V2X Collaborative Perception

Jiahao Huang

Networked Intelligence for Comprehensive Efficiency (NICE) Lab
College of Information Science and Electronic Engineering
Zhejiang University
http://nice.rongpeng.info/

Sep. 18, 2025

zjunice tutorial 2025

Content

The Foundations of Collaborative Perception

General Overview

Why Collaborative Perception?

- Standalone perception has fundamental physical limits. V2X collaborative perception is the key to breaking these barriers and enabling next-level autonomous safety.
  - Standalone Intelligence: Inherent Blind Spots & Hazards[a].
  - Enabling Technologies: 5G & Semantic Communication[b].
  - The V2X Vision: Achieving Collective Traffic Efficiency.

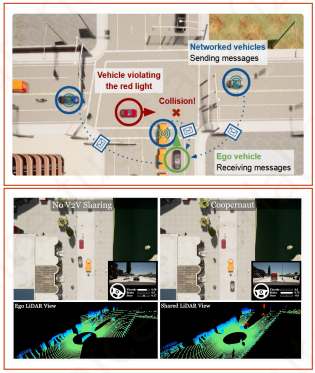


Figure: Coopernaut: End-to-End Driving with Cooperative Perception

[a] J. Cui et al., "Coopernaut: End-to-end driving with cooperative perception for networked vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 252–17 262.

[b] S. Chen et al., "Vehicle-to-everything (v2x) services supported by lte-based systems and 5g," *IEEE communications standards magazine*, vol. 1, no. 2, pp. 70–76, 2017 .

Paradigm of Collaborative Perception: Fusion[1]

How to fuse information?

- **Intermediate Fusion** is Widely adopted, best trade-off between high perception performance and low communication overhead.
  - **Early Fusion**: Agents share **raw sensor data** (e.g., LiDAR point clouds, camera images), requiring high communication bandwidth.
  - **Late Fusion**: Agents exchange only **final detection results** (e.g., bounding boxes), constrained by individual agent capabilities.
  - **Intermediate Fusion**: Agents transmit **neural network-extracted features** (often BEV-formatted), balancing performance and bandwidth.

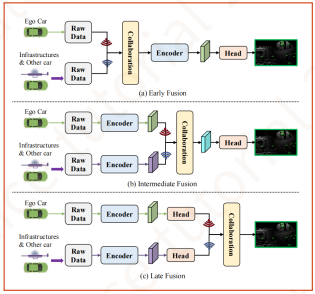


Figure: Fusion Methods

[1] S. V. Balkus et al., "A survey of collaborative machine learning using 5g vehicular communications," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1280–1303, 2022 .

Paradigm of Collaborative Perception: Pipeline

Gold-Standard

- Universal Pipeline: Extract → Share → Fuse
  - Input —*Raw Sensor Data: Point Cloud from Lidar, Pixels from Camera*→
  - Individual Feature Extraction —*Individual Perception*→
  - Feature Compression and Broadcasting[a][b] —*Intermediate Feature*→
  - Information Aggregation —*Fused Intermediate Feature*→
  - Perception and Prediction Heads —*Bounding Box*→ Output

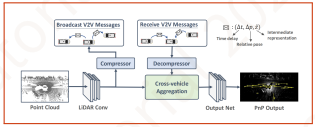


Figure: Pipeline

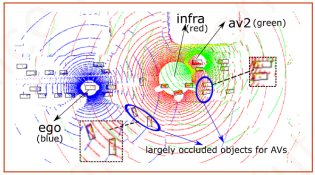


Figure: Collaboration

[a] T.-H. Wang et al., "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *European conference on computer vision*, Springer, 2020, pp. 605–621.

[b] R. Xu et al., "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," in *European conference on computer vision*, Springer, 2022, pp. 107–124.

Paradigm of Collaborative Perception: Pipeline

Gold-Standard

- Universal Pipeline: Extract → Share → Fuse
  - Input —*Raw Sensor Data: Point Cloud from Lidar, Pixels from Camera*→
  - Individual Feature Extraction —*Individual Perception*→
  - Feature Compression and Broadcasting[a][b] —*Intermediate Feature*→
  - Information Aggregation —*Fused Intermediate Feature*→
  - Perception and Prediction Heads —*Bounding Box*→ Output

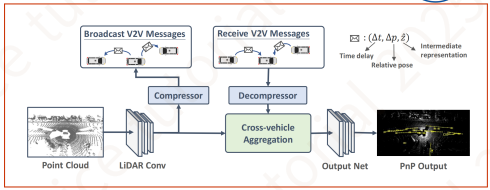


Figure: Pipeline

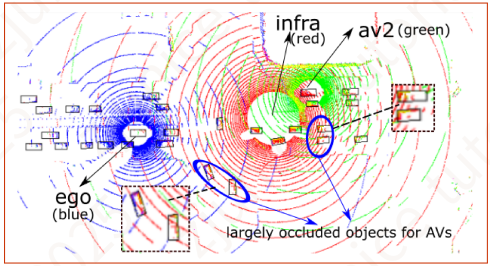


Figure: Collaboration

[a] T.-H. Wang et al., "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *European conference on computer vision*, Springer, 2020, pp. 605–621.

[b] R. Xu et al., "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," in *European conference on computer vision*, Springer, 2022, pp. 107–124.

Neural Networks for V2X Feature Extraction

Backbone: PointPillars

- A foundational backbone for its efficiently in generating BEV features.
  - **PointPillars**[a]: Utilizes vertical columns of points (pillars) to create a 2D pseudo-image, enabling efficient 2D CNN processing while retaining 3D spatial information.
  - **CenterPoint**[b]: Detects object centers as keypoints and then regresses other properties like size, orientation, and velocity from center-point features (Anchor-Free).

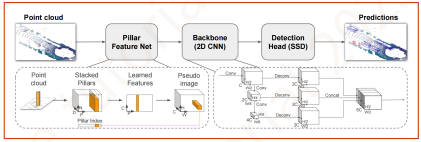


Figure: PointPillars

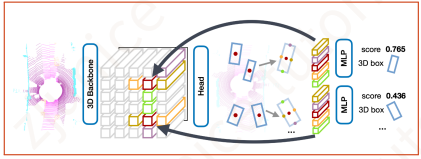


Figure: CenterPoint

[a] A. H. Lang et al., "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 697–12 705.

[b] T. Yin et al., "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 784–11 793.

Object Detection or Semantic Segmentation

Why do we use Object Detection?[a]:

- Object Detection is preferred over Segmentation due to its superior **communication efficiency** and direct relevance to **downstream tasks** like tracking and prediction.
  - **Object Detection**: Localize and classify distinct object instances using **bounding boxes**.
  - **Segmentation**: Assign a class **label to every pixel** in an image.

[a] J. Mao et al., "3d object detection for autonomous driving: A comprehensive survey," *International Journal of Computer Vision*, vol. 131, no. 8, pp. 1909–1963, 2023 .

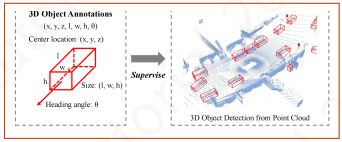


Figure: Object Detection

Figure: Segmentation

Measurement[2]

- Average Precision (AP) at a given IoU threshold is the gold-standard metric for evaluating both the accuracy (Precision) and completeness (Recall).
- Precision and Recall: Tilting Board
  - Precision = TP / (TP + FP): Which one is right?
  - Recall = TP / (TP + FN): How many right ones are found?
- AP: Measure correctness with both Precision and Recall
  - P-R curve: Varying the confidence threshold.
  - AP: Area under the P-R curve.
- AP over IoU: Classic Metric
  - IoU: Intersection over Union.
  - AP over IoU: From right or wrong to how right.

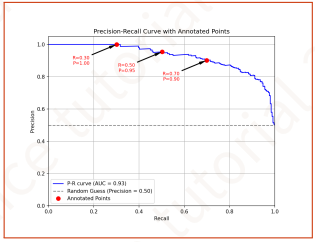


Figure: P-R Curve

[2] T.-H. Wang et al., "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *European conference on computer vision*, Springer, 2020, pp. 605–621.

Datasets: From V2V to V2Rsu and UAV2UAV

- Datasets are evolving from vehicle-to-infrastructure to air-to-ground, driving research towards more **complex, integrated scenarios**.
  - DAIR-V2X(CVPR 2022)[a]: Real World, Road-infrastructure Collaboration. It provides unmatched realism and **data fidelity**, but annotated data is **extremely scarce**.
  - UAV3D(NIPS 2024)[b]: Virtual World(AirSim), UAV Collaboration. It offers massive **scalability** with perfect **ground truth** and meets **"sim-to-real" domain gap**.

Figure: DAIR-V2X

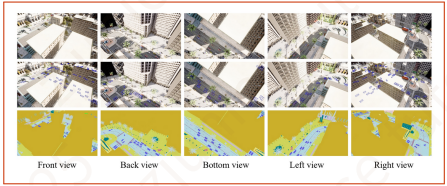


Figure: UAV3D

[a] H. Yu et al., "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 21 361–21 370.

[b] R. Sunderraman, J. S. Ji, et al., "Uav3d: A large-scale 3d perception benchmark for unmanned aerial vehicles," *Advances in Neural Information Processing Systems*, vol. 37, pp. 55 425–55 442, 2024 .

Mainstream Works

Mainstream Works

Main Challenges:

- Research focuses on three core challenges: 1. Communication constraints (bandwidth, latency), 2. Perception limits (occlusions), and 3. System imperfections (heterogeneity, errors).
  - Communication constraints: **Bandwidth**[ab] and **Temporal Asynchrony**.
  - Perception limits: **Occlusions** and Data Sparsity[c].
  - System imperfections: Localization Error, **Heterogeneity**[de] and Security.

[a] Y. Hu et al., "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *Advances in neural information processing systems*, vol. 35, pp. 4874–4886, 2022 .

[b] Y. Li et al., "Beyond bev: Optimizing point-level tokens for collaborative perception," *arXiv preprint arXiv:2508.19638*, 2025 .

[c] J. Hao et al., "Is intermediate fusion all you need for uav-based collaborative perception?" *arXiv preprint arXiv:2504.21774*, 2025 .

[d] Y. Lu et al., "An extensible framework for open heterogeneous collaborative perception," *arXiv preprint arXiv:2401.13964*, 2024 .

[e] X. Gao et al., "Stamp: Scalable task and model-agnostic collaborative perception," *arXiv preprint arXiv:2501.18616*, 2025 .

Where2Comm[3]: Overview (NIPS 2022)

- TLDR: Where2Comm pioneers on-demand communication: it intelligently asks for and shares only the **most critical information**—data from blind spots—to **maximize bandwidth efficiency**.

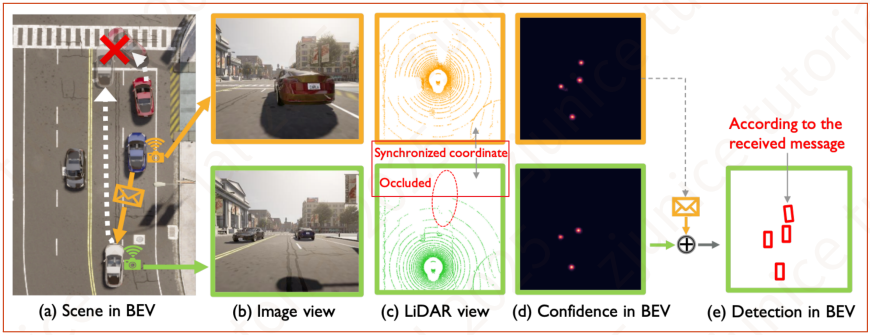


Figure: Overview of Where2Comm

[3] Y. Hu et al., "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *Advances in neural information processing systems*, vol. 35, pp. 4874–4886, 2022 .

Where2Comm: Solution (NIPS 2022)

- Former Solution[4]: learn to **wait** for valuable but delayed information.
- Solution: **Pragmatic Communication**
  - Qunatify the reliability via **confidence threshold**.
  - Multi-round broadcast for **request map**.
  - Pack and send most important **spatial information**.

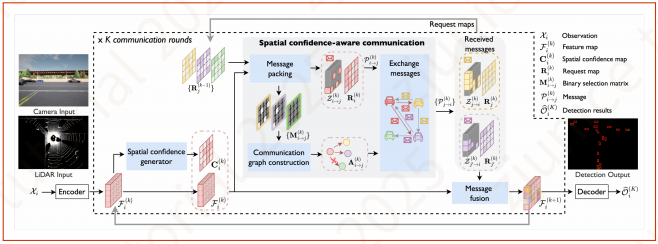


Figure: Architecture of Where2Comm

[4] Y.-C. Liu et al., "When2com: Multi-agent perception via communication graph grouping," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2020, pp. 4106–4115.

LIF[5]: Overview (arXiv 2025.7)

- TLDR: LIF introduces a novel hybrid late-intermediate fusion for air-ground collaboration, transmitting lightweight boxes and converting them to virtual BEV features to overcome **extreme UAV bandwidth limits**.

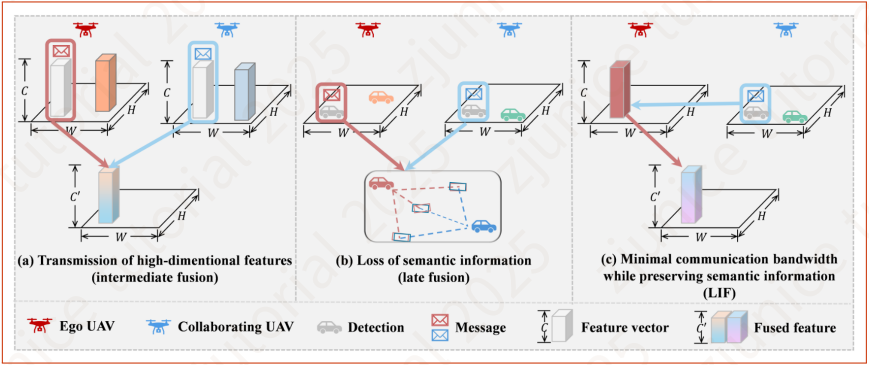
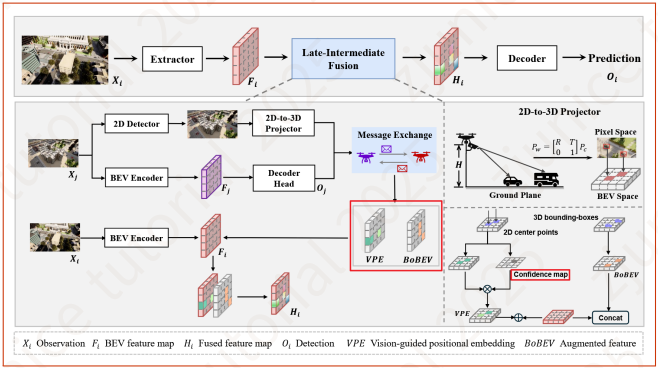


Figure: Overview of LIF

[5] J. Hao et al., "Is intermediate fusion all you need for uav-based collaborative perception?" *arXiv preprint arXiv:2504.21774, 2025* .

LIF: Solution (arXiv 2025.7)

- Former Solution[6]: Transmitting intermediate features is still too costly for UAV.
- Solution: Transmit **lightweight detection results** but solves information loss.
  - Box-based Virtual Feature: Virtualizes received bounding box into pseudo-map and **inject into ego's intermediate BEV map**.
  - **Uncertainty-driven** communication scheme: only transmit information when collaborator is highly confident.



STAMP[7]: Overview (arXiv 2025.1)

- TLDR: STAMP enables true **heterogeneous collaboration** by using **feature translator** modules, allowing agents with different sensors and models to effectively understand each other.

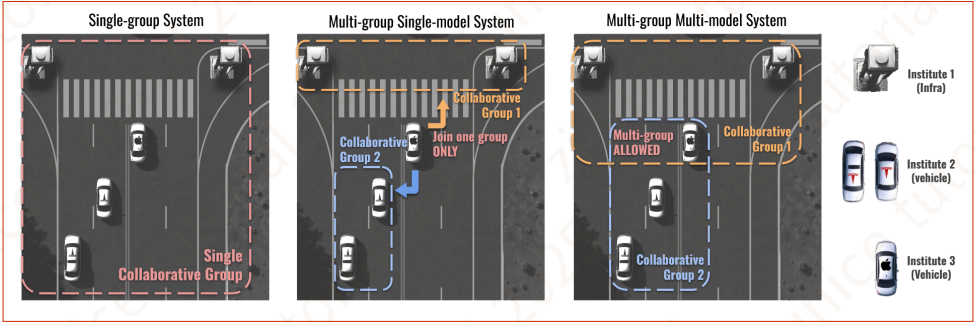


Figure: Overview of Heterogeneity

[7] X. Gao et al., "Stamp: Scalable task and model-agnostic collaborative perception," *arXiv preprint arXiv:2501.18616*, 2025 .

STAMP: Solution (arXiv 2025.1)

- Former Solution[8]: Requires significant retraining and adaptation.
- Solution: **Plug-and-play** collaboration with **slight communication overhead**.
  - Adaptor-reverter: Transform unique BEV features into a unified **protocal domain**.

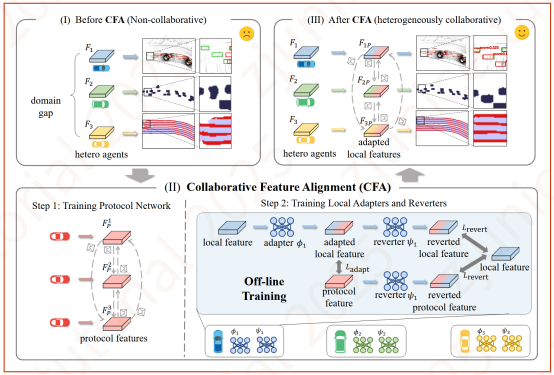


Figure: Overview of STAMP

[8] Y. Lu et al., "An extensible framework for open heterogeneous collaborative perception," *arXiv preprint arXiv:2401.13964*, 2024 .

CoPLOT[9]: Overview (arXiv 2025.8)

- TLDR: CoPLOT challenges the dominant BEV-based paradigm by proposing **communication with point-level tokens**, aiming to preserve more 3D information and improve efficiency.

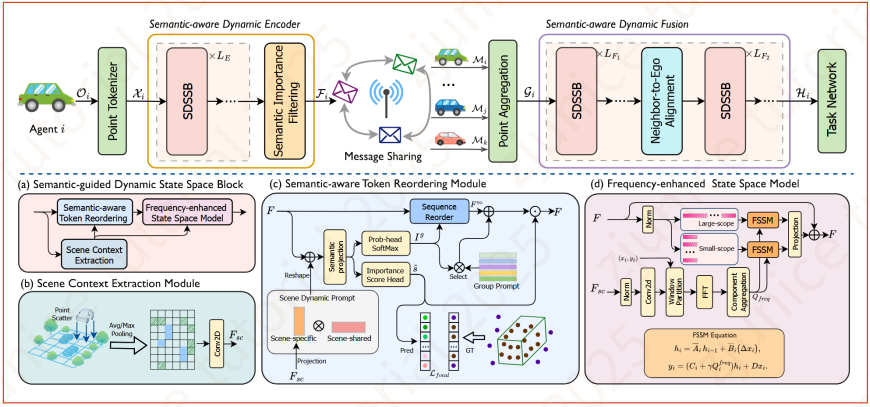


Figure: Overview of CoPLOT

CoPLOT: Solution (arXiv 2025.8)

- Former Solution[10]: Grid-based BEV features suffer from computational overhead.
- Solution: Treat original points as token.
  - **State Space Models(Mamba)** to process sequence of point-level tokens.
  - **Semantic-Aware Reordering**: Injects frequency-domain features into the SSM to better distinguish foreground objects from background clutter.

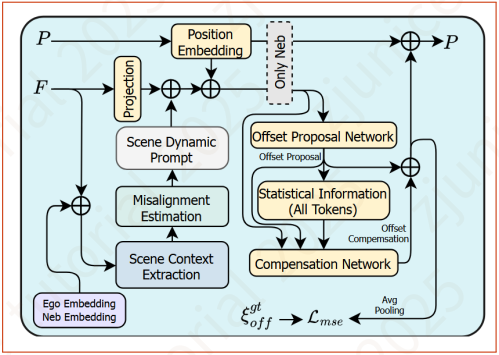


Figure: Alignment Module in CoPLOT

Conclusion

Conclusions

Foundations:

- Pipeline: Extract → Share → Fuse
- Platform: OpenCOOD, modular design, from dataloader, processor to fuser.
- Backbone: PointPillars, extract 2D features while retaining 3D spatial info.
- Datasets: Sim to real gap, complex scenarios.

Latest Advances:

- Where2Comm: Transmit only necessary region for communication efficiency.
- LIF: 2D → 3D projector to retain ground information, Uncertainty-driven Filter.
- STAMP: Light-weight aligner for heterogeneous **sensor, model and tasks**.
- CoPLOT: From BEV fusion to Point-level Token communication.

Thank you

Jiahao Huang
Networked Intelligence for Comprehensive Efficiency (NICE) Lab
College of Information Science and Electronic Engineering
Zhejiang University
https://nice.rongpeng.info

zjunice tutorial 2025

Appendix

Paradigm of Collaborative Perception: Implementation

Gold-Standard

- OpenCOOD[a] (ICRA 2022): Standard Simulation Platform.
  - Input —*Sensor Data with Logs: Relative pos, Sensor Params, etc*→
  - BaseDataset (Dataloader) —*Sensor Data indexed with timestamp*→
  - Preprocessor —*Intermediate Feature(Tensor of Voxel)*→
  - Information Aggregation —*Fused Intermediate Feature*→
  - Perception and Prediction Heads —*Bounding Box*→ Output

[a] R. Xu et al., "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, 2022, pp. 2583–2589.

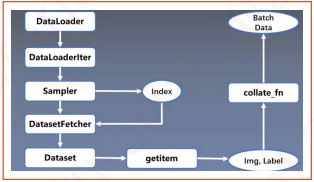


Figure: TorchLoader

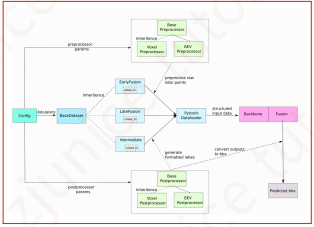


Figure: OpenCOOD

Where2Comm: Details

- **Confidence & Request Map**: First, the vehicle generates a confidence map from its own sensor data to find uncertain areas, then creates a request map from these ”blind spots” to ask for help.
- **Message Packing**: Next, a responding vehicle uses this request map as a mask on its own feature map, **packing only the specifically requested information into a compact message.**
- **Feature Fusion**: Finally, the original vehicle fuses the received features with its own using an attention mechanism to intelligently weigh the data, filling its blind spots for a more complete perception.

$$\mathbf{C}_i^{(k)} = \Phi_{\text{generator}}(\mathcal{F}_i^{(k)}) \in [0,1]^{H\times W},$$

$$\mathbf{R}_i^{(k)} = 1 - \mathbf{C}_i^{(k)} \in \mathbb{R}^{H\times W},$$

Figure: Confidence and Request

$$\mathbf{M}_{i\to j}^{(k)} = \begin{cases} \Phi_{\text{select}}(\mathbf{C}_i^{(k)}) \in \{0,1\}^{H\times W}, & k=0; \\ \Phi_{\text{select}}(\mathbf{C}_i^{(k)} \odot \mathbf{R}_j^{(k-1)}), \in \{0,1\}^{H\times W}, & k>0; \end{cases}$$

$$\mathbf{A}_{i,j}^{(k)} = \begin{cases} 1, \\ \max_{h\in\{0,1,\ldots,H-1\},w\in\{0,1,\ldots,W-1\}} \left(\mathbf{M}_{i\to j}^{(k)}\right)_{h,w} \in \{0,1\}, \end{cases}$$

Figure: Message Packing

$$\mathbf{W}_{j\to i}^{(k)} = \text{MHA}_{\text{W}}\left(\mathcal{F}_i^{(k)}, \mathcal{Z}_{j\to i}^{(k)}, \mathcal{Z}_{j\to i}^{(k)}\right) \odot \mathbf{C}_j^{(k)} \in \mathbb{R}^{H\times W},$$

$$\mathcal{F}_i^{(k+1)} = \text{FFN}\left(\sum_{j\in\mathcal{N}_i\cup\{i\}} \mathbf{W}_{j\to i}^{(k)} \odot \mathcal{Z}_{j\to i}^{(k)}\right) \in \mathbb{R}^{H\times W\times D},$$

Figure: Feature Fusion

LIF: Details

- **2D-3D Projection**: First, 2D detection results are transformed from the image plane to the 3D coordinate system. Second, the framework employs an uncertainty-driven communication mechanism to share high-quality detection results.
- **Vision-guided Positional Embedding (VPE)**: Transform 2D detection results into the ego-agent's BEV and generate a learnable positional embedding map to mark the locations of these targets.
- **Box-based virtual augmented BEV feature (BoBEV).**: Encode the geometric attributes (size, heading) and confidence scores of the received 3D bounding boxes into an augmented feature map.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{z_c} \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix}$$

$$P(r) = r \cdot \left(\frac{u - c_x}{f_x}, \frac{v - c_y}{f_y}, 1 \right)$$

$$P_w = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \cdot P_c$$

$$P_w(r) = T + r \cdot R \cdot \left(\frac{u - c_x}{f_x}, \frac{v - c_y}{f_y}, 1 \right)$$

Figure: Transformation

$$VPE_{m,n} = \begin{cases} Q \in \mathbb{R}^C, & (m,n) \text{ in } P_{bev} \\ 0, & \text{else} \end{cases}$$

$$BoBEV_{m,n} = \begin{cases} b_{3d}^{whl\theta s}, & (m,n) \text{ in } P_{bev} \\ 0, & \text{else} \end{cases}$$

Figure: VPE and CoBEV

STAMP: Details

- **Core Idea: Local Adaptation and Reversion**: First, each agent is equipped with a **local adapter** that maps its unique B3V feature to a unified, **shared protocal feature**. After receiving broadcasted features, each agent uses its **local reverter** to map the incoming protocol features back into its **own feature domain**.
- **For Adapter**: The objective is to minimize the gap between its adapted feature to a ground-truth protocal feature.
- **For Reverter**: Similarly trained to map features from the protocal domain back to agents' local feature.

$$\text{Adaptation:} \quad F_{iP} = \phi_i(F_i), \quad \forall i \in \{1, 2, \ldots, N\}$$

$$\text{Reversion:} \quad F_{ij} = \begin{cases} \psi_j(F_{iP}), & \text{if } i \neq j \\ F_i, & \text{if } i = j \end{cases} \quad \forall i, j \in \{1, 2, \ldots, N\}$$

Figure: Overview

$$\phi_i = \arg\min_{\phi_i} L_{\phi_i}(F_{iP}^{1:K}, F_P^{1:K}) \quad \text{where} \quad F_{iP}^k = \phi_i(F_i^k)$$

$$\psi_i = \arg\min_{\psi_i} \left(L_{\psi_i}(F_{Pi}^{1:K}, F_i^{1:K}) + L_{\psi_i}(F_{ii}^{1:K}, F_i^{1:K}) \right)$$

$$\text{where} \quad F_{Pi}^k = \psi_i(F_P^k), \; F_{ii}^k = \psi_i(F_{iP}^k)$$

Figure: Objective

$$L_{\phi_i}^d = \mathcal{L}_P(D_P \circ U_P(F_{iP}^{1:K}), \, \mathrm{GT}_P)$$
$$L_{\psi_i}^d = \mathcal{L}_i(D_i \circ U_i(F_{Pi}^{1:K}), \, \mathrm{GT}_i) + \mathcal{L}_i(D_i \circ U_i(F_{ii}^{1:K}), \, \mathrm{GT}_i)$$

Figure: Loss Function