LiDAR MOT-DETR: A LiDAR-based Two-Stage Transformer for 3D Multiple Object Tracking

Martha Teiko Teye[1], Ori Maoz[2], Matthias Rottmann[3]
[1] *University of Wuppertal,* [2] *Aptiv Services Deutschland,* [3] *Osnabrück University*

Introduction

- LiDAR MOT-DETR: a modular, detector-agnostic two-stage framework (smoother + tracker) for both online and offline 3D Multiple Object Tracking.
- Both modules build on the DETR transformer, adapted for LiDAR-based smoothing and tracking.
- Achieves SOTA results on nuScenes test set — aMOTA 0.724 / aMOTP 0.475 (online) and aMOTP 0.445 (offline).

Method

Smoother

⇒ Given detected objects from an off-the-shelf detector $\mathcal{D}^\tau = \{d_1^\tau, \ldots, d_{n^\tau}^\tau\}$, the smoother f processes a temporal window of frames $\mathcal{D}^{t,\ldots,t'} = (\mathcal{D}^t, \ldots, \mathcal{D}^{t'})$ and outputs smoothed detections $\mathcal{S}^\tau = f(\mathcal{D}^{t,\ldots,t'})$. We use a symmetric window of $2k{+}1$ frames, $t = \tau - k$, $t' = \tau + k$, centered on key frame τ.

Tracker

⇒ The tracker g takes previous track queries $\mathcal{Q}'^\tau$, smoothed detections $\mathcal{S}^\tau$, and point cloud features $\mathcal{F}^{\tau-k+1,\ldots,\tau}$ to produce tracked objects $\mathcal{T}^\tau$.
It is formulated as an attention block:

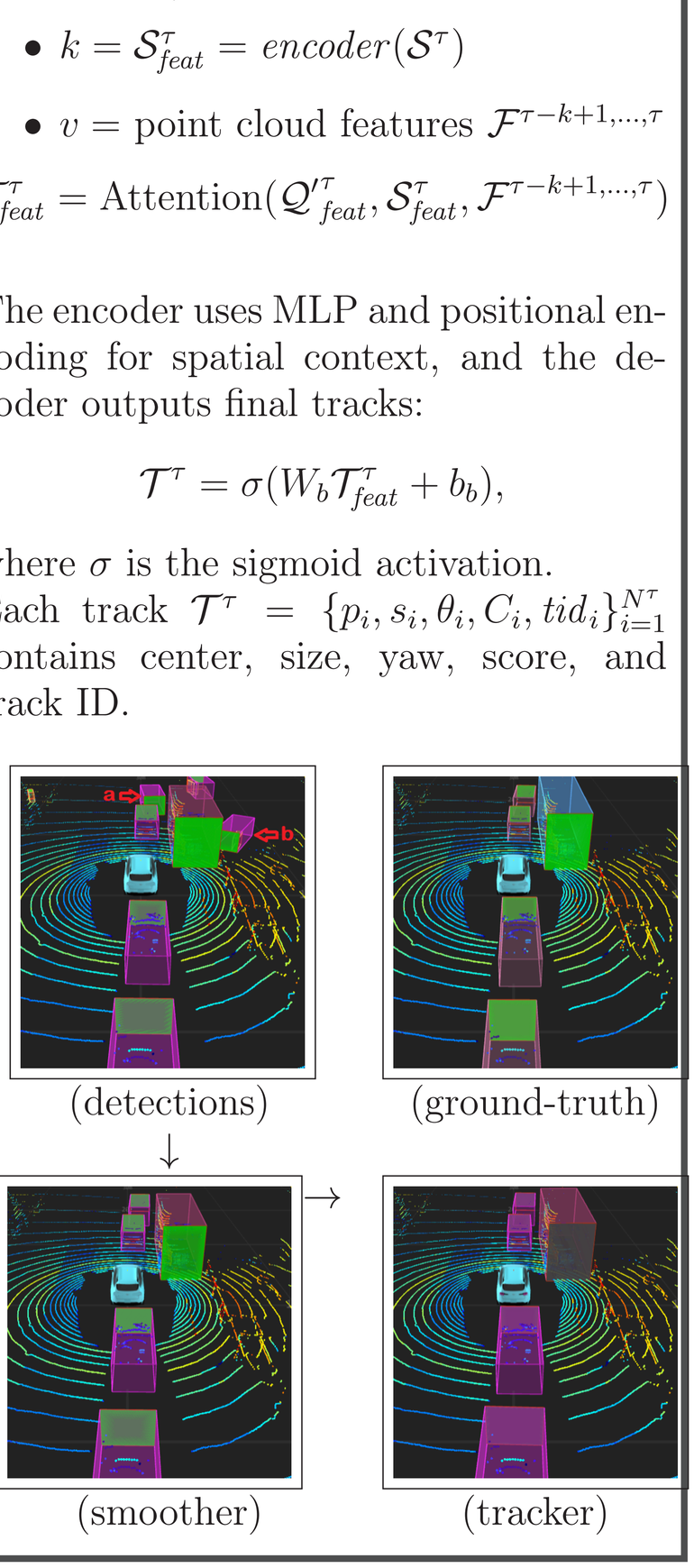
- $q = \mathcal{Q}'^\tau_{feat} = encoder(\mathcal{Q}'^\tau)$
- $k = \mathcal{S}^\tau_{feat} = encoder(\mathcal{S}^\tau)$
- v = point cloud features $\mathcal{F}^{\tau-k+1,\ldots,\tau}$

$$\mathcal{T}^\tau_{feat} = \text{Attention}(\mathcal{Q}'^\tau_{feat}, \mathcal{S}^\tau_{feat}, \mathcal{F}^{\tau-k+1,\ldots,\tau})$$

The encoder uses MLP and positional encoding for spatial context, and the decoder outputs final tracks:

$$\mathcal{T}^\tau = \sigma(W_b \mathcal{T}^\tau_{feat} + b_b),$$

where σ is the sigmoid activation.
Each track $\mathcal{T}^\tau = \{p_i, s_i, \theta_i, C_i, tid_i\}_{i=1}^{N^\tau}$ contains center, size, yaw, score, and track ID.



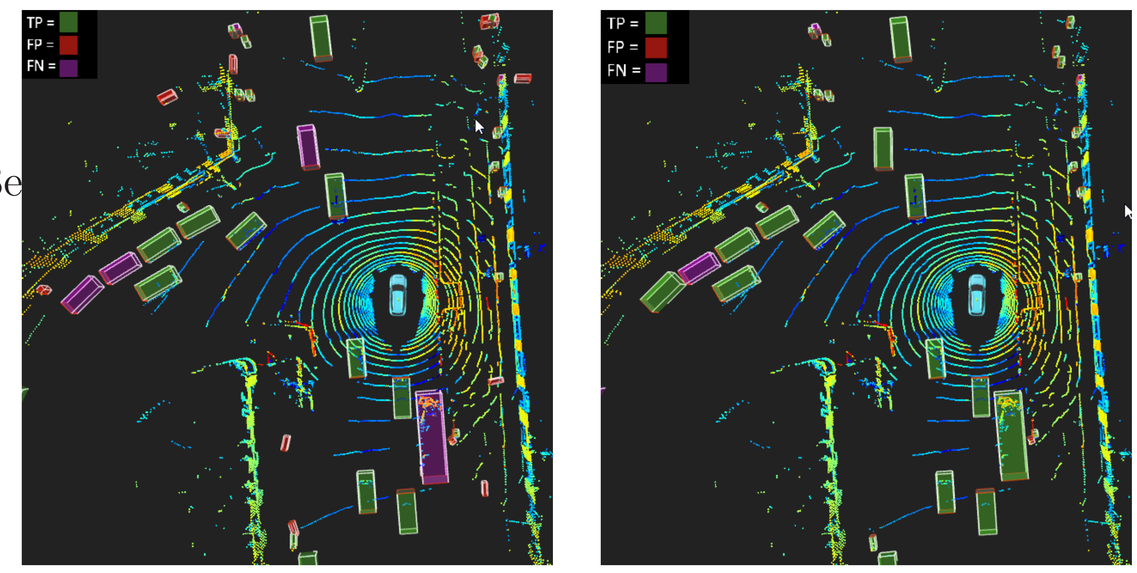
Comparision of LiDAR-only results on Nuscenes Test Set

Method	aMOTA↑	aMOTP↓	MOTA↑	MOTP ↑	TP↑	FN↓	IDS↓	FRAG↓
SimpleTrack [1]	0.668	0.550	0.537	0.284	95877	22928	575	529
3DMOTFormer [2]	0.682	0.496	0.556	0.297	95790	23357	438	529
ShaSTA [3]	0.696	0.540	0.578	0.295	**97799**	21293	473	<u>356</u>
FocalFormer3D [4]	<u>0.715</u>	0.549	**0.601**	0.309	97535	**21142**	888	810
LiDAR MOT-DETR (online)	0.724	0.475	0.594	0.310	97315	21846	404	528
LiDAR MOT-DETR (offline)	**0.726**	**0.445**	<u>0.592</u>	**0.312**	<u>97593</u>	21495	577	462

Best value highlighted in bold and second place underlined.

Qualitative Results on Nuscenes Validation Set

LiDAR MOT-DETR vs. 3DMOTFormer [2] on a nuScenes scene (01452fbfbf4543af8acdfd3e using CenterPoint [5] detections.



3DMOTFormer | LiDAR MOT-DETR

Smoother Results with Different Temporal Window (Frames)

Results are generated using CenterPoint[5], on nuScenes validation set.

Frames	aMOTA↑	aMOTP↓	IDS↓	FRAG↓	mAP↑
8	0.682	0.559	442	506	0.632
11	0.705	0.556	425	500	0.657
13	0.726	0.538	419	480	0.670
15	**0.735**	**0.523**	**407**	**474**	**0.671**

(a) offline mode

Frames	aMOTA↑	aMOTP↓	IDS↓	FRAG↓	mAP↑
8	0.677	0.557	457	502	0.629
11	0.698	0.542	428	498	0.635
13	0.713	0.532	424	497	0.634
15	**0.729**	**0.529**	**410**	**462**	**0.637**

(b) online mode

Conclusion

Achieves +0.7 pp aMOTA gains on nuScenes, without manual tuning—showing transformers capture temporal dependencies effectively.

References

[1] Ziqi Pang, Zhichao Li, and Naiyan Wang. Simpletrack: Understanding and rethinking 3d multi-object tracking. In *European Conference on Computer Vision*, pages 680–696. Springer, 2022.

[2] Shuxiao Ding, Eike Rehder, Lukas Schneider, Marius Cordts, and Juergen Gall. 3dmotformer: Graph transformer for online 3d multi-object tracking. In *Proceedings of the IEEE/CVF ICCV*, pages 9784–9794, October 2023.

[3] Tara Sadjadpour, Jie Li, Rares Ambrus, and Jeannette Bohg. Shasta: Modeling shape and spatio-temporal affinities for 3d multi-object tracking. *IEEE Robotics and Automation Letters*, 2023.

[4] Yilun Chen, Zhiding Yu, Yukang Chen, Shiyi Lan, Anima Anandkumar, Jiaya Jia, and Jose M Alvarez. Focalformer3d: focusing on hard instance for 3d object detection. In *Proceedings of the IEEE/CVF ICCV*, pages 8394–8405, 2023.

[5] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11784–11793, June 2021.