Learned and Handcrafted Affinities Are Complementary for V2I LiDAR Calibration

Junhyeok Lee, Suhwan Bae, Moonbeom Kim, and Jeongyeup Paek
Department of Computer Science and Engineering, Chung-Ang University, Seoul, Republic of Korea
ljh102715@cau.ac.kr, bbiddak99@cau.ac.kr, mbkim@cau.ac.kr, jpaek@cau.ac.kr

***Abstract*—Vehicle-to-infrastructure (V2I) cooperative perception is critical for safe autonomous driving because it extends sensing beyond the vehicle's field of view. Object-level V2I LiDAR calibration aligns vehicle and infrastructure observations by estimating their relative pose from correspondences between detected objects. These correspondences can be scored using either handcrafted geometric cues or affinity learned by a graph neural network from object attributes and relational geometry. However, comparing the two as substitutes or complements is difficult because calibration accuracy also depends on downstream match selection and pose estimation. We therefore fix these stages to isolate the affinity stage and compare handcrafted-only, learned-only, and hybrid affinities. On DAIR-V2X-C, hybrid affinities outperform handcrafted-only and learned-only affinities by 9.1 and 6.8 percentage points, respectively, reaching 69.5±0.5% success under 2 m and 2° error thresholds.**

***Index Terms*—Vehicle-to-infrastructure, LiDAR extrinsic calibration, Cooperative perception, Graph neural networks**

I. Introduction

In vehicle-to-infrastructure (V2I) cooperative perception, sharing detections from roadside infrastructure extends a vehicle's observable scene, but accurate fusion requires detections from the vehicle and infrastructure LiDARs to be expressed in a common coordinate frame. As illustrated in Fig. 1, object-level V2I calibration uses two sets of 3D detection boxes from the same traffic scene [1], one in the vehicle LiDAR frame and the other in the infrastructure LiDAR frame, and matches boxes that correspond to the same physical objects. This process consists of three stages: affinities are computed for cross-frame box pairs, correspondences between common objects are selected based on these affinities, and the relative sensor pose is estimated from the selected correspondences.

The first stage, affinity scoring, assigns a compatibility score to each cross-frame box pair, indicating how likely the two boxes are to represent the same physical object. Representative methods such as V2I-Calib [2] and V2X-Reg++ [3] compute these scores using predefined geometric cues, which we refer to as handcrafted affinity. In particular, overall-IoU (oIoU) evaluates scene-level consistency based on the overlap between aligned 3D boxes, whereas overall-distance (oDist) evaluates it based on the distances between aligned boxes. In contrast, recent matching methods use learned models to estimate compatibility scores, termed learned affinity. These methods represent objects and their relationships as a graph and use graph neural networks to estimate affinity from object attributes and relational context [4], [5].

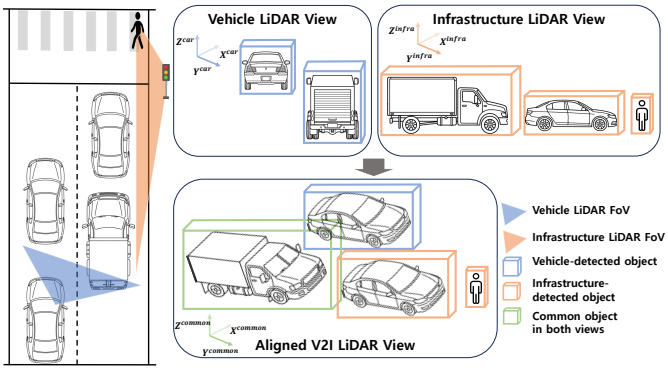

Fig. 1: Overview of object-level V2I LiDAR calibration, where corresponding vehicle- and infrastructure-side detections are matched and aligned in a common coordinate frame

This raises a central question for V2I calibration: whether learned affinity should replace handcrafted geometric affinity or whether the two provide complementary information. However, final calibration accuracy is determined not only by affinity scoring but also by the subsequent stages that select object correspondences and estimate the relative pose. We therefore isolate the affinity stage by fixing these downstream stages and varying only the affinity design, allowing a controlled comparison of handcrafted, learned, and hybrid affinities.

Our contributions are summarized as follows:

- Modularize object-level V2I LiDAR calibration into three stages: affinity scoring, match selection, and pose estimation, providing a unified analysis pipeline.
- Isolate affinity design by fixing match selection and pose estimation, and compare handcrafted, learned, and hybrid affinities with only the affinity varied.
- Demonstrate that learned and handcrafted affinities are complementary, with hybrid affinities outperforming both learned-only and handcrafted-only affinities and improving calibration accuracy and robustness.

This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the National Program for Excellence in SW), supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation) in 2026 (2025-0-00032), by IITP-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2026-RS-2022-00156353, 33%), and also by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00359450). Jeongyeup Paek is the corresponding author.

II. Related Work

Here, we review prior work on affinity design and downstream processing in object-level V2I LiDAR calibration.

Handcrafted affinity in object-level calibration. V2I-Calib [2] scores candidate 3D-box pairs using handcrafted oIoU affinity, which measures geometric consistency based on box overlap. Its journal extension, V2X-Reg++ [3], adopts oDist affinity, which instead measures consistency based on inter-box distance. We consider these representative targetless, initialization-free V2I LiDAR calibration methods based on handcrafted geometric affinity. V2V-origin methods such as CoAlign [6] require a coarse initial pose or different geometry and are therefore less direct comparators for targetless V2I calibration.

Learned affinity in graph-based matching. SuperGlue [4] popularized GNN-based matching with Sinkhorn assignment, while FreeAlign [5] applies object-graph matching to collaborative perception. These methods show that affinity can be learned from object attributes and relational structure rather than defined solely by handcrafted geometric cues. We use a standard SE(2)-invariant graph matcher to instantiate learned affinity, focusing on comparison with handcrafted affinity rather than on a new matching architecture.

Match selection and pose estimation. Affinity scores are converted into object correspondences before estimating the relative pose. Existing pipelines use assignment methods such as Hungarian or Sinkhorn-based matching [2], [4], followed by SVD-based or robust pose estimation [2], [3], [7]. Because these downstream choices also affect calibration accuracy, we control them throughout our evaluation to isolate the effect of affinity design.

III. Method

This section details the object-level V2I calibration pipeline, including the affinity formulations and the common downstream backend used for controlled comparison.

Problem and decomposition. Given 3D detection boxes obtained independently from the vehicle-side and infrastructure-side LiDAR observations, we consider two object sets captured in different coordinate frames, with potentially partial overlap between them. To represent the two object sets in a common coordinate frame, object-level V2I calibration estimates the extrinsic transform $(R, t) \in SE(3)$ through a calibration pipeline without assuming a given initial pose. We formulate this calibration pipeline as three stages and write it as

$$\Pi = (A, M, P),$$

where an affinity function A scores candidate vehicle-infrastructure box pairs, a match-selection module M extracts a partial one-to-one assignment, and a pose estimator P recovers (R, t) from matched box centers. V2I-Calib's native pipeline uses $(\mathrm{oIoU}, \mathrm{Hungarian}, \mathrm{weighted\ SVD})$, while V2X-Reg++'s published pipeline uses an oDist affinity. We reproduce these native pipelines under our held-out protocol as reference points, not as controlled affinity ablations. In the controlled comparison, we reuse only their affinity definitions; all handcrafted-only, learned-only, and hybrid variants use the same M and P, as illustrated in Fig. 2.

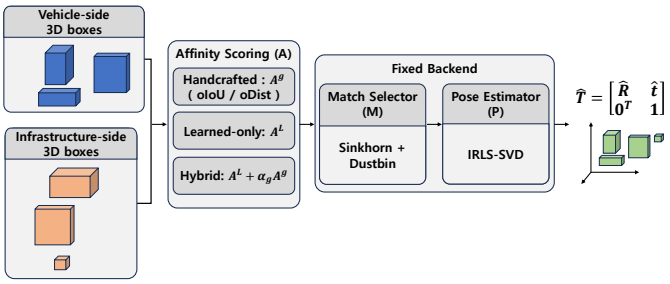

Fig. 2: Overview of the controlled V2I calibration pipeline.

Handcrafted and learned affinities. For the handcrafted affinities, we use two representative geometric cues from prior object-level V2I calibration methods: V2I-Calib's oIoU [2] and V2X-Reg++'s oDist [3]. Both are transform-conditioned: each candidate box pair forms a pose hypothesis, and the score measures how well that pose aligns the two object sets. oIoU uses transformed box overlap, while oDist uses an affinity based on center- and vertex-distance consistency. By contrast, the learned affinity is transform-free: it scores pairs from relational context before any pose is known. For the learned affinity, we build separate graphs for the vehicle-side and infrastructure-side object sets. Node features are box dimensions $[l, w, h]$ computed from oriented box corners plus a category embedding. Edge features describe SE(2)-invariant relative geometry: distance, bearing in the source box heading frame, and relative yaw. A two-layer GNN produces embeddings $h_i^{\mathrm{veh}}, h_j^{\mathrm{inf}} \in \mathbb{R}^H$, where H is the embedding dimension. The learned affinity is a scaled dot product between the two box embeddings,

$$A_{ij}^{\mathrm{L}} = \langle h_i^{\mathrm{veh}}, h_j^{\mathrm{inf}} \rangle / \sqrt{H},$$

so boxes with similar relational embeddings receive larger scores. Invalid cross-category pairs are masked. During training, the learned-only and hybrid variants use the same Sinkhorn-assignment loss with ground-truth correspondences.

Hybrid affinity. Handcrafted scores are reliable when a candidate transform makes the two object sets geometrically consistent, but can be ambiguous when few objects are shared. The learned affinity can disambiguate matches using relational context, but it does not directly measure the transform-conditioned geometric consistency captured by handcrafted affinities. Motivated by these different characteristics, we define the hybrid affinity as

$$A_{ij}^{\mathrm{hyb}(g)} = A_{ij}^{\mathrm{L}} + \alpha_g \, \widetilde{g}_{ij},$$

where $g \in \{\mathrm{oIoU}, \mathrm{oDist}\}$. Before normalization, both oIoU and oDist are represented as larger-is-better geometric affinity scores. For each frame, $\widetilde{g}$ is obtained by dividing the category-valid geometric-score matrix by its positive maximum; all-zero

matrices are left unchanged. The handcrafted-only variants use the same normalized geometric cue without the learned dot-product term. The scalar α_g is cue-specific and learned jointly with the GNN on training data; no held-out test frames are used to fit or tune this weight. The learned-only baseline uses no $\widetilde{g}$ term, so the hybrid comparison evaluates whether adding normalized handcrafted geometry helps under the same backend.

Fixed match selection and pose estimation. In the controlled comparison, all variants share the same match-selection module M and pose estimator P. Match selection uses Sinkhorn with a dustbin for unmatched boxes [4]. We decode the non-dustbin assignment block by keeping mutual row and column maxima whose assignment probability exceeds a threshold τ. The threshold τ is selected on validation frames for each affinity setting and then fixed for held-out evaluation. Pose estimation uses robust IRLS-SVD on matched box centers. With solver settings fixed across all controlled variants, IRLS-SVD iteratively computes residuals, removes high-residual matches, and re-solves weighted Procrustes. Thus, within the controlled comparison, calibration differences are not due to using different match-selection or pose-estimation methods.

Evaluation beyond match F1. Match F1 does not fully explain calibration accuracy. Rigid pose estimation can be strongly affected by a few incorrect correspondences. In Procrustes-based pose estimation, a wrong match far from the center of the matched object set can bias the estimated pose and create a large yaw error. Therefore, two affinities with similar F1 can have different calibration results if one of them keeps such outlier matches. For this reason, we report final pose success, not only match F1, and analyze large-error failures in §IV.

IV. Experiments

In this section, we first describe the evaluation setup, and then present the experimental results and analysis.

A. Experimental Setup

We evaluate on DAIR-V2X-C [8], a real-world cooperative perception dataset with synchronized vehicle- and infrastructure-side 3D detection boxes and ground-truth extrinsics. The learned affinity model is trained on frames $[0, 1500)$, validated on $[1500, 2000)$, and tested on a held-out split of 500 frames, $[2000, 2500)$. Following V2I-Calib, we retain the top $K{=}15$ boxes from each side. The primary metric is Success@2m,2°: a frame is successful only if both translation and rotation errors are below their respective thresholds. We use all test frames as the denominator, counting frames with no accepted matches as calibration failures. We also report matching F1 against ground-truth box correspondences to distinguish matching quality from final pose accuracy. We follow the EASY/HARD partition of DAIR-V2X-C used by V2I-Calib [2], where HARD frames generally contain fewer shared objects and lower cross-view object overlap than EASY frames. For learned and hybrid variants, results are averaged over multiple random seeds where available; confidence intervals are computed by bootstrapping the same held-out test frames.

TABLE I: Results on 500 held-out DAIR-V2X-C frames.

Setting	Affinity	ALL (%)	HARD (%)	F1 (%)
Native	V2I-Calib/oIoU [2]	53.6	39.4	76.0
	V2X-Reg++/oDist [3]	59.2	46.8	–
Fixed Backend	oIoU-only	60.0	46.1	–
	oDist-only	60.4±0.1	46.4±0.2	78.4±0.5
	Learned-only	62.7±0.8	50.4	82.5
	Learned+oIoU	66.8±0.6	54.1	90.1
	Learned+oDist	**69.5±0.5**	**57.2±0.7**	**91.5±0.3**

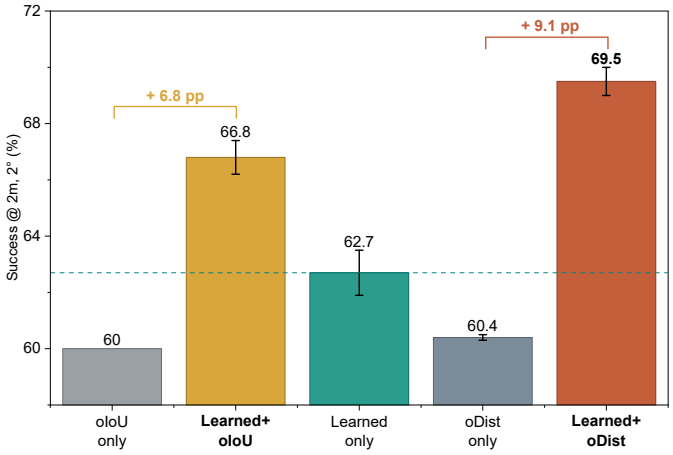

Fig. 3: Fixed-backend affinity comparison. Adding either oIoU or oDist improves the learned affinity, with learned+oDist achieving the highest calibration success.

B. Results

Main comparison. Table I reports DAIR-V2X-C held-out results under the all-frame Success@2m,2° metric. The native rows reproduce the published pipelines under our held-out protocol, while the fixed-backend rows share the same Sinkhorn match-selection module and IRLS-SVD pose estimator across affinity variants. Under this fixed backend, both hybrid variants achieve higher calibration success than their learned-only and corresponding handcrafted-only components. Learned+oIoU improves over learned-only and oIoU-only by 4.1 pp and 6.8 pp, respectively, while learned+oDist improves over learned-only and oDist-only by 6.8 pp and 9.1 pp, respectively. The strongest variant, learned+oDist, also exceeds the native V2X-Reg++/oDist reference by 10.3 pp. Together, these results provide direct evidence of complementarity between the two affinity families. Because the downstream modules are held constant, the consistent gains from combining the learned affinity with either oIoU or oDist indicate that learned relational context and transform-conditioned geometry contribute complementary information to calibration. Fig. 3 visualizes this comparison. We therefore treat the native V2X-Reg++/oDist result as an end-to-end reference and use fixed-

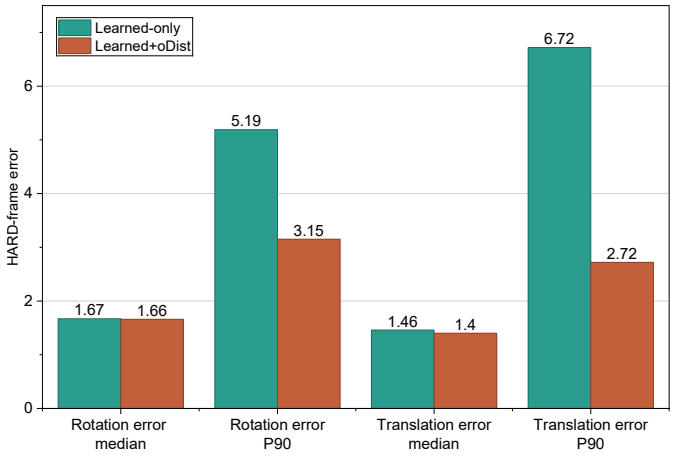

Fig. 4: For the strongest hybrid, learned+oDist, median HARD-frame errors remain similar to learned-only, while the high-error tail is reduced.

backend oDist-only as the direct handcrafted baseline for the 9.1 pp improvement of learned+oDist. Under the same fixed backend, oDist-only achieves 60.4% success, compared with 59.2% for the native V2X-Reg++/oDist pipeline, indicating that the controlled comparison does not disadvantage the geometric component.

Match coverage and quality. The learned-based variants achieve broader match coverage than the native handcrafted baselines. The native handcrafted baselines fail to produce selected matches on 16–23% of frames, whereas learned-only and learned+oDist produce matches on approximately 96.6% and 97.4% of frames, respectively. The F1 results further show that the hybrid variants improve correspondence quality, with learned+oDist achieving the highest F1 of 91.5%. These results indicate that learned affinity improves match coverage, while combining it with handcrafted geometry further improves correspondence accuracy.

Tail-error reduction. Fig. 4 focuses on the strongest hybrid, learned+oDist. On HARD frames, the median rotation and translation errors remain similar to those of learned-only, while their 90^{th}-percentile errors decrease from 5.19° and 6.72 m to 3.15° and 2.72 m, respectively. This indicates that the hybrid mainly reduces large pose failures rather than uniformly lowering errors across all frames. HARD frames contain fewer shared objects and lower cross-view overlap, whereas $|R|$ and $|t|$ are statistically indistinguishable between HARD and EASY frames. These observations suggest that the gain mainly comes from improved box-match disambiguation when shared object evidence is sparse.

Learned-affinity ablations. Table II evaluates the contribution of the learned affinity components. Removing the invariant graph edges and retraining reduces learned-only calibration success from 62.6% to 17.0% on ALL frames and from 51.8% to 14.2% on HARD frames, showing that relative-geometry context is critical to the learned affinity. Removing only the category embedding while retaining the same-category mask reduces learned-only performance to 58.8% on ALL frames and 46.8% on HARD frames, close to the fixed-backend oIoU-only results of 60.0% and 46.1%, respectively. Overall, these ablations show that relational geometry is critical, while category information provides additional benefit.

TABLE II: Ablation results for the learned affinity.

Variant	ALL (%)	EASY (%)	HARD (%)
Baseline (Learned-only)	62.6	85.1	51.8
No invariant edges	17.0	23.1	14.2
No category embedding	58.8	82.1	46.8

V. Conclusion

We studied object-level V2I LiDAR calibration as a three-stage pipeline of affinity scoring, match selection, and pose estimation. By fixing the downstream stages and varying only the affinity design, we conducted controlled comparisons on a held-out DAIR-V2X-C split. For two representative geometric cues, oIoU and oDist, hybrid affinities outperform their learned-only and corresponding handcrafted-only components under the same downstream backend. The strongest variant, learned+oDist, achieves 69.5±0.5% all-frame Success@2m,2°, improving over learned-only and fixed-backend oDist-only by 6.8 pp and 9.1 pp, respectively, while also reducing high-error tail failures. Overall, learned relational affinity and handcrafted geometric affinity act as complements rather than substitutes in object-level V2I LiDAR calibration. These findings suggest that combining learned relational cues with explicit geometric priors is a promising design principle for robust object-level V2I LiDAR calibration.

References

[1] J. Ryu and J. Paek, "Fast field-of-view expansion for collaborative object detection," in *The 22nd ACM International Conference on Mobile Systems, Applications, and Services (MobiSys'24)*. ACM, June 2024.

[2] Q. Qu, Y. Xiong, G. Zhang, X. Wu, X. Gao, X. Gao, H. Li, S. Guo, and G. Zhang, "V2I-Calib: A novel calibration approach for collaborative vehicle and infrastructure lidar systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 892–897.

[3] X. Zhang, Q. Qu, Y. Xiong, C. Xia, Z. Song, Q. Peng, K. Liu, J. Li, and K. Li, "V2X-Reg++: A Real-Time Global Registration Method for Multi-End Sensing System in Urban Intersections," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[4] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning Feature Matching with Graph Neural Networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[5] Z. Lei, Y. Lu, Y. Cheng, Y. Ren, J. Lin, Z. Wang, H. Li, and S. Chen, "Robust Collaborative Perception without External Localization and Clock Devices," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[6] Y. Lu, Q. Li, B. Liu, M. Dianati, C. Feng, S. Chen, and Y. Wang, "Robust Collaborative 3D Object Detection in Presence of Pose Errors," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[7] H. Yang, J. Shi, and L. Carlone, "TEASER: Fast and Certifiable Point Cloud Registration," 2020.

[8] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, and Z. Nie, "DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.