Multimodal Future Localization and Emergence Prediction for Objects in Egocentric View with a Reachability Prior
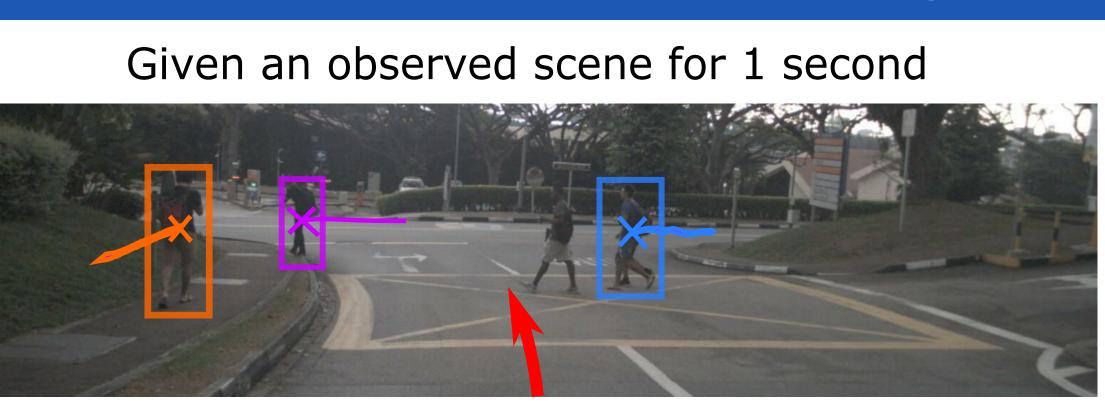
CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Osama Makansi
makansio@cs.uni-freiburg.de

Özgün Cicek
cicek@cs.uni-freiburg.de

Kevin Buchicchio
buchicchio@imra-europe.com

Thomas Brox
brox@cs.uni-freiburg.de

[O] Vision — COMPUTER VISION University of Freiburg

Motivation

Given an observed scene for 1 second — Predict the future in 3 seconds

left ① ② right

Challenges
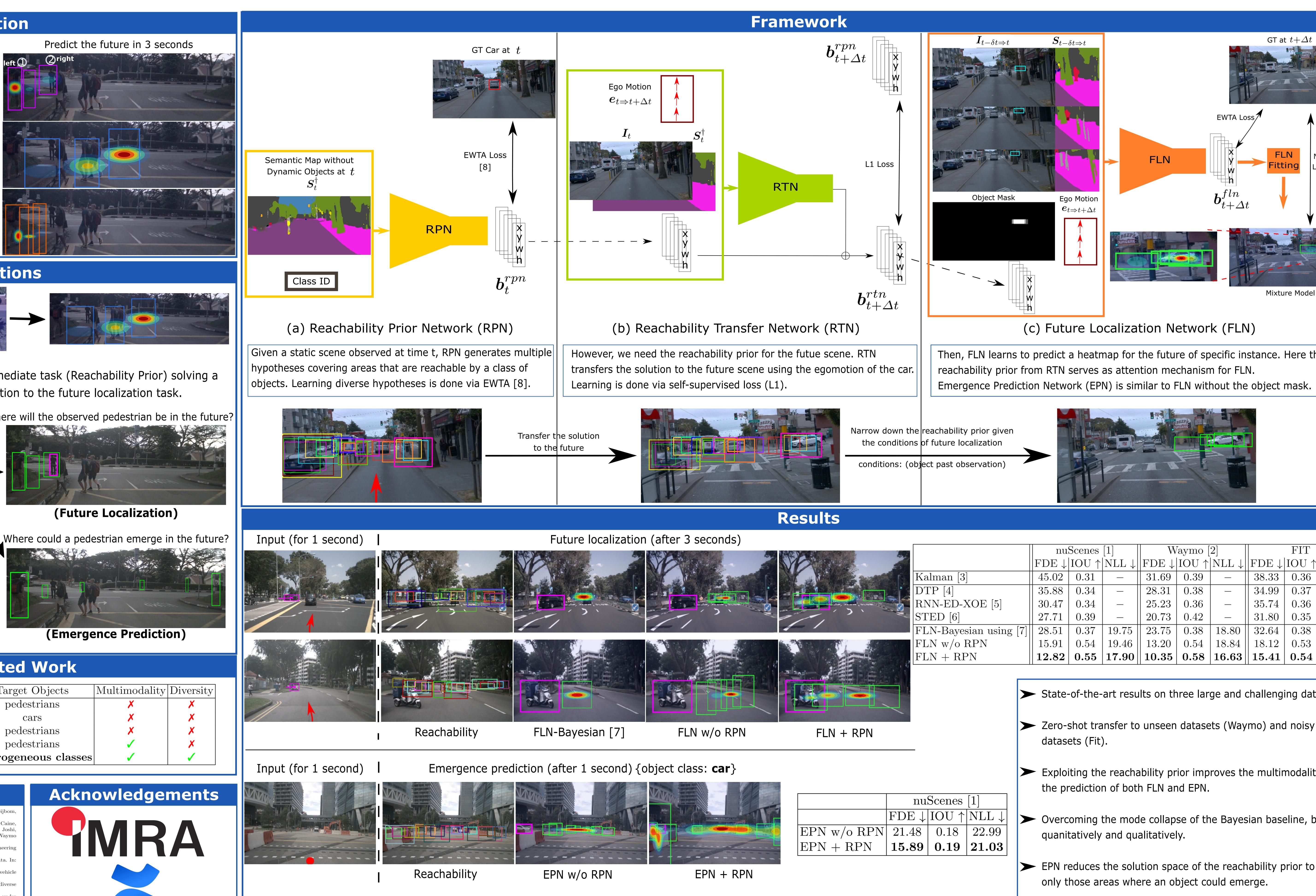
- Partial visibility due to the egocentric view.
- Large field-of-view change due to the egomotion.
- Multimodality and uncertainty of the future.

Input Conditions

- Scene context.　　- Object's location history.

Contributions

→ Bringing the multimodality of the future from top-view to egocentric view

from [8]

→ This is achieved by introducing a new intermediate task (Reachability Prior) solving a broader task, then narrowing down the solution to the future localization task.
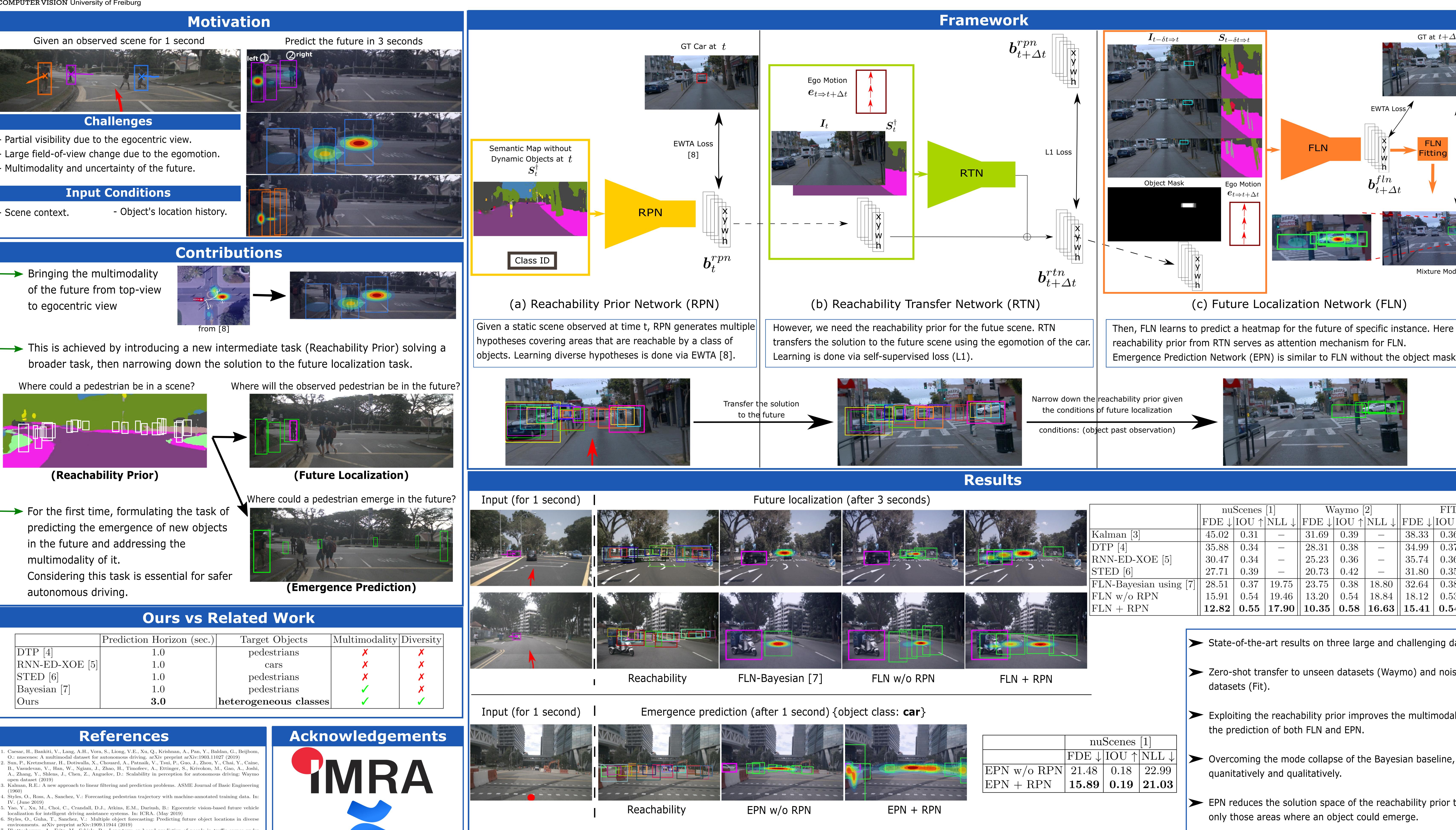
Where could a pedestrian be in a scene?
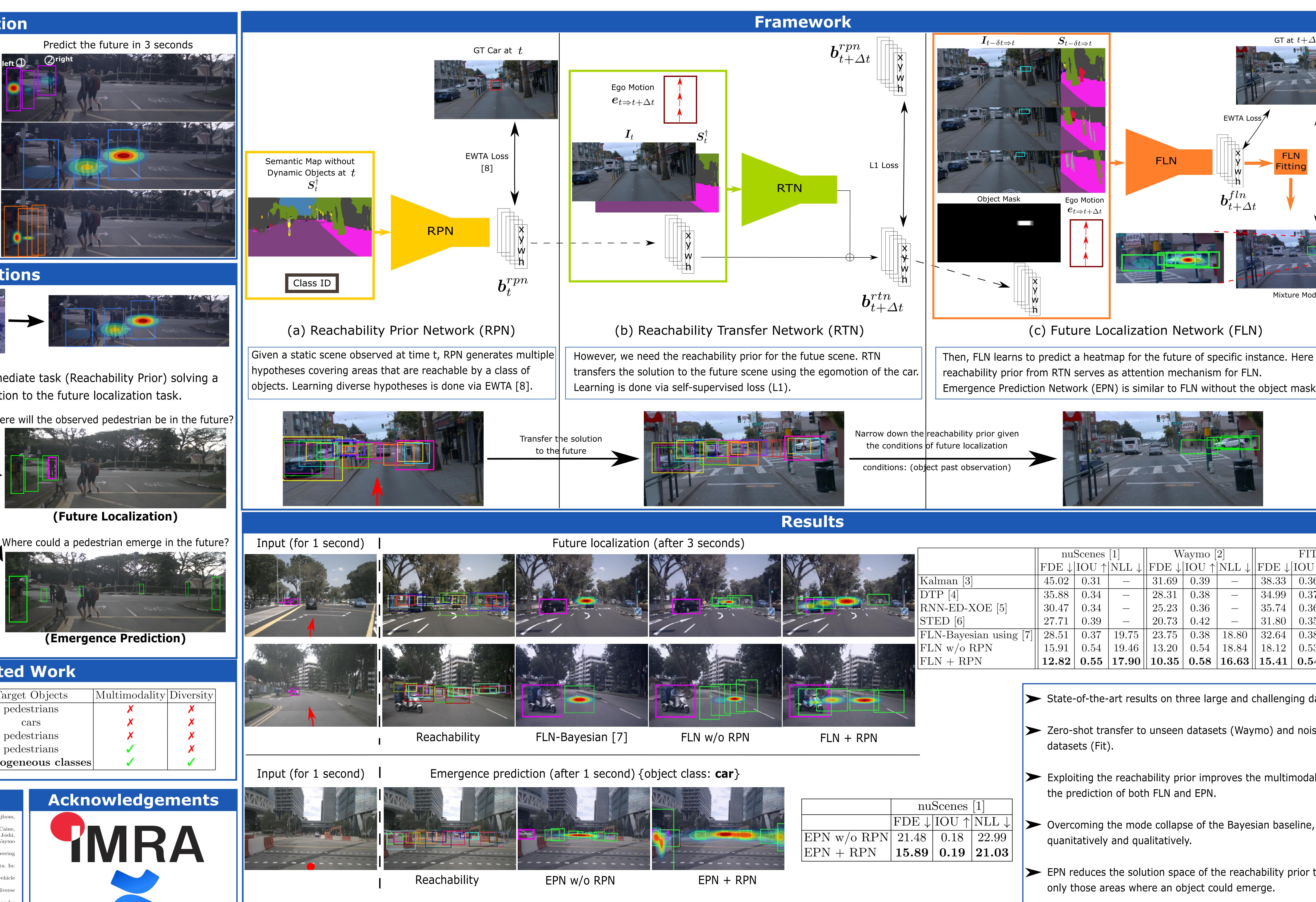**(Reachability Prior)**

Where will the observed pedestrian be in the future?
(Future Localization)

Where could a pedestrian emerge in the future?
(Emergence Prediction)

→ For the first time, formulating the task of predicting the emergence of new objects in the future and addressing the multimodality of it.
Considering this task is essential for safer autonomous driving.
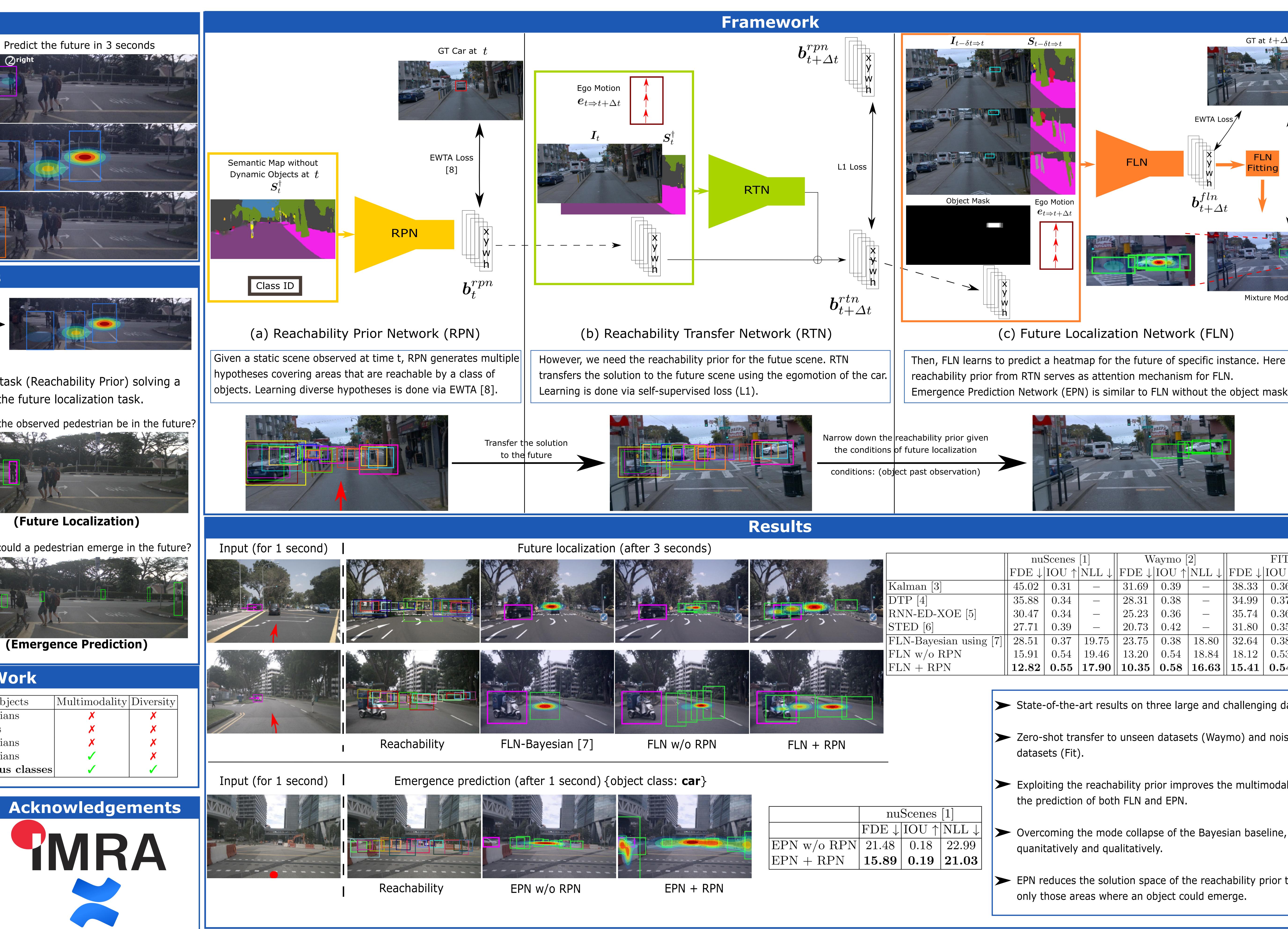
Ours vs Related Work

	Prediction Horizon (sec.)	Target Objects	Multimodality	Diversity
DTP [4]	1.0	pedestrians	✗	✗
RNN-ED-XOE [5]	1.0	cars	✗	✗
STED [6]	1.0	pedestrians	✗	✗
Bayesian [7]	1.0	pedestrians	✓	✗
Ours	**3.0**	**heterogeneous classes**	✓	✓

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027 (2019)
2. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset (2019)
3. Kalman, R.E.: A new approach to linear filtering and prediction problems. ASME Journal of Basic Engineering (1960)
4. Styles, O., Ross, A., Sanchez, V.: Forecasting pedestrian trajectory with machine-annotated training data. In: IV. (June 2019)
5. Yao, Y., Xu, M., Choi, C., Crandall, D.J., Atkins, E.M., Dariush, B.: Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In: ICRA. (May 2019)
6. Styles, O., Guha, T., Sanchez, V.: Multiple object forecasting: Predicting future object locations in diverse environments. arXiv preprint arXiv:1909.11944 (2019)
7. Bhattacharyya, A., Fritz, M., Schiele, B.: Long-term on-board prediction of people in traffic scenes under uncertainty. In: CVPR. (June 2018)
8. Makansi, O., Ilg, E., Cicek, O., Brox, T.: Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In: CVPR. (June 2019)

Acknowledgements

IMRA

Framework

GT Car at t

$b_{t+\Delta t}^{rpn}$

Semantic Map without Dynamic Objects at t $S_t^\dagger$

Class ID

RPN → x y w h — b_t^{rpn}

EWTA Loss [8]

Ego Motion $e_{t\Rightarrow t+\Delta t}$

I_t　$S_t^\dagger$

RTN

L1 Loss

$b_{t+\Delta t}^{rtn}$

$I_{t-\delta t \Rightarrow t}$　$S_{t-\delta t \Rightarrow t}$

Object Mask　Ego Motion $e_{t\Rightarrow t+\Delta t}$

FLN

EWTA Loss　FLN Fitting　NLL Loss

GT at $t+\Delta t$

$b_{t+\Delta t}^{fln}$

Mixture Model

(a) Reachability Prior Network (RPN)

Given a static scene observed at time t, RPN generates multiple hypotheses covering areas that are reachable by a class of objects. Learning diverse hypotheses is done via EWTA [8].

(b) Reachability Transfer Network (RTN)

However, we need the reachability prior for the future scene. RTN transfers the solution to the future scene using the egomotion of the car. Learning is done via self-supervised loss (L1).

(c) Future Localization Network (FLN)

Then, FLN learns to predict a heatmap for the future of specific instance. Here the reachability prior from RTN serves as attention mechanism for FLN. Emergence Prediction Network (EPN) is similar to FLN without the object mask.

Transfer the solution to the future →

Narrow down the reachability prior given the conditions of future localization — conditions: (object past observation)

Results

Input (for 1 second) | Future localization (after 3 seconds)

Reachability | FLN-Bayesian [7] | FLN w/o RPN | FLN + RPN

	nuScenes [1]			Waymo [2]			FIT		
	FDE ↓	IOU ↑	NLL ↓	FDE ↓	IOU ↑	NLL ↓	FDE ↓	IOU ↑	NLL ↓
Kalman [3]	45.02	0.31	–	31.69	0.39	–	38.33	0.36	–
DTP [4]	35.88	0.34	–	28.31	0.38	–	34.99	0.37	–
RNN-ED-XOE [5]	30.47	0.34	–	25.23	0.36	–	35.74	0.36	–
STED [6]	27.71	0.39	–	20.73	0.42	–	31.80	0.35	–
FLN-Bayesian using [7]	28.51	0.37	19.75	23.75	0.38	18.80	32.64	0.38	20.56
FLN w/o RPN	15.91	0.54	19.46	13.20	0.54	18.84	18.12	0.53	20.38
FLN + RPN	**12.82**	**0.55**	**17.90**	**10.35**	**0.58**	**16.63**	**15.41**	**0.54**	**19.08**

Input (for 1 second) | Emergence prediction (after 1 second) {object class: **car**}

Reachability | EPN w/o RPN | EPN + RPN

	nuScenes [1]		
	FDE ↓	IOU ↑	NLL ↓
EPN w/o RPN	21.48	0.18	22.99
EPN + RPN	**15.89**	**0.19**	**21.03**

➤ State-of-the-art results on three large and challenging datasets.

➤ Zero-shot transfer to unseen datasets (Waymo) and noisy datasets (Fit).

➤ Exploiting the reachability prior improves the multimodality of the prediction of both FLN and EPN.

➤ Overcoming the mode collapse of the Bayesian baseline, both quantatively and qualitatively.

➤ EPN reduces the solution space of the reachability prior to cover only those areas where an object could emerge.