TDT: Teaching Detectors to Track without Fully Annotated Videos Supplementary Materials

Shuzhi Yu[1*] Guanhang Wu[2] Chunhui Gu[2] Mohammed E. Fathy[2]
[1]Duke University [2] Google LLC
shuzhiyu@cs.duke.edu {guanhangwu,chunhui,msalem}@google.com

A. Additional implementation details

We trained our joint model on the BaseTrainSet and CrowdHuman datasets for 200K iterations with batch size 32 and evaluated it on the MOT16 and MOT17 private test sets. We further fine-tuned this model on the MOT20 training set for another 30K iterations with batch size 16 and evaluated it on the MOT20 test set. All of our models are trained from scratch without pre-training on any datasets.

We set ϵ as 0.001 and the momentum as 0.997 for our batch normalization layer. The γ and β are learnable.

The L2 weight decay was set 0.0001. We applied the cosine learning rate decay during training and the initial learning rate was 0.15. We warmed up the training process with a small learning rate 0.001 for the first 2K iterations. During fine-tuning, we set the initial learning rate as 0.015 with the same warming up procedure.

The detection threshold was 0.5 for MOT16 and MOT17 test sets and 0.3 for MOT20 test set.

Our system was implemented in TensorFlow. Our current implementation is not optimal and the running speed would be faster with more careful design and implementation.

B. More ablation studies on parameters

Figure 1 shows the ablation studies on some other hyper-parameters of our proposed system, namely the training time, pyramid levels of FPN, and the weight of embedding loss. From the figure, we see that doubling the training time only marginally helps the detection and tracking performance (comparing group 1 and 2). In addition, our system is robust to the weight of the embedding loss. Both the detection and tracking have similar performance between setting $\alpha_e = 2$ and $\alpha_e = 10$ (comparing group 1 and 3). Using features of larger resolutions from the Feature Pyramid Network does not have much impact either (comparing group 3 and 4).

*This work was done when Shuzhi Yu was an intern at Google

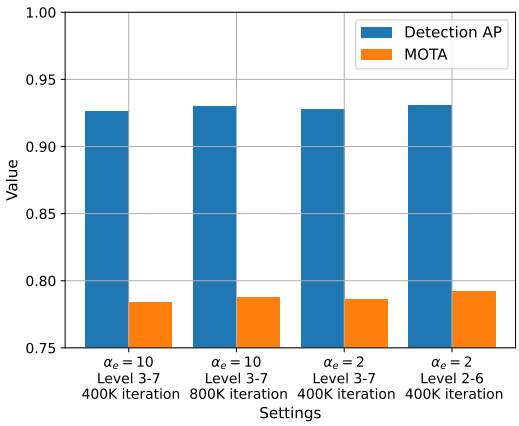

Figure 1. Comparison of detection (Average Precision) and tracking (MOTA) among different hyper-parameter settings. All these models have the same architecture. The backbone architecture is ResNet-34.

C. Inference speed of two-stage tracker

One of the advantages of the one-shot tracker is its faster inference speed than the two-stage tracker. Since the joint model generates an embedding for every anchor box in one forward pass, the running time stays the same regardless of the number of objects in the frame. However, it takes two-stage trackers increasing time with more objects in the scene. For example, the inference speed of the counterpart two-stage tracker of our TDT-tracker is 3.03 FPS for scenes with less people (*e.g.* MOT15) and 0.98 FPS for a scene with more people (*e.g.* sequence MOT20-05). In comparison, our one-shot tracker used very similar running time around 10 FPS regardless of the types of the scene.

D. Detailed performance on the benchmark datasets

Figure 2, Fig. 3, and Fig. 4 show the detailed tracking performance of our TDT-tracker on different video sequences of MOT16, MOT17, and MOT20 respectively.

These resutls were evaluated by the private benchmark server. In general, our TDT-tracker does well on sequences with clear people but poorly on those crowded sequences or those with small people. State-of-the-art trackers have better performance due to their previlege of training on the fully annotated tracking datasets that are similar to these test datasets. Our TDT-tracker would largely improve if our teacher embedder sees similar samples during training.

E. Qualitative analysis

Figure 5, Fig. 6, and Fig. 7 show three qualitative examples of our TDT-tracker. The detailed analysis is in the caption of each figure.

References

[1] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv:2003.09003[cs]*, Mar. 2020. arXiv: 2003.09003. 4

[2] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831. 4

[3] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 4

	MOTA	MOTP	IDF1	IDP	IDR	TP	FP	FN	Rcll	Prcn	MTR	PTR	MLR	MT	PT	ML	IDSW	FAR	FM
seq																			
MOT16-01	59.69	79.82	62.86	82.70	50.70	3886	34	2509	60.77	99.13	39.13	43.48	17.39	9	10	4	35	0.08	136
MOT16-03	70.00	77.91	65.09	68.14	62.30	84629	10972	19927	80.94	88.52	59.46	32.43	8.11	88	48	12	470	7.31	909
MOT16-06	59.19	78.08	59.72	70.25	51.93	7766	763	3772	67.31	91.05	36.65	38.91	24.43	81	86	54	174	0.64	315
MOT16-07	65.40	79.65	57.53	66.27	50.83	11679	840	4643	71.55	93.29	42.59	51.85	5.56	23	28	3	164	1.68	353
MOT16-08	48.88	79.01	47.06	55.05	41.09	10478	2016	6259	62.60	83.86	28.57	66.67	4.76	18	42	3	281	3.23	439
MOT16-12	60.70	80.19	63.86	72.35	57.15	5818	735	2477	70.14	88.78	32.56	56.98	10.47	28	49	9	48	0.82	170
MOT16-14	49.04	77.81	54.94	75.38	43.22	9941	658	8542	53.78	93.79	18.29	57.32	24.39	30	94	40	219	0.88	493
OVERALL	64.05	78.30	61.51	68.09	56.10	134197	16018	48129	73.60	89.34	36.50	47.04	16.47	277	357	125	1391	2.71	2815

Figure 2. Detailed tracking performance of our TDT-tracker on each testing sequence in MOT16.

	MOTA	MOTP	IDF1	IDP	IDR	TP	FP	FN	Rcll	Prcn	MTR	PTR	MLR	MT	PT	ML	IDSW	FAR	FM
seq																			
MOT17-01-DPM	59.15	79.79	62.53	82.70	50.26	3885	35	2565	60.23	99.11	33.33	45.83	20.83	8	11	5	35	0.08	136
MOT17-01-FRCNN	59.15	79.79	62.53	82.70	50.26	3885	35	2565	60.23	99.11	33.33	45.83	20.83	8	11	5	35	0.08	136
MOT17-01-SDP	59.15	79.79	62.53	82.70	50.26	3885	35	2565	60.23	99.11	33.33	45.83	20.83	8	11	5	35	0.08	136
MOT17-03-DPM	70.47	77.95	65.31	68.41	62.48	84918	10683	19757	81.13	88.83	59.46	32.43	8.11	88	48	12	471	7.12	884
MOT17-03-FRCNN	70.47	77.95	65.31	68.41	62.48	84918	10683	19757	81.13	88.83	59.46	32.43	8.11	88	48	12	471	7.12	884
MOT17-03-SDP	70.47	77.95	65.31	68.41	62.48	84918	10683	19757	81.13	88.83	59.46	32.43	8.11	88	48	12	471	7.12	884
MOT17-06-DPM	60.36	78.18	59.43	70.78	51.21	7911	615	3873	67.13	92.79	36.94	40.54	22.52	82	90	50	183	0.52	328
MOT17-06-FRCNN	60.36	78.18	59.43	70.78	51.21	7911	615	3873	67.13	92.79	36.94	40.54	22.52	82	90	50	183	0.52	328
MOT17-06-SDP	60.36	78.18	59.43	70.78	51.21	7911	615	3873	67.13	92.79	36.94	40.54	22.52	82	90	50	183	0.52	328
MOT17-07-DPM	64.67	79.78	56.30	66.12	49.02	11811	714	5082	69.92	94.30	38.33	51.67	10.00	23	31	6	172	1.43	373
MOT17-07-FRCNN	64.67	79.78	56.30	66.12	49.02	11811	714	5082	69.92	94.30	38.33	51.67	10.00	23	31	6	172	1.43	373
MOT17-07-SDP	64.67	79.78	56.30	66.12	49.02	11811	714	5082	69.92	94.30	38.33	51.67	10.00	23	31	6	172	1.43	373
MOT17-08-DPM	48.50	79.40	43.31	58.42	34.41	11512	929	9612	54.50	92.53	23.68	63.16	13.16	18	48	10	338	1.49	539
MOT17-08-FRCNN	48.50	79.40	43.31	58.42	34.41	11512	929	9612	54.50	92.53	23.68	63.16	13.16	18	48	10	338	1.49	539
MOT17-08-SDP	48.50	79.40	43.31	58.42	34.41	11512	929	9612	54.50	92.53	23.68	63.16	13.16	18	48	10	338	1.49	539
MOT17-12-DPM	58.89	80.18	62.48	72.66	54.81	5845	692	2822	67.44	89.41	31.87	53.85	14.29	29	49	13	49	0.77	173
MOT17-12-FRCNN	58.89	80.18	62.48	72.66	54.81	5845	692	2822	67.44	89.41	31.87	53.85	14.29	29	49	13	49	0.77	173
MOT17-12-SDP	58.89	80.18	62.48	72.66	54.81	5845	692	2822	67.44	89.41	31.87	53.85	14.29	29	49	13	49	0.77	173
MOT17-14-DPM	49.15	77.81	54.98	75.52	43.22	9941	638	8542	53.78	93.97	18.29	57.32	24.39	30	94	40	219	0.85	493
MOT17-14-FRCNN	49.15	77.81	54.98	75.52	43.22	9941	638	8542	53.78	93.97	18.29	57.32	24.39	30	94	40	219	0.85	493
MOT17-14-SDP	49.15	77.81	54.98	75.52	43.22	9941	638	8542	53.78	93.97	18.29	57.32	24.39	30	94	40	219	0.85	493
OVERALL	63.83	78.38	60.89	68.59	54.75	407469	42918	156759	72.22	90.47	35.41	47.26	17.32	834	1113	408	4401	2.42	8778

Figure 3. Detailed tracking performance of our TDT-tracker on each testing sequences in MOT17.

	MOTA	MOTP	IDF1	IDP	IDR	TP	FP	FN	Rcll	Prcn	MTR	PTR	MLR	MT	PT	ML	IDSW	FAR	FM
seq																			
MOT20-04	56.84	80.99	52.75	71.72	41.71	158291	1110	115793	57.75	99.30	12.86	72.65	14.50	86	486	97	1404	0.53	10653
MOT20-06	38.31	77.59	37.56	50.00	30.07	66365	13488	66392	49.99	83.11	20.66	47.97	31.37	56	130	85	2018	13.38	3755
MOT20-07	69.24	79.26	55.90	61.82	51.01	25416	1897	7685	76.78	93.05	56.76	40.54	2.70	63	45	3	599	3.24	724
MOT20-08	23.47	75.93	32.85	40.78	27.50	35873	16367	41611	46.30	68.67	13.09	52.88	34.03	25	101	65	1321	20.31	2109
OVERALL	47.88	79.41	46.02	60.36	37.19	285945	32862	231481	55.26	89.69	18.52	61.35	20.13	230	762	250	5342	7.34	17241

Figure 4. Detailed tracking performance of our TDT-tracker on each testing sequence in MOT20.