TrackFormer: Multi-Object Tracking with Transformers
Supplementary Materials

Tim Meinhardt[1]* Alexander Kirillov[2] Laura Leal-Taixé[1] Christoph Feichtenhofer[2]

[1]Technical University of Munich [2]Facebook AI Research (FAIR)

Abstract

This section provides additional material for the main paper: §A contains further implementation details for TrackFormer (§A.1), a visualization of the Transformer encoder-decoder architecture (§A.3), and parameters for multi-object tracking (§A.4). §B contains a discussion related to public detection evaluation (§B.1), and detailed per-sequence results for MOT17 and MOTS20 (§B.2).

A. Implementation details

A.1. Backbone and training

We provide additional hyperparameters for TrackFormer. This supports our implementation details reported in Section 4.2 of the main paper. The Deformable DETR [23] encoder and decoder both apply 6 individual layers with multi-headed self-attention [17] with 8 attention heads. We do not use the "DC5" (dilated conv_5) version of the backbone as this will incur a large memory requirement related to the larger resolution of the last residual stage. We expect that using "DC5" or any other heavier, or higher-resolution, backbone to provide better accuracy and leave this for future work. Furthermore, we also apply the refinement of deformable reference point coined as *bounding box refinement* in [23].

Our training hyperparameters follow deformable DETR [23]. The weighting parameters of the cost and their corresponding loss terms are set to $\lambda_{\text{cls}} = 2$, $\lambda_{\ell_1} = 5$ and $\lambda_{\text{iou}} = 2$. The probabilities for the track augmentation at training time are $p_{\text{FN}} = 0.4$ and $p_{\text{FP}} = 0.1$ Furthermore, every MOT17 [13] frame is jittered by 1% with respect to the original image size similar to the adjacent frame simulation.

A.2. Dataset splits

All experiments evaluated on dataset splits (ablation studies and MOTS20 training set in Table 2) apply the same private training pipeline presented in Section 4.2 to each split. For our ablation on the MOT17 [13] training set, we separate the 7 sequences into 2 splits and report results from training on the first 50% and evaluating on the last 50% of frames. For MOTS20 we average validation metrics over all splits and report the results from a single epoch (which yields the best mean MOTA / MOTSA) over all splits, *i.e.*, we do not take the best epoch for each individual split. Before training each of the 4 MOTS20 [18] splits, we pre-train the model on all MOT17 sequences excluding the corresponding split of the validation sequence.

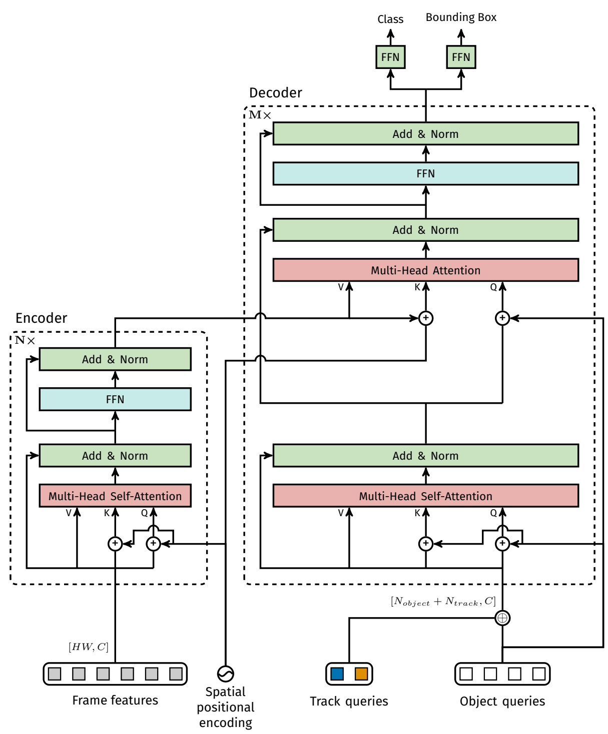

Figure A.1. The TrackFormer encoder-decoder architecture. We indicate the tensor dimensions in squared brackets.

*Work done during an internship at Facebook AI Research (FAIR).

A.3. Transformer encoder-decoder architecture

To foster the understanding of TrackFormer's integration of track queries within the decoder self-attention block, we provide a simplified visualization of the encoder-decoder architecture in Figure A.1. In comparison to the original illustration in [4], we indicate *track identities* instead of spatial encoding with *color-coded* queries. The frame features (indicated in grey) are the final output of the CNN feature extractor and have the same number of channels as both query types. The entire Transformer architecture applies $N = 6$ and $M = 6$ independently supervised encoder and decoder layers, with feature and object encoding as in [4]. To improve, tracking consistency we stack the feature maps of the previous and current frame and apply a spatial positional and temporal encoding as in [19] Track queries are fed *autoregressively* from the *previous frame* output embeddings of the last decoding layer (before the final feed-forward class and bounding box networks (FFN)). The object encoding is achieved by adding the initial object queries to the key (K) and query (Q) of the corresponding embeddings at each decoder layer.

A.4. Multi-object tracking parameters

In Section 3.2, we explain the process of track initialization and removal over a sequence. The corresponding hyperparameters were optimized by a grid search on the MOT17 validation split. The grid search yielded track initialization and removal thresholds of $\sigma_{\text{detection}} = 0.4$ and $\sigma_{\text{track}} = 0.4$, respectively. TrackFormer benefits from an NMS operation for the removal of strong occlusion cases with an intersection over union larger than $\sigma_{\text{NMS}} = 0.9$.

For the track query re-identification, our search proposed an optimal inactive patience and score of $T_{\text{track-reid}} = 5$ and $\sigma_{\text{track-reid}} = 0.4$, respectively.

B. Experiments

B.1. Public detections and track filtering

TrackFormer implements a new tracking-by-attention paradigm which requires track initializations to be filtered for an evaluation with public detections. Here, we provide a discussion on the comparability of TrackFormer with earlier methods and different filtering schemes.

Common tracking-by-detection methods directly process the MOT17 public detections and report their mean tracking performance over all three sets. This is only possible for methods that perform data association on a bounding box level. However, TrackFormer and point-based methods such as CenterTrack [22] require a procedure for filtering track initializations by public detections in a comparable manner. Unfortunately, MOT17 does not provide a standardized protocol for such a filtering. The authors of CenterTrack [22] filter detections based on bounding box center distances (CD). Each public detection can possibly initialize a single track but only if its center point falls in the bounding box area of the corresponding track.

Method	IN	IoU	CD	MOTA ↑	IDF1 ↑
			Offline		
MHT_DAM [11]	×			50.7	47.2
jCC [10]	×			51.2	54.5
FWT [8]	×			51.3	47.6
eHAF [15]	×			51.8	54.7
TT [21]	×			54.9	63.1
MPNTrack [3]	×			58.8	61.7
Lif_T [9]	×			60.5	65.6
			Online		
MOTDT [5]	×			50.9	52.7
FAMNet [6]	×			52.0	48.7
Tracktor++ [1]	×			56.3	55.1
GSM_Tracktor [12]	×			56.4	57.8
TMOH [16]	×			62.1	62.8
CenterTrack [22]		×		60.5	55.7
TrackFormer		×		62.3	57.6
CenterTrack [22]			×	61.5	59.6
TrackFormer			×	63.4	60.0

Table A.1. Comparison of modern multi-object tracking methods evaluated on the **MOT17** [13] test set for different **public detection processing**. Public detections are either directly processed as input (IN) or applied for filtering of track initializations by center distance (CD) or intersection over union (IoU). We report mean results over the three sets of public detections provided by [13] and separate between online and offline approaches. The arrows indicate low or high optimal metric values.

In Table A.1, we revisit our MOT17 test set results but with this public detections center distance (CD) filtering, while also inspecting the CenterTrack per-sequence results in Table A.5. We observe that this filtering does not reflect the quality differences in each set of public detections, *i.e.*, DPM [7] and SDP [20] results are expected to be the worst and best, respectively, but their difference is small.

We hypothesize that a center distance filtering is not in accordance with the common public detection setting and propose a filtering based on Intersection over Union (IoU). For IoU filtering, public detections only initialize a track if they have an IoU larger than 0.5. The results in Table A.1 show that for TrackFormer and CenterTrack IoU filtering performs worse compared to the CD filtering which is expected as this is a more challenging evaluation protocol. We believe IoU-based filtering (instead of CD-based) provides a fairer comparison to previous MOT methods which directly process public detections as inputs (IN). This is validated by the per-sequence results in Table A.4, where IoU filtering shows differences across detectors that are more meaningfully correlated with detector performance, compared to the relatively uniform performance across detections with the CD based method in Table A.5 (where DPM, FRCNN and SDP show *very similar* performance).

Consequently, we follow the IoU-based filtering protocol to compare with CenterTrack in our main paper. While our gain over CenterTrack seems similar across the two filtering techniques for MOTA (see Table A.1), the gain in IDF1 is significantly larger under the more challenging IoU-based protocol, which suggests that CenterTrack benefits from the less challenging CD-based filtering protocol, while TrackFormer does not rely on the filtering for achieving its high IDF1 tracking accuracy.

B.2. MOT17 and MOTS20 sequence results

In Table A.3 and Table A.4, we provide per-sequence MOT17 [13] test set results for private and public detection filtering via Intersection over Union (IoU), respectively. Futhermore, we present per-sequence TrackFormer results on the MOTS20 [18] test set in Table A.2.

Evaluation metrics In Section 4.1 we explained two compound metrics for the evaluation of MOT results, namely, Multi-Object Tracking Accuracy (MOTA) and Identity F1 score (IDF1). [2] However, the MOTChallenge benchmark implements all CLEAR MOT [2] evaluation metrics. In addition to MOTA and IDF1, we report the following CLEAR MOT metrics:

MT: Ground truth tracks covered for at least 80%.

ML: Ground truth tracks covered for at most 20%.

FP: False positive bounding boxes not corresponding to any ground truth.

FN: False negative ground truth boxes not covered by any bounding box.

ID Sw.: Bounding boxes switching the corresponding ground truth identity.

sMOTSA: Mask-based Multi-Object Tracking Accuracy (MOTA) which counts true positives instead of only masks with IoU larger than 0.5.

Acknowledgements: We are grateful for discussions with Jitendra Malik, Karttikeya Mangalam, and David Novotny.

Sequence	sMOTSA ↑	IDF1 ↑	MOTSA ↑	FP ↓	FN ↓	ID Sw. ↓
MOTS20-01	59.8	68.0	79.6	255	364	16
MOTS20-06	63.9	65.1	78.7	595	1335	158
MOTS20-07	43.2	53.6	58.5	834	4433	75
MOTS20-12	62.0	76.8	74.6	549	1063	29
ALL	54.9	63.6	69.9	2233	7195	278

Table A.2. We present TrackFormer tracking and segmentation results on each individual sequence of the **MOTS20** [18] test set. MOTS20 is evaluated in a private detections setting. The arrows indicate low or high optimal metric values.

Sequence	Public detection	MOTA ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓	ID Sw. ↓
MOT17-01	DPM [7]	57.9	49.7	11	4	477	2191	45
MOT17-03	DPM	88.6	79.6	124	3	2469	9365	122
MOT17-06	DPM	59.8	60.8	104	27	1791	2775	173
MOT17-07	DPM	65.5	49.5	24	5	1030	4671	118
MOT17-08	DPM	54.5	42.5	24	9	1461	7861	279
MOT17-12	DPM	51.8	63.0	43	14	1880	2258	42
MOT17-14	DPM	47.4	54.9	41	20	2426	7138	164
MOT17-01	FRCNN [14]	57.9	49.7	11	4	477	2191	45
MOT17-03	FRCNN	88.6	79.6	124	3	2469	9365	122
MOT17-06	FRCNN	59.8	60.8	104	27	1791	2775	173
MOT17-07	FRCNN	65.5	49.5	24	5	1030	4671	118
MOT17-08	FRCNN	54.5	42.5	24	9	1461	7861	279
MOT17-12	FRCNN	51.8	63.0	43	14	1880	2258	42
MOT17-14	FRCNN	47.4	54.9	41	20	2426	7138	164
MOT17-01	SDP [20]	57.9	49.7	11	4	477	2191	45
MOT17-03	SDP	88.6	79.6	124	3	2469	9365	122
MOT17-06	SDP	59.8	60.8	104	27	1791	2775	173
MOT17-07	SDP	65.5	49.5	24	5	1030	4671	118
MOT17-08	SDP	54.5	42.5	24	9	1461	7861	279
MOT17-12	SDP	51.8	63.0	43	14	1880	2258	42
MOT17-14	SDP	47.4	54.9	41	20	2426	7138	164
All	All	74.1	68.0	1113	246	34602	108777	2829

Table A.3. We report **private TrackFormer** results on each individual sequence evaluated on the **MOT17** [13] test set. To follow the official MOT17 format, we display the same results per public detection set. The arrows indicate low or high optimal metric values.

Sequence	Public detection	MOTA ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓	ID Sw. ↓
MOT17-01	DPM [7]	49.9	43.0	5	8	258	2932	40
MOT17-03	DPM	74.0	66.5	85	18	1389	25396	374
MOT17-06	DPM	53.6	51.8	63	75	711	4575	180
MOT17-07	DPM	52.6	48.1	12	16	258	7663	88
MOT17-08	DPM	32.5	31.9	10	32	288	13838	128
MOT17-12	DPM	51.3	57.7	21	31	606	3565	53
MOT17-14	DPM	38.1	42.0	15	63	627	10505	314
MOT17-01	FRCNN [14]	50.9	42.3	8	6	308	2813	48
MOT17-03	FRCNN	75.3	67.0	84	16	1434	24040	335
MOT17-06	FRCNN	57.2	54.8	73	48	960	3856	226
MOT17-07	FRCNN	52.4	47.9	12	11	499	7437	106
MOT17-08	FRCNN	31.1	31.7	10	36	285	14166	102
MOT17-12	FRCNN	47.7	56.7	19	32	702	3785	45
MOT17-14	FRCNN	37.8	41.8	17	56	1300	9795	406
MOT17-01	SDP [20]	53.7	45.3	10	5	556	2386	47
MOT17-03	SDP	79.6	65.8	95	13	2134	18632	545
MOT17-06	SDP	56.4	54.0	82	57	1017	3889	228
MOT17-07	SDP	54.6	47.8	16	11	590	6965	121
MOT17-08	SDP	35.0	33.0	12	27	443	13152	144
MOT17-12	SDP	48.9	57.5	22	28	850	3527	54
MOT17-14	SDP	40.4	42.4	17	49	1376	9206	434
ALL	ALL	62.3	57.6	688	638	16591	192123	4018

Table A.4. We report **TrackFormer** results on each individual sequence and set of public detections evaluated on the **MOT17** [13] test set. We apply our minimum **Intersection over Union (IoU)** public detection filtering. The arrows indicate low or high optimal metric values.

Sequence	Public detection	MOTA ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓	ID Sw. ↓
MOT17-01	DPM [7]	41.6	44.2	5	8	496	3252	22
MOT17-03	DPM	79.3	71.6	94	8	1142	20297	191
MOT17-06	DPM	54.8	42.0	54	63	314	4839	175
MOT17-07	DPM	44.8	42.0	11	16	1322	7851	147
MOT17-08	DPM	26.5	32.2	11	37	378	15066	88
MOT17-12	DPM	46.1	53.1	16	45	207	4434	30
MOT17-14	DPM	31.6	36.6	13	78	636	11812	196
MOT17-01	FRCNN [14]	41.0	42.1	6	9	571	3207	25
MOT17-03	FRCNN	79.6	72.7	93	7	1234	19945	180
MOT17-06	FRCNN	55.6	42.9	57	59	363	4676	190
MOT17-07	FRCNN	45.5	41.5	13	15	1263	7785	156
MOT17-08	FRCNN	26.5	31.9	11	36	332	15113	89
MOT17-12	FRCNN	46.1	52.6	15	45	197	4443	30
MOT17-14	FRCNN	31.6	37.6	13	77	780	11653	202
MOT17-01	SDP [20]	41.8	44.3	7	8	612	3112	27
MOT17-03	SDP	80.0	72.0	93	8	1223	19530	181
MOT17-06	SDP	55.5	43.8	56	61	354	4712	181
MOT17-07	SDP	45.2	42.4	13	15	1332	7775	147
MOT17-08	SDP	26.6	32.3	11	36	350	15067	91
MOT17-12	SDP	46.0	53.0	16	45	221	4426	30
MOT17-14	SDP	31.7	37.1	13	76	749	11677	205
All	All	61.5	59.6	621	752	14076	200672	2583

Table A.5. We report the original per-sequence **CenterTrack** [22] **MOT17** [13] test set results with **Center Distance (CD)** public detection filtering. The results do not reflect the varying object detection performance of DPM, FRCNN and SDP, respectively. The arrows indicate low or high optimal metric values.

References

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *Int. Conf. Comput. Vis.*, 2019. 2

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008, 2008. 3

[3] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[4] Nicolas Carion, F. Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *Eur. Conf. Comput. Vis.*, 2020. 2

[5] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *Int. Conf. Multimedia and Expo*, 2018. 2

[6] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *Int. Conf. Comput. Vis.*, 2019. 2

[7] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009. 2, 4, 5

[8] Roberto Henschel, Laura Leal-Taixé, Daniel Cremers, and Bodo Rosenhahn. Improvements to frank-wolfe optimization for multi-detector multi-object tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[9] Andrea Hornakova, Roberto Henschel, Bodo Rosenhahn, and Paul Swoboda. Lifted disjoint paths with application in multiple object tracking. In *Int. Conf. Mach. Learn.*, 2020. 2

[10] Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox, and Bernt Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018. 2

[11] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M. Rehg. Multiple hypothesis tracking revisited. In *Int. Conf. Comput. Vis.*, 2015. 2

[12] Qiankun Liu, Qi Chu, Bin Liu, and Nenghai Yu. Gsm: Graph similarity model for multi-object tracking. In *Int. Joint Conf. Art. Int.*, 2020. 2

[13] Anton Milan, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv:1603.00831*, 2016. 1, 2, 3, 4, 5

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inform. Process. Syst.*, 2015. 4, 5

[15] H. Sheng, Y. Zhang, J. Chen, Z. Xiong, and J. Zhang. Heterogeneous association graph fusion for target association in multiple object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 2

[16] Daniel Stadler and Jurgen Beyerer. Improving multiple pedestrian tracking by track management and occlusion handling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10958–10967, June 2021. 2

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017. 1

[18] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 3

[19] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[20] Fan Yang, Wongun Choi, and Yuanqing Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2, 4, 5

[21] Y. Zhang, H. Sheng, Y. Wu, S. Wang, W. Lyu, W. Ke, and Z. Xiong. Long-term tracking with deep tracklet association. *IEEE Trans. Image Process.*, 2020. 2

[22] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *ECCV*, 2020. 2, 5

[23] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *Int. Conf. Learn. Represent.*, 2021. 1