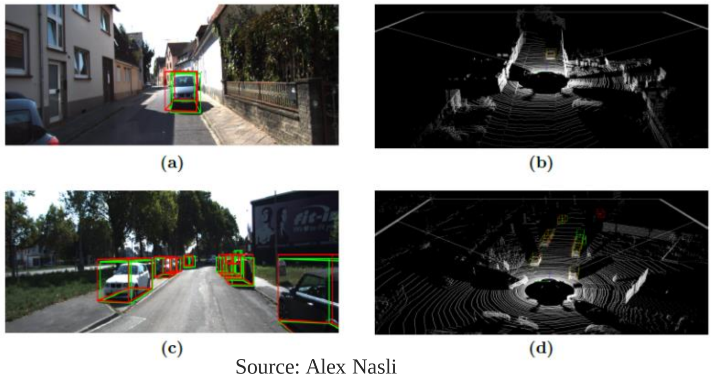
A Cost-Effective Approach to Modeling Object Interactions on the Edge

Arun CS Kumar
Engineer @ NEMO-Ridecell

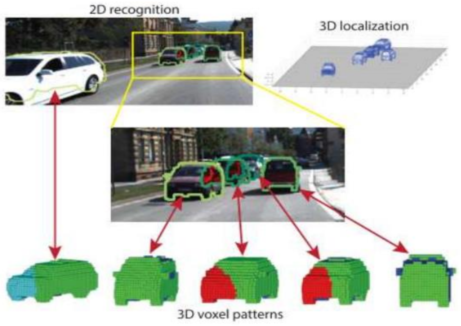
Perception on the Edge - 3D Vision

Common approaches used in 3D perception systems



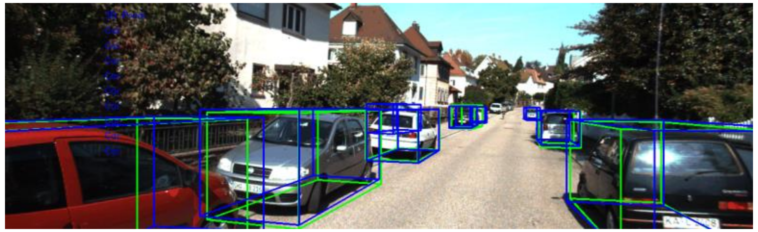
Source: Alex Nasli

Lidar 3D Object Detection



(Figure from Xiang *et al*. 2015)

Camera 3D Object Reconstruction



Source: Harishankar V

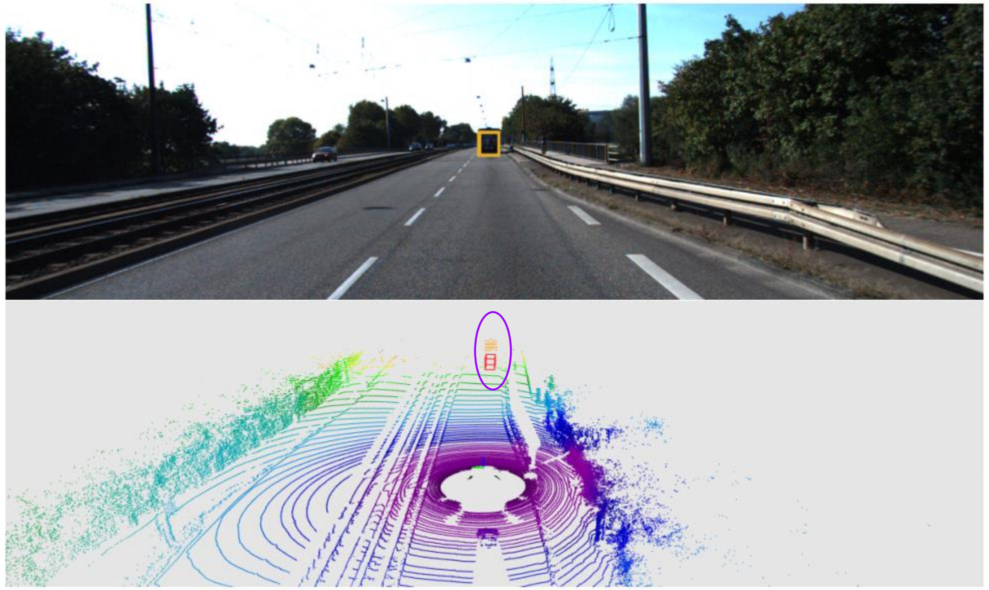
Camera 3D Object Detection

Applications of scalable 3D perception

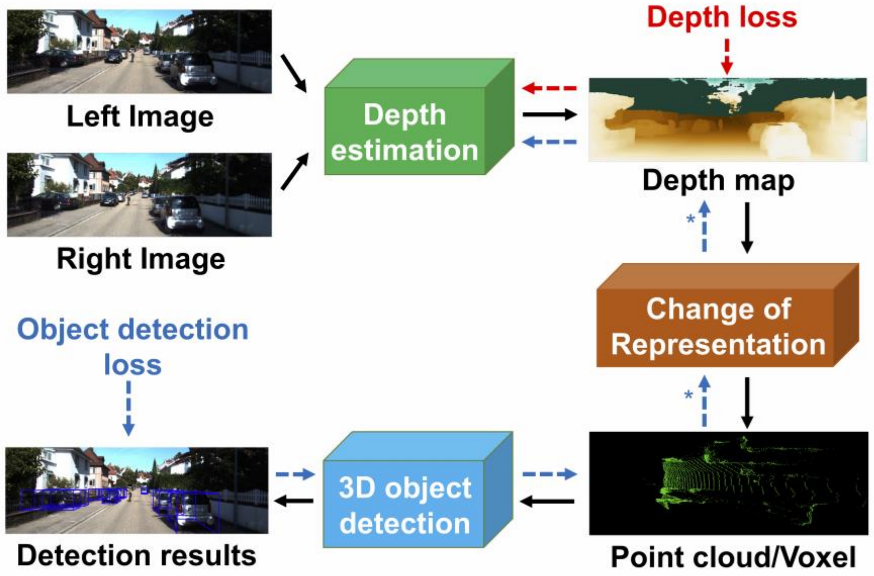
- ➢ Model road object interactions (automotive & auto-insurance industry)
- ➢ Model interactions of human - robot co-working in warehouses
- ➢ Detect, track & model human motions across surveillance systems
- ➢ Query raw data for interactions for offline/off-board (edge) applications

RIDECELL

Shortcomings with the State-of-the-Art



Source: Ma et al 2021 [2]



Source: Qian et al 2020 [3]

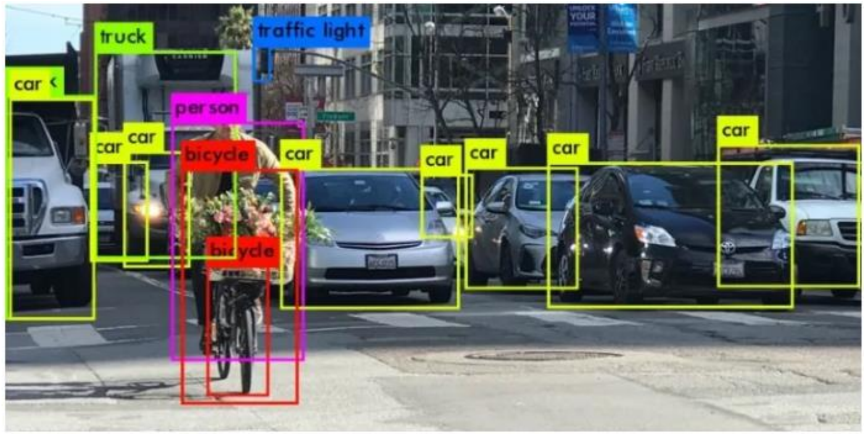
Direct 3D Object Detection/Prediction

- ➢ Object scale / depth prediction error

Two-Stage 3D Object Detection (PseudoLidar)

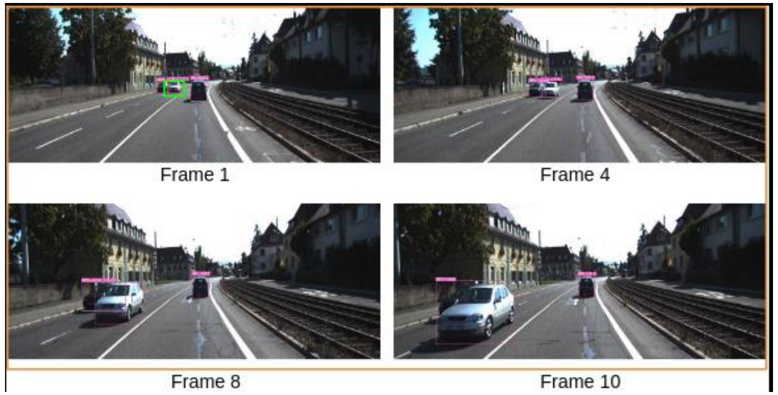
- ➢ Computationally intensive process (large memory footprint)
- ➢ Smaller objects are often missed by depth prediction algorithms

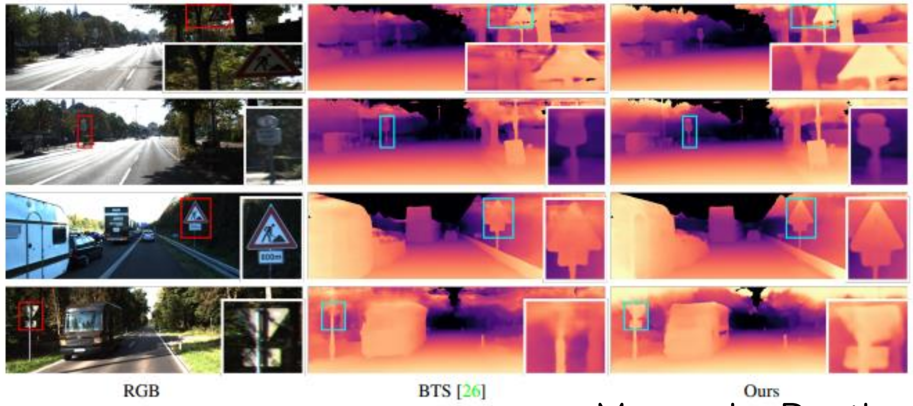
A Novel Object Representation



Source: robocademy

2D Object Detection



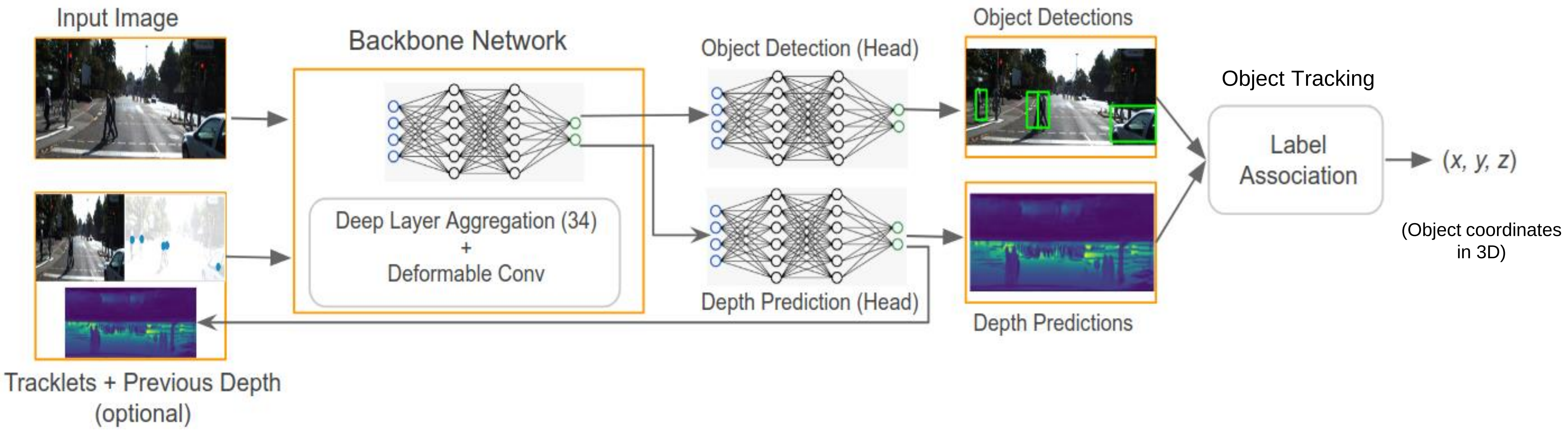


Source: Adabins 2021 [1]

Monocular Depth Prediction

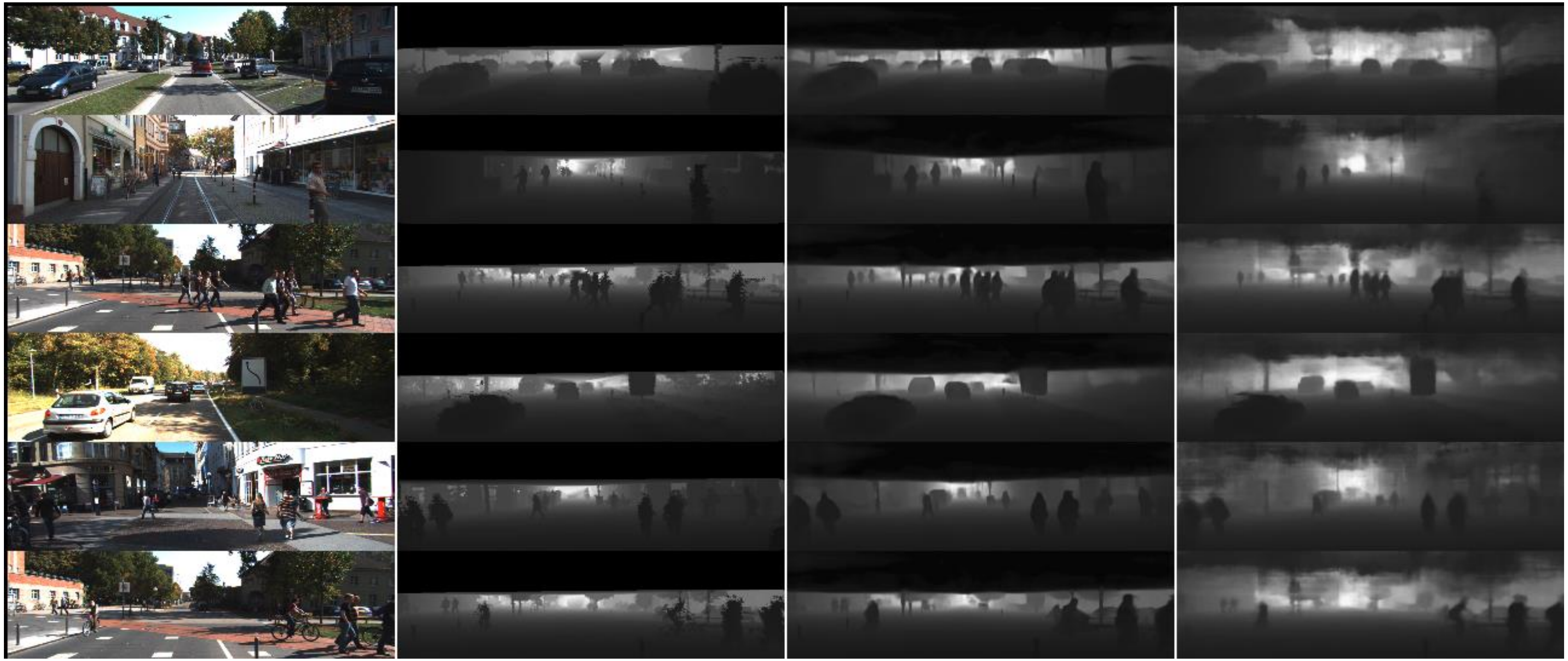
- ➢ Objects as spatio-temporal 3D points in birds eye view
- ➢ 2D object detection + depth prediction = 3D object points
- ➢ Track-able across video frames
- ➢ Computationally inexpensive ~10x faster than [5]
- ➢ No need for laborious & expensive 3D annotations

Proposed Architecture



- ➢ Fu et al, Deep Layer Aggregation 2019 - 34 Layered Convnet with hierarchical aggregation
- ➢ Deformable Convolutional Networks 2017 - Dai et al 2017, deforming convolutions for enhancing transformations
- ➢ Tracklets: Short Track of object in 3D over a small number of frames

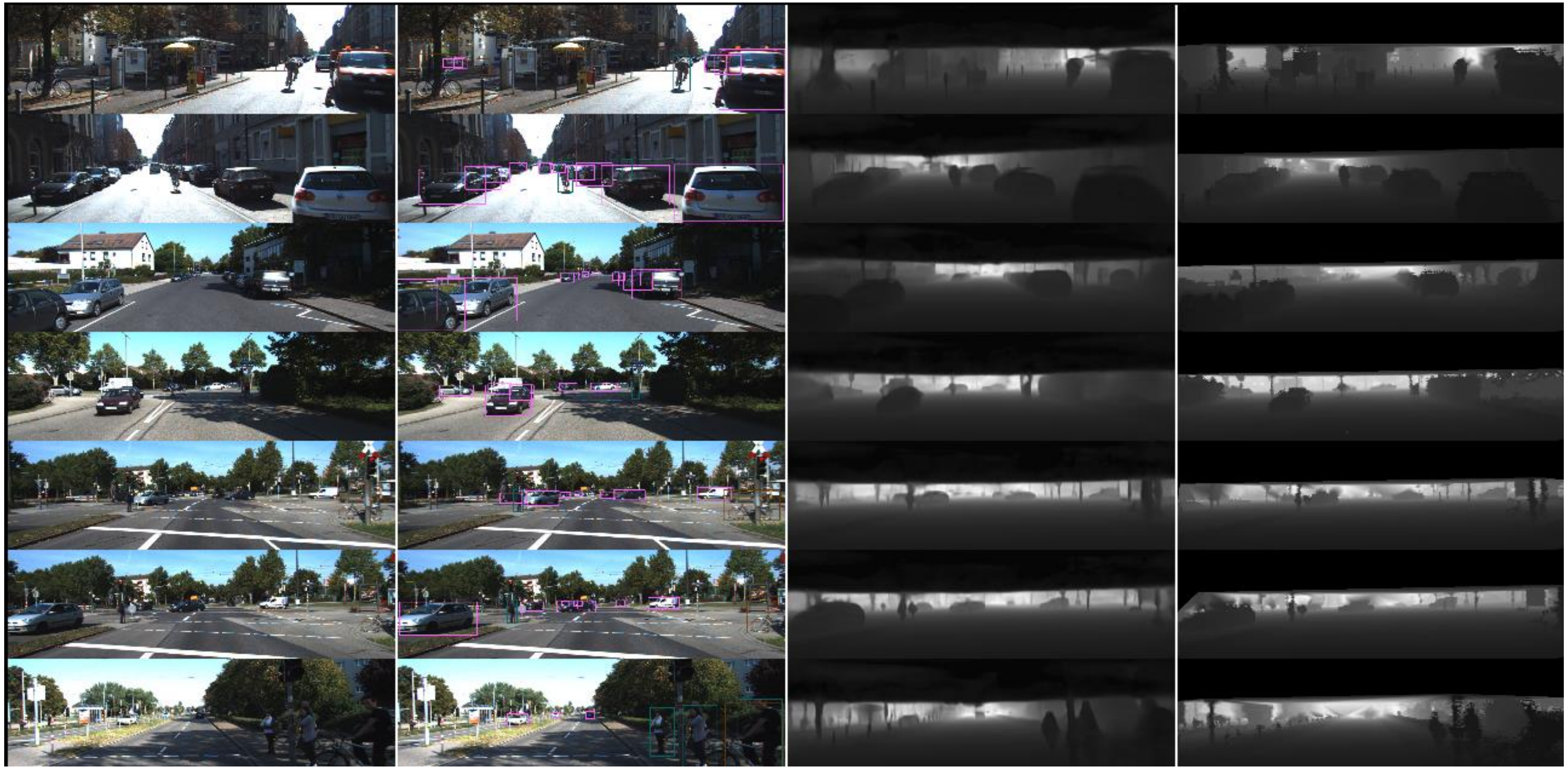
Qualitative Results - Depth Prediction



Input Image | Ground Truth Depths | Our Predictions | DORN [4]

* At inference, monocular depth prediction can be computed using only the input frame and does not need the adjacent frames

Qualitative Results - Overall



Input Image | Our Object Detections | Our Depth Predictions | Ground Truth Depths

RIDECELL

Quantitative Results - Depth Prediction

Current state-of-the-art vs. (Ours)

θ	Supervision			Error Metric				Accuracy Metric		
	Depth	Pose	Unsupervised	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen *et al.* [6] (Coarse)	✓			0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen *et al.* [6] (Fine)	✓			0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu *et al.* [22]	✓			0.202	1.614	6.523	0.275	0.678	0.895	0.965
DORN [8] (50m cap)	✓			0.071	0.268	2.271	0.116	0.936	0.985	0.995
BTS [20]	✓			0.056	0.169	1.925	0.087	0.964	0.994	0.999
Ours*	✓			0.102	0.750	4.137	0.169	0.898	0.967	0.986
Godard *et al.* [12]		✓		0.148	1.344	5.927	0.247	0.803	0.922	0.964
Garg *et al.* [9] (50m cap)		✓		0.169	1.080	5.104	0.273	0.740	0.904	0.962
PackNet-SfM [13] (640 x 192 res.)		✓		0.078	0.420	3.485	0.121	0.931	0.986	0.996
Zhou *et al.* [46](w/o exp. mask)			✓	0.221	2.226	7.527	0.294	0.676	0.885	0.954
Zhou *et al.* [46]			✓	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Zhou *et al.* [46](50m cap)			✓	0.208	1.551	5.452	0.273	0.695	0.900	0.964
Kuznietsov et al. [17]	✓	✓(stereo)		0.113	0.741	4.621	0.189	0.875	0.964	0.988
Kuznietsov et al. [17]		✓(stereo)		0.308	9.367	8.700	0.367	0.752	0.904	0.952

Comparison of Monocular depth prediction results on KITTI dataset

Threshold: % of y_i s.t. $\max(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}) = \delta < thr$

Abs Relative difference: $\frac{1}{|T|}\sum_{y\in T}|y - y^*|/y^*$

Squared Relative difference: $\frac{1}{|T|}\sum_{y\in T}||y - y^*||^2/y^*$

RMSE (linear): $\sqrt{\frac{1}{|T|}\sum_{y\in T}||y_i - y_i^*||^2}$

RMSE (log): $\sqrt{\frac{1}{|T|}\sum_{y\in T}||\log y_i - \log y_i^*||^2}$

RMSE (log, scale-invariant): The error Eqn. 1

RIDECELL

Quantitative Results - Object Tracking

	Time(ms)	MOTA ↑	MOTP ↑	MT ↑	ML ↓	IDSW ↓	FRAG ↓
AB3D	4+D	83.84	85.24	66.92	11.38	9	224
BeyondPixel	300+D	84.24	85.73	73.23	2.77	468	944
3DT	30+D	84.52	85.64	73.38	2.77	377	847
mmMOT	10+D	84.77	85.21	73.23	2.77	284	753
MOTSFusion	440+D	84.83	85.21	3.08	2.77	275	759
MASS	10+D	85.04	85.53	74.31	2.77	301	744
Centertrack	82	89.44	85.05	82.31	2.31	116	334
Centertrack*	**30.47**	81.63	82.96	**85.25**	**2.87**	44	157
Ours*	31.70	**83.54**	**84.67**	79.86	3.59	**27**	**138**

Comparison of Object Tracking results on KITTI dataset (D - Detection time)

Object Tracking Sub-Module

Object Detections (Bounding box 2D) → Label Association Estimation → Tracked Objects (x, y, z)

MOTA: Multi-Object Tracking Accuracy

MOTP: Multi-Object Tracking Precision

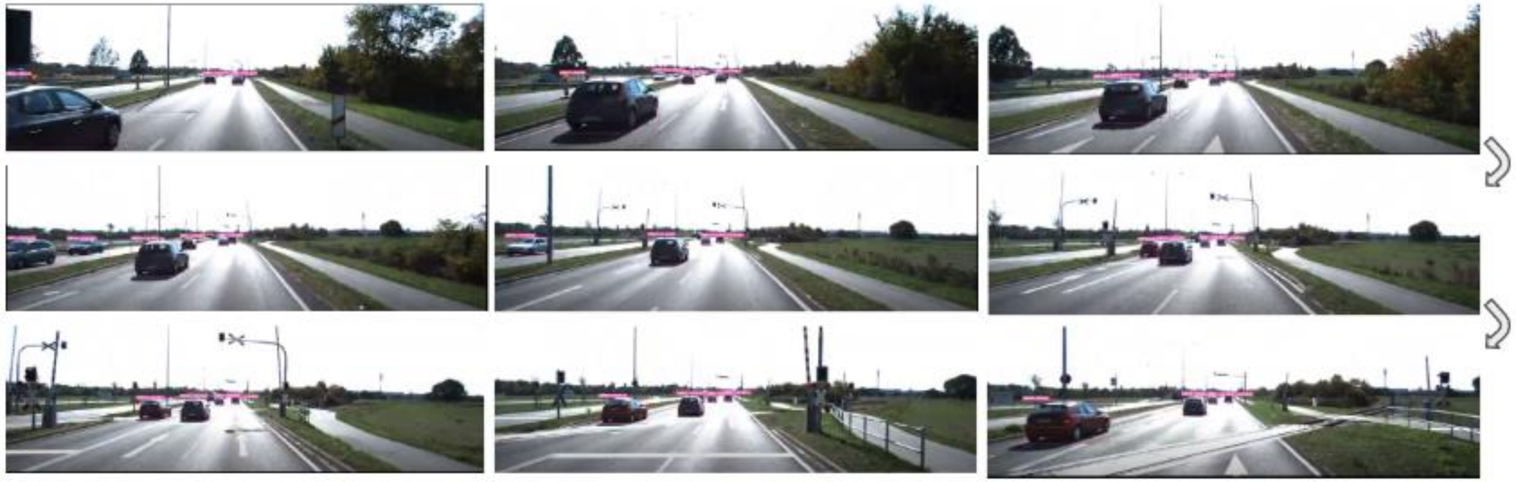
MT: Most Tracked objects ratio (> 80% time)

ML: Most Lost objects ratio (< 20% time)

IDSW: Number of Identity Switch

FRAG: Track Fragmentation

Qualitative Results - Object Interactions



Qualitative representation interactions; right between ego vehicle (blue) and an actor of interest (marked in red) - frames on left

RIDECELL

Conclusion & Future Work

- Novel 3D object representation for off-board applications.
- Objects as spatio-temporal 3D points
- Unified learning framework that is computationally 10x faster than SOTA
- Eliminates need for expensive 3D annotations and data collection setup
- Inexpensive & efficient direction for off-board perception applications
- Efficient for modeling object onteractions in 3D
- Replaceable network components – compatible for edge compute
- Model object interactions end-to-end – future work

Contributors

Paridhi Singh,
Computer Vision Engineer

Gaurav Singh,
Systems Architect & Perception Lead

Arun CS Kumar,
Engineer

References

[1] Bhat, S. F., Alhashim, I., & Wonka, P. (2021). Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4009-4018).

[2] Ma, X., Zhang, Y., Xu, D., Zhou, D., Yi, S., Li, H., & Ouyang, W. (2021). Delving into localization errors for monocular 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4721-4730).

[3] Qian, R., Garg, D., Wang, Y., You, Y., Belongie, S., Hariharan, B., ... & Chao, W. L. (2020). End-to-end pseudo-lidar for image-based 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5881-5890).

[4] Fu, H., Gong, M., Wang, C., Batmanghelich, K., & Tao, D. (2018). Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2002-2011).

[5] Wang, Y., Chao, W. L., Garg, D., Hariharan, B., Campbell, M., & Weinberger, K. Q. (2019). Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8445-8453).