SiamMOT: Siamese Multi-Object Tracking

Bing Shuai Andrew Berneshawi Xinyu Li Davide Modolo Joseph Tighe
Amazon Web Service (AWS)
{bshuai, bernea, xxnl, dmodolo, tighej }@amazon.com

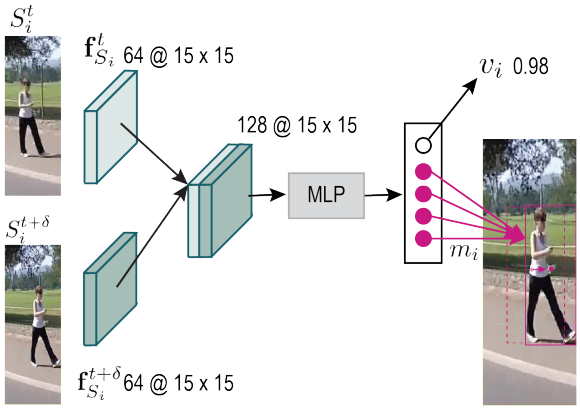

Figure 1: *Network architecture of Implicit Motion Model (IMM).*

1. Implicit Motion Model

We show the graphic illustration of our Implicit Motion Model (IMM) in Fig. 1. Please refer to the main paper for definition of mathematical notation. In general, IMM learns the relative location / scale changes (encoded in m_i) of person instances with visual features of both frames. We empirically set the shape of $\mathbf{f}_{S_i}^{t+\delta}$ to be $c \times 15 \times 15$, and we observe diminished performance gain when we increase it to $c \times 30 \times 30$. Under current configurations, IMM has already entailed significantly more ($400\times$) learnable parameters than EMM in the parameterization of Siamese tracker.

2. Explicit Motion Model

During inference, we empirically set $\lambda = 0.4$ in generating penalty map ($\boldsymbol{\eta}_i$) by default. Due to the large person motion in CRP videos, we use $\lambda = 0.1$, which does not heavily penalize a matched candidate region if it is far away from the target's location in previous frame.

3. Caltech Roadside Pedestrians (CRP)

We use CRP for ablation analysis mainly because videos in CRP are long and people are moving very fast, which presents a different tracking scenario comparing to existing dataset including MOT17 and TAO. As CRP is not widely used for multi-person tracking, we adopt the following evaluation protocol: we only evaluate on frames where ground truth is available and we do not penalize detected instances that overlap with background bounding boxes (instance id = 0). As background bounding boxes are not annotated tightly, we enforce a very loose IOU matching, i.e. a detected bounding box is deemed matched to a background one if their IOU overlap is larger then 0.2.

Sampled triplets	Caltech Roadside Pedestrians				
	MOTA ↑	IDF1 ↑	FP ↓	FN ↓	IDsw ↓
P + H	76.1	81.3	2679	2595	1266
P + N	74.6	79.0	2428	2768	1758
P + H + N	76.4	81.1	2548	2575	1311

Table 1: *Effects of sampled triplets for training forward tracker in SiamMOT. P / N / H are positive / negative / hard training triplet. P+H triplets are usually used in single-object tracking.*

Training in SiamMOT. We present the ablation experiments in Tab. 1. Overall, we observe similar trend as that in MOT17, but we don't observe that FP (in MOTA metric) is reduced as significant as in MOT when negative triplets are added (+N) during training. We find this is mainly because 1), detection in CRP is very accurate and 2), CRP is not exhaustively annotated, so large percentage of FP results from tracking un-annotated person in the background rather from real false detection. Note how hard examples (+H) is important to reduce id switches (i.e. false matching).

Inference in SiamMOT. We find that $\tau > 1$ (frame) has negligible effect in CRP. This is mainly because person moves too fast in CRP videos, so the tracker in SiamMOT fails to track them forward beyond 2 frames in CRP.

4. MOT17

We use public detection to generate our results on test set. We follow recent practices [1, 4] that re-scores the pro-

Sequence	Det	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN ↓	IDsw↓
MOT17-01	DPM	53.3	47.1	33.3%	37.5%	150	2830	34
MOT17-03	DPM	76.5	71.7	57.4%	11.5%	1359	23137	131
MOT17-06	DPM	54.9	52.7	31.9%	30.2%	1089	4043	178
MOT17-07	DPM	59.9	52.5	23.3%	18.3%	651	6034	86
MOT17-08	DPM	40.1	35.1	21.1%	31.6%	443	12094	125
MOT17-12	DPM	56.1	62.8	36.3%	31.9%	436	3349	21
MOT17-14	DPM	43.9	49.0	15.9%	29.3%	947	9077	340
MOT17-01	FRCNN	52.5	45.6	33.3%	37.5%	198	2836	27
MOT17-03	FRCNN	76.8	74.9	56.8%	10.1%	1428	22787	123
MOT17-06	FRCNN	58.2	54.8	37.8%	18.0%	1283	3412	227
MOT17-07	FRCNN	58.2	54.0	23.3%	15.0%	740	6264	65
MOT17-08	FRCNN	36.4	35.5	21.1%	39.5%	399	12933	99
MOT17-12	FRCNN	50.1	59.2	27.5%	41.8%	512	3796	19
MOT17-14	FRCNN	44.2	49.7	16.5%	28.7%	1352	8542	414
MOT17-01	SDP	55.4	47.8	33.3%	33.3%	237	2601	37
MOT17-03	SDP	82.5	74.5	68.2%	8.10%	1846	16283	183
MOT17-06	SDP	57.6	54.7	41.0%	23.9%	1304	3469	219
MOT17-07	SDP	62.7	52.6	33.3%	11.7%	984	5228	89
MOT17-08	SDP	42.1	36.7	25.0%	28.9%	527	11559	152
MOT17-12	SDP	54.8	63.6	37.4%	35.2%	665	3233	24
MOT17-14	SDP	48.9	63.5	18.3%	23.2%	1548	7448	447
All		65.9	63.5	34.6%	23.9%	18098	170955	3040

Table 2: *Detailed result summary on MOT17 test videos.*

vided public detection by using the detector in SiamMOT. This is allowed in public detection protocol. We report detailed video-level metrics in Tab. 2.

5. HiEve

We use public detection to generate our results on test videos, the same practice as that in MOT17. Please refer to the following link in official leaderboard for detailed video-level metrics as well as visualized predictions. http://humanineventes.org/tracker.html?tracker=1&id=200

6. TAO-person

Performance per dataset. We report performance of different subset in TAO-person in Tab. 3. This dataset-wise performance gives us understanding how SiamMOT performs on different tracking scenarios. Overall, SiamMOT performs very competitive on self-driving street scenes, e.g. BDD and Argoverse as well as on movie dataset Charades.

Federated MOTA. For reference, we also report MOT Challenge metric [3] on Tao-person validation set in Tab. 4. We find that SiamMOT also significantly outperforms Tracktor++ [2] on those metrics.

7. Sensitivity analysis of parameters

We present the sensitivity analysis of parameters α and β that is used in inference, as we observe that the tracking performance is relatively more sensitive to their value changes. To elaborate, α indicates the detection confidence threshold

Subset in TAO	SiamMOT(ResNet-101)		SiamMOT(DLA-169)	
	TAP@0.5	TAP@0.75	TAP@0.5	TAP@0.75
YFCC100M	41.3%	18.3%	40.8%	20.0%
HACS	33.1%	17.3%	35.1%	18.2%
BDD	72.3%	41.3%	73.8%	42.8%
Argoverse	66.3%	39.5%	71.7%	42.7%
AVA	41.2%	25.8%	41.8%	26.8%
LaSOT	28.4%	14.9%	28.7%	16.7%
Charades	74.8%	68.2%	85.7%	68.4%
All	41.1%	23.0%	42.1%	24.3%

Table 3: *dataset-wise performance on TAO-person.*

Model	Backbone	MOTA ↑	IDF1 ↑	MT ↑	ML↓	FP ↓	FN ↓	IDsw ↓
Tracktor++ [2]	ResNet-101	66.6	64.8	1529	411	12910	2821	3487
SiamMOT	ResNet-101	74.6	68.0	1926	204	7930	4195	1816
SiamMOT	DLA-169	75.5	68.3	1941	190	7591	4176	1857
SiamMOT+	DLA-169	76.7	70.9	1951	190	7845	3561	1834

Table 4: *MOT Challenge metric on TAO-person validation.*

α	β	MOTA ↑	IDF1 ↑	FP ↓	FN ↓	IDsw ↓
0.4	0.4	63.8	58.5	6105	33876	707
0.4	0.6	63.0	54.4	4973	35707	922
0.4	0.8	59.7	51.1	2595	41686	975
0.6	0.4	63.3	58.4	5726	34833	671
0.6	0.6	62.4	54.5	4330	37034	869
0.6	0.8	59.6	51.1	2322	42167	918
0.8	0.4	61.8	58.3	4742	37611	588
0.8	0.6	60.9	54.8	3169	40030	729
0.8	0.8	58.7	51.6	1842	43730	793

Table 5: *Sensitity analysis of α and β on MOT17 dataset. The experiment settings are exactly the same as that in ablation analysis.*

that we use to start a new trajectory, and β is the visibility confidence threshold that is used to determined whether a trajectory needs to be continued. We do a grid search of α ($[0.4 : 0.8 : 0.2]$) and β ($[0.4 : 0.8 : 0.2]$), and we present their results on MOT17 in Tab. 5. As expected, large values of α and β makes the solver too cautious, which leads to high FN. A good balance is achieved when $\beta = 0.4$, and $\alpha = 0.6$ is used in the rest of paper to avoid the solver overfitting specifically to MOT17.

References

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, 2019.

[2] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *European Conference on Computer Vision*, 2020.

[3] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[4] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *ECCV*, 2020.