Deep Fusion of Multi-Object Densities Using Transformer

Lechi Li, Chen Dai, Yuxuan Xia, Lennart Svensson

Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden

Decentralized multi-sensor fusion

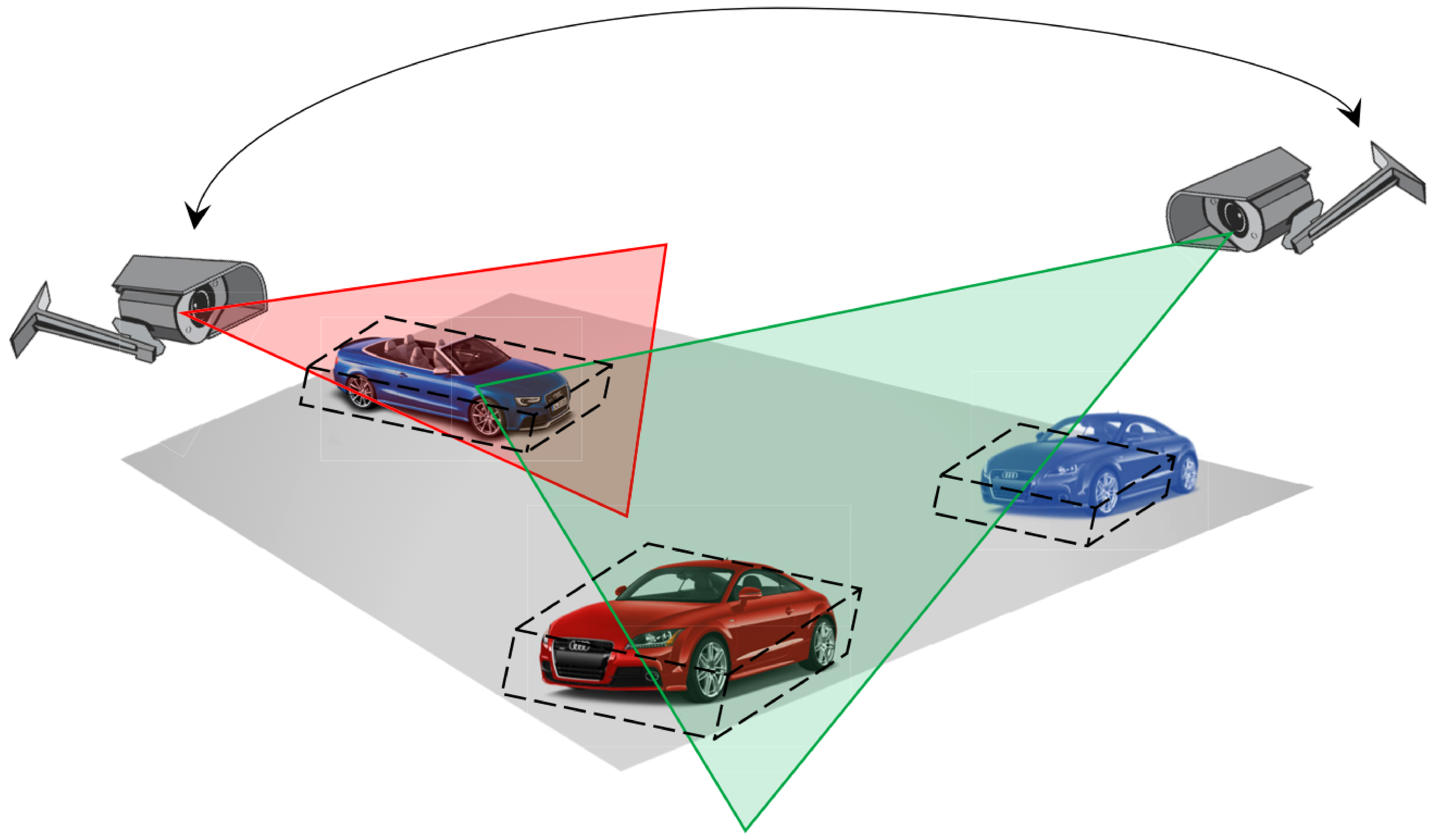

Figure 1. Scenario with three objects observed by two sensors with partially overlapping field-of-views. Image taken from [1].

- Multiple sensors with possibly different field-of-views (FoVs), where tracking is performed locally at each sensor, are utilized to estimate the moving objects in the region of interest.
- To leverage the information available at local sensors, the local multi-object densities (MODs) obtained at each sensor need to be fused to obtain a global MOD.

Limitations of model-based methods

- Current model-based methods, in theory, do not provide Bayesian optimal fusion results.
- Current fusion methods only use the local multi-object posterior densities at the current time while ignoring information regarding the objects' previous states.
- It is difficult to develop a model-based method that leverages all the uncertainties captured in the MODs on sets of trajectories [2].

Contributions

- **The first deep learning-based solution to the fusion of MODs.**
- MT3 [3], a transformer-based deep neural network, is used to fuse local MODs on sets of trajectories to obtain a global MOD that describes the set of current objects.

Problem formulation

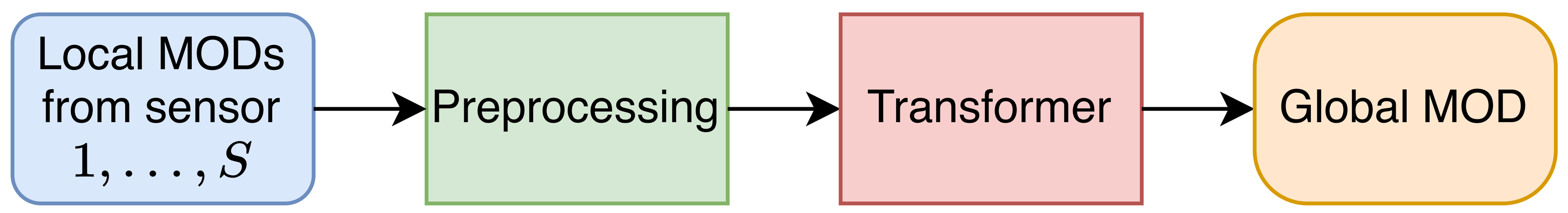

Figure 2. Diagram of fusion of MODs using transformer.

- The local MOD $f^s(\mathbf{X}_T \mid \mathbf{z}^s_{1:T})$ at the s-th sensor represents the posterior density of the set $\mathbf{X}_T$ of trajectories in the time interval $1:T$ in the form of multi-Bernoulli (MB) [2].
- The global MOD $f(\mathbf{x}_T \mid \mathbf{z}^1_{1:T}, \ldots, \mathbf{z}^S_{1:T})$ represents the set of detected objects $\mathbf{x}_T$ at time step T in the form of MB, where an existence probability and a Gaussian single-object density characterize each Bernoulli component.

Deep MOD fusion using transformer

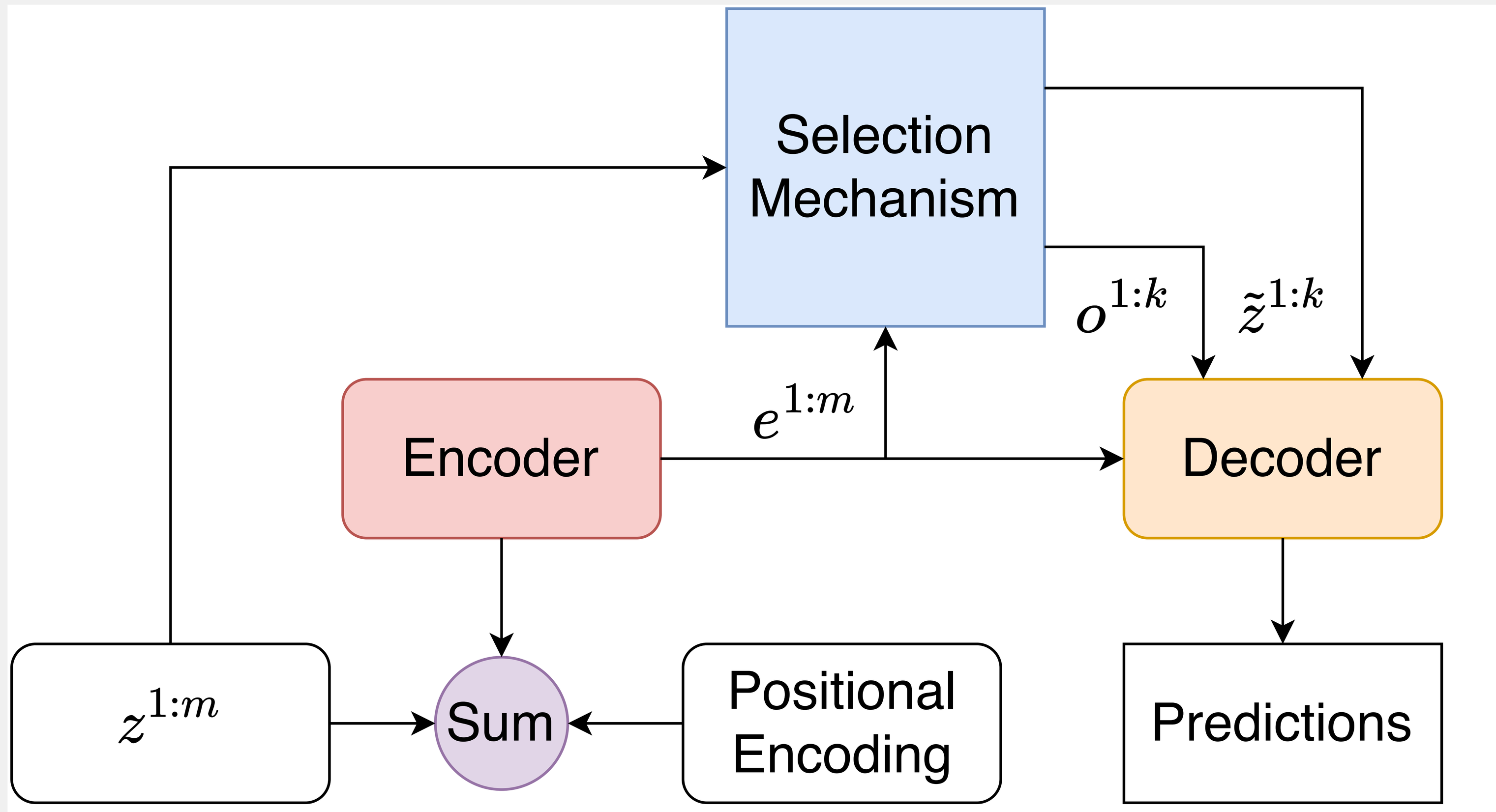

Figure 3. The structure of MT3 [3].

- **Preprocessing** Concatenate parameters of the i-th Bernoulli at the s-th sensor:
  - Existence probability $r^{s,i}$.
  - Trajectory start time $t^{s,i}$.
  - Probability mass function of trajectory length $w^{s,i}_j$, $j \in \{1, \ldots, \ell^{s,j}\}$.
  - Mean of state sequence $x^{s,i}_{1:\ell^{s,i}} \in \mathbb{R}^{4\ell^{s,i}}$.
  - (Reduced) covariance of state sequence $P^{s,i}_{1:\ell^{s,i}} \in \mathbb{R}^{4\ell^{s,i} \times 4\ell^{s,i}}$.

  Include sensor position information when sensors are mobile.
- **Positional encoding** Encode time step $t \in \{t^{s,i}, \ldots, t^{s,i} + \ell^{s,i} - 1\}$, trajectory index $i \in \{1, \ldots, n^s\}$ and sensor index $s \in \{1, \ldots, S\}$ into the input embedding.
- **Loss** Negative log-likelihood (NLL) of the MB density evaluated at the ground truth [4].

Scenario setup

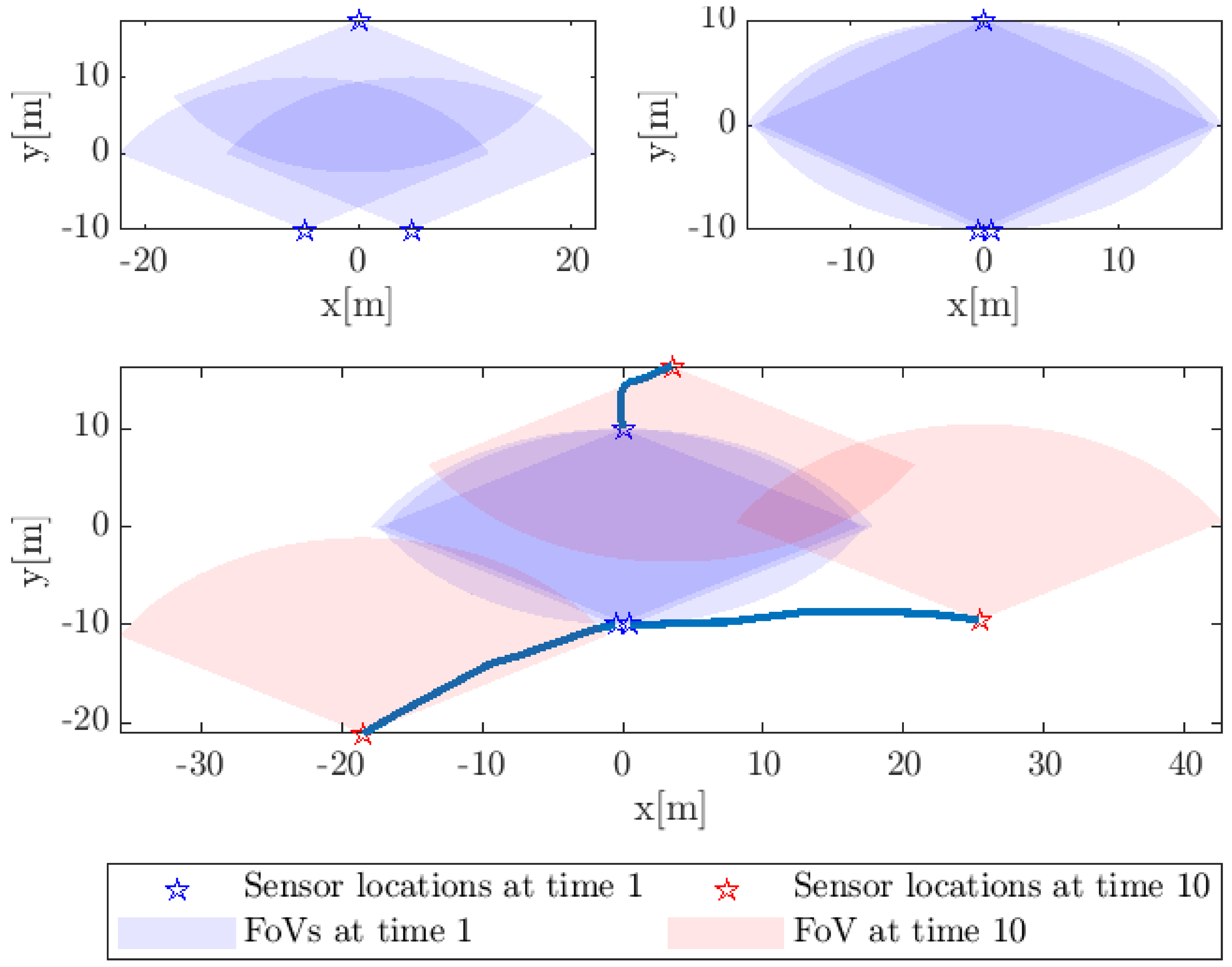

Figure 4. An illustration of scenario 1 (top left), scenario 2 (top right), and a sample plot of scenario 3 (bottom, sensor trajectories are shown in solid lines).

Results on synthetic data

Scenario	1		2		3	
Method	Bayesian	MT3	Bayesian	MT3	Bayesian	MT3
GOSPA-total	1.618	**1.373**	1.440	**1.413**	1.294	**0.885**
GOSPA-loc	0.596	0.959	0.693	1.040	0.256	0.597
GOSPA-miss	0.058	0.051	0.076	0.031	0.003	0.025
GOSPA-false	0.964	0.363	0.671	0.342	1.035	0.263
NLL-total	12.207	**1.195**	2.695	**1.833**	13.105	**0.429**
NLL-loc	11.766	0.572	2.103	1.169	12.656	0.089
NLL-miss	0.320	0.304	0.253	0.411	0.428	0.137
NLL-false	0.121	0.319	0.340	0.253	0.021	0.202

Table 1. Performance comparison with [1] in terms of GOSPA [5] and NLL [4].

Results show that MT3 [3] outperforms the model-based Bayesian fusion method [1] in terms of both GOSPA and NLL in all the scenarios.

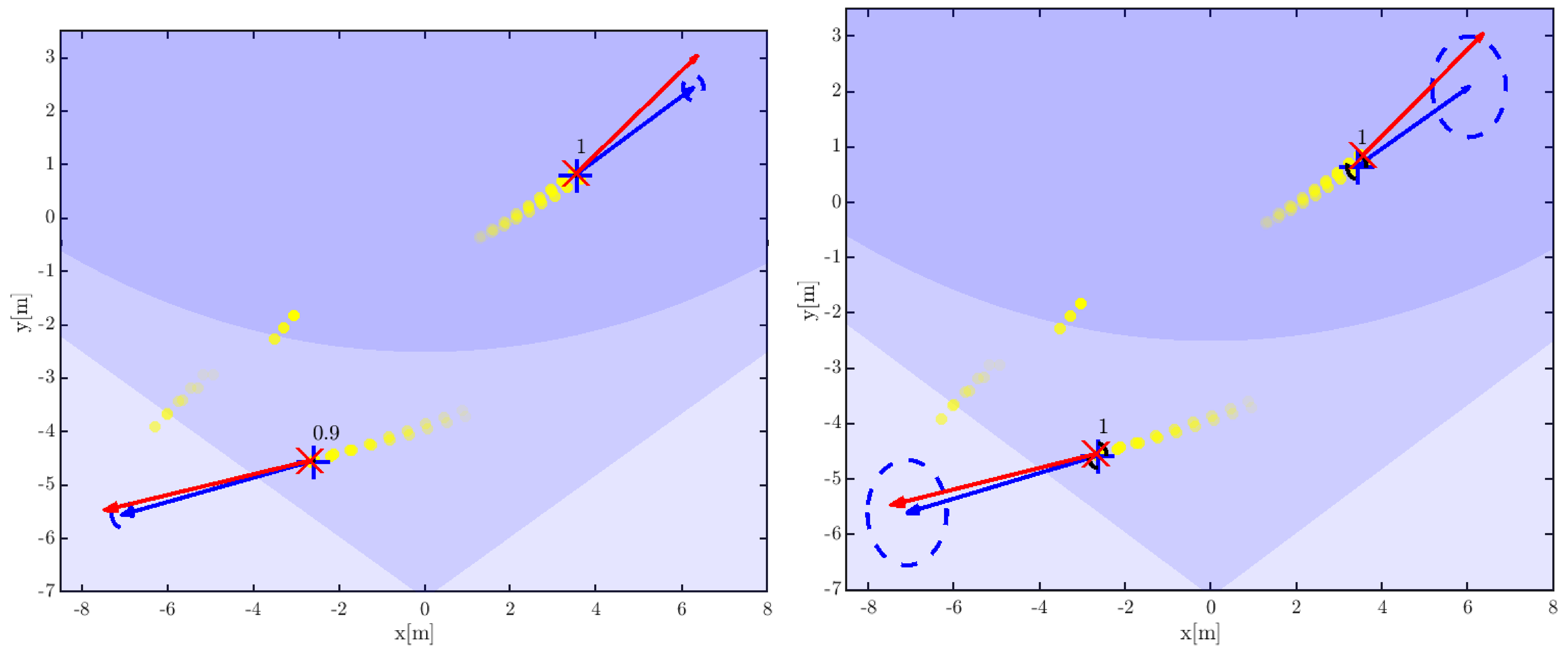

Figure 5. Sample plot of Bayesian method (Left) and MT3 (Right). Yellow-filled circles indicate estimated positions obtained from local filters. The ground truth positions/velocities at the current time are shown in red crosses/arrows, respectively, while predicted positions/velocities are shown in blue plus signs/arrows. The blue/black dished ellipses represent the 3-σ level of predicted velocities/positions.

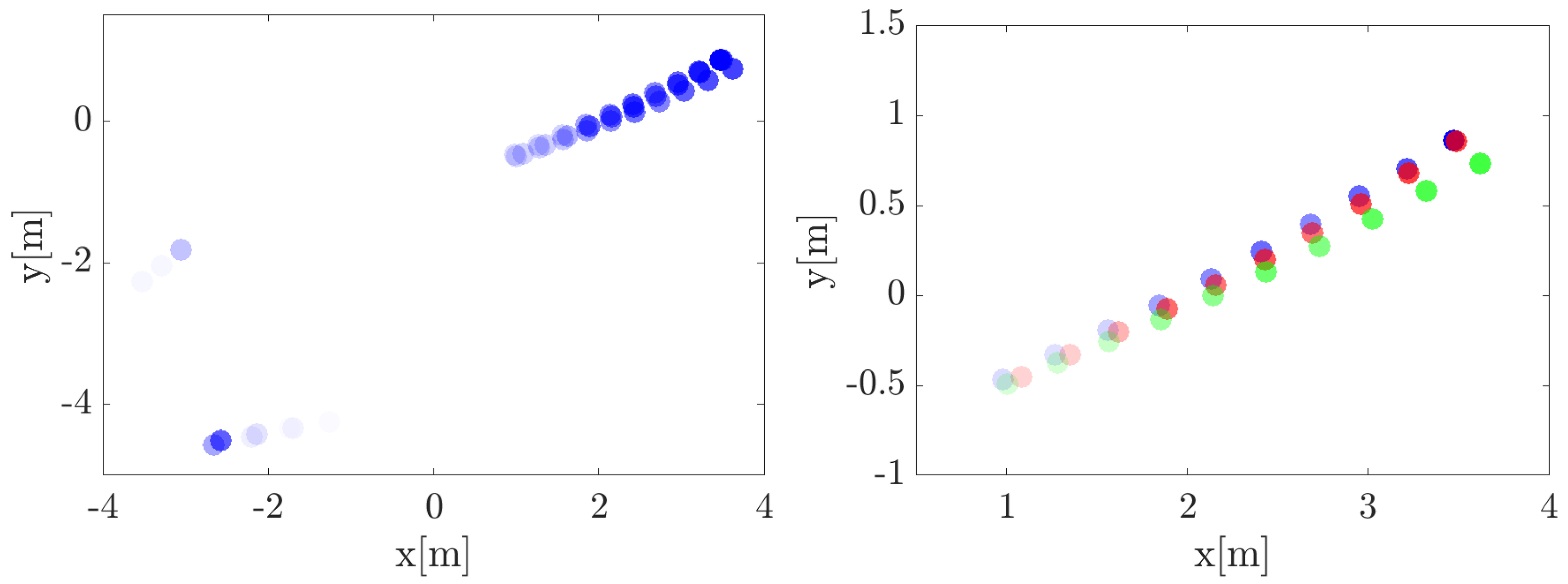

Figure 6. Left: attention maps of the predictions of MT3 in blue-filled circles. Right: corresponding trajectory estimates from different sensors, indicated using different colours.

[1] M. Fröhle, K. Granström, and H. Wymeersch, "Decentralized Poisson multi-Bernoulli filtering for vehicle tracking," *IEEE Access*, vol. 8, pp. 126414–126427, 2020.

[2] Á. F. García-Fernández, L. Svensson, J. L. Williams, Y. Xia, and K. Granström, "Trajectory Poisson multi-Bernoulli filters," *IEEE Trans. on Signal Process.*, vol. 68, pp. 4933–4945, 2020.

[3] J. Pinto, G. Hess, W. Ljungbergh, Y. Xia, L. Svensson, and H. Wymeersch, "Next generation multitarget trackers: Random finite set methods vs transformer-based deep learning," in *24th International Conference on Information Fusion*, IEEE, 2021.

[4] J. Pinto, Y. Xia, L. Svensson, and H. Wymeersch, "An uncertainty-aware performance measure for multi-object tracking," *IEEE Signal Process Lett.*, vol. 28, pp. 1689–1693, 2021.

[5] A. S. Rahmathullah, Á. F. García-Fernández, and L. Svensson, "Generalized optimal sub-pattern assignment metric," in *20th International Conference on Information Fusion*, IEEE, 2017.

yuxuan.xia@chalmers.se