Paper : **V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer**

Runsheng Xu , Hao Xiang , Zhengzhong Tu2, Xin Xia,
Ming-Hsuan Yang, and Jiaqi Ma

8th Aug -2022

Presented by
Naay, Danny
GNR 650

Goal

Perceiving the complex driving environment precisely is crucial to the safety of autonomous vehicles (Avs)

Goal is to develop a robust fusion system to enhance the vehicle’s perception capability and handle real world challenges in a unified end-to-end fashion.

Motivation

Single-vehicle perception systems => Demonstrated significant improvement in several tasks such as semantic segmentation, depth estimation and object detection and tracking

- Challenges remain. Single-agent perception system tends to suffer from occlusion and sparse sensor observation at a far distance, which can potentially cause catastrophic consequences.

Vehicle-to-Vehicle (V2V) collaboration => Visual information from multiple nearby AVs are shared for a complete and accurate understanding of the environment.

- Challenge remains : This ignores a critical collaborator – roadside infrastructure

Deploying a robust V2X perception system is non-trivial as it forms heterogeneous graph formed by infrastructure and AVs, unlike homogeneous graphs in V2V.

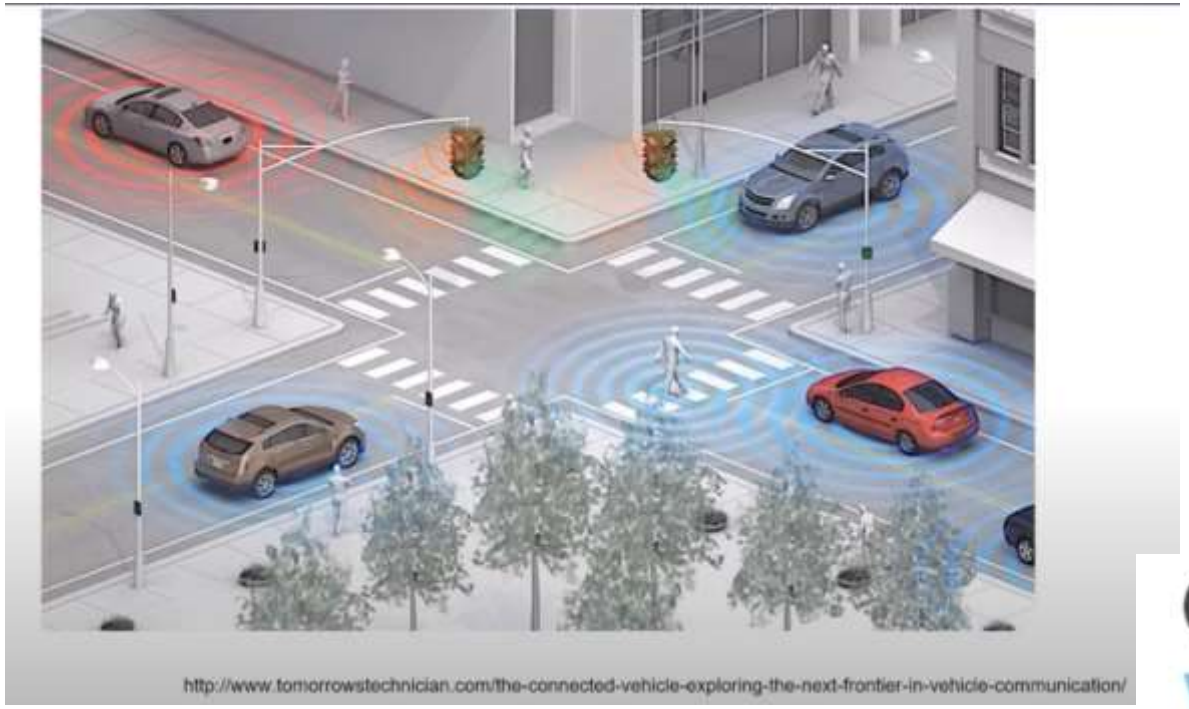
V2X Challenge

Configuration discrepancies between infrastructure and vehicle sensors, such as types, noise levels, installation height, and even sensor attributes and modality.

GPS localization noises and the asynchronous sensor measurements of AVs and infrastructure can introduce inaccurate coordinate transformation and lagged sensing information.

Asynchronous information sharing, pose errors, and heterogeneity of V2X components

V2X



http://www.tomorrowstechnician.com/the-connected-vehicle-exploring-the-next-frontier-in-vehicle-communication/

V2V

Vehicle-to-vehicle (direct communications between vehicles, primarily for safety)

V2I

Interaction between vehicles and infrastructure such as traffic signals

V2P

Interaction between vehicles and pedestrians (and cyclists)

V2X

Vehicle-to-everything (all of the above)

What is cooperative perception framework

How to efficiently fuse visual cues from neighboring agents.

Based on its message sharing strategy it is divided into 3 categories :

1. Early fusion : Where raw data is shared and gathered to form a holistic view.
2. Intermediate fusion : Where intermediate neural features are extracted based on each agent’s observation and then transmitted.
3. Late fusion : Where detection outputs are shared.

Intermediate fusion has attracted increasing attention because of its good balance between accuracy and

What is cooperative perception framework .. cont

Intermediate fusion methods for **V2V perception :**
OPV2V -> implements a single-head self-attention module to fuse features,
F-Cooper -> employs maxout fusion operation.
V2VNet -> proposes a spatial-aware message passing mechanism to jointly reason detection and prediction.

Here, regresses vehicles localization errors with consistent pose constraints

Most V2X methods explored late fusion strategies to aggregate information from infrastructure and vehicles

What is cooperative perception framework .. Alternati

LiDAR-based 3D object detection:

Numerous methods have been explored to extract features from raw points, voxels, bird-eye-view (BEV) images, and their mixtures.

PointRCNN proposes a two-stage strategy based on raw point clouds, which learns rough estimation in the first stage and then refines it with semantic attributes.

Despite having high accuracy, their inference speed and memory consumption are difficult to optimize due to reliance on 3D convolutions.

Solution - Scope

V2X - VIT :

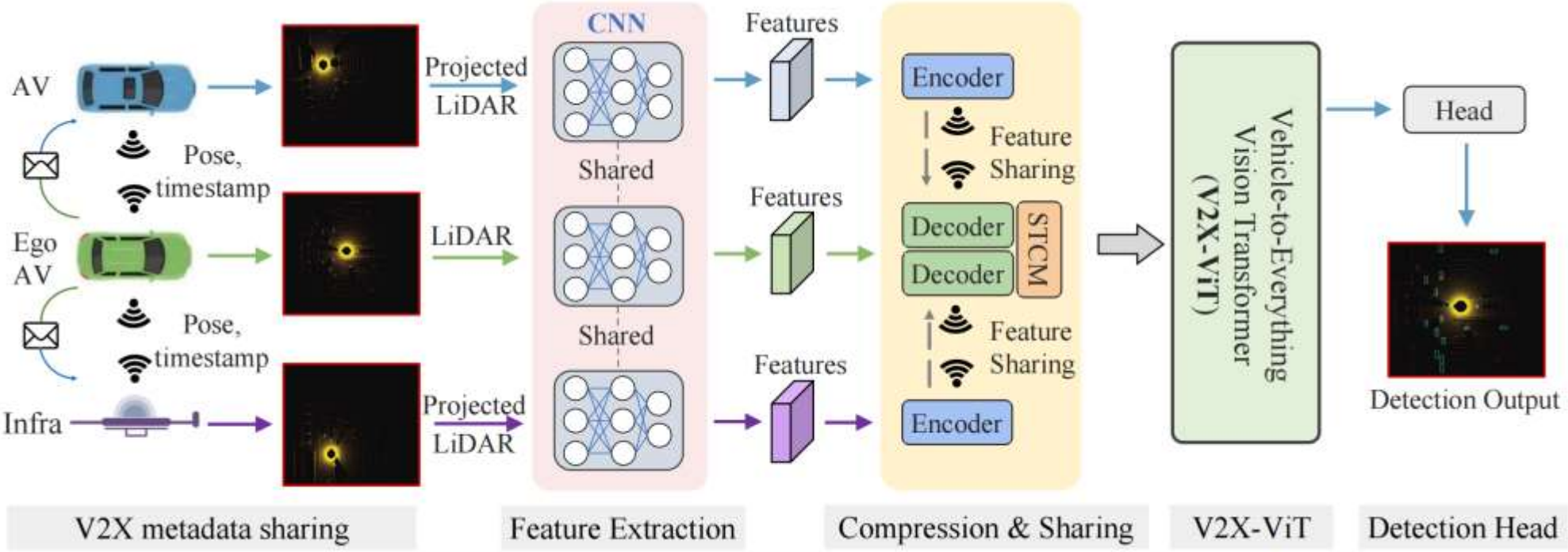
Here, AVs and Infrastructure capture, encode, compress, and send intermediate visual features with each other, while the ego vehicle (i.e., receiver) employs V2X-Transformer to perform information fusion for object detection

1) A customized heterogeneous multi-agent self-attention module that explicitly considers agent types (vehicles and infrastructure) and their connections when performing attentive fusion.

2) A multi-scale window attention module that can handle localization errors by using multi-resolution windows in parallel.

3) Also integrate a delay-aware positional encoding to handle the time
delay uncertainty further.

V2X -ViT



Given all these definitions, the entire V2X-ViT model can be formulated as:

$$z_i = \text{PointPillar}(x_i), \qquad x_i \in \mathbb{R}^{P\times 4}, z_i \in \mathbb{R}^{H\times W\times C} \quad \text{for agent } i \tag{20}$$

$$\mathbf{z}_0 = \text{STCM}([z_0, ..., z_M]) + \text{DPE}([\Delta t_0, ..., \Delta t_M]), \quad \text{for ego AV} \tag{21}$$

$$\mathbf{z}'_\ell = \mathbf{z}_{\ell-1} + \text{MSwin}(\text{HSMA}(\text{LN}(\mathbf{z}_0))), \quad \mathbf{z}_0 \in \mathbb{R}^{M\times H\times W\times C} \quad \ell = 1, ..., L \tag{22}$$

$$\mathbf{z}_\ell = \mathbf{z}'_\ell + \text{MLP}(\text{LN}(\mathbf{z}'_\ell)), \quad \ell = 1, ..., L \tag{23}$$

$$\mathbf{y} = \text{Head}(\mathbf{z}_L), \tag{24}$$

Why NOT ViT

The full self attention in ViT ,

> having global interaction,

> locality into self-attention, such as Swin, CSwin, Twins, window, and sparse attention.

However, their efficacy to represent heterogeneous graphs has been less studied.

Scope

Build a robust cooperative perception framework with V2X communication using a novel Vision Transformer. Build a holistic attention model, namely V2X-ViT, to effectively fuse information across on-road agents namely AV and Infrastructure

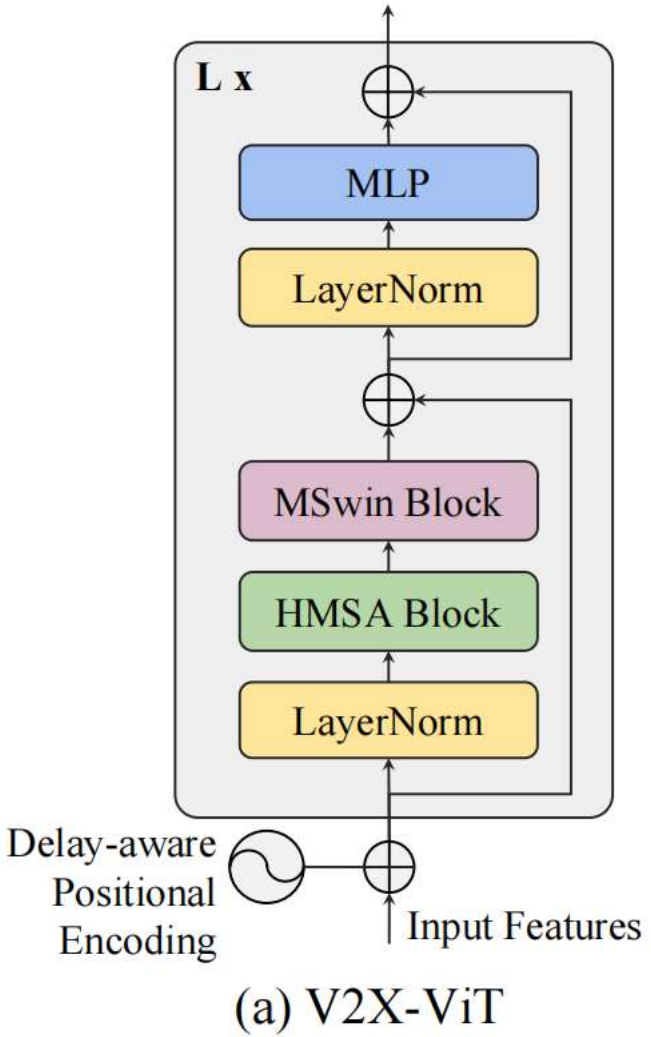
V2X-ViT consists of alternating layers of:

- > heterogeneous multiagent self-attention and
- > multi-scale window self-attention,

Inspired by Hetrogeneous Graph Transfomer, build a customized heterogeneous multi-head self-attention module tailored for graph attribute-aware multi-agent 3D visual feature fusion, which is able to capture the heterogeneity of V2X systems.

Architecture At A Glan

5 components
1) Metadata sharing,
2) Feature extraction,
3) Compression and sharing
4) V2X vision Transformer a
5) Detection head



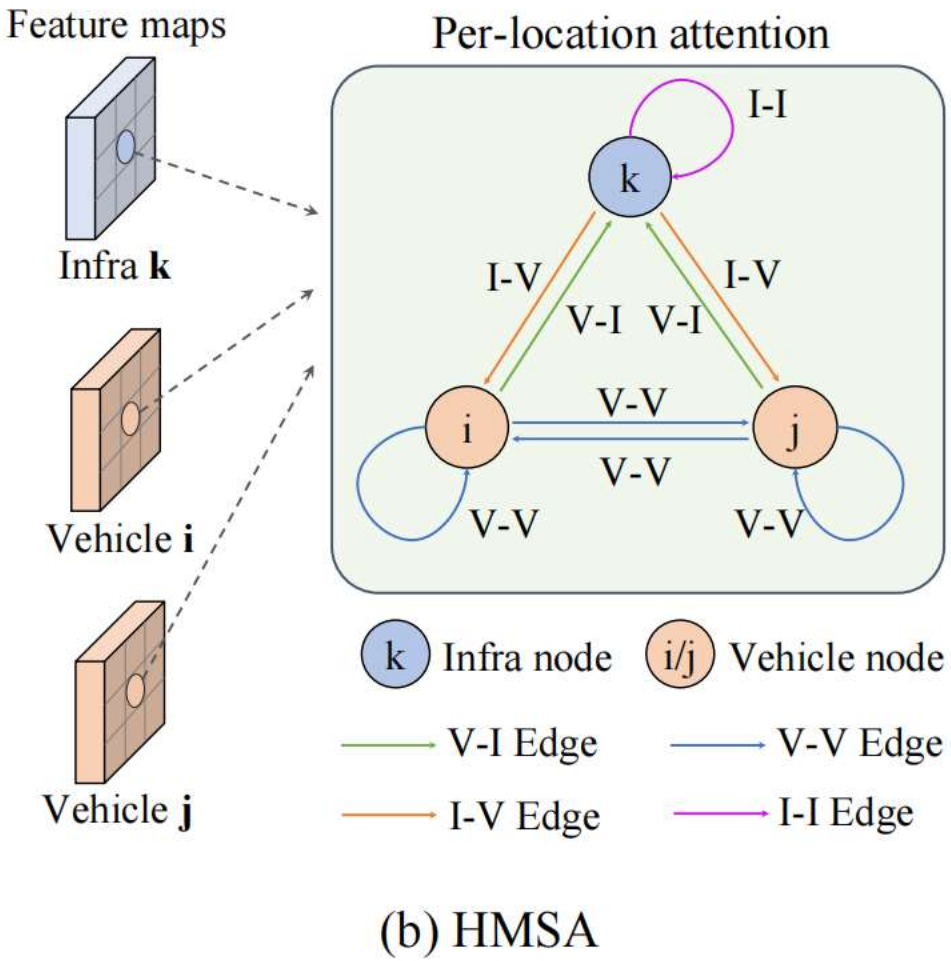
(a) V2X-ViT

V2X – ViT Architecture

Heterogeneous multi-agent self-attention :: HMSA

Learns different relationships based on node and edge types to effectively capture the heterogeneous graph representation

We have two types of nodes {I, V} and four types of edges {V-V, V-I, I-V, I-I}
Unlike traditional attention where the node features are treated as a vector.



(b) HMSA

V2X - ViT Architecture - HSMA

$$\mathbf{H}_i = \underset{\forall j \in N(i)}{\mathsf{Dense}_{c_i}} \left(\mathbf{ATT}\left(i,j\right) \cdot \mathbf{MSG}\left(i,j\right) \right)$$

$$\mathbf{ATT}\left(i,j\right) = \underset{\forall j \in N(i)}{\mathsf{softmax}} \left(\underset{m \in [1,h]}{\Vert} \mathrm{head}^m_{\mathrm{ATT}}\left(i,j\right) \right) \tag{2}$$

$$\mathrm{head}^m_{\mathrm{ATT}}\left(i,j\right) = \left(\mathbf{K}^m\left(j\right) \mathbf{W}^{m,\mathrm{ATT}}_{\phi(e_{ij})} \mathbf{Q}^m\left(i\right)^T \right) \frac{1}{\sqrt{C}} \tag{3}$$

$$\mathbf{K}^m\left(j\right) = \mathsf{Dense}^m_{c_j}\left(\mathbf{H}_j\right) \tag{4}$$

$$\mathbf{Q}^m\left(i\right) = \mathsf{Dense}^m_{c_i}\left(\mathbf{H}_i\right) \tag{5}$$

$$\mathbf{MSG}\left(i,j\right) = \underset{m \in [1,h]}{\Vert} \mathrm{head}^m_{\mathrm{MSG}}\left(i,j\right)$$

$$\mathrm{head}^m_{\mathrm{MSG}}\left(i,j\right) = \mathsf{Dense}^m_{c_j}\left(\mathbf{H}_j\right) \mathbf{W}^{m,\mathrm{MSG}}_{\phi(e_{ij})}.$$

To incorporate the semantic meaning of edges we multiple with weight
Separate Dense model for each vehicle
and infrastructure

`Matrix to` is used to project the features based on the edge type between source node and target node

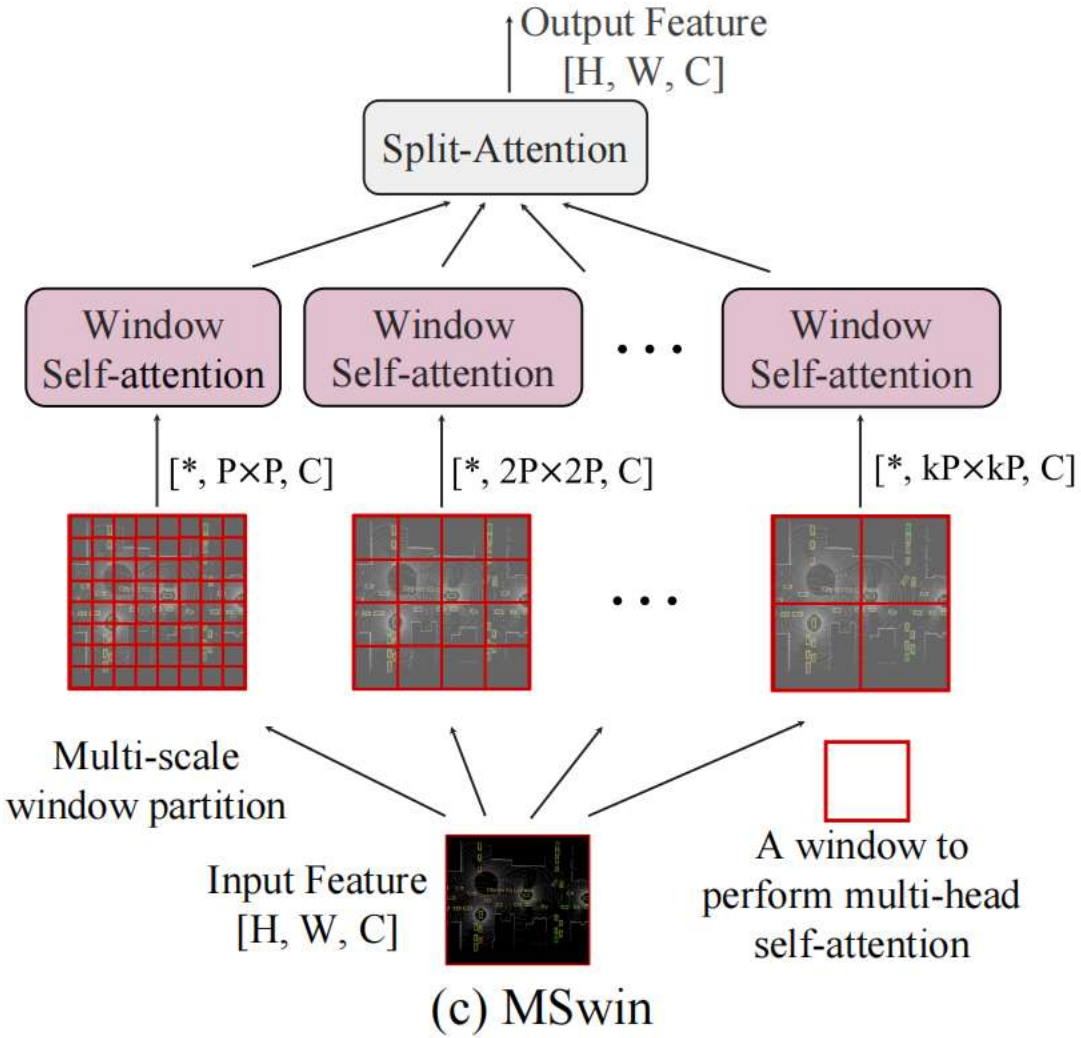
`Linear Aggregator/Projectors indexed by node : Dense [ci]`
`Attention calculates important weights between pair of nodes based on edges : ATT`

V2X – ViT Architecture

Multi-scale window attention : (Mswin)

New type of attention mechanism tailored for efficient long-range spatial interaction on high-resolution detection.

MSwin uses multiple window sizes to aggregate spatial information, which greatly improves the detection robustness against localization errors.



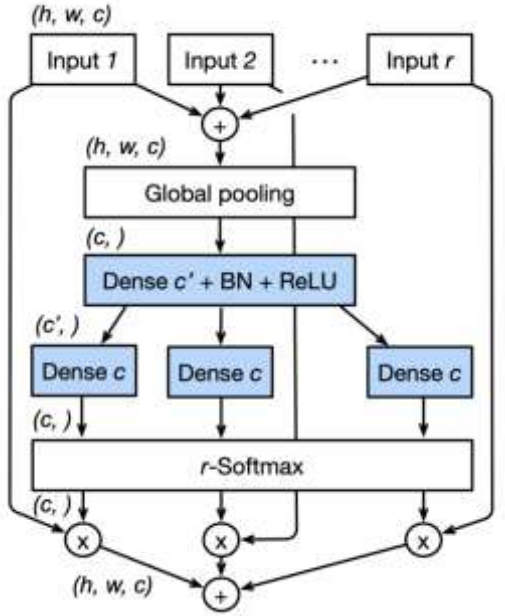
V2X – ViT Architecture - MSWIM

$$\mathbf{H} = [\mathbf{H}^1, \mathbf{H}^2, ..., \mathbf{H}^{HW/(P_j)^2}], \qquad \text{for branch } j$$
$$\hat{\mathbf{H}}_m^i = \text{Attention}(\mathbf{H}^i\mathbf{W}_m^Q, \mathbf{H}^i\mathbf{W}_m^K, \mathbf{H}^i\mathbf{W}_m^V), \qquad i = 1, ..., HW/(P_j)^2$$
$$\mathbf{Y}_m = [\hat{\mathbf{H}}_m^1, \hat{\mathbf{H}}_m^2, ..., \hat{\mathbf{H}}_m^{HW/(P_j)^2}], \qquad m = 1, ..., h_j$$
$$\mathbf{Y}^j = [\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_{h_j}],$$
$$\mathbf{Y} = \text{SplitAttention}(\mathbf{Y}^1, \mathbf{Y}^2, ...\mathbf{Y}^k),$$

H be an input feature of a single agent
H [i,m] is the self attention for each branch
Y [*m]* is the output of the *m*-th head for branch *j*.
Y [j] the outputs for all heads 1, 2, ..., *hj* are concatenated

The Attention is alike Swin, which denotes relative self-a

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}((\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} + \mathbf{B})\mathbf{V})$$



The Split Attention block enables attention across feature-map groups.

V2X - ViT Architecture

Delay-aware positional encoding:
To encode this temporal information, we leverage an adaptive delay-aware positional encoding (DPE), composed of a linear projection and a learnable embedding.

We initialize it with sinusoid functions conditioned on time delay and channel and then a linear projection will further warp the learnable embedding so it can generalize better for unseen time delay.

We add this projected embedding to each agents' feature before feeding into the Transformer so that the features are temporally aligned beforehand

V2X - ViT Architecture

$$\mathbf{p}_c\left(\Delta t_i\right)=\begin{cases}\sin\left(\Delta t_i/10000^{\frac{2c}{C}}\right), & c=2k\\ \cos\left(\Delta t_i/10000^{\frac{2c}{C}}\right), & c=2k+1\end{cases}$$

$$\mathrm{DPE}\left(\Delta t_i\right)=f\left(\mathbf{p}\left(\Delta t_i\right)\right)$$

$$\mathbf{H}_i=\mathbf{H}_i+\mathrm{DPE}\left(\Delta t_i\right)$$

`Function f` will further warp the learnable embedding so it can generalize better for unseen time delay. We add this projected embedding to each agents' feature **H***i* before feeding to the Transformer

Test Data under 2 settings

Novel large-scale V2X perception dataset using that explicitly considers real-world noises during V2X communication, using the high-fidelity simulator CARLA & Cooperative driving automation simulation tool OpenCDA

In total, there are 11,447 frames in our dataset (33,081 samples if we count frames per agent in the same scene), and the train/validation/test splits are 6,694/1,920/2,833, respectively

Adam optimizer with an initial learning rate of 0.001 and steadily decay it every 10 epochs using a factor of 0.1. All models are trained on Tesla V100.

1) Perfect Setting, where the pose is accurate, and everything is synchronized across agents;

Test Data under 2 settings .. cont

The communication range of each agent is set as 70 m, whereas all the agents out of this broadcasting radius of ego vehicle is ignored.

During Training - A random AV is selected as the ego vehicle,
During Testing - Evaluate on a fixed ego vehicle for all the compared models.

For the PointPillar backbone, we set the voxel resolution to 0.4 m for
both height and width

Our V2X-ViT has 3 encoder layers with 3 window sizes in MSwin: 4, 8, and 16

Results

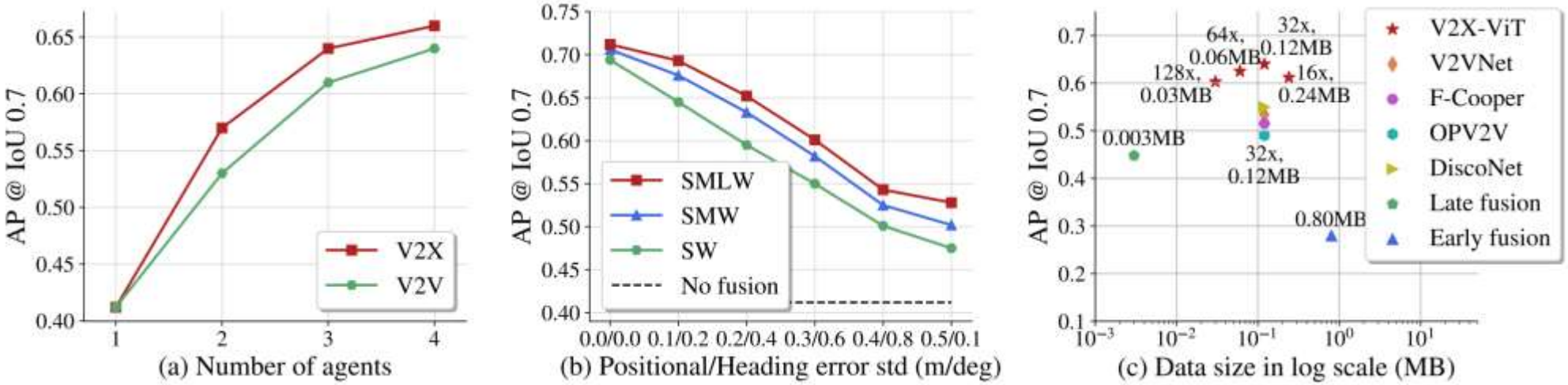
Experimental results demonstrate that V2XViT sets new state-of-the-art performance for 3D object detection and achieves robust performance even under harsh, noisy environments.

The detection performance is measured with Average Precisions (AP) at Intersection-over-Union (IoU) thresholds of 0.

Table 1: **3D detection performance comparison on V2XSet.** We show Average Precision (AP) at IoU=0.5, 0.7 on *Perfect* and *Noisy* settings, respectively.

	Perfect		Noisy	
Models	AP0.5	AP0.7	AP0.5	AP0.7
No Fusion	0.606	0.402	0.606	0.402
Late Fusion	0.727	0.620	0.549	0.307
Early Fusion	0.819	0.710	0.720	0.384
F-Cooper [5]	0.840	0.680	0.715	0.469
OPV2V [50]	0.807	0.664	0.709	0.487
V2VNet [45]	0.845	0.677	0.791	0.493
DiscoNet [25]	0.844	0.695	0.798	0.541
V2X-ViT (Ours)	**0.882**	**0.712**	**0.836**	**0.614**

Quantitative evaluation



Note :

b:
i) using a single small window branch (SW),
ii) using a small and a middle window (SMW), and
iii) using all three window branches (SMLW).

c:
With the default compression rate (32x), our V2X-ViT outperforms other intermediate fusion methods substantial
Even under a 128x compression rate, our model can still maintain high performance.
[V2X-ViT is insensitive to large compression rate.]

Ablation Studies

MSwin	SpAttn	HMSA	DPE	AP0.5 / AP0.7
				0.719 / 0.478
✓				0.748 / 0.519
✓	✓			0.786 / 0.548
✓	✓	✓		0.823 / 0.601
✓	✓	✓	✓	**0.836** / **0.614**

Component ablation study

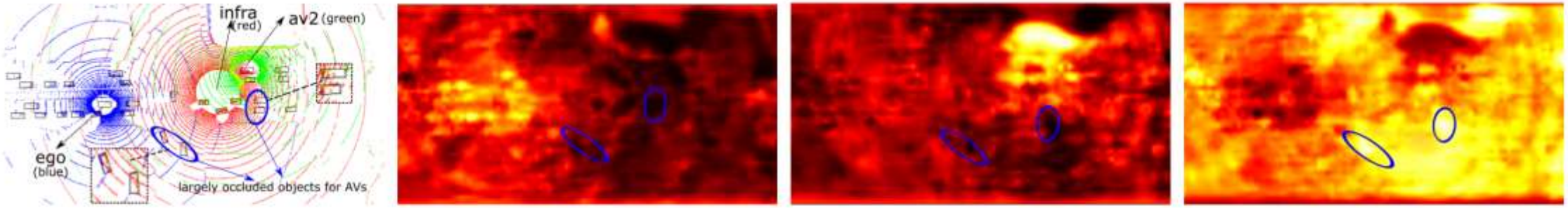
Delay/Model	w/o DPE	w/ DPE
100 ms	0.639	0.650
200 ms	0.558	0.572
300 ms	0.496	0.514
400 ms	0.458	0.478

Effect of DPE w.r.t. time delay

Model	Time	AP0.7(prf/nsy)
V2X-ViT$_S$	28ms	0.696 / 0.591
V2X-ViT	57ms	0.712 / 0.614

Inference Time measurement
1 encoder
3 encoders

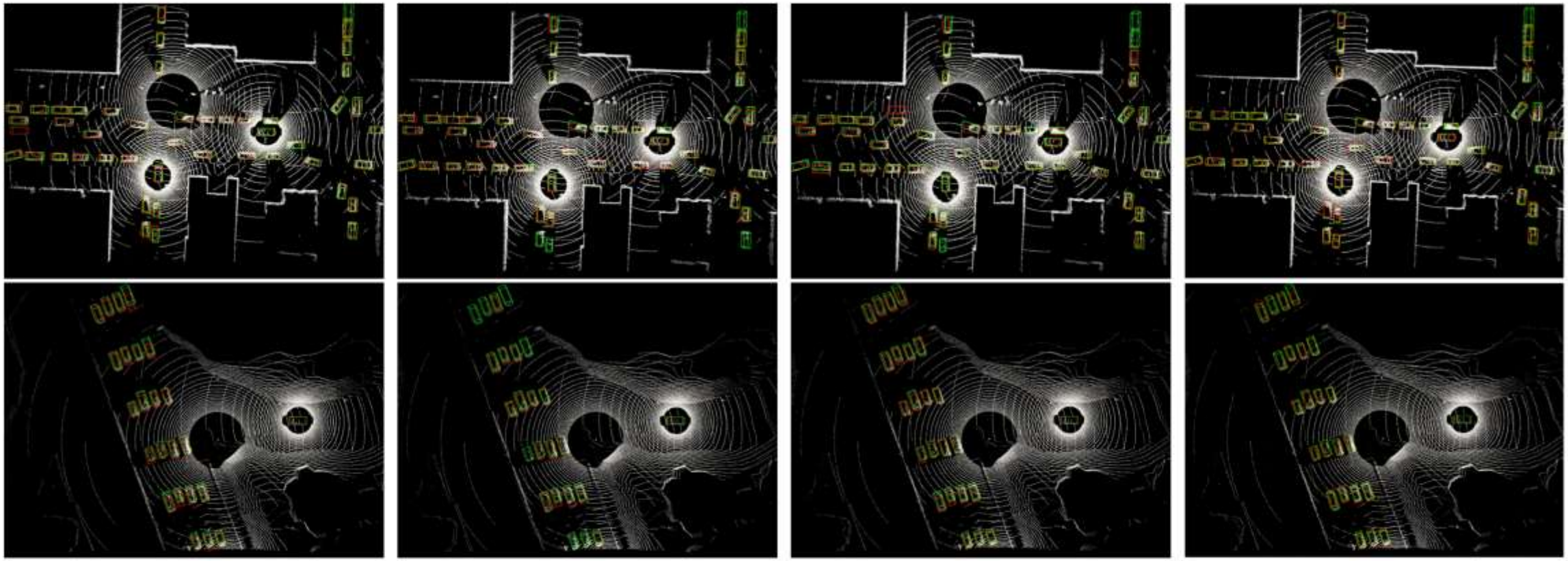
Qualitative evaluation



(a) LiDAR points (better zoom-in) (b) attention weights ego paid to ego (c) attention weights ego paid to av2 (d) attention weights ego paid to infra

Aggregated LiDAR points and attention maps for ego.

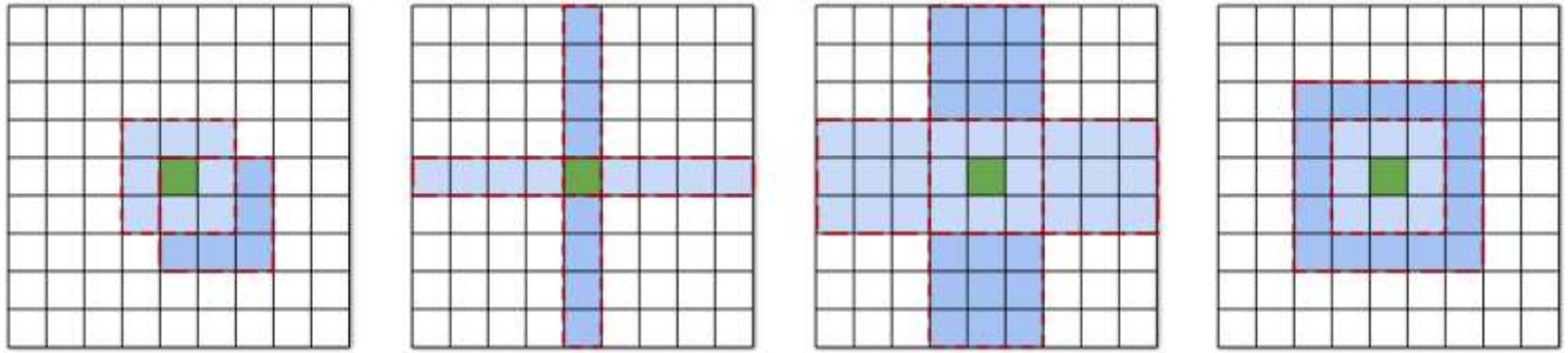
Qualitative evaluation



(a) OPV2V [50] (b) V2VNet [45] (c) DiscoNet [25] (d) V2X-ViT (ours)

Computation Complexity

Attention Models	Complexity
ViT [9]	$\mathcal{O}(4HWC^2 + 2(HW)^2C) \sim \mathcal{O}(N^2)$
Axial [44]	$\mathcal{O}(HWC(4C + H + W)) \sim \mathcal{O}(N\sqrt{N})$
Swin [30]	$\mathcal{O}(4HWC^2 + 2P^2HWC) \sim \mathcal{O}(N)$
CSwin [8]	$\mathcal{O}(HWC(4C + sH + sW)) \sim \mathcal{O}(N\sqrt{N})$
MSwin (ours)	$\mathcal{O}(\frac{1}{3}k^3P^2HWC + \frac{2HWC^2k^2}{h}) \sim \mathcal{O}(N)$



(a) Swin (b) Axial (b) CSwin (c) MSwin

Fig. 8: Visualizations of approximated receptive fields (blue shaded pixels) for the green pixel for (a) Swin [30] (b) Axial [44], (c) CSwin [8] and (d) MSwin attention. MSwin obtains multi-scale long-range interactions at linear complexity.

Research Gap

- Assumed the sensor setups among the same category of agents are identical in HMSA.
- The time delay is set to 100 ms for all the evaluated models to have a fair comparison of their robustness against asynchronous message propagation.
- Models trained on simulated datasets, there are known issues on data bias and generalization ability to real-world scenarios.
- Although the design choice of our communication approach (i.e., project LiDAR to others at the beginning) has an advantage of accuracy, its scalability is limited.
- Future research on fairness, privacy, and robustness in

Conclusion

- HMSA for adaptive information fusion between heterogeneous agents.
- MSwin that simultaneously captures local and global spatial feature interactions in parallel
- V2XSet, a new large-scale open simulation dataset for V2X perception, which explicitly accounts for imperfect real-world conditions.
- Experiments show that the proposed V2X-ViT significantly advances the performance on V2X LiDAR-based 3D object detection, achieving a 21.2% gain of AP compared to single-agent baseline and performing favorably against leading intermediate fusion methods by at least 7.3%
- Compared with existing datasets, V2XSet incorporates both V2X cooperation and realistic noise simulation.

Detailed Architecture

Table T2: **Detailed architectural specifications for V2X-ViT.**

	Output size	V2X-ViT framework
PointPillar Encoder	$M \times 352 \times 96 \times 256$	[Voxel samp. reso. 0.4m, Scatter, 64] [Conv3x3, 64, stride 2, BN, ReLU] $\times$ 3 [Conv3x3, 128, stride 2, BN, ReLU] $\times$ 5 [Conv3x3, 256, stride 2, BN, ReLU] $\times$ 8 [ConvT3x3, 128, stride 1, BN, ReLU] $\times$ 1 [ConvT3x3, 128, stride 2, BN, ReLU] $\times$ 1 [ConvT3x3, 128, stride 4, BN, ReLU] $\times$ 1
	$M \times 176 \times 48 \times 256$	[Concat3, 384] [Conv3x3, 256, stride 2, ReLU Conv3x3, 256, stride 1, ReLU] $\times$ 1
Delay-aware Pos. Encoding	$M \times 176 \times 48 \times 256$	[sin-cos pos. encoding] [Linear, 256] $\times$ 1
Transformer Backbone	$M \times 176 \times 48 \times 256$	[HSMA, dim 256, head 8 MSwin, dim 256, head {16, 8, 4}, ws. $\{4 \times 4, 8 \times 8, 16 \times 16\}$ MLP, dim 256] $\times$ 3
Detection Head	$176 \times 48 \times 16$	Cls. head: [Conv1x1, 2, stride 1] Regr. head: [Conv1x1, 14, stride 1]

Spatial-Temporal Attention Module

$$\begin{bmatrix} x_s \\ y_s \end{bmatrix} = \Gamma_\xi \left(\begin{bmatrix} x_t \\ y_t \\ 1 \end{bmatrix} \right) = \begin{bmatrix} R_{11} & R_{12} & \delta x \\ R_{21} & R_{22} & \delta y \end{bmatrix} \begin{bmatrix} x_t \\ y_t \\ 1 \end{bmatrix}$$

Due to the time delay, the ego vehicle observes the data at a different time *t[e]*. To correct this misalignment between the received features and ego-vehicle's features, a global transformation in terms of a differential 2D transformation is applied, to warp the intermediate features spatially.