# *BB-Align*: A Lightweight Pose Recovery Framework for Vehicle-to-Vehicle Cooperative Perception

Lixing Song∗, William Valentine∗, Qing Yang†, Honggang Wang‡, Hua Fang△, Ye Liu°

∗ Rose-Hulman Institute of Technology, IN, USA † University of North Texas, TX, USA. ‡ Yeshiva University, NJ, USA △ University of Massachusetts Dartmouth, MA, USA ° Macau University of Science and Technology, Macao, China



© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Background & Motivation

- Traditional autonomous driving systems can be limited by the inherent constraints of single-vehicle perception systems, such as:
  - Short range
  - Occlusions (blocking of the line of sight)
- By integrating **distributed computing** into autonomous driving, **cooperative perception** offers a viable solution to address these limitations

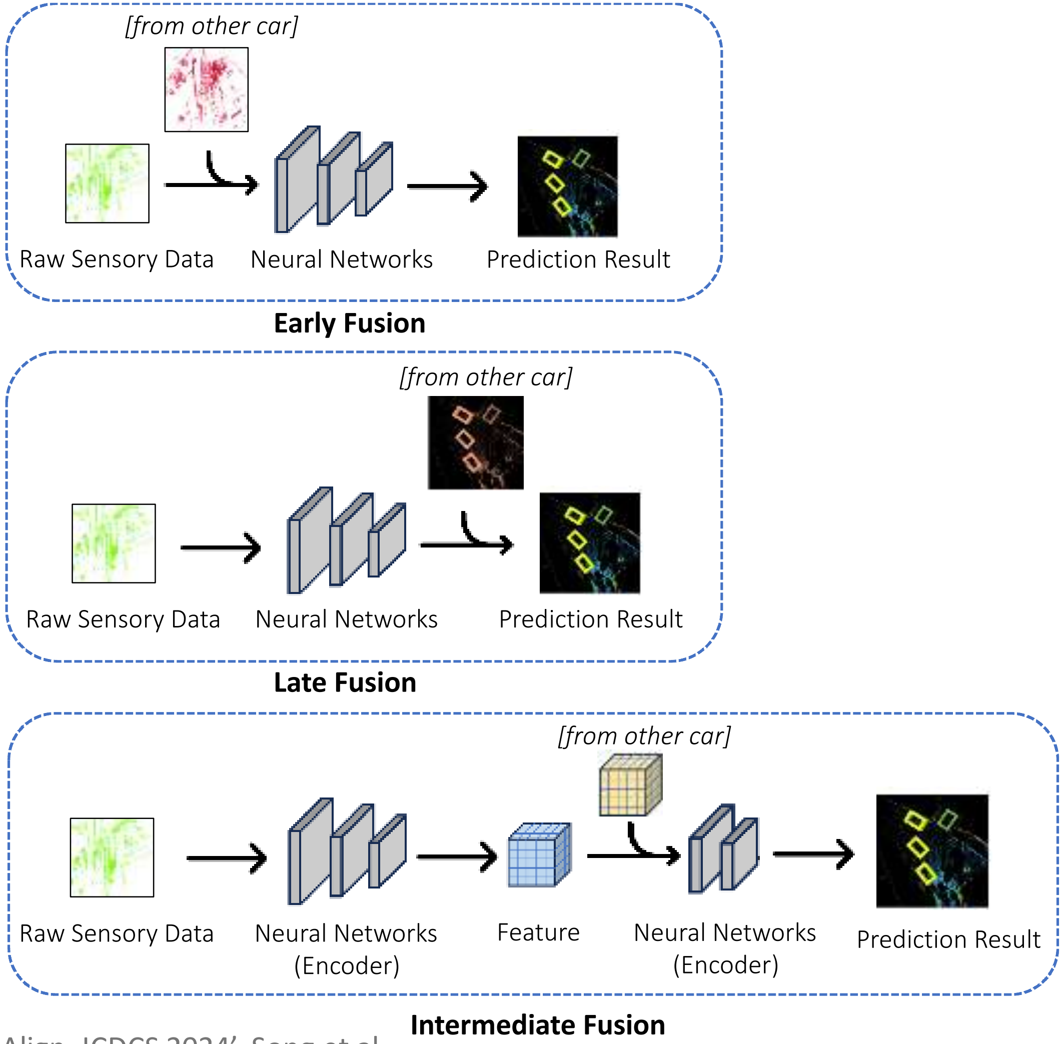


**Credit to Coopernaut (CVPR 2022)*


© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Background & Motivation (cnt.)

- Cooperative Perception Fusion Mechanisms



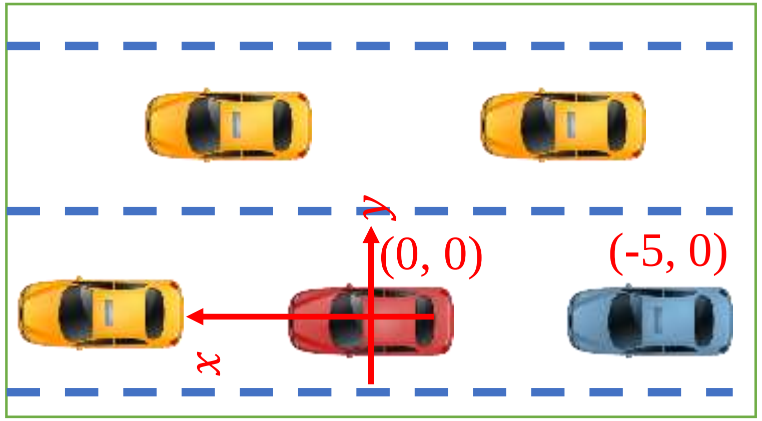


Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

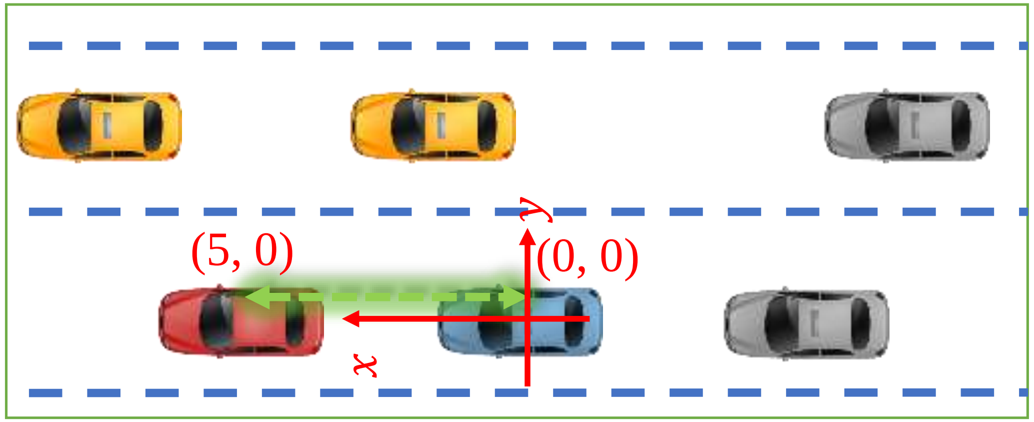
# Background & Motivation (cnt.)

- Fusing shared data from other vehicle(s) requires accurate pose information (location, orientation) to adjust point of view(s).

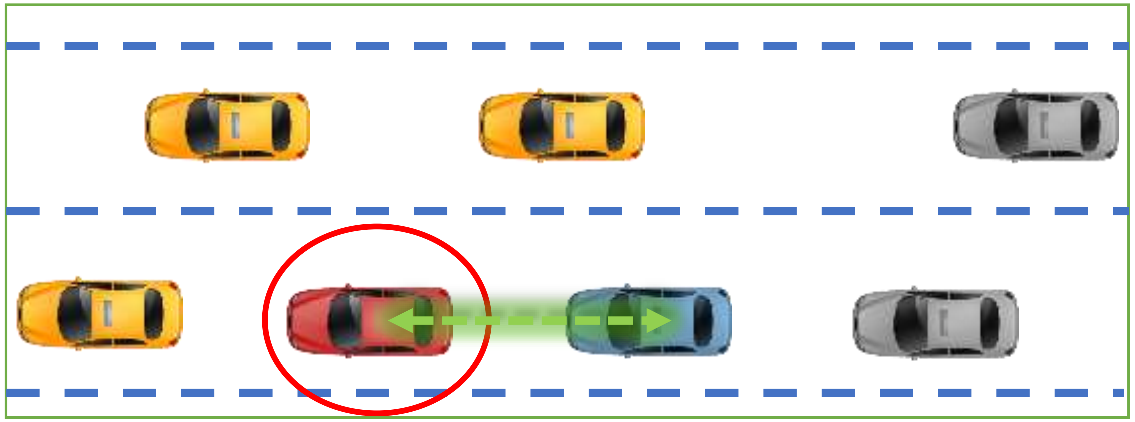
Ego Car's View




Other Car's View



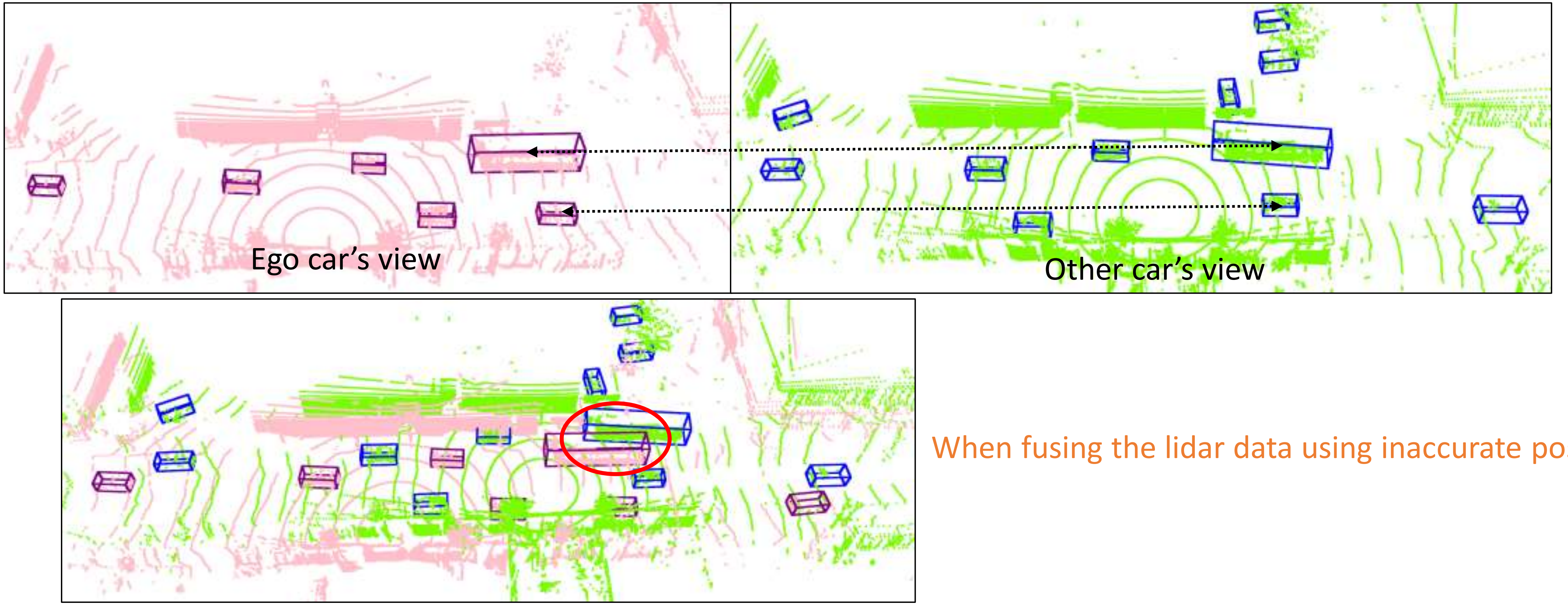

Fused View from Ego Car




*© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.*

# Background & Motivation (cnt.)

- Fusing with corrupted pose information can lead to false detection thus hampering driving policy




When fusing the lidar data using inaccurate pose



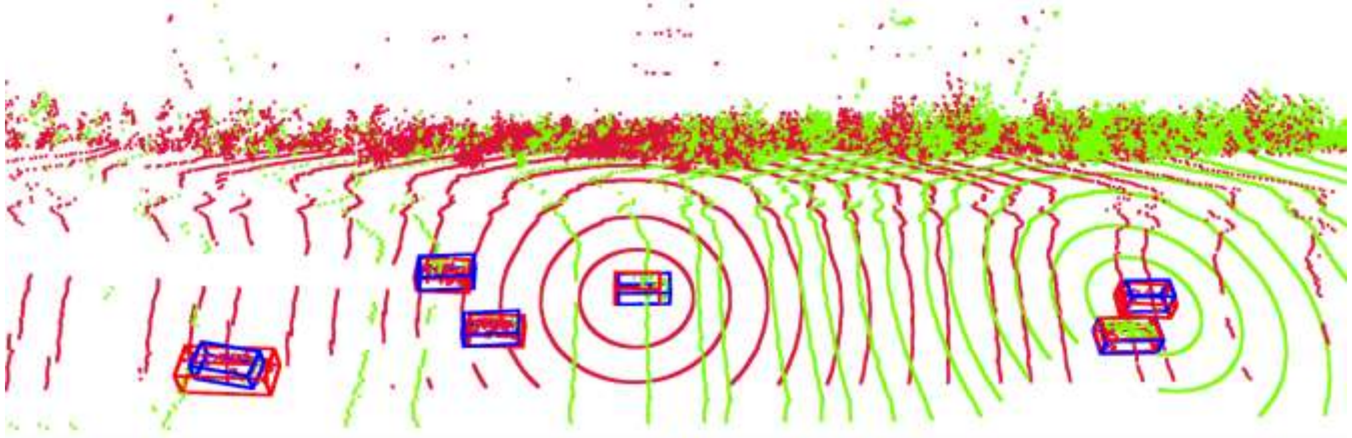
© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Objective

- Input: Two sets of Lidar point clouds captured from two vehicles
- Output: Relative pose, **transformation matrix**, between the two vehicles, i.e., distance and orientation
- Cost: Minimal amount of data shared/transmitted between two vehicles

For Ground Vehicles

$$T = \left( \begin{array}{c|c} R(\boldsymbol{\alpha}, \beta, \gamma) & \begin{array}{c} \mathbf{t_x} \\ \mathbf{t_y} \\ t_z \end{array} \\ \hline \begin{array}{ccc} 0 & 0 & 0 \end{array} & 1 \end{array} \right)$$



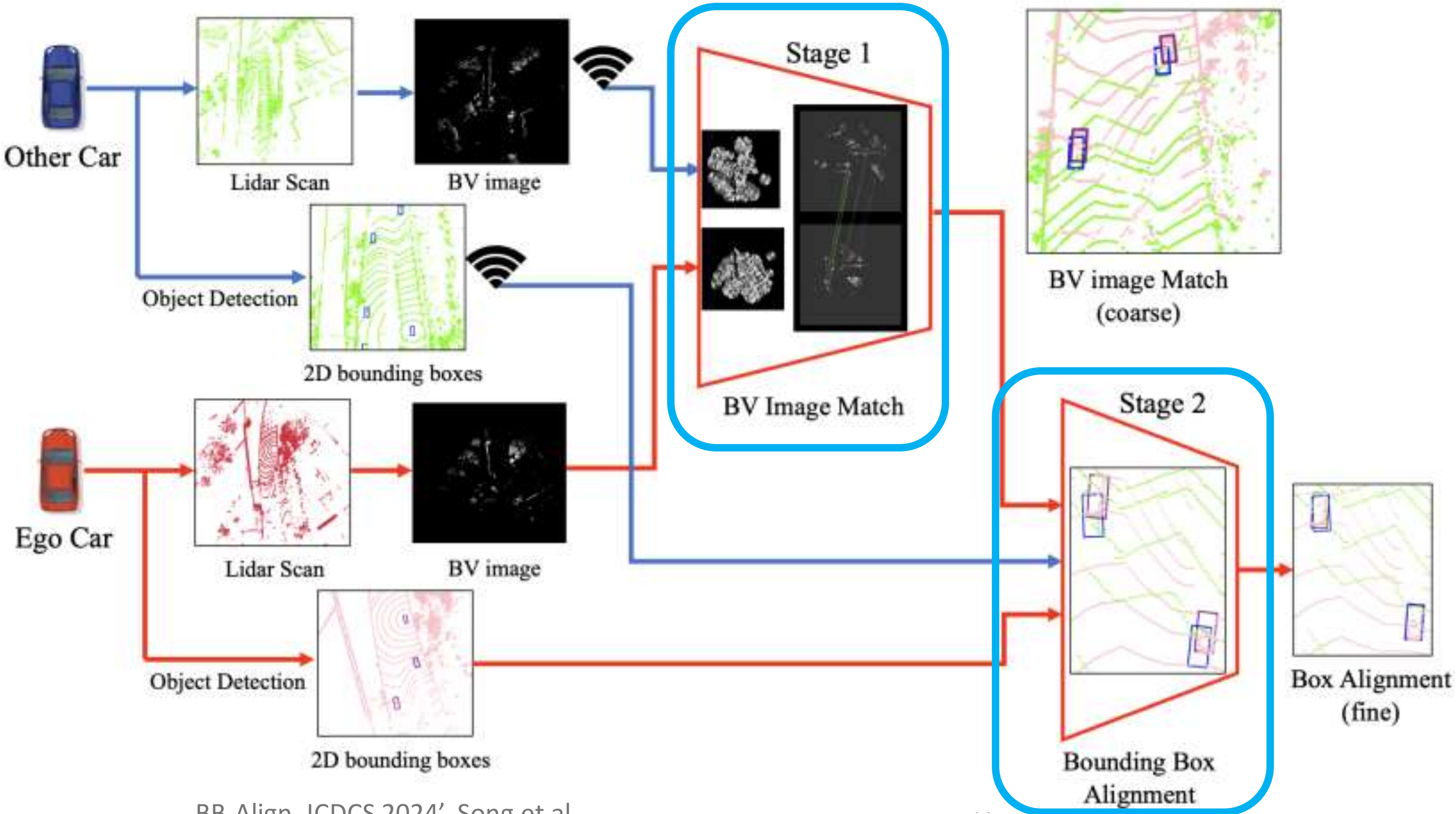
$$\hat{P} = (\underbrace{\boxed{\hat{x}, \hat{y}, \hat{z}}}_{\text{destination}}) = ((\underbrace{\boxed{x, y, z}}_{\text{source}}, 1) \times T^T)[:3]$$



© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Proposed Method (BB-Align)

- A two-staged design:
  - 1) Lidar **B**ird's-eye View (BV) images match
  - 2) Object Bounding **B**oxes alignment

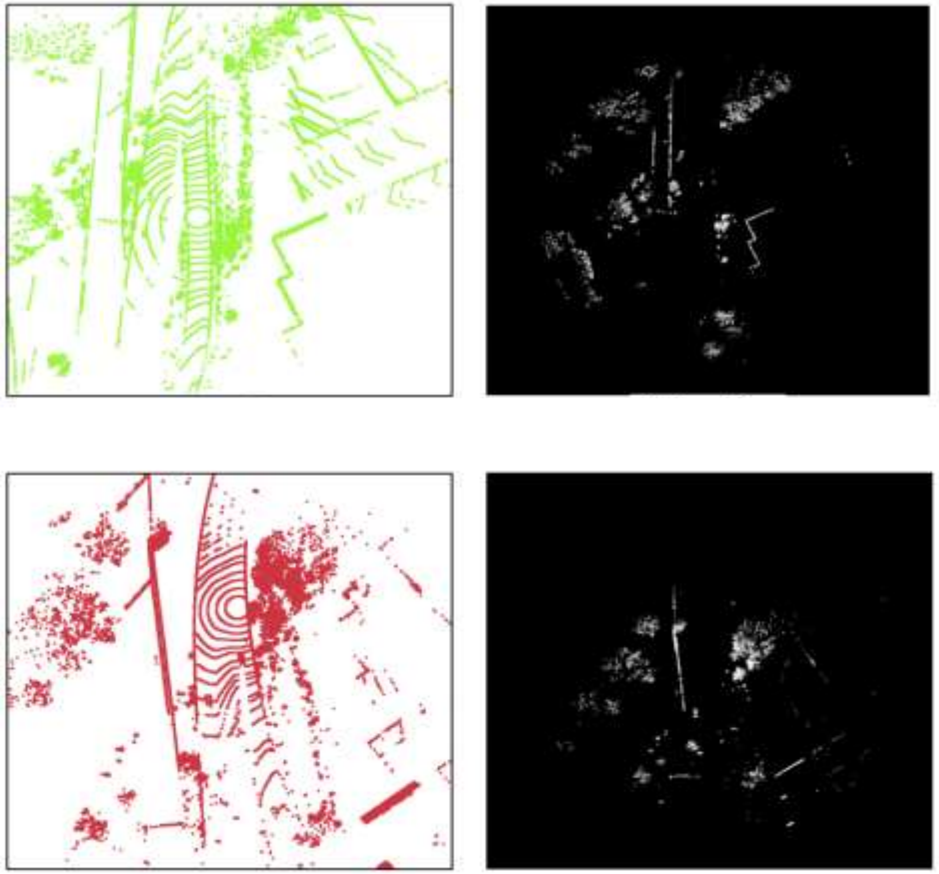




grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Stage 1: BV Image Matching

- Given Lidar point cloud, generate a BV image as a height map
- Apply image matching techniques to find relative pose between two BV images
  - Detecting keypoints (corners, edges)
  - Computing descriptors using surrounding pixels for each keypoint
  - Use paired keypoints to calculate transformation



However, the extreme sparsity of Lidar BV images poses significant challenges, particularly in computing effective descriptors.


© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Stage 1: BV Image Matching (cnt.)

- Log-Gabor filter-based representation

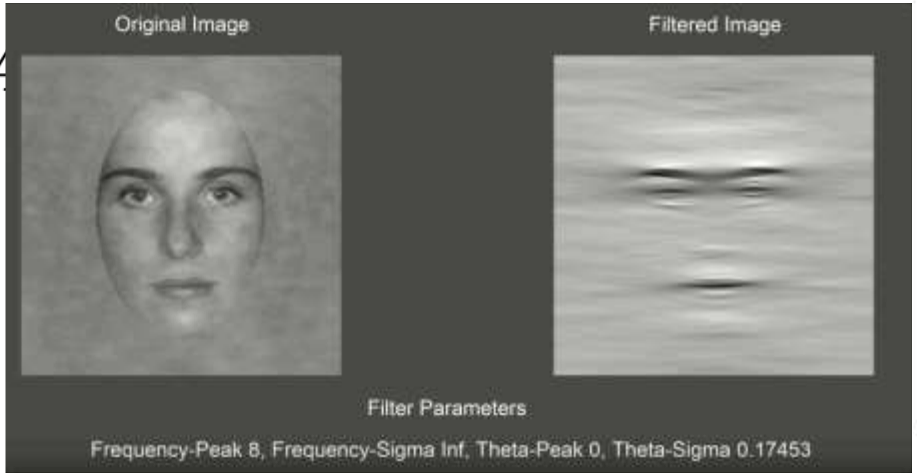
BV image $\mathcal{B} = \{B_{uv} \mid u, v = 1, .., H\}$ $\rightarrow$ $\boxed{\begin{aligned}\rho &= \sqrt{u^2 + v^2}, \\ \theta &= \arctan 2(v, u).\end{aligned}}$ $\rightarrow$ $B_{\rho\theta}$

2-D Log-Gabor filter with parameter $s, o$ :

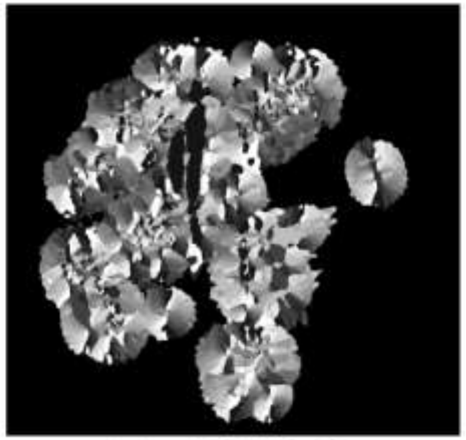
$$L(\rho, \theta, \boxed{s, o}) = \exp\left(-\frac{(\rho - \boxed{R[s]})^2}{2\sigma_\rho^2}\right) \cdot \exp\left(-\frac{(\theta - \boxed{O[o]})^2}{2\sigma_\theta^2}\right)$$

Pass $B_{\rho\theta}$ through a bank of filters :

Generate Maximum Index Map (MIM):




(b) $\mathcal{B}_0$



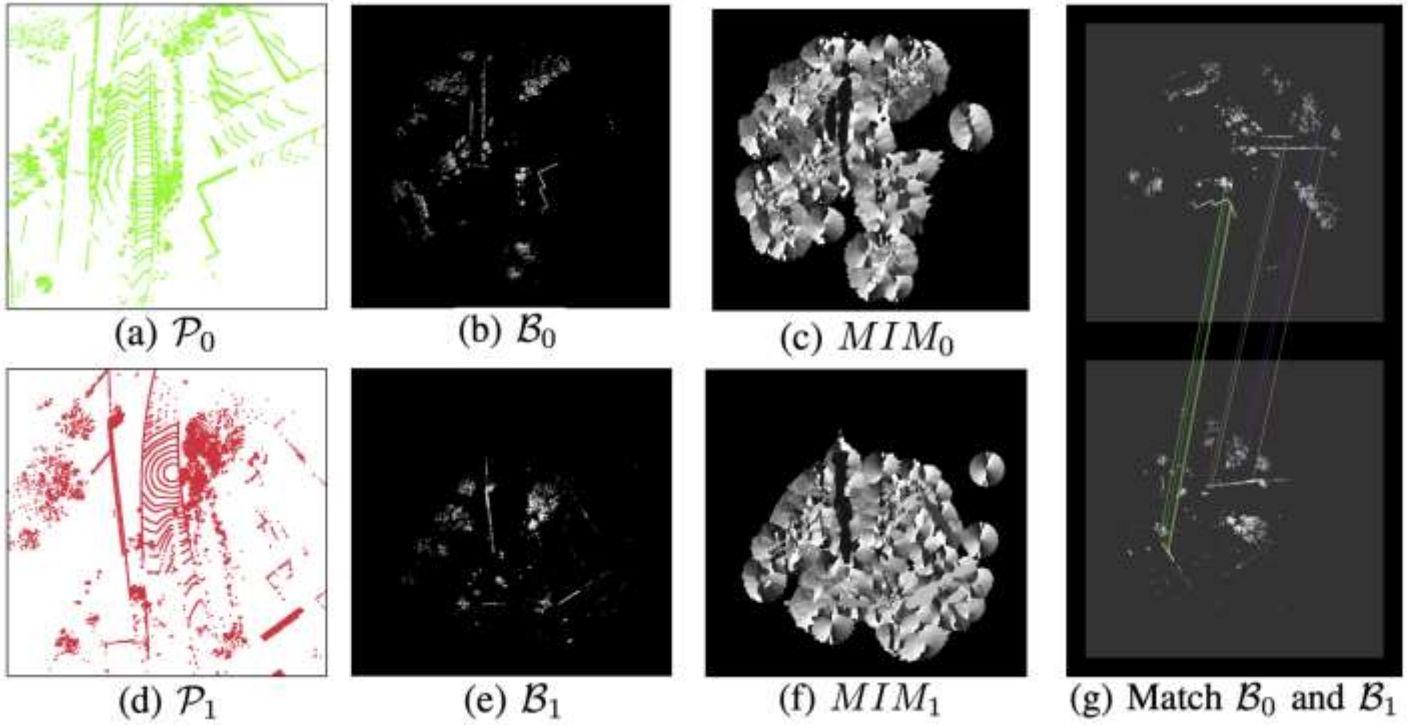
(c) $MIM_0$

Credit to https://peterscarfe.com/logGaborFilter.html


© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Stage 1: BV Image Matching (cnt.)

- Given the feature map MIM, we can computer Bird’s-eye View Feature Transform (BVFT) descriptors [1] for all keypoints (similar to SIFT).
- With the paired keypoints in pairs, we employ the RANdom SAmple Consensus (RANSAC) algorithm to estimate the relative pose between the two images.



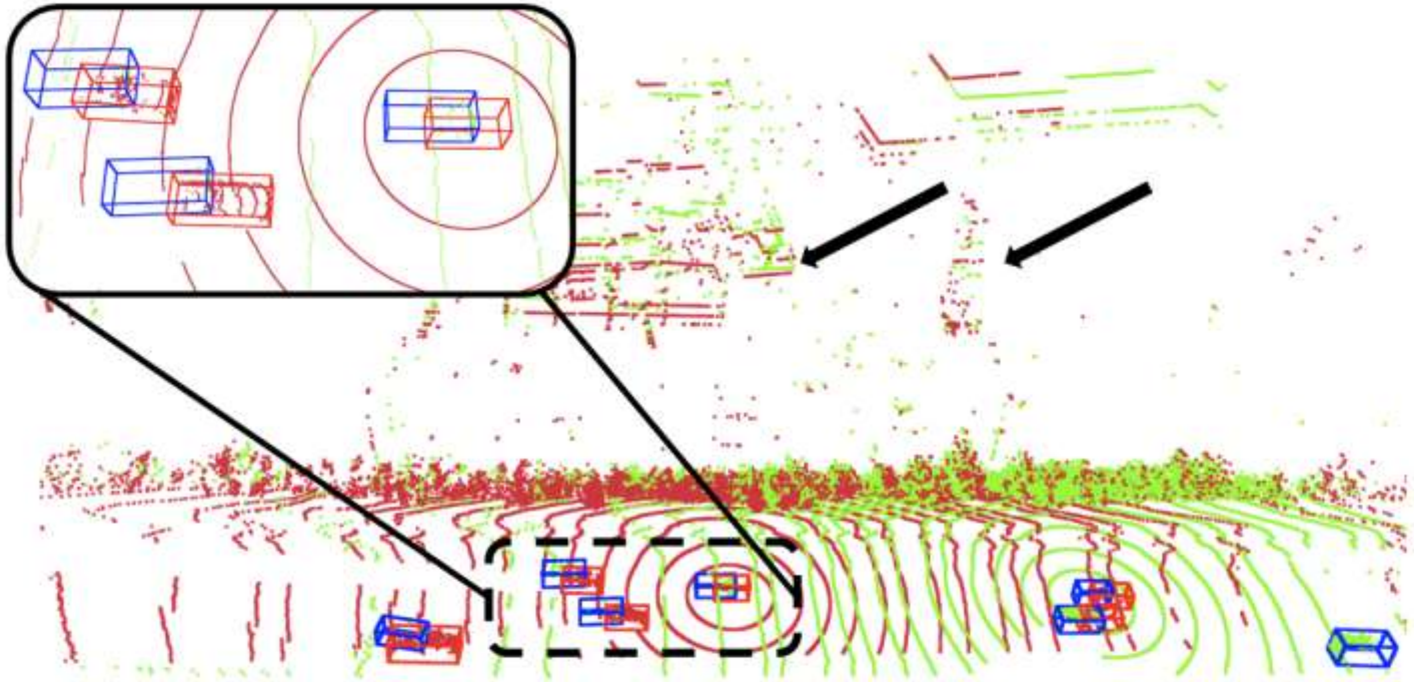

*[1] L. Luo, S. Cao, B. Han, H.-L. Shen, and J. Li, “Bvmatch: Lidar-based place recognition using bird’s-eye view images,” IEEE Robotics and Automation Letters, 2021.*



© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Stage 2: Motivation

- LiDAR self-motion distortion: When the car is moving, each point is not measured at the same location, thus causing distortion.



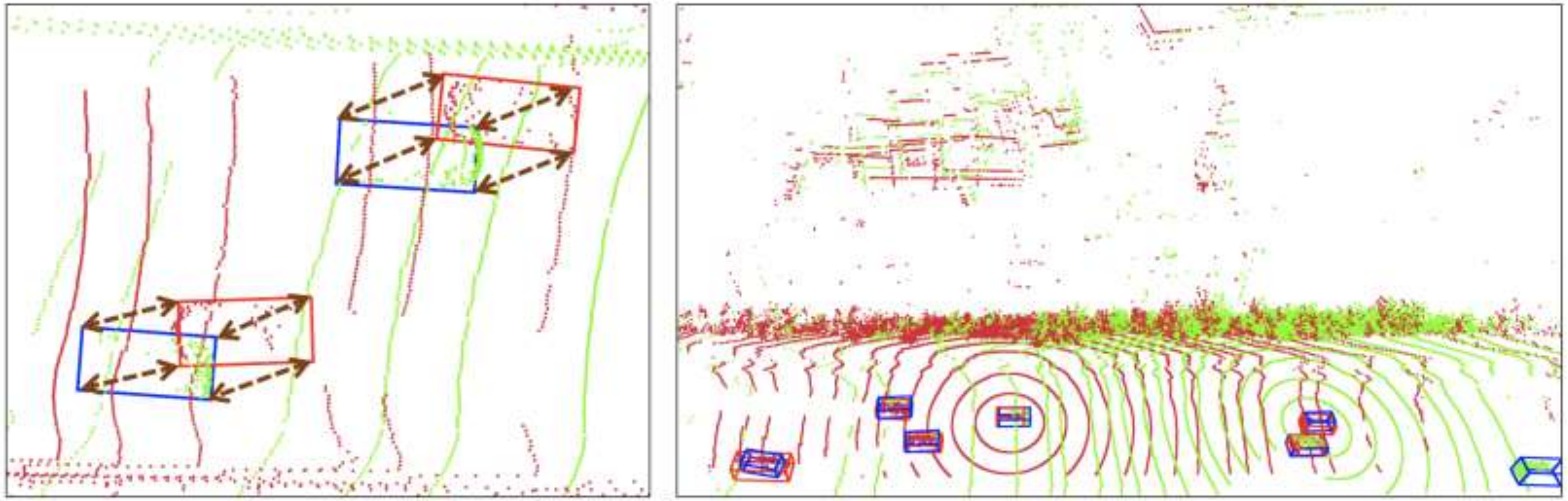
The 3-D bounding boxes, indicated in blue and red, highlight objects (cars) detected by different cars.

Large static landmarks (buildings, trees) are aligned, but the moving objects (vehicles) are not.


© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Stage 2: Object Bounding Box Alignment

- Given the coarsely aligned images, we use the **vertices** of the detected objects (cars) as common observations for further alignment by running RANSAC again.



(a) (b)


© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Performance Evaluation

- Dataset: the only real-world V2V dataset, V2V4Real. We selected 12K frames out of the total 20K, focusing on those where at least **two common cars are observed** by both vehicles
- Model setup:
  - BV image match in written C++ integrated into codebase of V2V4Real.
  - Object detection models: PointPillar-based ***F-Cooper*** and the self-attention-enhanced ***coBEVT***
- Metrics:
  - **Translation Error**: the absolute error of positional shift $t_x, t_y$,
  - **Rotation Error:** the absolute angular difference α.


© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Accuracy Study

- Compared to VIPS[1]: The only other non-training, plug-and-play method, which is based on graph matching.

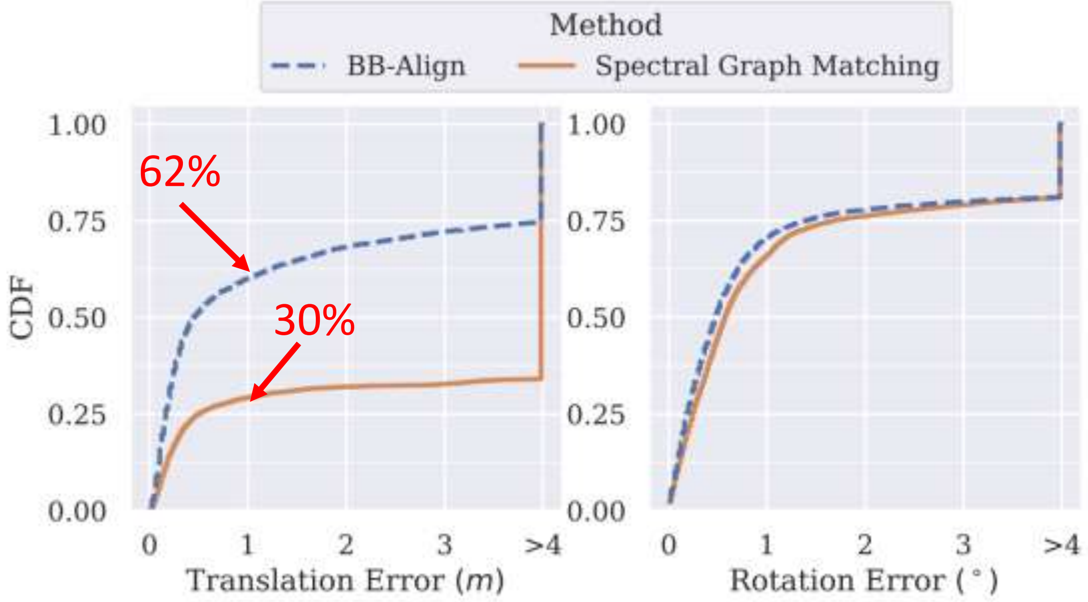



Fig. 7: Pose recovery accuracy comparison.

*S. Shi, J. Cui, Z. Jiang, Z. Yan, G. Xing, J. Niu, and Z. Ouyang, "Vips: real-time perception fusion for infrastructure-assisted autonomous driving," in Proceedings of the 28th Annual International Conference on Mobile Computing And Networking, MobiCom '22.*




© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Performance Impact Factors

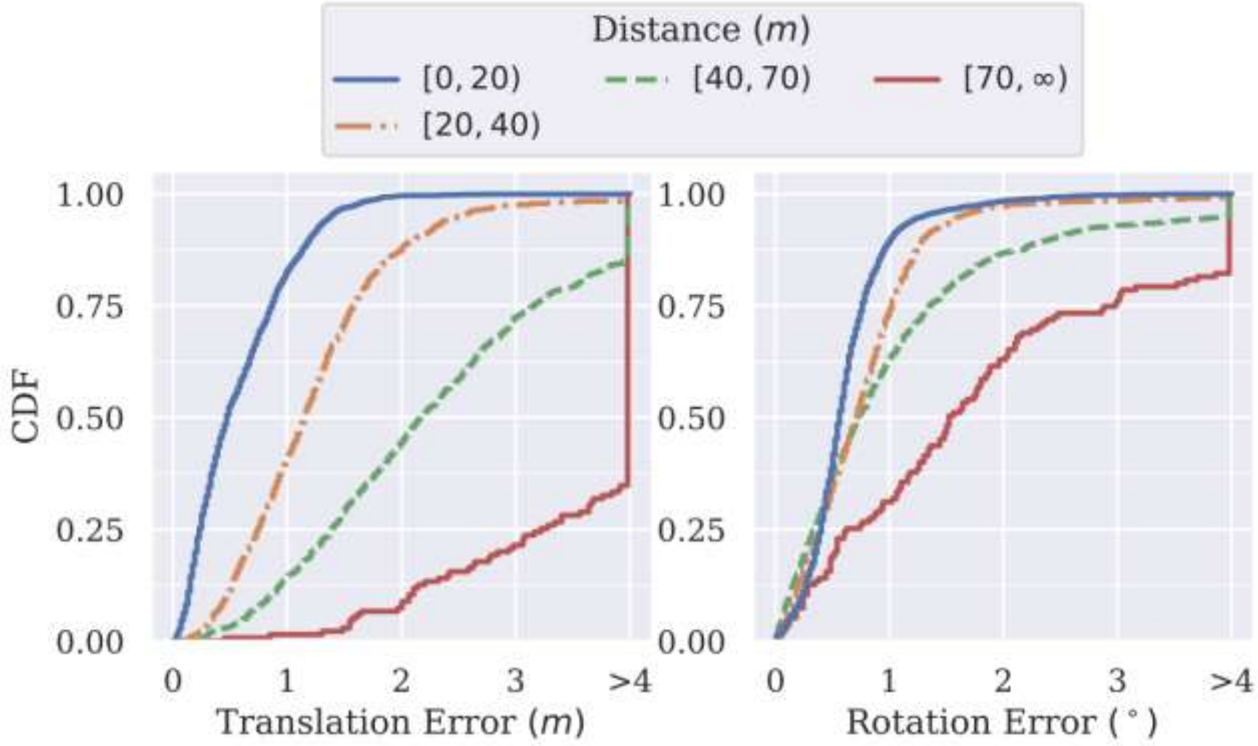



Fig. 11: Accuracy of BV image matching w.r.t. distance ($m$).

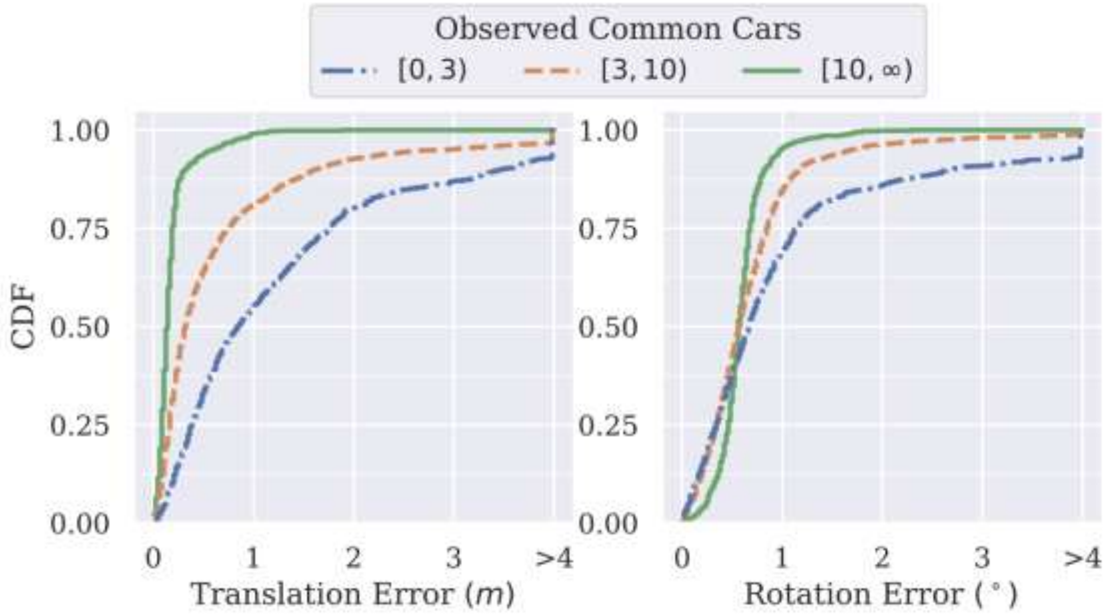



Fig. 12: Accuracy of box alignment (upon BV image matching) w.r.t. the number of commonly observed cars between the two vehicles.

Stage 1 (BV Image Matching) is sensitive to distance. Stage 2(Box Alignment) is largely determined by co-visible cars.


© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Objection Detection Improvement

- We incorporate the proposed method into various fusion techniques and examine the differences compared to not using it.

| Method | AP@IoU=0.5/0.7 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\sigma_t = 2m, \sigma_\theta = 2°$ | | | | **Pose Recovered** | | | |
| | Overall | 0-30m | 30-50m | 50-100m | Overall | 0-30m | 30-50m | 50-100m |
| Early Fusion | 21.2/8.9 | 34.4/14.8 | 19.6/9.9 | 3.5/0.9 | 39.6/18.0 | 67.1/36.5 | 30.5/13.0 | 7.1/1.3 |
| Late Fusion | 18.7/9.3 | 33.1/18.9 | 16.8/7.9 | 2.5/0.6 | 33.9/12.9 | 63.0/28.3 | 27.0/9.2 | 4.7/0.7 |
| F-Cooper | 26.5/14.3 | 43.0/25.0 | 23.5/12.3 | 3.6/1.3 | 40.8/18.1 | 70.6/35.7 | 29.6/11.8 | 7.1/1.1 |
| coBEVT | 31.1/17.8 | 52.6/32.0 | 27.2/15.6 | 4.7/1.9 | 38.9/14.7 | 71.5/29.4 | 28.6/11.4 | 5.2/0.9 |

TABLE I: Comparison of object detection results under corrupted pose, with and without our pose recovery framework.

The improvement is significant in all cases, with nearly a 2x gain in the early fusion case.

Notably, the improvement in the close-range scenarios (0-30m) is even more exciting, with AP@IoU=0.5 scores across all methods exceeding 60.0, and some reaching above 70.0.


© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.

# Summary and Future Work

- We introduce BB-Align, a lightweight, two-stage pose recovery framework tailored for V2V cooperative perception.
- Utilizing Bird's-eye View (BV) images and object bounding boxes, the framework accurately estimates the relative pose between two cars while minimizing communication costs.
- Designed as a non-training-based, plug-and-play module, BB-Align integrates seamlessly with existing V2V systems.
- Future work includes exploring enhancements in time efficiency.

**Questions?**

Lixing Song song3@rose-hulman.edu


© The work is supported by the National Science Foundation grants CNS-2231519, OAC-2017564, and ECCS-2010332.