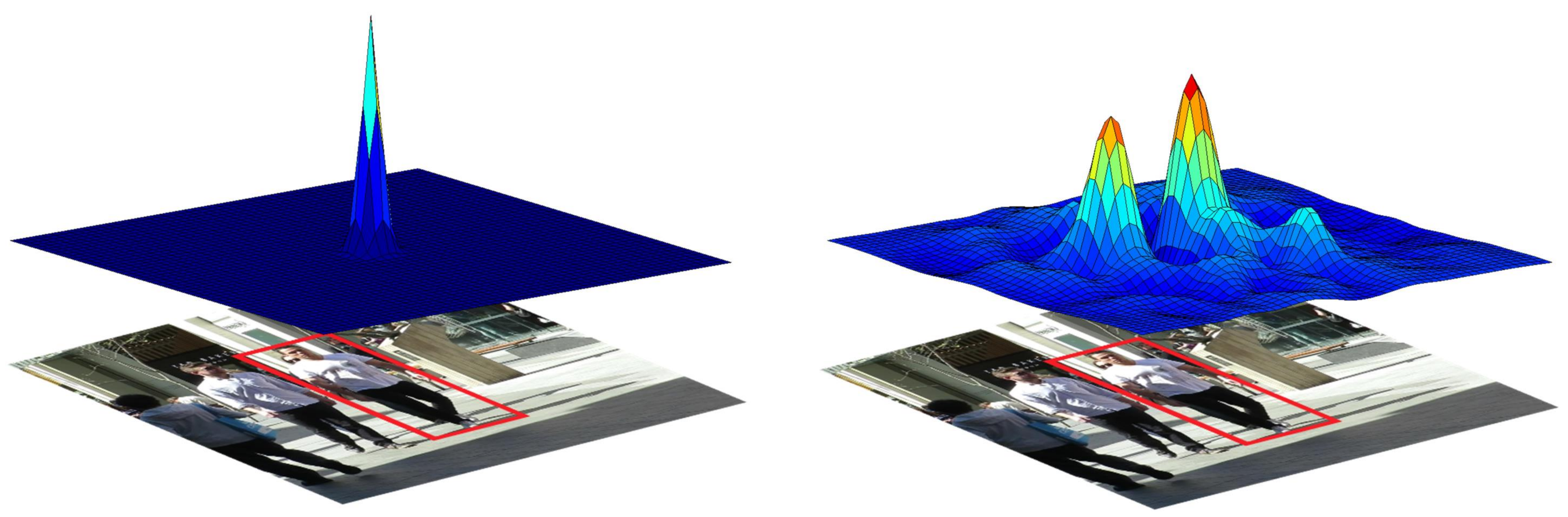
Online MOT Pipeline





Motivations

➢ Data imbalance in tracking



➢ Noisy samples with misalignment, missing part, occlusion

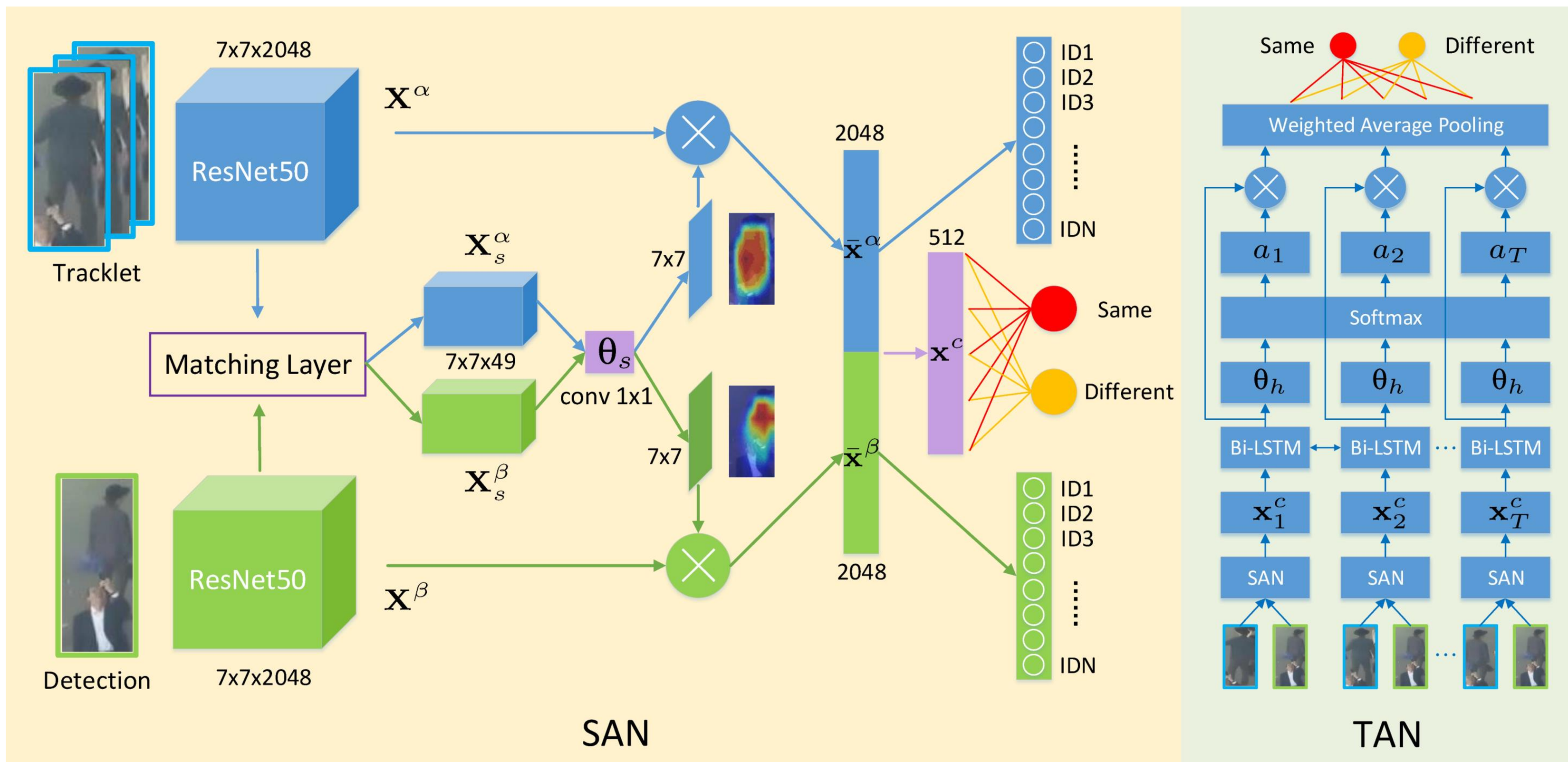
Approach

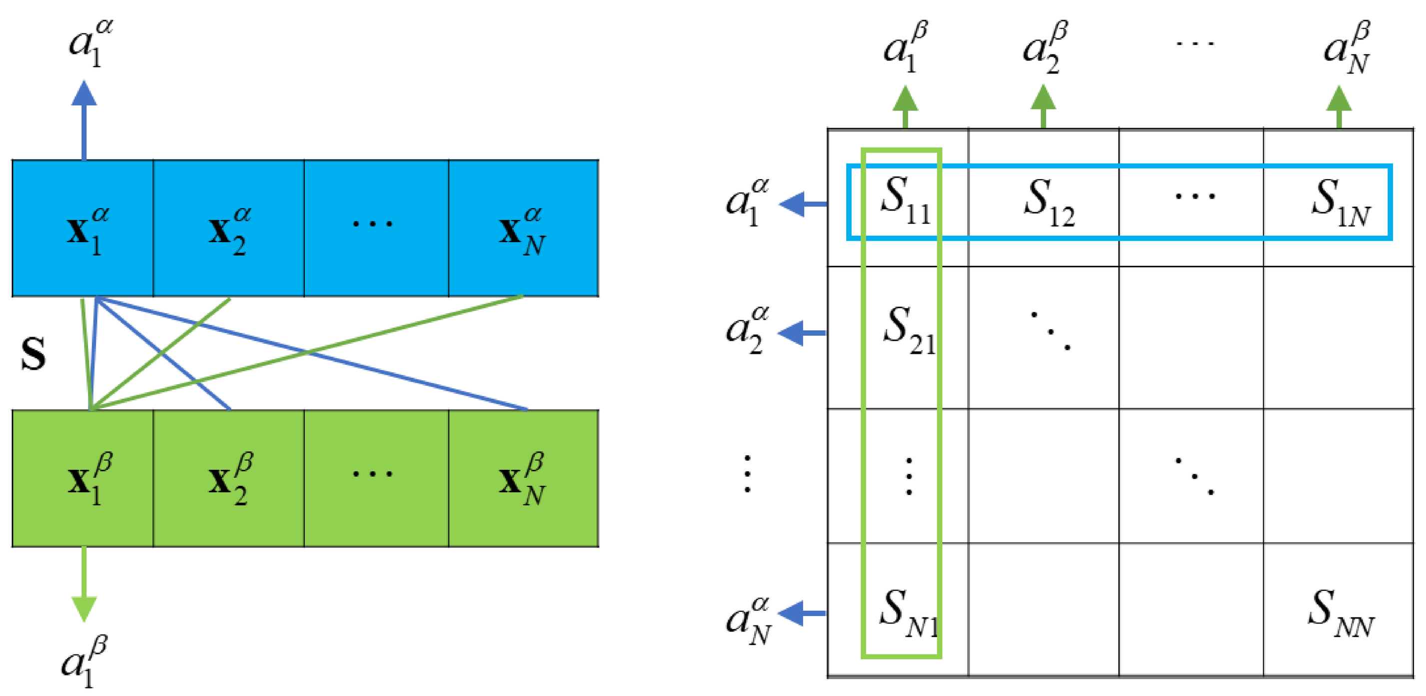
➢ Cost-sensitive loss function for tracking

$$E(f)=\sum_{j=1}^{M}\alpha_j \left\| q(t)(S_f\{\mathbf{x}_j\}(t)-y_j(t)\right\|_{L^2}+\sum_{d=1}^{D}\left\| w(t)f^d(t)\right\|_{L^2}$$

$$q(t)=\left|\frac{S_f\{\mathbf{x}_j\}(t)-y_j(t)}{\max_t |S_f\{\mathbf{x}_j\}(t)-y_j(t)|}\right|^2$$

➢ Spatial-temporal attention networks for data association



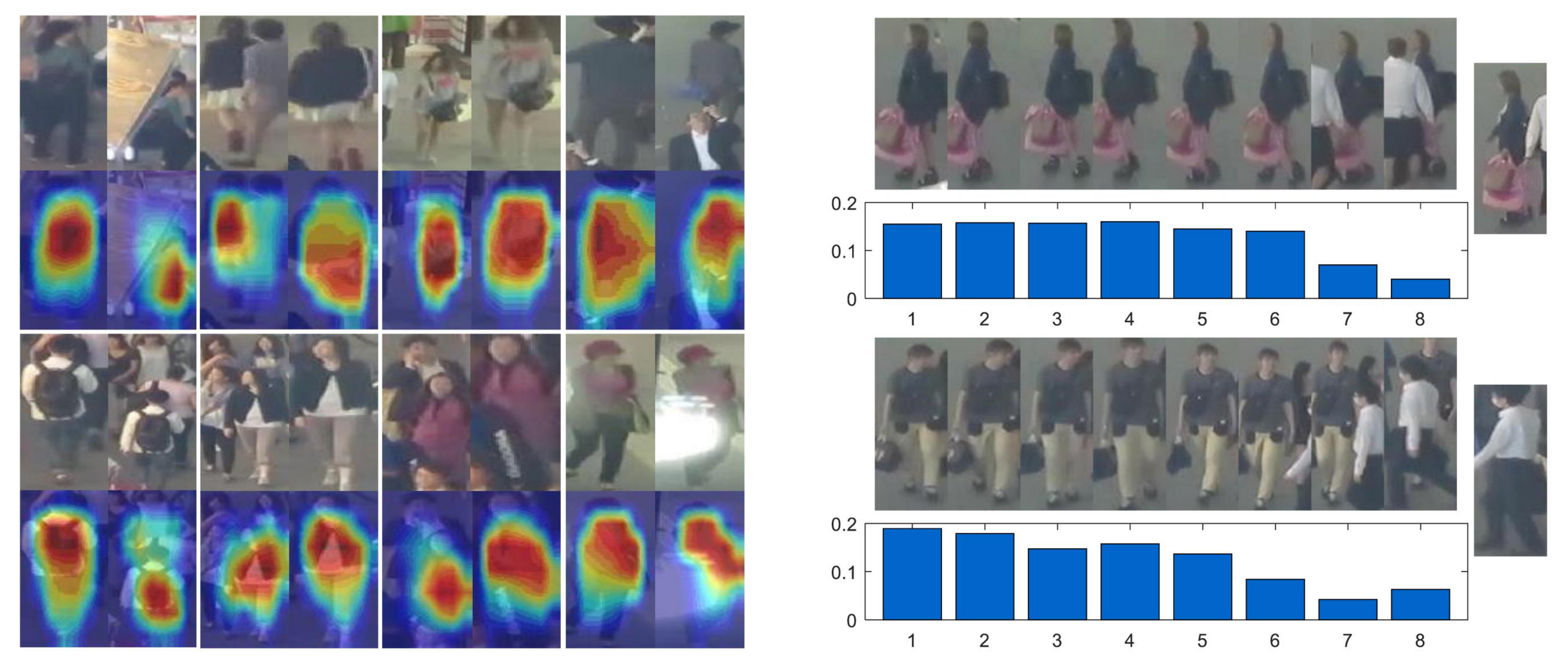


$$\mathbf{X}^\alpha,\mathbf{X}^\beta\in\mathbb{R}^{H\times W\times C},\ \mathbf{X}^\alpha=\{\mathbf{x}_1^\alpha,\cdots,\mathbf{x}_N^\alpha\},\ \mathbf{x}_N^\alpha\in\mathbb{R}^C,\ N=H\times W$$

$$\mathbf{S}=\begin{bmatrix}(\mathbf{x}_1^\alpha)^{\mathrm{T}}\\ \vdots\\ (\mathbf{x}_N^\alpha)^{\mathrm{T}}\end{bmatrix}\cdot\left[\mathbf{x}_1^\beta,\cdots,\mathbf{x}_N^\beta\right]=\begin{bmatrix}(\mathbf{s}_1)^{\mathrm{T}}\\ \vdots\\ (\mathbf{s}_N)^{\mathrm{T}}\end{bmatrix} \qquad a_i^\alpha=\frac{\exp(\boldsymbol{\theta}_s^{\mathrm{T}}\mathbf{s}_i)}{\sum_{i=1}^{N}\exp(\boldsymbol{\theta}_s^{\mathrm{T}}\mathbf{s}_i)}$$

Experiments

➢ Visualization of spatial and temporal attention



➢ Performance on the MOT benchmark datasets

Table 1. Tracking performance on the MOT16 dataset.

Mode	Method	MOTA ↑	MOTP ↑	IDF ↑	IDP ↑	IDR ↑	MT ↑	ML ↓	FP ↓	FN ↓	IDS ↓	Frag ↓	AR ↓
Online	OVBT [3]	38.4	75.4	37.8	55.4	28.7	7.5%	47.3%	11,517	99,463	1,321	2,140	49.8
	EAMTT [43]	38.8	75.1	42.4	65.2	31.5	7.9%	49.1%	8,114	102,452	965	1,657	37.4
	oICF [22]	43.2	74.3	49.3	73.3	37.2	11.3%	48.5%	6,651	96,515	**381**	**1,404**	33.3
	CDA_DDAL [2]	43.9	74.7	45.1	66.5	34.1	10.7%	44.4%	6,450	95,175	676	1,795	31.8
	STAM [10]	46.0	74.9	50.0	71.5	38.5	14.6%	43.6%	6,895	91,117	473	1,422	29.6
	AMIR [42]	**47.2**	**75.8**	46.3	68.9	34.8	14.0%	**41.6%**	**2,681**	92,856	774	1,675	21.8
	Ours	46.1	73.8	**54.8**	**77.2**	**42.5**	**17.4%**	42.7%	7,909	**89,874**	532	1,616	**19.3**
Offline	QuadMOT [45]	44.1	76.4	38.3	56.3	29.0	14.6%	44.9%	6,388	94,775	745	1,096	31.9
	EDMT [7]	45.3	75.9	47.9	65.3	37.8	17.0%	39.9%	11,122	87,890	639	946	20.3
	MHT_DAM [23]	45.8	76.3	46.1	66.3	35.3	16.2%	43.2%	6,412	91,758	590	781	23.7
	JMC [47]	46.3	75.7	46.3	66.3	35.6	15.5%	**39.7%**	6,373	90,914	657	1,114	21.1
	NOMT [9]	46.4	76.6	**53.3**	**73.2**	**41.9**	18.3%	41.4%	9,753	87,565	**359**	**504**	16.3
	MCjoint [21]	47.1	76.3	52.3	73.9	40.4	**20.4%**	46.9%	6,703	89,368	370	598	18.6
	NLLMPa [29]	47.6	78.5	47.3	67.2	36.5	17.0%	40.4%	**5,844**	89,093	629	768	16.8
	LMP [48]	**48.8**	**79.0**	51.3	71.1	40.1	18.2%	40.1%	6,654	**86,245**	481	595	**14.8**

Table 2. Tracking performance on the MOT17 dataset.

Mode	Method	MOTA ↑	MOTP ↑	IDF ↑	IDP ↑	IDR ↑	MT ↑	ML ↓	FP ↓	FN ↓	IDS ↓	Frag ↓	AR ↓
Online	GM_PHD [16]	36.4	76.2	33.9	54.2	24.7	4.1%	57.3%	23,723	330,767	4,607	11,317	23.0
	GMPHD_KCF [26]	39.6	74.5	36.6	49.6	29.1	8.8%	43.3%	50,903	284,228	5,811	7,414	23.5
	E2EM	47.5	**76.5**	48.8	68.4	37.9	16.5%	**37.5%**	**20,655**	272,187	3,632	12,712	13.1
	Ours	**48.2**	75.9	**55.7**	**75.9**	**44.0**	**19.3%**	38.3%	26,218	**263,608**	**2,194**	**5,378**	**11.4**
Offline	IOU [5]	45.5	76.9	39.4	56.4	30.3	15.7%	40.5%	**19,993**	281,643	5,988	7,404	16.4
	EDMT [7]	50.0	77.3	**51.3**	**67.0**	**41.5**	**21.6%**	**36.3%**	32,279	**247,297**	**2,264**	3,260	**9.9**
	MHT_DAM[23]	**50.7**	**77.5**	47.2	63.4	37.6	20.8%	36.9%	22,875	252,889	2,314	**2,865**	10.8

Conclusions

- Introduce a cost-sensitive tracking loss for single object tracking.
- Propose a spatial attention network which generates dual attention maps to focus on matching regions between the paired images.
- Design a temporal attention network to adaptively allocate different degrees of attention to different observations in the trajectory.
- Achieve favorable performance against the state-of-the-art online and offline MOT methods in terms of identity-preserving metrics.

ECCV 2018
European Conference on Computer Vision