Contrastive Learning for Multi-Object Tracking with Transformers

Supplementary Material

A. Influence of the memory length

Tracking embeddings are kept in the memory for a given number of previous frames T, and they allow object re-identification without the need for learnable parameters to update tracks. The influence of the maximum memory length on the tracking metrics for BDD100K and MOT17 is displayed in Table 7 and Table 8. A history of one previous frame leads to low results on both datasets. The optimal memory length varies from around 9 previous frames on BDD100K to around 30 previous frames on MOT17. Since the parameter T only influences the association task, it has to be sufficiently high to recover IDs after occlusions or miss detections but not too high to avoid picking up an ID of an object that has left the scene. We state the difference between BDD100K and MOT17 mostly comes from the frame rates, which are respectively 5Hz and 30 Hz[2]. In seconds, the optimal memory durations are respectively 1.8s and 1s.

length	mHOTA↑	mMOTA↑	mIDF1↑	IDSw↓
1	34.6	33.1	38.8	15404
5	38.0	35.1	45.1	6827
9	**38.4**	35.3	**45.8**	6186
12	38.3	**35.4**	45.4	**6092**
15	38.2	**35.4**	45.2	6153

Table 7. Influence of the memory length on MOT metrics on BDD100K validation set.

length	HOTA↑	MOTA↑	IDF1↑	IDSw↓
1	57.0	72.2	63.6	1102
10	62.3	**73.6**	73.7	368
20	63.0	**73.6**	75.7	340
30	**63.5**	**73.6**	**76.4**	**331**
40	63.2	**73.6**	76.2	343

Table 8. Influence of the memory length on MOT metrics on MOT17 validation set.

B. Qualitative results

Figure 5 shows predictions from ContrasTR on the validation set of BDD100K. Since our method learns instance-level features by exploiting different views of the same object, it is robust in case of large camera movement (Figure 5a). It also performs well in case of (partial) object occlusion (Figure 5c). Furthermore, the method predicts discriminative tracking embeddings even under night conditions (Figure 5b) and in case of motion blur (Figure 5d). Nevertheless, the method sporadically swaps or re-assigns IDs from disappeared pedestrians in heavily crowded scenes (Figure 5e), or it assigns an ID from an occluded object to a newly appeared one. In Figure 5f, the car highlighted with a red circle (first frame) is occluded in the second frame, and its ID is assigned to a car further away. After the occlusion, the car further away kept its ID, and the first car got a new ID. Incorporating into the model trajectory and temporal information could be beneficial in mitigating these failure cases.

[2]On MOT17, videos 1-4 and 7-12 have a frame rate of 30 Hz, whereas videos 5-6 and 13-14 have a frame rate of respectively 14Hz and 25 Hz.

C. Additional tracking embeddings visualization

Figure 6 shows a t-SNE projection (left) and the average cosine similarity (right) for the predicted tracking embeddings on three different videos from BDD100K. Tracking IDs are assigned with DETR's bipartite matching. Therefore, every ground truth object will be assigned to the tracking embedding with minimum object detection matching cost. One can see that most tracking embeddings are clustered per ground-truth ID, even during night conditions (Figure 6a and Figure 6c).

D. Hyper-parameters

We report the hyper-parameters used for the experiments on MOT17 and BDD100K in Table 9.

E. Additional details on benchmarks results

MOT17. We report in Table 10 the results of our method obtained on the validation set. The validation set has been selected following [50] and the training setup follows the one presented in Section 4.1. We also include a detailed Table 11 that includes more sub-metrics and video-level performance on the test set of MOT17.

BDD100K. We also report detailed tables that include sub-metrics on the validation set and test set of BDD100K in Table 12 and in Table 13 respectively. The training setup follows the one presented in Section 4.1.

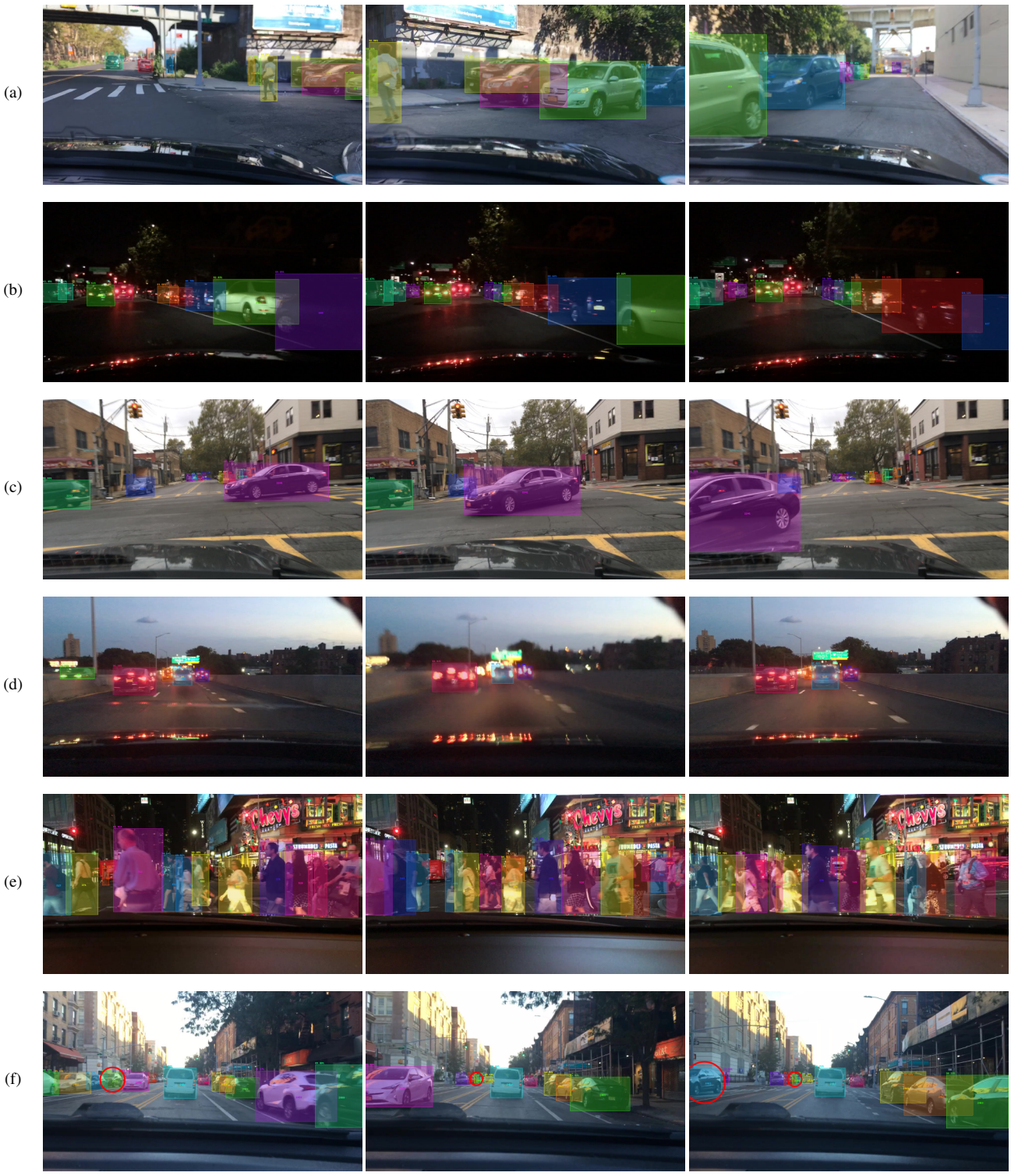

Figure 5. Predictions and failure cases of our model on the validation set of BDD100K.

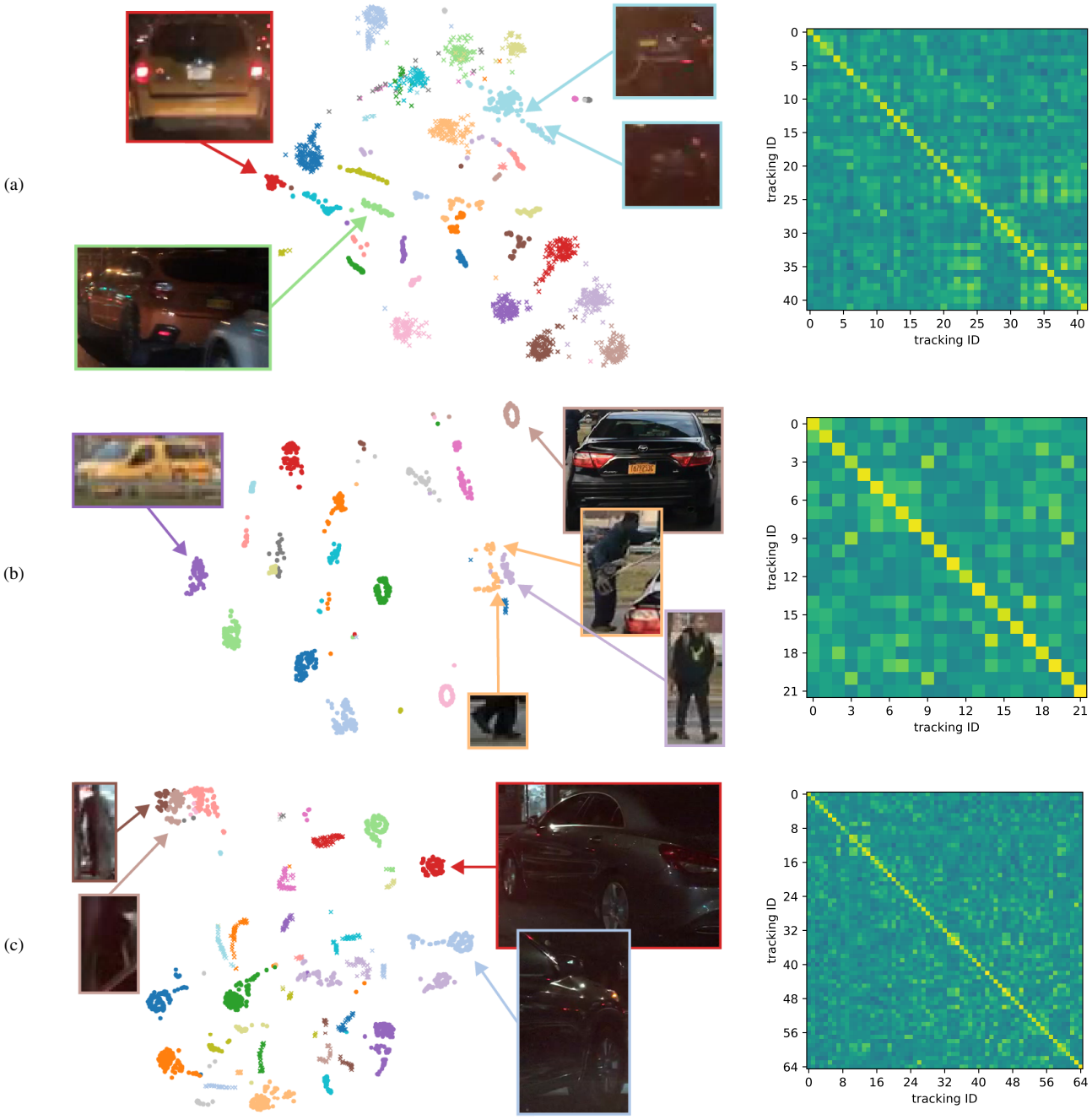

Figure 6. Visualization of predicted tracking embeddings on 3 different sequences from the BDD100K validation set. On the left, a t-SNE projection of the embeddings for the first 40 ground-truth objects, where each color-symbol pair represents a ground-truth tracking ID assigned with DETR's bipartite matching. On the right, the average cosine similarity of predicted objects associated with the same ground-truth instance ID.

	MOT17	BDD100K	BDD100K
model	Deformable-DETR with all refinements from DINO		
backbone	ResNet-50	ResNet-50	Swin-L
classification head	linear layer	linear layer	linear layer
localization head	3 layers MLP	3 layers MLP	3 layers MLP
tracking head	3 layers MLP	3 layers MLP	3 layers MLP
hidden dimension	256	256	256
dim feedforward (heads)	256	256	256
dim feedforward (transformer)	1024	1024	1024
num heads	8	8	8
num queries	300	300	300
weight decay	0.05	0.05	0.05
dropout	0	0	0
clip max norm	0.1	0.1	0.1
cls loss coef	2	2	2
bbox loss coef	5	5	5
giou loss coef	2	2	2
focal alpha	0.25	0.25	0.25
contrastive loss temperature τ	0.1	0.1	0.1
Pre-training			
dataset	CrowdHuman	BDD100k detection	BDD100k detection
epochs	50	36	36
LR drop (epoch)	None	None	None
LR	$2 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	$2 \cdot 10^{-4}$
LR backbone	$2 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	$2 \cdot 10^{-5}$
LR linear proj mult	0.1	0.1	0.1
batch size	16	48	24
contrastive loss coef	2	2	2
Training			
epochs	15	10	10
LR drop (epoch)	10	8	8
LR	$2 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	$2 \cdot 10^{-5}$
LR backbone	$2 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$2 \cdot 10^{-6}$
LR linear proj mult	0.1	0.1	0.1
batch size	16	40	32
contrastive loss coef	2	1	1
num frames per video N_f	8	10	8
objectness threshold	0.5	0.4	0.4
memory length T	20	9	9
new instance id threshold	0.5	0.5	0.5

Table 9. Full set of hyper-parameters.

Sequence	HOTA↑	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	Rcll↑	Prcn↑	ID Sw.↓	Frag↓
MOT17-02	45.0	50.1	54.6	12	12	413	4421	55.3	93.0	91	254
MOT17-04	73.3	85.5	87.8	47	3	755	2680	88.9	96.6	80	368
MOT17-05	48.3	72.3	59.3	37	9	147	741	78.0	94.7	41	98
MOT17-09	61.7	78.4	71.8	17	1	25	572	80.1	98.9	26	40
MOT17-10	57.7	70.8	74.4	15	2	188	1480	75.0	95.9	59	240
MOT17-11	62.9	67.1	72.6	21	9	492	984	78.2	87.8	12	70
MOT17-13	58.3	68.2	75.9	29	1	442	540	82.9	85.5	22	108
Overall	63.5	73.6	76.4	178	37	2462	11418	78.8	94.5	331	1178

Table 10. Detailed results on MOT17 validation split.

Sequence	HOTA↑	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	Rcll↑	Prcn↑	ID Sw.↓	Frag↓
MOT17-01	50.3	53.9	62.0	8	8	329	2602	59.7	92.1	42	118
MOT17-03	67.8	88.9	83.3	129	1	3420	7978	92.4	96.6	218	1354
MOT17-06	47.6	61.7	58.7	79	60	734	3668	68.9	91.7	116	352
MOT17-07	47.1	64.7	56.8	22	11	618	5214	69.1	95.0	123	419
MOT17-08	39.1	48.3	42.9	21	17	288	10439	50.6	97.4	195	491
MOT17-12	55.7	60.0	66.5	37	21	636	2781	67.9	90.2	52	198
MOT17-14	40.0	46.0	53.0	19	48	809	9042	51.1	92.1	127	653
Overall	58.9	73.7	71.8	945	498	20502	125172	77.8	95.5	2619	10755

Table 11. Detailed results on MOT17 test split.

	TETA↑	HOTA↑	MOTA↑	MOTP↑	IDF1↑	FP↓	FN↓	IDSw↓	MT↑	PT↓	ML↓	FM↓
pedestrian	58.4	46.3	55.6	-	56.0	-	-	-	-	-	-	-
rider	52.8	44.1	43.8	-	59.8	-	-	-	-	-	-	-
car	74.4	64.0	72.3	-	72.4	-	-	-	-	-	-	-
truck	62.3	51.3	44.3	-	59.3	-	-	-	-	-	-	-
bus	66.9	58.9	50.6	-	67.2	-	-	-	-	-	-	-
train	23.3	1.8	0.0	-	2.5	-	-	-	-	-	-	-
motorcycle	52.1	46.4	34.7	-	58.1	-	-	-	-	-	-	-
bicycle	52.8	41.1	32.7	-	47.9	-	-	-	-	-	-	-
Average	55.4	44.2	41.8	83.4	52.9	24580	113632	6360	8935	5862	3248	12707
Overall	71.5	60.8	67.4	86.1	69.2	24580	113632	6360	8935	5862	3248	12707

Table 12. Detailed results with a Swin-L backbone on BDD100K validation split.

	TETA↑	HOTA↑	MOTA↑	MOTP↑	IDF1↑	FP↓	FN↓	IDSw↓	MT↑	PT↓	ML↓	FM↓
pedestrian	58.5	46.7	55.4	-	58.9	-	-	-	-	-	-	-
rider	54.6	46.2	45.3	-	61.3	-	-	-	-	-	-	-
car	74.6	64.7	73.4	-	73.8	-	-	-	-	-	-	-
truck	60.1	49.0	39.8	-	58.1	-	-	-	-	-	-	-
bus	63.4	54.6	44.4	-	61.4	-	-	-	-	-	-	-
train	29.2	21.7	7.2	-	26.6	-	-	-	-	-	-	-
motorcycle	52.2	42.8	38.4	-	56.0	-	-	-	-	-	-	-
bicycle	52.9	43.0	38.7	-	56.1	-	-	-	-	-	-	-
Average	55.7	46.1	42.8	80.6	56.5	48894	204063	10793	16917	10261	4953	22975
Overall	71.4	61.1	67.7	85.7	70.5	48894	204063	10793	16917	10261	4953	22975

Table 13. Detailed results with a Swin-L backbone on BDD100K test split.