

LiDAR–Camera Calibration under Lossy V2X for RSU–Vehicle Cooperative Perception

Meixuan Bao, Tao Yu, Kui Wang, and Kei Sakaguchi

Institute of Science Tokyo, Tokyo, Japan

Email: {bao, yutao, kuiw, sakaguchi}@mobile.ee.titech.ac.jp

Abstract—Accurate LiDAR–camera calibration is essential for roadside unit (RSU)–vehicle cooperative perception. Unlike conventional onboard calibration, where the LiDAR and camera are rigidly mounted on the same platform, an RSU–vehicle system involves a fixed infrastructure LiDAR and a moving vehicle-mounted camera. Their relative transformation is therefore time-varying and must be refined online from a noisy localization-based pose prior. In addition, when vehicle-side visual features are transmitted to the RSU through a latency-critical vehicle-to-everything (V2X) uplink, packet loss may make the received image-token representation incomplete. This paper studies LiDAR–camera calibration under lossy V2X feature transmission. We propose an RSU-side calibration network that fuses projection-conditioned LiDAR point tokens with received image tokens using cross-modal attention. To mitigate packet-loss-induced feature incompleteness, the network further maintains lightweight streaming memories for visual and LiDAR-side temporal context. A digital-twin dataset generated in AWSIM is used for evaluation. Under a 50 cm / 5° pose-prior error with packet-loss-induced feature degradation, the proposed temporal model reduces both translation and rotation errors compared with the single-frame baseline. Object-level fusion evaluation further shows that calibration correction improves LiDAR–camera association consistency.

Index Terms—LiDAR–camera calibration, lossy V2X, packet loss, RSU–vehicle cooperative perception, sensor fusion

I. INTRODUCTION

Cooperative perception is an important direction for intelligent transportation systems [1], [2]. Roadside units (RSUs) can complement vehicle-side perception by observing traffic scenes from infrastructure viewpoints and sharing information through vehicle-to-infrastructure (V2I) communication. In such systems, accurate LiDAR–camera alignment is essential because calibration errors cause projected LiDAR points to be associated with incorrect image regions.

Most existing LiDAR–camera calibration methods focus on sensors mounted on the same vehicle, where the extrinsic matrix is physically fixed [5]. Representative methods such as LCCNet [3] and CalibAnything [4] follow this fixed-rig setting. The RSU–vehicle case is different: the LiDAR is fixed on infrastructure, while the camera moves with the vehicle, making the relative transformation time-varying. A coarse

LiDAR-to-camera pose can be derived from RSU and vehicle localization, but it may contain localization bias, frame-wise noise, and temporal inconsistency. Therefore, the problem becomes online pose-prior refinement rather than fixed extrinsic calibration.

Another difficulty comes from vehicle-to-everything (V2X) communication. Latency-critical perception often favors User Datagram Protocol (UDP)-like non-retransmission transmission, where retransmission delay is avoided but transmitted features may be incomplete. Feature-level cooperative perception has been studied to reduce communication cost [6], [7], but calibration under lossy feature reception remains less explored. To compensate for V2X-induced feature loss, temporal information can be reused so that previously received visual–geometric correspondences support calibration when the current frame is incomplete.

This paper studies LiDAR–camera calibration under lossy V2X feature transmission for RSU–vehicle cooperative perception. The method uses the current pose prior to project RSU LiDAR points into the camera image, constructs projection-conditioned point tokens, and fuses them with received image tokens using bidirectional cross-modal attention. Streaming temporal memories are introduced as a mechanism to compensate for incomplete current-frame feature reception.

The main contributions are summarized as follows:

- We formulate RSU–vehicle LiDAR–camera calibration as an online pose-prior refinement problem under lossy V2X feature transmission.
- We model the effect of non-retransmission V2X feature transmission as packet-loss-induced image-token loss at the RSU, and introduce streaming temporal memories to compensate for this feature loss during online calibration.
- We construct a digital-twin evaluation pipeline and show that calibration correction improves both pose accuracy and object-level LiDAR–camera association consistency.

II. PROBLEM SETTING

A. System Scenario

We consider an infrastructure-assisted V2X perception system consisting of a fixed roadside unit (RSU) and a moving vehicle. The RSU is equipped with a LiDAR sensor and a roadside computing unit, while the vehicle is equipped with a forward-facing camera. In the considered deployment, the

*This work was supported by MONET Technologies Inc. (Responsible person: Toshiki Demizu, Shunsuke Kawasaki, Takahiro Ishihara) and by the Ministry of Land, Infrastructure, Transport and Tourism under the Small/Startup Business Innovation Research (SBIR) program, and by the Science Tokyo–RWTH Aachen SPF 2026 Seed Grant.

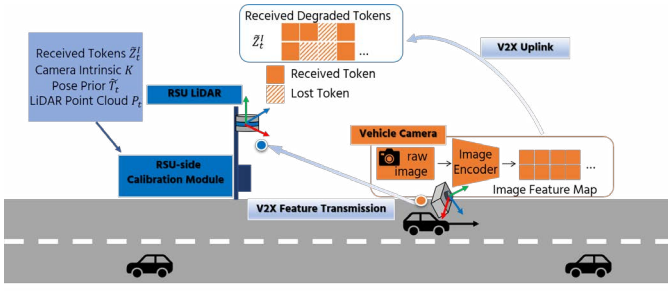


Fig. 1. System overview of the RSU-vehicle calibration framework under lossy V2X feature transmission. The vehicle sends image feature tokens to the RSU, where packet loss may cause incomplete token reception. The RSU fuses the received image tokens with local LiDAR data and refines the time-varying LiDAR-to-camera transformation from a noisy pose prior.

vehicle first extracts a visual feature map from the camera image and transmits its tokenized representation to the RSU through a V2X uplink. All subsequent computation, including LiDAR-camera fusion, temporal memory update, and calibration refinement, is performed on the RSU side.

The RSU LiDAR coordinate frame is fixed in the infrastructure, whereas the camera coordinate frame moves with the vehicle. Therefore, the relative LiDAR-to-camera transformation is time-varying and must be refined online from a noisy pose prior. The two sensors observe overlapping parts of the same scene from different positions and orientations, which produces viewpoint discrepancy, limited overlap, and occlusion-dependent correspondence errors. Fig. 1 illustrates the overall system scenario, lossy V2X transmission, and the motivation for temporal memory.

B. Objective

Let P_t be the RSU LiDAR point cloud and I_t be the vehicle camera image at time t . The vehicle-side encoder extracts image tokens Z_t^I from I_t and transmits them to the RSU. The target transformation is the LiDAR-to-camera transformation $T_t^{gt} \in SE(3)$ from the RSU LiDAR frame to the vehicle camera frame. Given a coarse pose prior $\tilde{T}_t$ from infrastructure and vehicle localization, the network predicts a residual correction $\Delta\hat{T}_t$:

$$\hat{T}_t = \Delta\hat{T}_t \tilde{T}_t \quad (1)$$

The objective is to minimize the translation and rotation errors between $\hat{T}_t$ and T_t^{gt} during online RSU-side inference.

C. Feature Transmission under Lossy V2X

The system targets latency-critical V2X perception, where retransmission may introduce undesirable delay. We therefore consider a UDP-like non-retransmission feature-transmission scheme, in which packet loss may occur after vehicle-side image encoding and before RSU-side reception. In this study, the effect of V2X communication is modeled as packet loss during the transmission of vehicle-side image features to the RSU after image encoding. As a result, the RSU may receive incomplete image feature tokens due to packet loss.

Let Z_t^I denote the image-token representation extracted at the vehicle side. The received representation at the RSU is modeled as

$$\tilde{Z}_t^I = M_t^{\text{frame}} \cdot (M_t^{\text{feat}} \odot Z_t^I) \quad (2)$$

where M_t^{feat} is a frame-internal packet-loss mask over image feature tokens, and $M_t^{\text{frame}} \in \{0, 1\}$ indicates whether the whole visual feature frame is received. In the lossy setting, both the frame-internal feature-token loss rate and the whole-frame loss probability are set to 0.3.

Under lossy V2X transmission, a single-frame model must estimate the residual pose from incomplete current-frame features. To compensate for this packet-loss-induced feature incompleteness, we introduce temporal context so that previously received visual and geometric information can be reused through streaming memories.

D. Coordinate Systems

The calibration uses the world frame $\mathcal{W}$, RSU LiDAR frame $\mathcal{L}$, vehicle camera frame $\mathcal{C}_t$, and image frame $\mathcal{I}$. The ground-truth LiDAR-to-camera transformation is computed from simulator poses as

$${}^{\mathcal{C}_t}T_{\mathcal{L}} = ({}^{\mathcal{W}}T_{\mathcal{C}_t})^{-1} {}^{\mathcal{W}}T_{\mathcal{L}} \quad (3)$$

A 3D LiDAR point $x_{\mathcal{L}}$ is projected to the image by

$$u \sim K {}^{\mathcal{C}_t}T_{\mathcal{L}} x_{\mathcal{L}} \quad (4)$$

where K is the camera intrinsic matrix.

III. PROPOSED METHOD

A. Network Architecture

At the RSU side, the proposed network estimates a residual LiDAR-camera calibration from the received image tokens $\tilde{Z}_t^I$, the RSU LiDAR point cloud P_t , camera intrinsics K , and a coarse pose prior $\tilde{T}_t$. The network predicts a residual transformation $\Delta\hat{T}_t$ and refines the pose prior according to Eq. (1).

As shown in Fig. 2, the RSU-side network consists of a projection-conditioned point-token encoder, a bidirectional cross-modal attention module, dual streaming temporal memories, and a recursive pose-prior update. The received image tokens are integrated into a gated image-token memory, while the LiDAR stream maintains a compact bird's-eye-view (BEV) memory to summarize temporal geometric context.

The RSU LiDAR branch keeps the point cloud as sparse point tokens. For each sampled point, the current pose prior $\tilde{T}_t$ transforms the point into the camera frame and projects it onto the image plane. Each point token contains LiDAR-frame coordinates, camera-frame coordinates, normalized image coordinates, depth, and an in-image validity flag. Therefore, the point representation explicitly encodes how the current pose prior aligns the LiDAR structure with the received visual tokens.

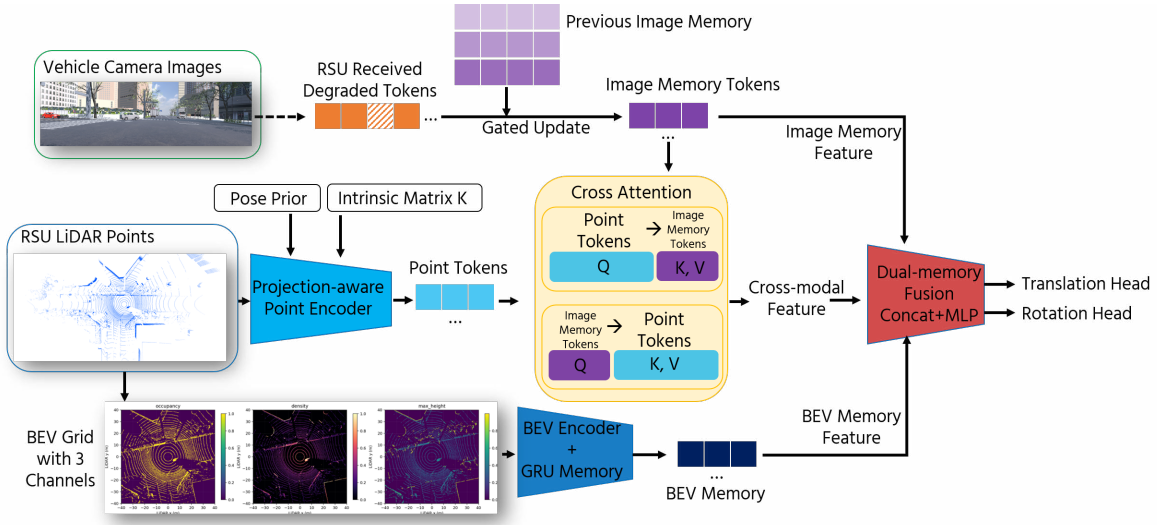


Fig. 2. Architecture of the proposed RSU-side calibration network. Received image tokens are updated by streaming image memory and fused with projection-conditioned LiDAR point tokens through cross-modal attention. The fused feature and LiDAR-side BEV memory are used for recursive pose refinement.

B. Projection-Conditioned Cross-Modal Attention

Let $\tilde{Z}_t^I$ denote the received image tokens and Z_t^P denote the projection-conditioned point tokens. Before cross-modal fusion, the received image tokens are integrated into a streaming image-token memory M_t^I . The point tokens then attend to the image memory, rather than relying only on the currently received incomplete tokens. Following the multi-head attention mechanism [8], the bidirectional cross-modal attention is written as

$$\begin{aligned}\bar{Z}_t^P &= \text{Attn}(Z_t^P, M_t^I, M_t^I) \\ \bar{M}_t^I &= \text{Attn}(M_t^I, \bar{Z}_t^P, \bar{Z}_t^P)\end{aligned}\quad (5)$$

where the first direction lets LiDAR point tokens query the image-memory tokens, and the second direction lets image-memory tokens incorporate sparse geometric cues from the RSU LiDAR. Residual connections and layer normalization are applied after each attention operation. The attended point and image-memory tokens are mean-pooled and fused to obtain a cross-modal feature F_t^{ca} .

Since the point-token projection depends on $\tilde{T}_t$, the attention module receives pose-dependent geometric evidence. When the pose prior is inaccurate, the projected point locations and received visual tokens become inconsistent, allowing the network to learn calibration-relevant residuals.

C. Dual Streaming Temporal Memory

To mitigate packet-loss-induced incompleteness in the received image tokens, the model maintains a token-level image memory. Given received image tokens $\tilde{Z}_t^I$ and previous memory M_{t-1}^I , a gated update is applied:

$$M_t^I = \alpha_t \odot \tilde{Z}_t^I + (1 - \alpha_t) \odot M_{t-1}^I \quad (6)$$

where the gate α_t is predicted from the received image tokens and the previous memory. This memory allows the RSU to

reuse visual context from earlier frames when the current feature-token reception is incomplete.

In parallel, the model uses a LiDAR-side BEV memory. The raw point cloud is rasterized into a BEV grid with occupancy, density, and maximum-height channels. A lightweight BEV convolutional neural network (CNN) extracts a compact BEV feature, which is updated by a gated recurrent unit (GRU) memory [9]:

$$h_t^B = \text{GRU}(E_{\text{BEV}}(P_t), h_{t-1}^B) \quad (7)$$

where $E_{\text{BEV}}(\cdot)$ denotes the BEV encoder and h_t^B is the LiDAR-side temporal memory state.

The final calibration feature is obtained by fusing three features with a multilayer perceptron (MLP):

$$F_t = \text{MLP}([F_t^{ca}; \text{Pool}(M_t^I); h_t^B]) \quad (8)$$

where F_t^{ca} is the cross-modal feature, $\text{Pool}(M_t^I)$ is the image-memory summary, and h_t^B is the BEV-memory summary.

D. Recursive Pose-Prior Update and Online Inference

The method processes frames sequentially and does not require future frames or full-sequence optimization. At the first frame of a streaming chunk, the model uses the noisy initial pose as the pose prior:

$$\tilde{T}_1 = T_1^{\text{init}} \quad (9)$$

For later frames, the current pose prior is obtained by propagating the previous refined pose using the relative motion implied by the noisy initial poses:

$$\tilde{T}_t = (T_t^{\text{init}}(T_{t-1}^{\text{init}})^{-1}) \hat{T}_{t-1} \quad (10)$$

The network then predicts a residual correction $\Delta\hat{T}_t$ and updates the refined pose using Eq. (1). This recursive design allows the model to use temporal continuity while still performing online frame-by-frame inference.

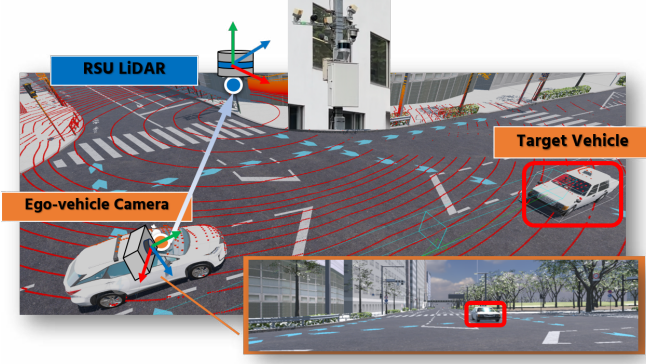


Fig. 3. Digital-twin evaluation scenario used for RSU-vehicle LiDAR-camera calibration.

The residual head predicts a translation residual $\Delta \hat{\mathbf{t}}_t$ and a quaternion residual $\Delta \hat{\mathbf{q}}_t$. During training with recursive initialization, the supervised residual target is computed with respect to the actual pose prior used at that frame:

$$\Delta T_t^{gt} = T_t^{gt} \tilde{T}_t^{-1} \quad (11)$$

The model is trained to regress this residual correction using a SmoothL1 translation loss and a quaternion rotation-distance loss. Auxiliary motion loss, projection loss, and translation-direction loss are disabled in the reported experiments. Feature-token packet loss is applied during training and evaluation to represent incomplete V2X feature reception under lossy transmission.

IV. EVALUATION

A. Digital-Twin Dataset Generation

Digital-twin platforms have been increasingly used to support smart-mobility and cooperative-driving validation [10], [11]. We evaluate the proposed method using a digital-twin dataset generated in AWSIM, an open-source simulator integrated with Autoware [12]. The scenario includes a fixed RSU LiDAR and a vehicle-mounted camera, whose simulator poses are used to compute ground-truth LiDAR-to-camera transformations using Eq. (3). Each sample contains a synchronized image, point cloud, camera intrinsics, noisy pose prior, and residual calibration target. Fig. 3 shows the evaluation scenario.

The dataset contains 194,778 synchronized camera-LiDAR frame pairs after timestamp matching between the camera and LiDAR streams. After valid-label filtering, 54,241 synchronized samples are used in the current evaluation, including 43,574 training samples, 7,569 validation samples, and 3,098 test samples. Each sample contains a camera image, an RSU LiDAR point cloud, camera intrinsics, a noisy pose prior, and the corresponding residual calibration target.

To emulate localization uncertainty, the pose prior is generated by perturbing the ground-truth relative pose with localization-like noise. In the standard setting, the maximum pose-prior error is 50 cm / 5°. Larger settings, including 50 cm

TABLE I
CALIBRATION ACCURACY OF THE PROPOSED METHOD UNDER THE LOSSY V2X FEATURE-TRANSMISSION MODEL.

Method	Prior error	Loss	Trans. / Rot. error
Ours single-frame	0.5 m / 5°	0	8.59 cm / 0.31°
Ours single-frame	0.5 m / 5°	0.3	9.58 cm / 0.50°
Ours temporal	0.5 m / 5°	0.3	5.76 cm / 0.42°
Ours temporal	0.5 m / 20°	0.3	15.92 cm / 4.64°
Ours temporal	1.5 m / 20°	0.3	34.00 cm / 4.81°
Ours temporal	3.0 m / 15°	0.3	39.31 cm / 2.97°

/ 20°, 1.5 m / 20°, and 3 m / 15°, are used to evaluate robustness to coarse vehicle localization.

To represent lossy V2X feature transmission, we apply the two-level packet-loss model in Eq. (2) to the transmitted image-token representation. The lossy setting uses a frame-internal token loss rate of 0.3 and a whole-frame visual feature loss probability of 0.3. This model captures incomplete visual feature reception at the RSU, while not simulating the full packet-level communication protocol, scheduling, or channel dynamics. We report translation error in centimeters and rotation error in degrees.

B. Calibration Accuracy on the Digital-Twin Dataset

Table I summarizes the calibration accuracy under different pose-prior errors and packet-loss settings. Under the 50 cm / 5° lossy V2X setting, temporal memory reduces the translation error from 9.58 cm to 5.76 cm and the rotation error from 0.50° to 0.42° compared with the single-frame model. Under larger pose-prior errors, the proposed framework still reduces the initial misalignment, although the remaining error increases as the prior becomes coarser.

C. Comparison with LCCNet and CalibAnything

We compare the proposed method with LCCNet [3], a learning-based calibration baseline designed for fixed vehicle-mounted LiDAR-camera rigs, and CalibAnything [4], a recent targetless single-frame calibration method. Since the RSU-vehicle relative pose is time-varying, CalibAnything is applied frame by frame, and its lossy result is reported with image-level masking as a sensitivity reference.

Table II shows that LCCNet performs reasonably in the small clean setting but degrades under packet loss and coarse pose-prior errors. The proposed method achieves lower translation error in all compared settings. Although LCCNet obtains competitive rotation accuracy in the small-prior setting, its translation error increases substantially under lossy feature transmission and coarse pose priors.

D. Effect of Temporal Memory

As shown in Table I, temporal memory reduces the error from 9.58 cm / 0.50° to 5.76 cm / 0.42° under the 50 cm / 5° lossy V2X setting. To further verify the contribution of the memory state, we evaluate the same temporal checkpoint with zero memory and full memory. Full memory reduces the

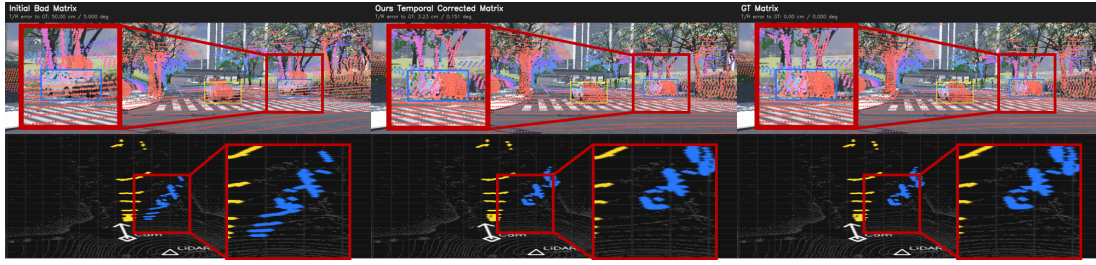


Fig. 4. Object-level fusion comparison under bad prior, our correction, and ground-truth calibration.

TABLE II

COMPARISON WITH CALIBRATION BASELINES ON THE RSU–VEHICLE DATASET. FOR LEARNING-BASED METHODS, LOSS=0.3 DENOTES THE LOSSY V2X FEATURE-TRANSMISSION MODEL; FOR CALIBANYTHING, 0.3[†] DENOTES IMAGE-LEVEL MASKING WITH THE SAME MISSING RATIO.

Prior error	Method	Loss	Trans. / Rot. error
0.5 m / 5°	CalibAnything	0	99.04 cm / 4.13°
	LCCNet	0	19.33 cm / 0.38°
	Ours single-frame	0	8.59 cm / 0.31°
	CalibAnything	0.3 [†]	111.48 cm / 4.85°
	LCCNet	0.3	33.61 cm / 0.54°
	Ours single-frame	0.3	9.58 cm / 0.50°
0.5 m / 20°	LCCNet	0.3	37.08 cm / 5.50°
	Ours temporal	0.3	15.92 cm / 4.64°
1.5 m / 20°	LCCNet	0.3	107.92 cm / 9.05°
	Ours temporal	0.3	34.00 cm / 4.81°
3.0 m / 15°	LCCNet	0.3	240.82 cm / 8.47°
	Ours temporal	0.3	39.31 cm / 2.97°

[†] For CalibAnything, 0.3 denotes image-level masking with the same missing ratio, not the feature-token loss used by learning-based methods.

TABLE III
OBJECT-LEVEL FUSION CONSISTENCY.

Metric	Bad prior	Ours corrected	Reduction
BEV center median error	1.16 m	0.05 m	95.7%
Depth median error	1.08 m	0.13 m	88.1%
Point-count median difference	6 points	1 point	83.3%

translation error from 6.06 cm to 4.79 cm, indicating that the stored temporal context contributes to calibration accuracy.

E. Improvement on Object-Level Fusion

Beyond direct pose error, we evaluate whether the corrected calibration improves object-level LiDAR–camera association. We detect 2D objects in the image and compare the LiDAR point associations obtained using a bad prior, our corrected calibration, and ground-truth calibration. All metrics are computed with respect to the ground-truth associations.

As shown in Table III, our correction is much closer to the ground truth than the bad prior. The BEV center median error is reduced from 1.16 m to 0.05 m, and the depth median error is reduced from 1.08 m to 0.13 m. This indicates that

the proposed calibration correction benefits downstream fusion consistency.

V. CONCLUSION

This paper presented an RSU–vehicle LiDAR–camera calibration framework under lossy V2X feature transmission. The proposed network refines a noisy pose prior online using projection-conditioned point tokens, cross-modal attention, and streaming temporal memories. Experiments on a digital-twin dataset show improved calibration accuracy under packet-loss-induced feature incompleteness and better object-level LiDAR–camera association consistency. The single-frame and temporal-memory models run at 29.09 FPS and 28.03 FPS, respectively, in online streaming inference. Future work will extend the evaluation to real-world RSU–vehicle validation and full object-level fusion detection benchmarks.

REFERENCES

- [1] Y. Li *et al.*, “V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10914–10921, 2022.
- [2] R. Xu *et al.*, “V2X-ViT: Vehicle-to-everything cooperative perception with Vision Transformer,” in *European Conference on Computer Vision (ECCV)*, 2022, pp. 107–124.
- [3] X. Lv *et al.*, “LCCNet: LiDAR and camera self-calibration using cost volume network,” in *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 2888–2895.
- [4] Z. Luo *et al.*, “Zero-training LiDAR-camera extrinsic calibration method using Segment Anything Model,” in *International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14472–14478.
- [5] Z. Tan *et al.*, “A review of deep learning-based LiDAR and camera extrinsic calibration,” *Sensors*, vol. 24, no. 12, 3878, 2024.
- [6] Q. Chen *et al.*, “F-Cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds,” in *Symposium on Edge Computing (SEC)*, 2019, pp. 88–100.
- [7] Y. Hu *et al.*, “Where2comm: Communication-efficient collaborative perception via spatial confidence maps,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [8] A. Vaswani *et al.*, “Attention is all you need,” in *Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.
- [9] K. Cho *et al.*, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.
- [10] T. Yu *et al.*, “Internet of federated digital twins: Connecting twins beyond borders for Society 5.0,” *IEEE Internet of Things Magazine*, vol. 7, no. 5, pp. 64–71, 2024.
- [11] K. Nonomura *et al.*, “Digital twin-empowered cooperative autonomous car-sharing services: Proof-of-concept,” in *IEEE Intelligent Vehicles Symposium (IV)*, 2025.
- [12] “AWSIM,” <https://github.com/autowarefoundation/AWSIM>