A Dynamic-confidence 3D Multi-Object Tracking Method Based on Spatio-Temporal Association

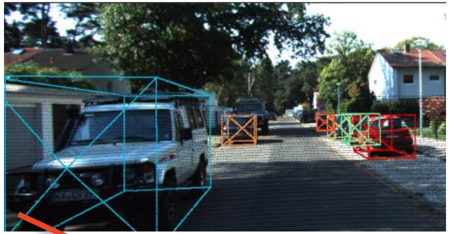
Ruihao ZENG and Mohsen RAMEZANI

The University of Sydney

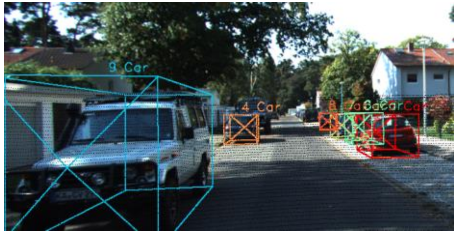
What is multi-object tracking

Environmental perception → AVs decision-making

Detection



Tracking



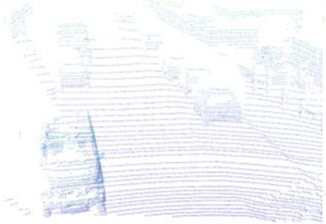
Bounding boxes

Camera-based

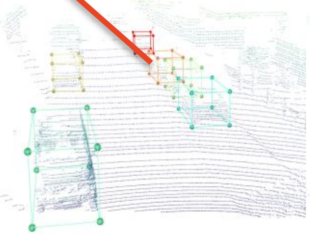
Detection → Tracking

LiDAR-based

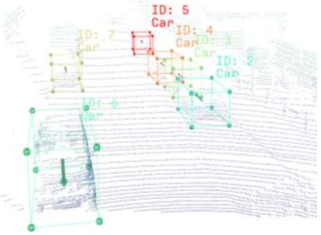
Point Cloud



Detection

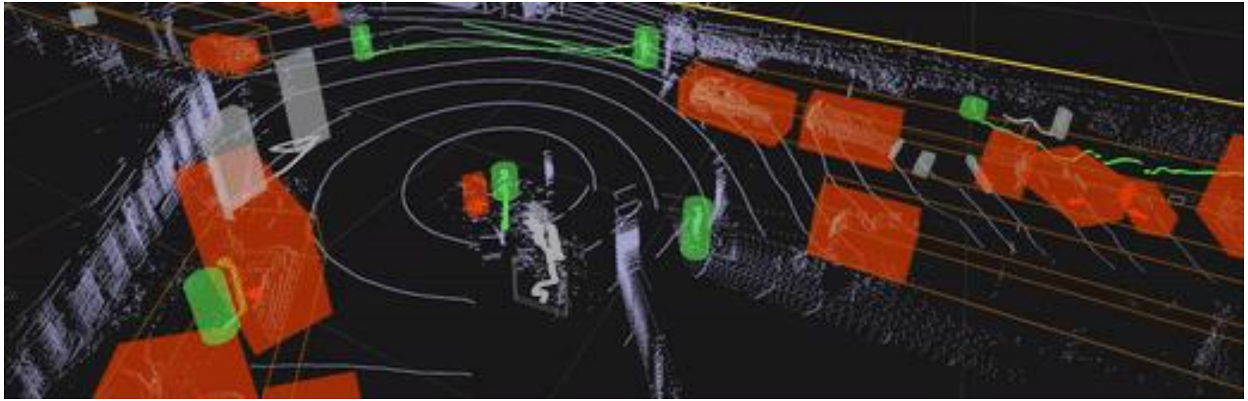


Tracking



Implementation

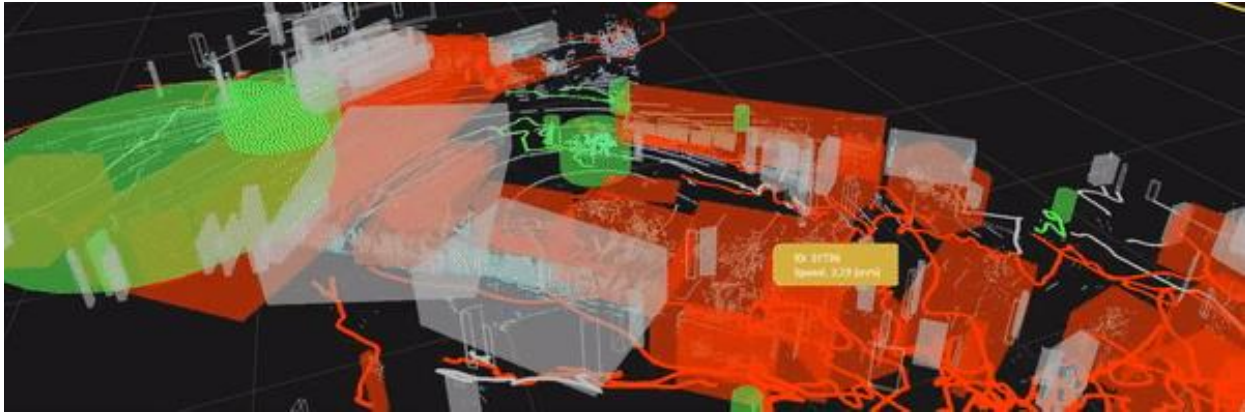
LiDAR - fixed



Collected on 30-March-2023 and 09-March-2023.

Implementation

LiDAR - moving



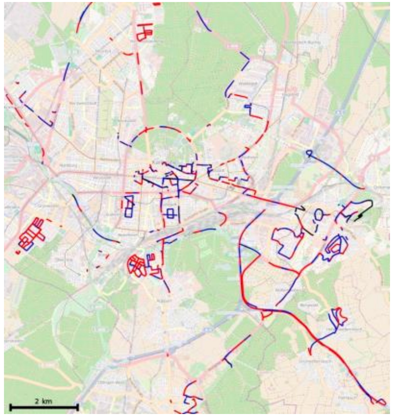
Most of the targets remain stationary.

Collected on 30-March-2023 and 09-March-2023.

Research Challenges

- **Reliable** data association while moving without color, shape, material information, etc.
- Complete tracking as fast as possible in **real time**.
- **Continuous** tracking capability when the object is obscured or missed temporarily.

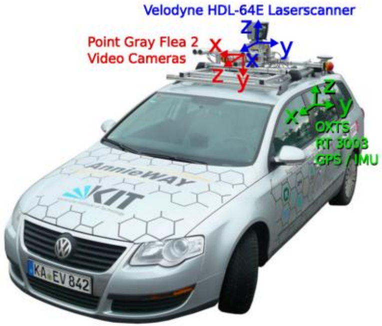
Dataset - KITTI



Recording zone. Metropolitan area of Karlsruhe, Germany.

Point cloud as raw data.

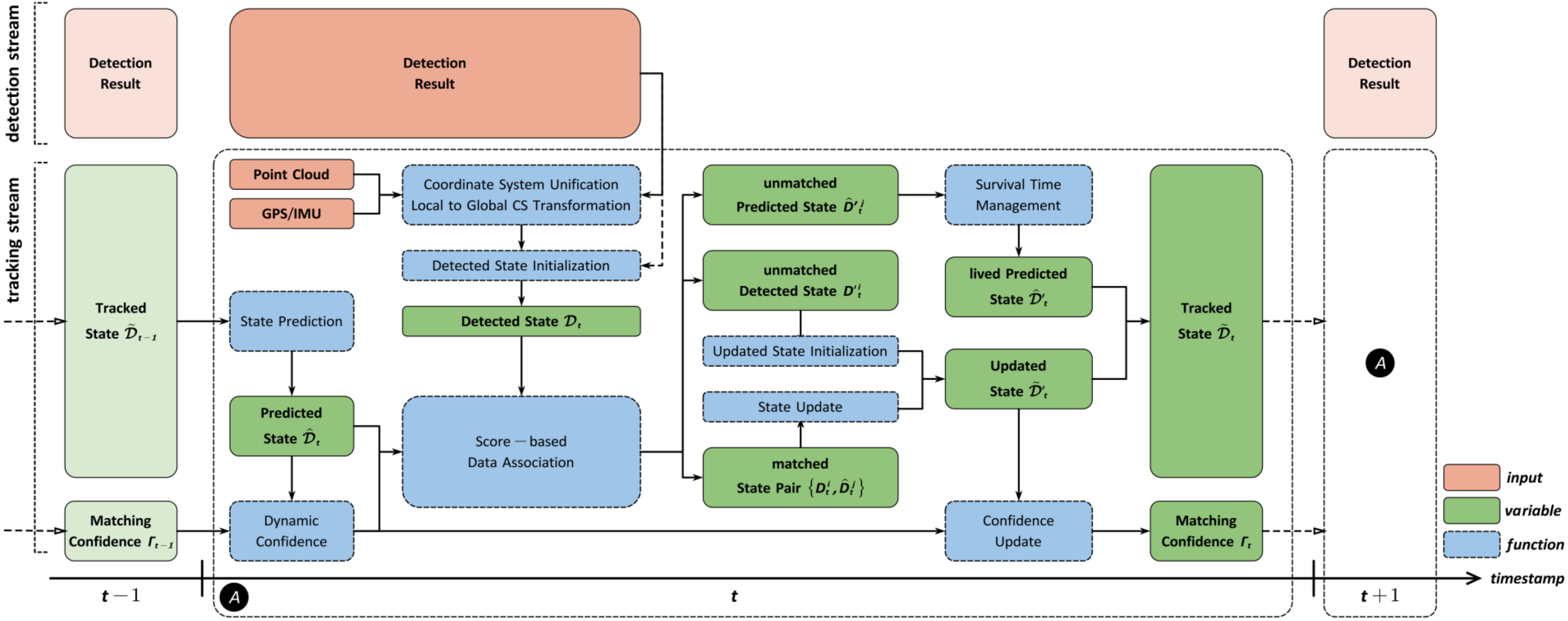
From “Vision meets Robotics: The KITTI Dataset” and the KITTI dataset.



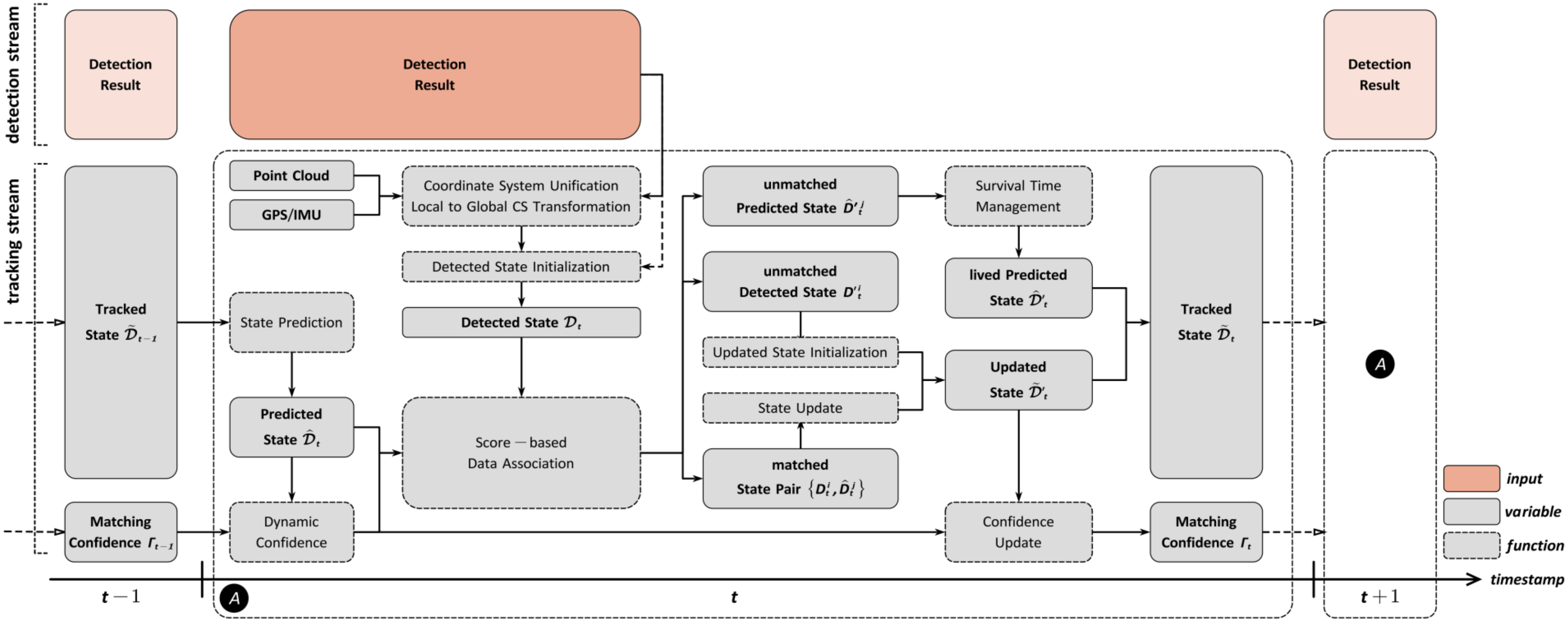
Recording platform. 1 LiDAR, 4 cameras, 1 GPS/IMU.

Image for ground truth (annotation).

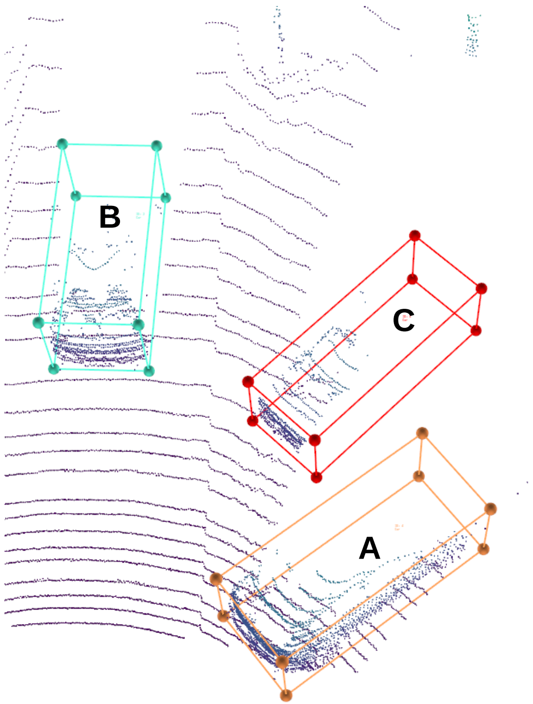
Proposed MOT framework



Detection stream



Detection result - example

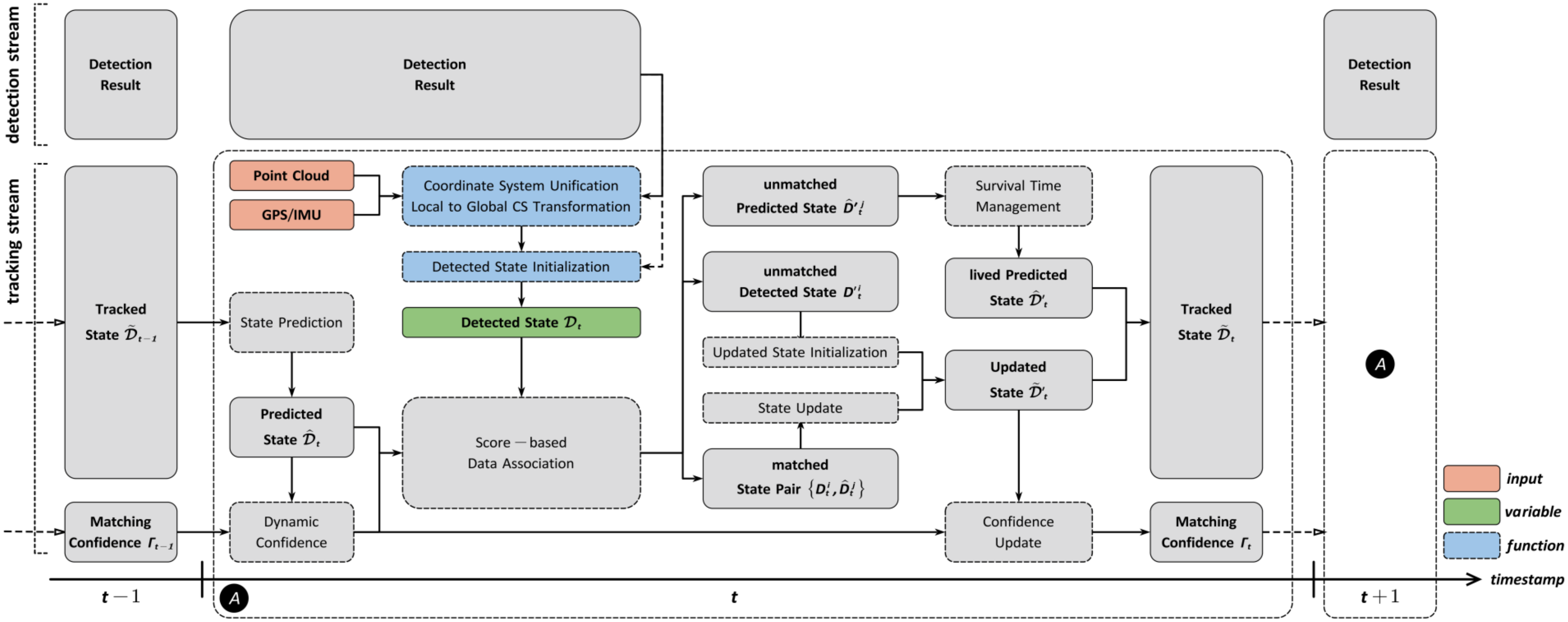


Part of detection results.

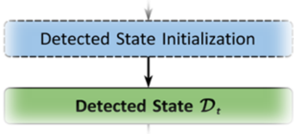
Parameter	Value		
	Object A	Object B	Object C
Label	Car	Car	Car
Truncated	-1	-1	-1
Occlude	-1	-1	-1
Observation angle *(radian measure)*	-7.5146	-7.8332	-7.4890
2D Bounding Box X_min *(camera-CS)*	890.1342	556.6973	814.2166
2D Bounding Box Y_min *(camera-CS)*	146.5483	173.5228	175.5845
2D Bounding Box X_max *(camera-CS)*	1241.0000	669.5584	1026.1429
2D Bounding Box Y_max *(camera-CS)*	374.0000	280.0149	293.4225
3D Bounding Box height *(meter)*	1.5791	1.5593	1.3754
3D Bounding Box width *(meter)*	1.6725	1.6592	1.5274
3D Bounding Box length *(meter)*	4.0309	3.6525	3.9645
3D Center Point X *(LiDAR-CS)*	4.3920	0.0286	4.5773
3D Center Point Y *(LiDAR-CS)*	6.6461	12.4317	10.4703
3D Center Point Z *(LiDAR-CS)*	1.4059	1.5722	1.4221
Yaw/Orientation *(radian measure)*	-6.9514	-7.8323	-7.0879
Detection score	7.0113	6.3238	5.2618

Be saved in *.txt* format.

CS transformation & Detected state initialization



Detected state initialization



Each of the included elements in the detected state represents the **detected information** of one bounding box.

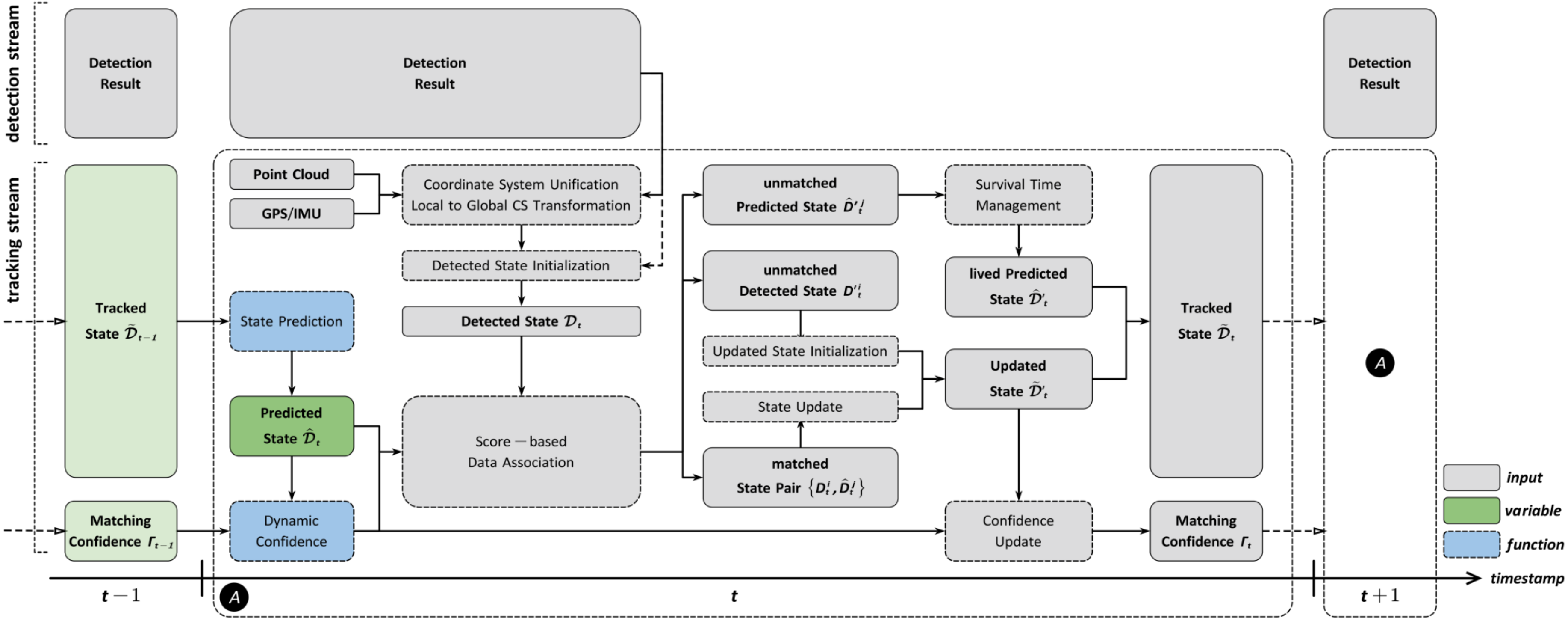
$$D_t^i = \left[x_t^i, y_t^i, z_t^i, \dot{x}_t^i, \dot{y}_t^i, \dot{z}_t^i, \ddot{x}_t^i, \ddot{y}_t^i, \ddot{z}_t^i, w_t^i, h_t^i, l_t^i, \theta_t^i, \dot{\theta}_t^i, \ddot{\theta}_t^i, f_t^{i,1}, \cdots, f_t^{i,\xi}\right]^T$$

All detection methods treat objects at various timestamps as **separate** and **unrelated** entities. Velocity and acceleration, which are derived from temporal changes in object positions, aren't included in the detection results.

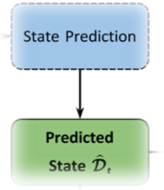
$$D{*}_t^i = \left[x_t^i, y_t^i, z_t^i, w_t^i, h_t^i, l_t^i, \theta_t^i\right]^T$$

$$\mathrm{Init}\left(D{*}_t^i\right) \rightarrow \begin{cases} \dot{x}_t^i, \dot{y}_t^i, \dot{z}_t^i = 0 \\ \ddot{x}_t^i, \ddot{y}_t^i, \ddot{z}_t^i = 0 \\ \dot{\theta}_t^i = \ddot{\theta}_t^i = 0 \end{cases}$$

State prediction & Dynamic confidence



CMTA prediction model



Constant Moving and Turning Acceleration:

Assumption 1 : In the prediction phase, accelerations and angular acceleration are **constant** in all directions.

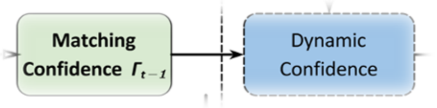
Assumption 2 : Variables and independent are following **Gaussian distribution**.

$$\left[\boxed{x_t^i, y_t^i, z_t^i, \dot{x}_t^i, \dot{y}_t^i, \dot{z}_t^i}, \ddot{x}_t^i, \ddot{y}_t^i, \ddot{z}_t^i, w_t^i, h_t^i, l_t^i, \boxed{\theta_t^i, \dot{\theta}_t^i}, \ddot{\theta}_t^i\right]^T$$

$$\downarrow$$

$$\left[\boxed{\hat{x}_{t+1}^i, \hat{y}_{t+1}^i, \hat{z}_{t+1}^i, \hat{\dot{x}}_{t+1}^i, \hat{\dot{y}}_{t+1}^i, \hat{\dot{z}}_{t+1}^i}, \hat{\ddot{x}}_{t+1}^i, \hat{\ddot{y}}_{t+1}^i, \hat{\ddot{z}}_{t+1}^i, \widehat{w}_{t+1}^i, \hat{h}_{t+1}^i, \hat{l}_{t+1}^i, \boxed{\hat{\theta}_{t+1}^i, \hat{\dot{\theta}}_{t+1}^i}, \hat{\ddot{\theta}}_{t+1}^i\right]^T$$

Dynamic confidence



The **more** predictions it makes, the **less** reliable the predictions are.

Predict:

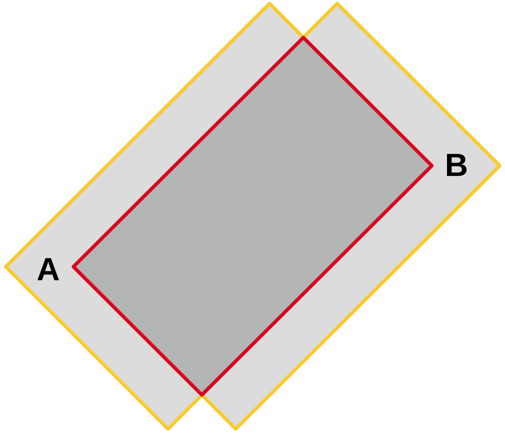
$$\hat{\gamma}^i_{t+1} = \begin{cases} 1 & t = 0 \text{ or } \widehat{D}^{*i}_t = \emptyset \\ I_OU(\widehat{D}^{*i}_t, D^{*i}_t) \cdot \hat{\gamma}^i_t & I_OU(\widehat{D}^{*i}_t, D^{*i}_t)) \geq 1 - \mu \\ (1 - \mu) \cdot \hat{\gamma}^i_t & \text{otherwise} \end{cases}$$

Update:

$$\gamma^i_{t+1} = \begin{cases} 1 & {\sigma^i_t}' = 0 \text{ and } {\sigma^i_{t+1}}' \neq 0 \\ \gamma^i_{t+1} & D^i_{t+1} = \emptyset \\ \hat{\gamma}^i_{t+1} + \left(1 - I_OU\left(\widetilde{D}'^i_t, D^i_t\right)\right) \cdot {\sigma^i_{t+1}}' & \text{otherwise} \end{cases}$$

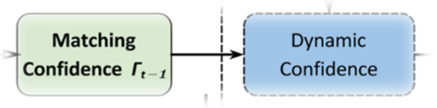
σ^i_t is the detected score.

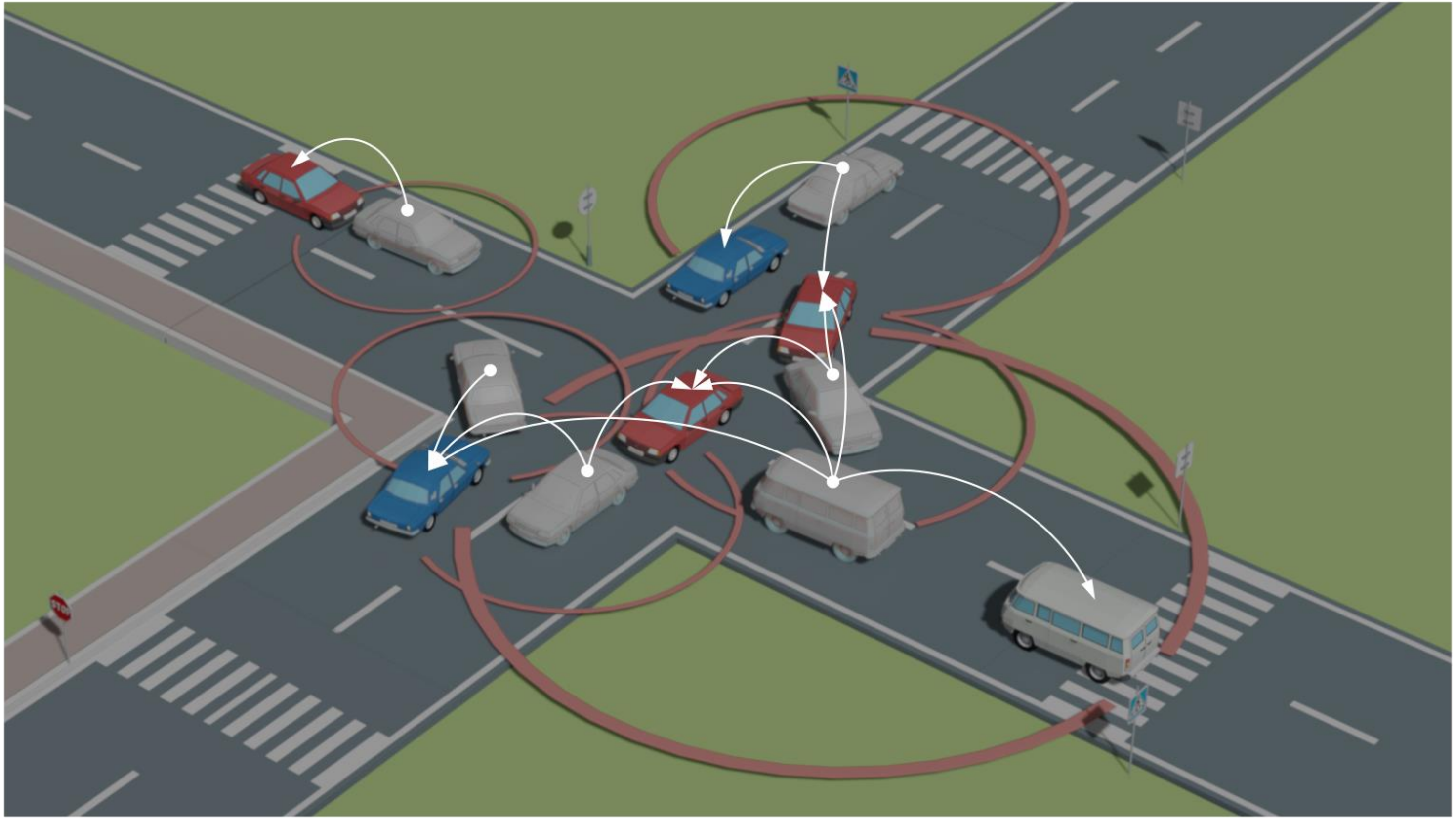
Intersection over union.



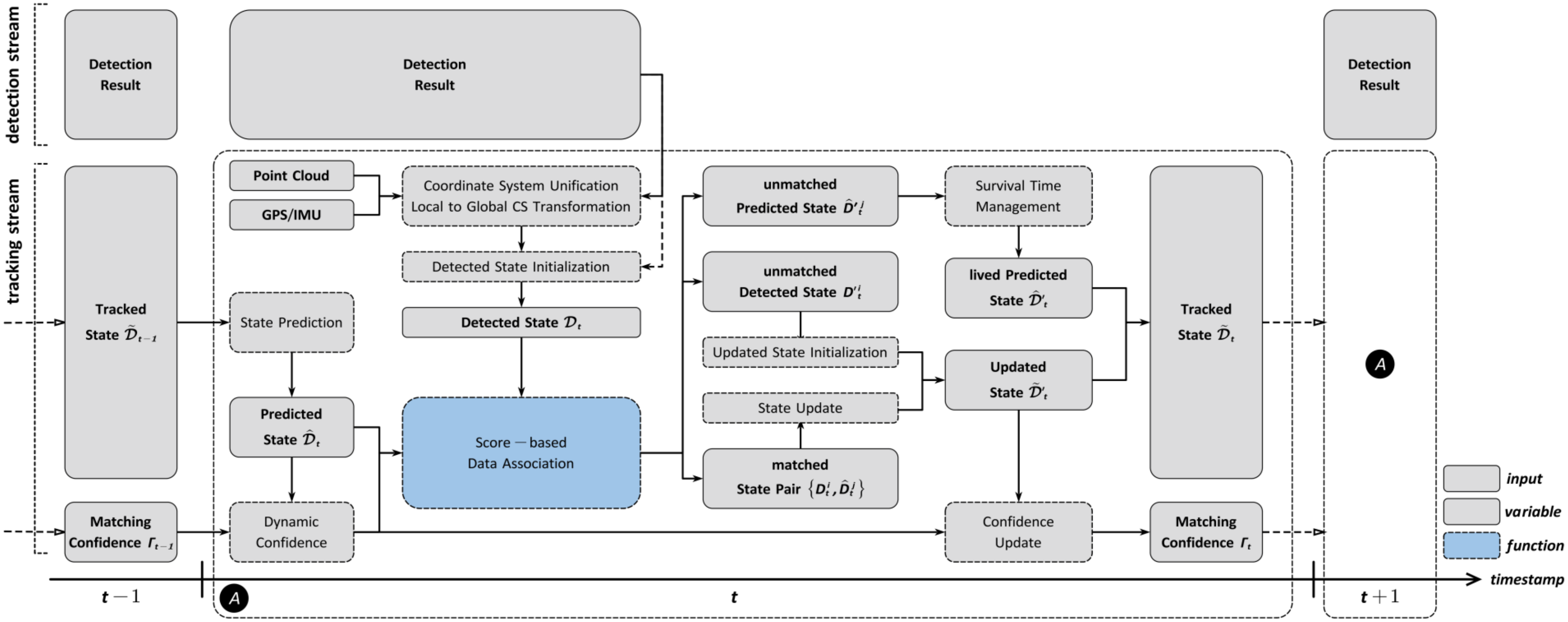
$$I_OU(A, B) = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{A \cap B}{A \cup B}$$

Dynamic confidence

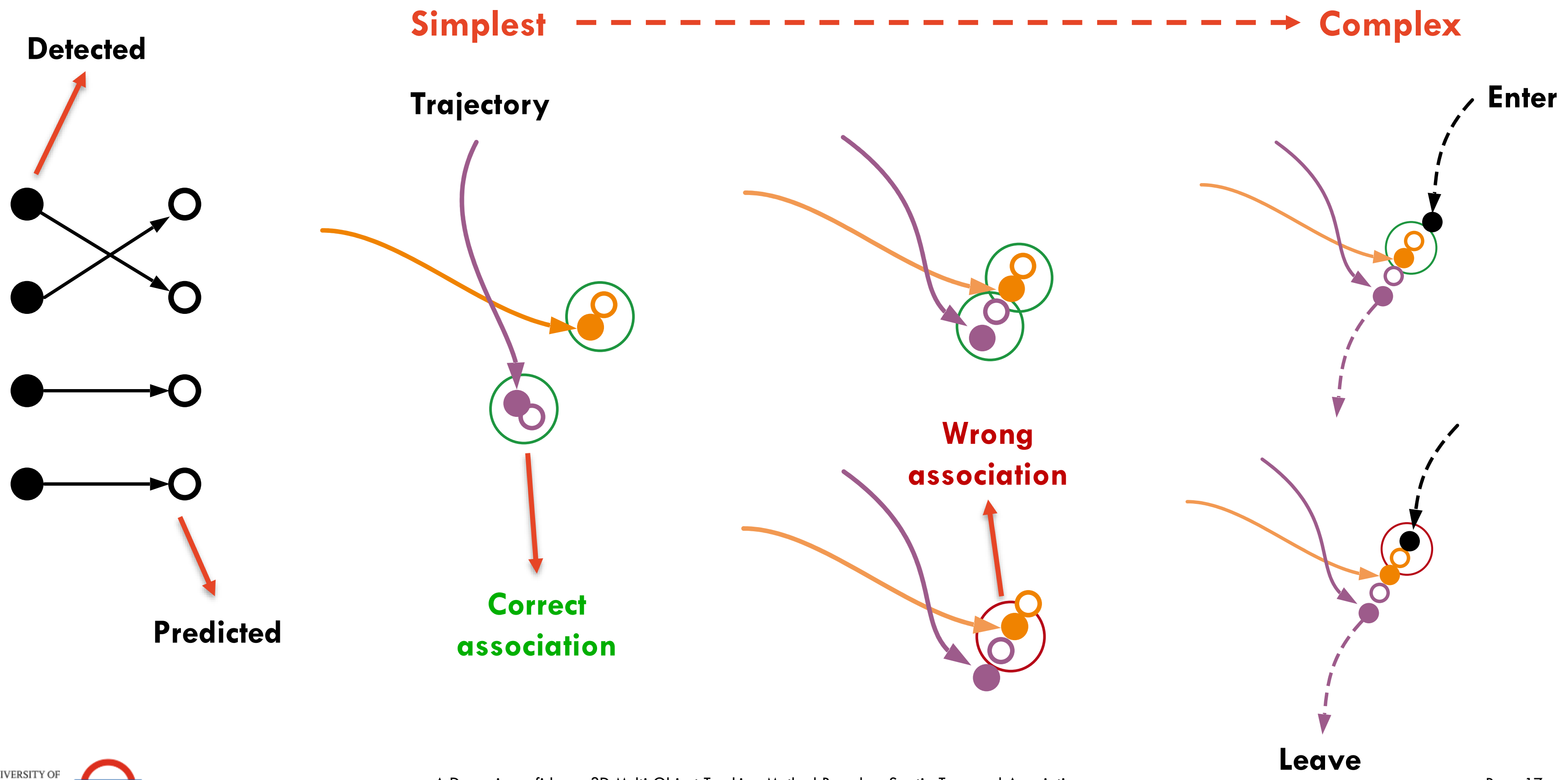




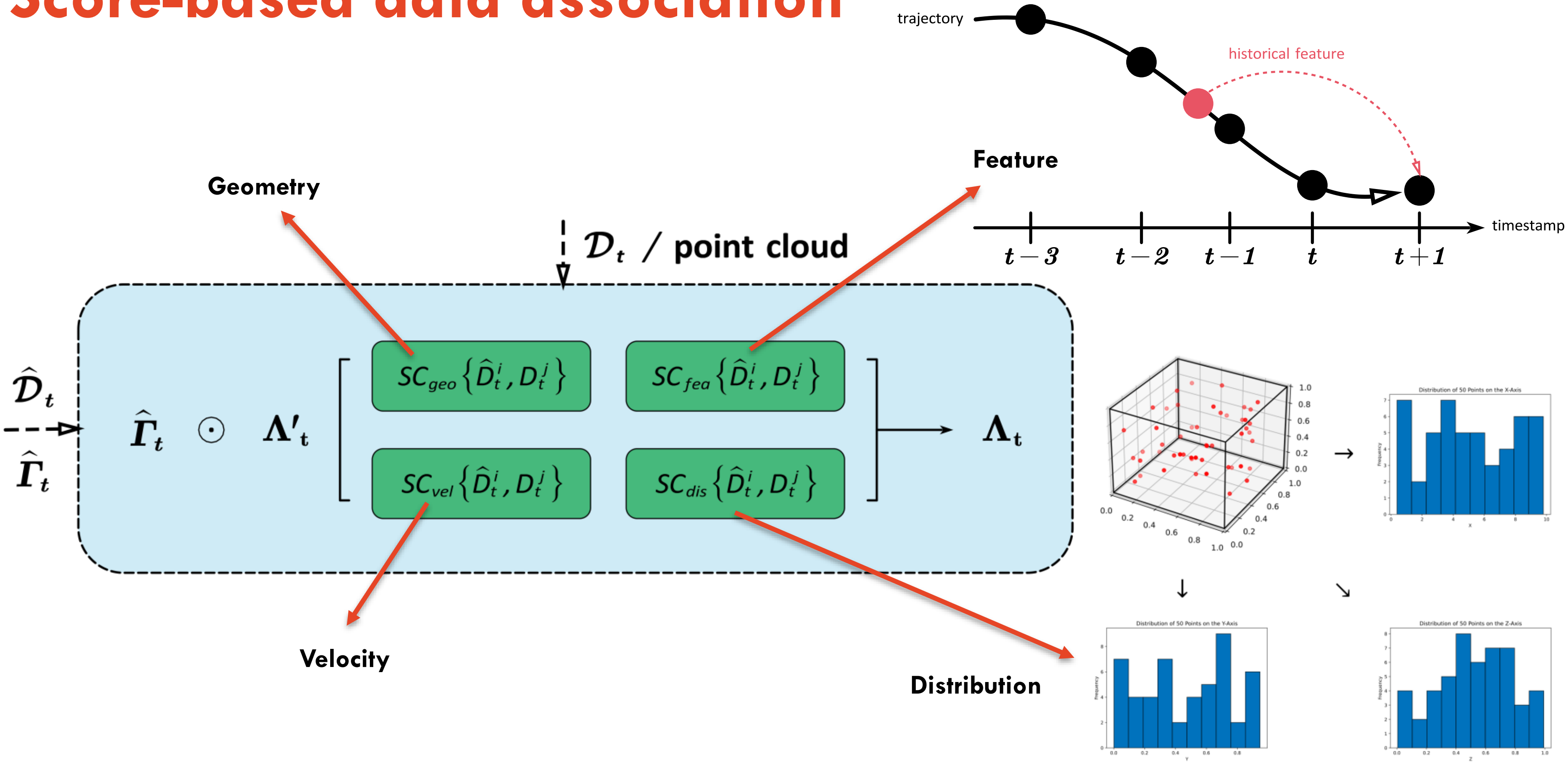
Data association



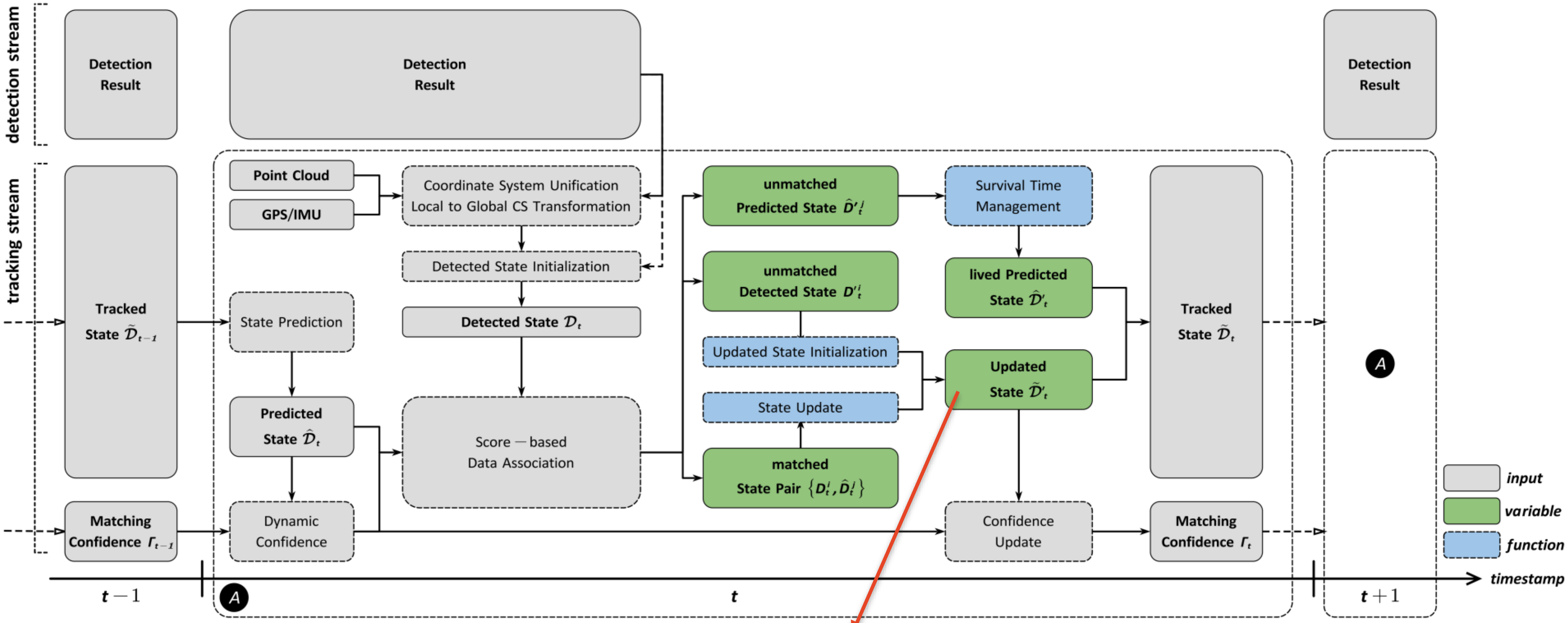
Data association



Score-based data association

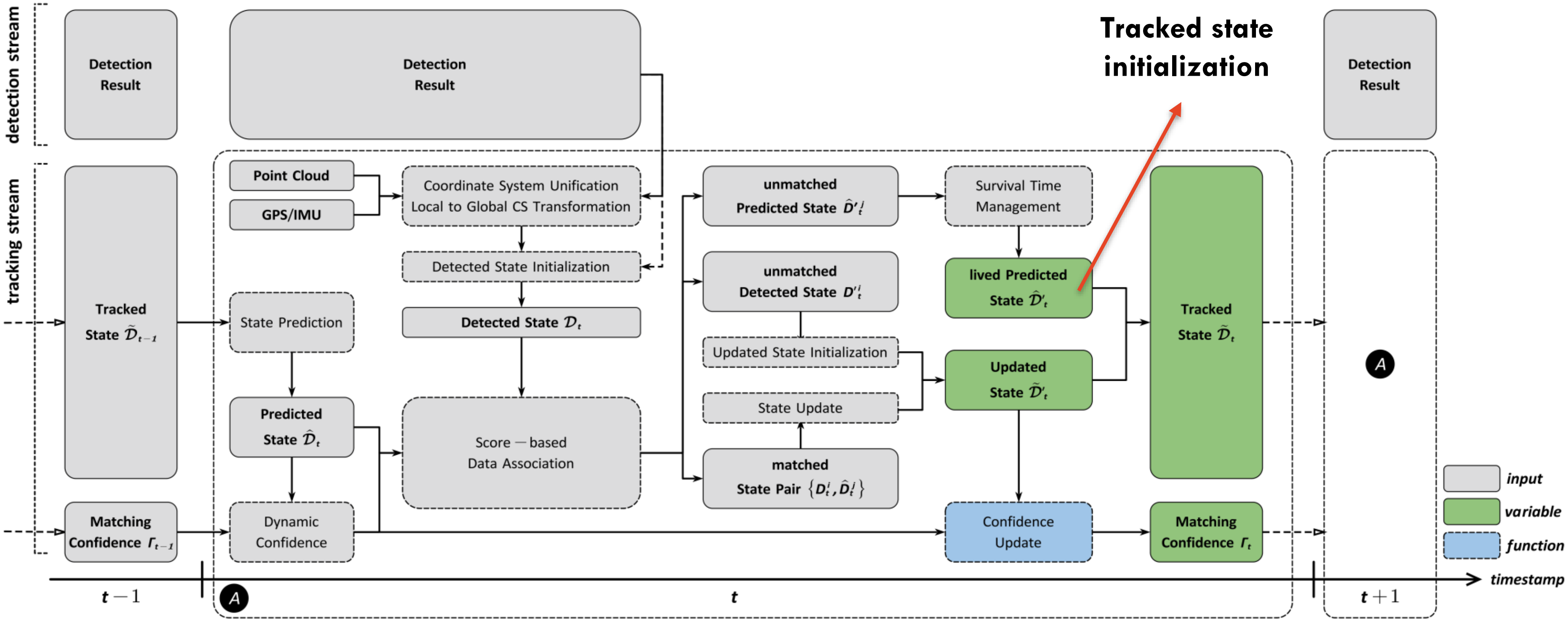


State update



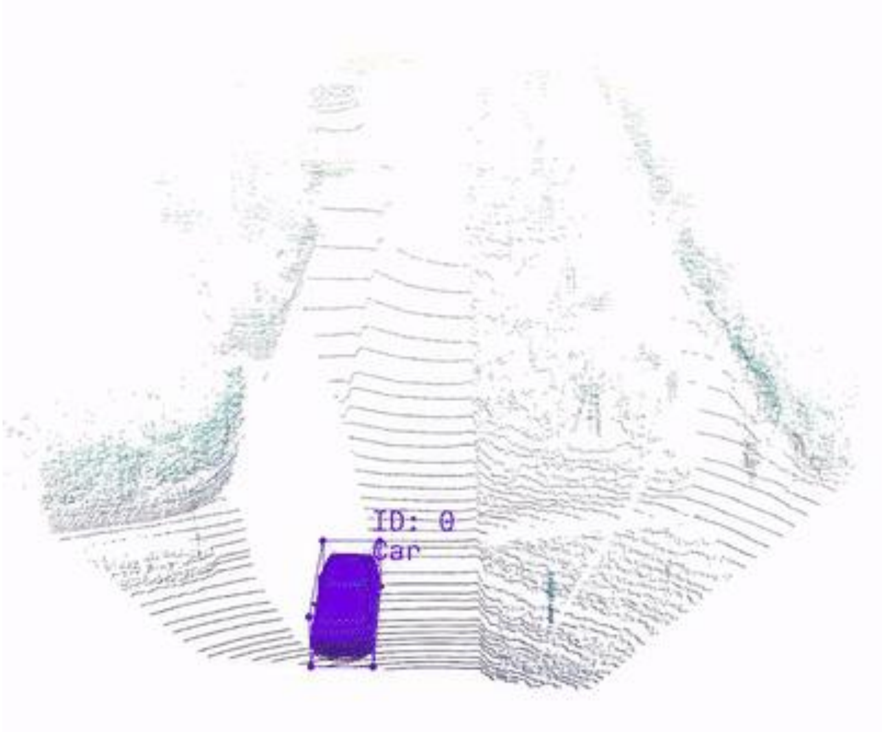
Kalman filter update

Tracked state

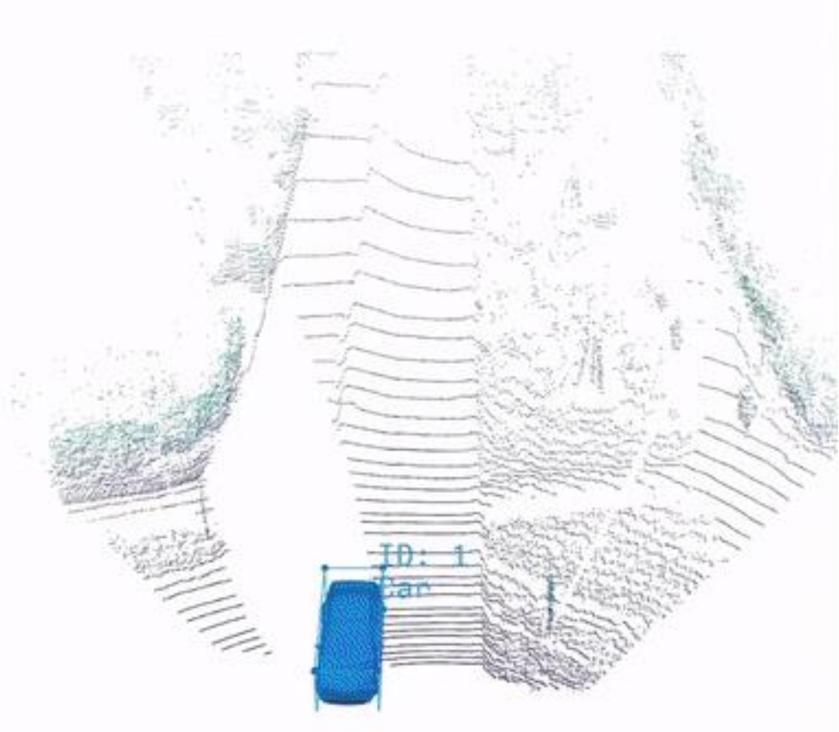


Performance visualization

Proposed method



FANTrack



Performance comparison

Year-Method	Sensor	Type	MOTA	MOTP	Recall	Precision	MT	ML	IDS
19-ComplexerYOLO [1]	Camera	3D	75.70%	78.46%	85.32%	95.18%	58.00%	5.08%	1186
19-FANTrack [2]	Camera/LiDAR	3D	77.72%	82.33%	83.66%	96.15%	62.62%	8.77%	150
18-PMBM [3]	Camera	3D	80.39%	82.33%	85.01%	96.93%	62.77%	6.15%	121
19-aUToTrack [4]	Camera/LiDAR	2D	82.25%	80.52%	89.36%	97.03%	56.77%	7.38%	1025
19-AB3DMOT [5]*	LiDAR	3D	83.84%	85.24%	88.32%	96.98%	66.92%	11.38%	9
19-3DT [6]	Camera	3D	84.52%	85.64%	88.81%	97.95%	73.38%	2.77%	377
19-mmMOT [7]	Camera/LiDAR	2D	84.77%	85.21%	88.81%	97.93%	73.23%	2.77%	284
20-JRMOT [8]	Camera/LiDAR	3D	85.70%	85.48%	89.51%	97.81%	71.85%	4.00%	98
23-ACKF-MOT [9]*	LiDAR	3D	88.73%	86.81%	-	-	85.62%	5.01%	8
22-CasA-MOT [10]*	LiDAR	3D	88.88%	84.37%	92.62%	97.75%	80.00%	8.31%	208
Ours* *(latest)*	LiDAR	3D	90.74%	89.46%	96.50%	95.37%	92.57%	3.82%	11
Ours* *(testing part)*	LiDAR	3D	90.34%	86.17%	-	-	-	-	-

* denotes using the same detector – pvRCNN **(also used as the baseline detector)**

xxxx indicates the performance has not been tested in our own experimental environment

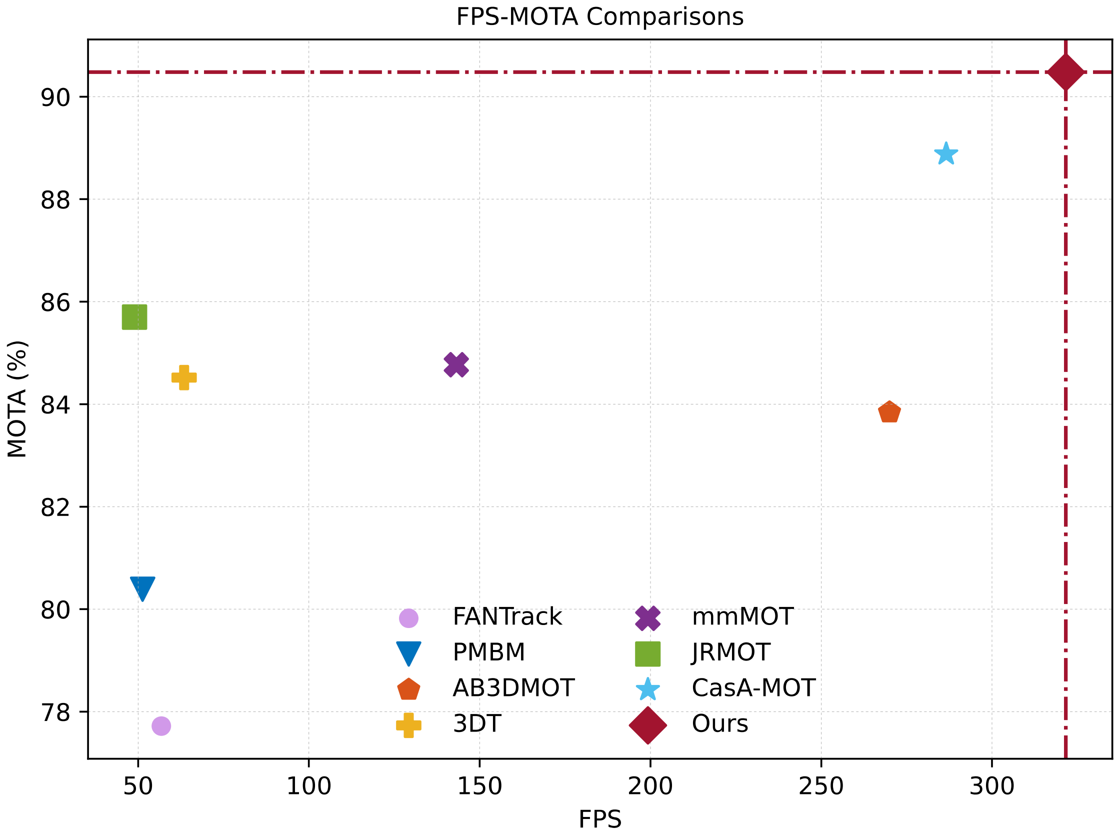
Performance comparison

Experimental configuration

Processor: *AMD Ryzen 5 5600X Desktop Processor*

GPU: *GeForce RTX 3070 VENTUS, 8GB*

Memory: *16GB (2x8GB) PC4-28800 3600MHz DDR4*



Summary

- **More efficient** and **accurate** method compared with SOTA methods.
- More **stable** tracking of temporarily disappearing objects.
- CPU-based.
- Better long-term tracking capabilities (to be further verified).
- Better response to multi-category objects (to be further verified).

Benchmark references

[1] Simon, Martin, Karl Amende, Andrea Kraus, Jens Honer, Timo Samann, Hauke Kaulbersch, Stefan Milz, and Horst Michael Gross. "Complexer-yolo: Real-time 3d object detection and tracking on semantic point clouds." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0-0. *2019*.

[2] Baser, Erkan, Venkateshwaran Balasubramanian, Prarthana Bhattacharyya, and Krzysztof Czarnecki. "Fantrack: 3d multi-object tracking with feature association network." In 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 1426-1433. IEEE, *2019*.

[3] Scheidegger, Samuel, Joachim Benjaminsson, Emil Rosenberg, Amrit Krishnan, and Karl Granström. "Mono-camera 3d multi-object tracking using deep learning detections and pmbm filtering." In 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 433-440. IEEE, *2018*.

[4] Burnett, Keenan, Sepehr Samavi, Steven Waslander, Timothy Barfoot, and Angela Schoellig. "autotrack: A lightweight object detection and tracking system for the sae autodrive challenge." In 2019 16th Conference on Computer and Robot Vision (CRV), pp. 209-216. IEEE, *2019*.

[5] Weng, Xinshuo, Jianren Wang, David Held, and Kris Kitani. "3d multi-object tracking: A baseline and new evaluation metrics." In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10359-10366. IEEE, *2020*.

[6] Hu, Hou-Ning, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krahenbuhl, Trevor Darrell, and Fisher Yu. "Joint monocular 3D vehicle detection and tracking." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5390-5399. *2019*.

[7] Zhang, Wenwei, Hui Zhou, Shuyang Sun, Zhe Wang, Jianping Shi, and Chen Change Loy. "Robust multi-modality multi-object tracking." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2365-2374. *2019*.

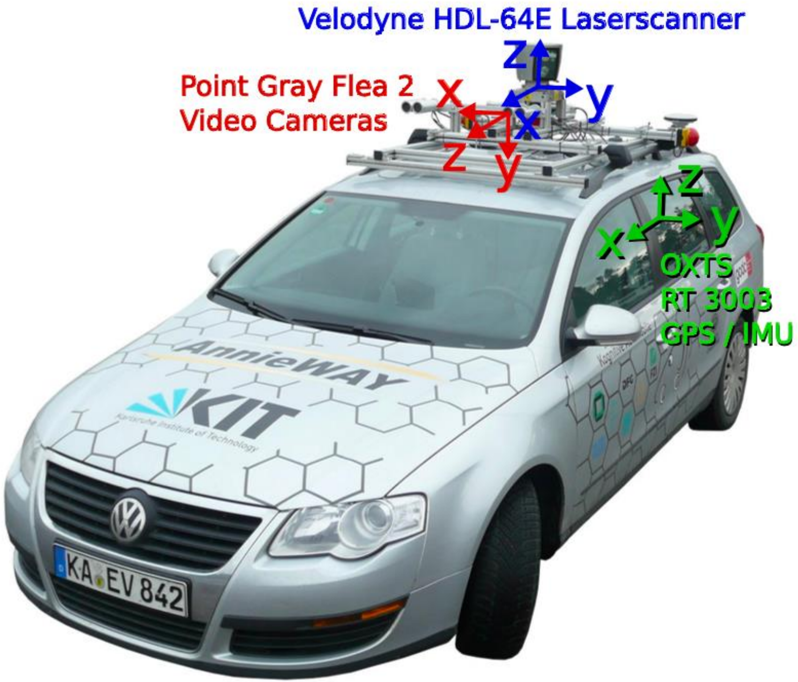
[8] Shenoi, Abhijeet, Mihir Patel, JunYoung Gwak, Patrick Goebel, Amir Sadeghian, Hamid Rezatofighi, Roberto Martin-Martin, and Silvio Savarese. "Jrmot: A real-time 3d multi-object tracker and a new large-scale dataset." In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10335-10342. IEEE, *2020*.

[9] Wu, Hai, Wenkai Han, Chenglu Wen, Xin Li, and Cheng Wang. "3d multi-object tracking in point clouds based on prediction confidence-guided data association." IEEE Transactions on Intelligent Transportation Systems 23, no. 6 (*2021*): 5668-5677.

[10] Guo, Ge, and Shijie Zhao. "3D multi-object tracking with adaptive cubature Kalman filter for autonomous driving." IEEE Transactions on Intelligent Vehicles (*2022*).

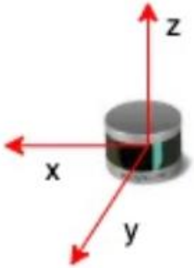
Questions?

Point cloud & GPS/IMU



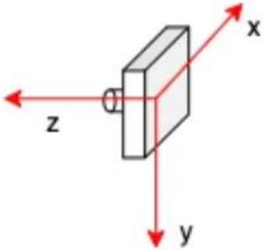
Recording platform.

From "Vision meets Robotics: The KITTI Dataset".



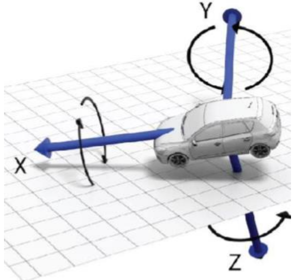
LiDAR - CS

Detection result



Camera - CS

Ground truth



GPS - CS

Global position

Data association

$$\boldsymbol{SC_{geo}^{i,j}} := \lambda_1 \cdot \mathcal{N}\left(\sum_{k\in\{w,h,l\}} \frac{\left|\hat{k}-k\right|}{\hat{k}+k}\right) + \lambda_2 \cdot \mathcal{N}(\|\hat{p}-p\|_2^2) + \lambda_3 \cdot \mathcal{N}\left(sin\left|\hat{\theta}-\theta\right|\right)$$

$$\boldsymbol{SC_{fea}^{i,j}} := \mathcal{N}\left(\exp\left(\left\|\hat{f}-\bar{f}\right\|_2^2\right)\right)$$

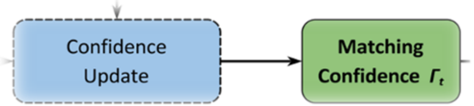
$$\left\{v_{t+1}\middle|D_{t+1}^i\right\} = \left\{v_t\middle|\widehat{D}_t^j\right\} + \left\{\overline{\Delta v_t}\middle|\widehat{D}_t^j\right\}$$

$$\boldsymbol{SC_{vel}^{i,j}} := \lambda_4 \cdot \mathcal{N}\left(\frac{\left|\overline{v_{t+1}^i} - \overline{v_{t+1}^j}\right|}{\overline{v_{t+1}^i} + \overline{v_{t+1}^j}}\right) + \lambda_5 \cdot \mathcal{N}\left(\sum_{m=1}^{n}\left(v_m - \overline{v_{t+1}^i}\right)^2 + \sum_{m=1}^{n}\left(v_m - \overline{v_{t+1}^j}\right)^2\right)$$

$$S_C(A,B) := \cos(\theta) = \frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

$$\begin{aligned}\boldsymbol{SC_{dis}^{i,j}} :=\ & \eta_1 \cdot S_C\left(ProjDis\left(D_x^i\right), ProjDis\left(\widehat{D}_x^j\right)\right)\\ & +\eta_2 \cdot S_C\left(ProjDis\left(D_y^i\right), ProjDis\left(\widehat{D}_y^j\right)\right)\\ & +\eta_3 \cdot S_C\left(ProjDis\left(D_z^i\right), ProjDis\left(\widehat{D}_z^j\right)\right)\end{aligned}$$

Confidence update



label | *truncated* | *occlude* | observation angle | **2D_bbs_Xmin** | **2D_bbs_Ymin** | **2D_bbs_Xmax** | **2D_bbs_Ymax** | **3D_bbs_height** | **3D_bbs_width** | **3D_bbs_length** | 3D_x | 3D_y | 3D_z | yaw $\boxed{\text{| detection score}}$

$$\sigma_{t+1}^{i}{}' = \mathrm{sigmoid}\left(\sigma_{t+1}^{i}\right) = \frac{1}{1 + e^{-\sigma_{t+1}^{i}}}$$

Detection score

$$\gamma_{t+1}^{i} = \begin{cases} 1 & \sigma_t^{i}{}' = 0 \text{ and } \sigma_{t+1}^{i}{}' \neq 0 \\ \gamma_{t+1}^{i} & D_{t+1}^{i} = \emptyset \\ \hat{\gamma}_{t+1}^{i} + \left(1 - I_O U\left(\widetilde{D}'^{i}_{t}, D_t^{i}\right)\right) \cdot \sigma_{t+1}^{i}{}' & \text{otherwise} \end{cases}$$

Performance indexes

Based on bounding boxes

gtDet - ground truth detection
prDet – predicted detection (tracked)

	Correct gtDet	Missed gtDet
Correct prDet	TP	-
Extra prDet	FP	FN

$$S(IoU_{Loc}) \geq 50\%$$

$$\text{MOTA} = 1 - \frac{|\text{FN}| + |\text{FP}| + |\text{IDSW}|}{|\text{gtDet}|}$$

$$\text{MOTP} = \frac{1}{|\text{TP}|} \sum_{\text{TP}} S$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\textbf{M}\text{ostly }\textbf{T}\text{racked} = \frac{"tracker\ output"}{"GT\ trajectorie"} \geq 80\%$$

$$\textbf{M}\text{ostly }\textbf{L}\text{ost} = \frac{"tracker\ output"}{"GT\ trajectorie"} \leq 20\%$$