CATNet: Collaborative Alignment and Transformation Network for Cooperative Perception

Supplementary Material

1. Further Implementation Details

Our method, implemented within the OpenCood framework [4], processes point cloud data by voxelizing it into a $0.4m \times 0.4m \times 0.4m$ 3D grid. This data is then encoded into 256-channel 2D BEV features using the PointPillars [1], followed by multi-scale processing with a ResNet backbone yielding [64, 128, 256] output channels for high-level semantic information. The model is trained for 30 epochs using the AdamW optimizer [3] (initial learning rate 0.002, weight decay $1e-4$, cosine annealing), employing a weighted combination of Focal Loss [2] and L1 Loss for object detection. LiDAR perception ranges are set specifically for OPV2V/V2XSet ([-140.8, -38.4, -3, 140.8, 38.4, 1]) and DAIR-V2X ([-100.8, -40, -3.5, 100.8, 40, 1.5]), with both datasets involving vehicle and RSU collaborative agents. To address asynchronous latency, the STSync module is introduced for dynamic temporal alignment, fusing recent temporal features guided by Ego features. All experiments are conducted on a server with one NVIDIA 4090 GB GPU, running PyTorch 1.13 + CUDA 11.7.

ST-Gate Implementation Details. The Spatio-Temporal Gate (ST-Gate) adaptively fuses a hidden state $\mathbf{H}$ and an aligned feature $\mathbf{F}$ by combining spatial and channel attention. First, the inputs are concatenated along the channel dimension to aggregate information:

$$\mathbf{Z} = \text{Concat}(\mathbf{H}, \mathbf{F}) \tag{1}$$

Next, a spatial attention map $\mathbf{W}_s$ is computed by pooling channel-wise information and applying a convolution, while a channel attention map $\mathbf{W}_c$ is derived by pooling space-wise information and applying an MLP:

$$\mathbf{W}_s = \sigma\left(\text{Conv}\left(\text{Concat}\left[\text{AvgPool}_c(\mathbf{Z}), \text{MaxPool}_c(\mathbf{Z})\right]\right)\right) \tag{2}$$

$$\mathbf{W}_c = \sigma\left(\text{MLP}\left(\text{AvgPool}_s(\mathbf{Z}) + \text{MaxPool}_s(\mathbf{Z})\right)\right) \tag{3}$$

The two attention maps are then combined to generate a final gating coefficient $\mathbf{z}$:

$$\mathbf{z} = \sigma(\mathbf{W}_s \odot \mathbf{W}_c)$$

Finally, this coefficient is used to perform a weighted fusion on the inputs to produce the output $\mathbf{S}$:

$$\mathbf{S} = (1 - \mathbf{z}) \odot \mathbf{H} + \mathbf{z} \odot \mathbf{F}$$

Here, σ denotes the Sigmoid function and $\odot$ represents element-wise multiplication.

Details of the Multi-Path Scanning Strategy. To comprehensively capture spatial dependencies and cross-subband correlations, we employ a multi-path scanning strategy that processes the wavelet subbands in four distinct orders. As illustrated in Figure 2, the input feature map is first divided into a grid of non-overlapping patches. These patches are then flattened into four different sequences and fed into the State Space Model (SSM). Let the sequence of patches in a standard raster scan order (top-to-bottom, left-to-right) be denoted as $P = (p_1, p_2, \ldots, p_N)$. The four scanning paths are defined as follows:

- **Interleaved Forward Scan (L^+_{inter}):** This path processes the patches in a standard raster scan order, from p_1 to p_N. At each step, features from all four subbands corresponding to a single patch are processed together. This scan captures local contextual information in a conventional forward direction. The sequence is:

$$(p_1, p_2, p_3, \ldots, p_N)$$

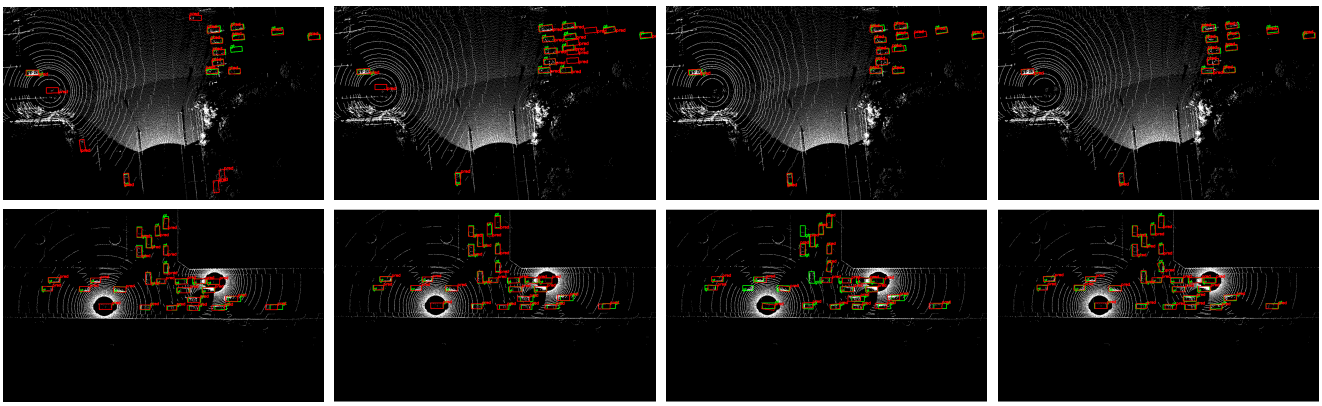
- **Interleaved Backward Scan (L^-_{inter}):** This path is the reverse of the interleaved forward scan, processing patches from p_N to p_1. This allows the model to capture dependencies from a reverse contextual perspective. The sequence is:

$$(p_N, p_{N-1}, p_{N-2}, \ldots, p_1)$$

- **Progressive Forward Scan (L^+_{prog}):** This path follows the strategy described in the main text, prioritizing the integration of features across subbands. It sequentially processes all patches from one subband before moving to the next, following a high-to-low frequency order ($F_{HH} \rightarrow F_{HL} \rightarrow F_{LH} \rightarrow F_{LL}$). Within each subband, patches are processed in a standard raster scan. This path is designed to first correct high-frequency details and then progressively incorporate lower-frequency context.
- **Progressive Backward Scan (L^-_{prog}):** This is the reverse of the progressive forward scan. It processes subbands in a low-to-high frequency order ($F_{LL} \rightarrow F_{LH} \rightarrow F_{HL} \rightarrow F_{HH}$) and reverses the spatial scan direction within each subband.

2. Hyperparameter Analysis.

Table 1 illustrates the trade-off between historical frame count and latency on DAIR-V2X. The experimental results demonstrate that the model consistently maintains high accuracy across different latency levels and historical data



(a) CoAlign (b) ERMVP (c) V2X-ViT (d) CATNet

Figure 1. Qualitative evaluation under latency and noise conditions, where green and red bounding boxes denote ground truth annotations and detection results respectively. The proposed approach demonstrates improved detection accuracy with better alignment and fewer detection errors.

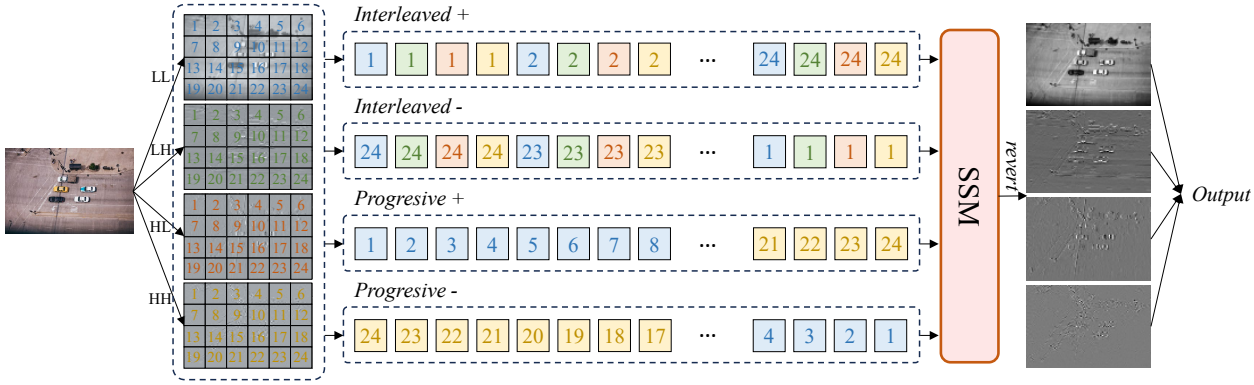


Figure 2. Scan Order of WTMamba.

Frames/Delay	200ms	300ms	400ms	500ms
AP@0.5				
1	0.7242	0.7180	0.7134	0.7089
2	0.7244	0.7178	0.7140	0.7093
3	0.7257	0.7200	0.7159	0.7071
4	0.7227	0.7190	0.7129	0.7054
5	0.7178	0.7114	0.7081	0.7032
AP@0.7				
1	0.5740	0.5748	0.5760	0.5747
2	0.5750	0.5751	0.5754	0.5751
3	0.5754	0.5761	0.5759	0.5747
4	0.5756	0.5765	0.5757	0.5750
5	0.5759	0.5753	0.5750	0.5745

Table 1. Comparison of accuracy under different latencies and historical data on the DAIR-V2X dataset.

lengths. Through systematic evaluation, we ultimately selected 3 frames of historical data as the optimal configuration.

3. Computational Efficiency Breakdown

We analyze the computational efficiency of CATNet under various historical frame settings ($K \in [2, 5]$). As detailed in Table 2, the computational cost (GFLOPs) and inference latency (ms) primarily scale with the STSync module, while the WTDen and AdpSel modules maintain constant overheads of 22.6 and 11.7 GFLOPs, respectively. Even at $K = 5$, CATNet achieves a real-time inference speed of 91.5 ms. Notably, at $K = 2$, our end-to-end (E2E) cost of 334.1 GFLOPs is significantly more efficient than *How2comm* (741.34 GFLOPs) and comparable to *V2VNet*, demonstrating a superior balance between spatio-temporal alignment accuracy and system efficiency.

3.1. 3. Qualitative Results

Figure 1 compares detection performance under communication loss and noise interference. While existing meth-

Table 2. Efficiency analysis of different modules under varying historical frame settings. E2E: End to end consumption.

Frames	Computational Cost (GFLOPs)					Inference Time (ms)			
	STSync	WTDen	AdpSel	E2E	*vs. SOTA*	STSync	WTDen	AdpSel	E2E
2	195.5	22.6	11.7	334.1	V2X-ViT: 281.0 V2VNet: 496.7 How2comm: 741.34	25.1	18.8	3.3	64.3
3	272.1	22.6	11.7	410.7		35.0	18.9	3.3	76.1
4	323.7	22.6	11.7	462.8		41.1	18.9	3.3	84.0
5	367.0	22.6	11.7	578.1		46.0	18.9	3.3	91.5

ods suffer from notable false positives/misses, CATNet improves robustness via temporal feature enhancement and adaptive noise suppression. Its selective feature enhancement module further sharpens critical region responses, boosting detection accuracy significantly.

References

[1] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12697–12705, 2019. 1

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 1

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 1

[4] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *International Conference on Robotics and Automation (ICRA)*, pages 2583–2589, 2022. 1