Collaborative Perception – An Introduction

EEE 6712

Presented by: Ghayoor Shah,
Mahdi Zaman

UCF

Collaborative Perception

- Sharing and fusion of sensory info (cameras, LiDAR) using communication (V2X) to enhance the overall perception.
- Single-agent perception systems face challenges:
  - Occlusion
  - Sparse sensor observations at far distances
  - Limited line-of-sight
- Multi-agent perception using cooperation can address these challenges:
  - Vehicle-vehicle communication (V2V)
  - Vehicle-infrastructure communication (V2I, I2V)
- Downstream Tasks of Collaborative Perception:
  - 3D Object Detection
  - Semantic Segmentation

Limitations of Single-agent Perception:

- Vehicles are constrained by cost and space limitations. Often equipped with low-precision sensors and low-power computing devices
- Single vehicle can only have limited sight-of-view due to obstruction of other vehicles and obstacles
- Long-range objects exhibit sparsity in sensor data, making it prone to erroneous perception

Collaborative Perception - Benefits

- Through obtaining extra perception info from other vehicles and infrastructures, vehicles can overcome the occlusion and long-range perception issues faced by individual perception and achieve beyond line-of-sight perception capability.
- Vehicles can leverage the powerful computing resources on the cloud platform by V2I/V2N to efficiently execute large-scale and regularly updated perception models.

Performance Comparison – Qualitative (DAIR-V2X Dataset - Real)

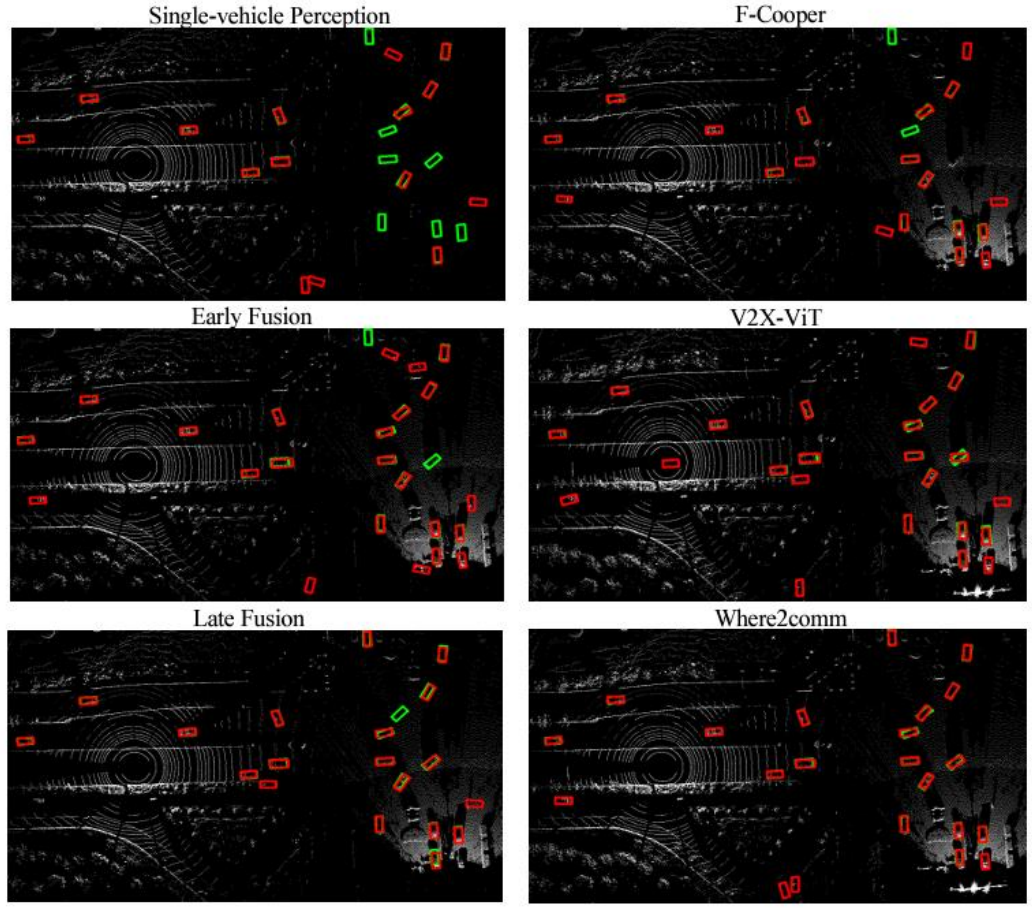


Fig. 8. Visualizations on DAIR-V2X-C.

Performance Comparison – Qualitative (V2XSet Dataset - Simulated)

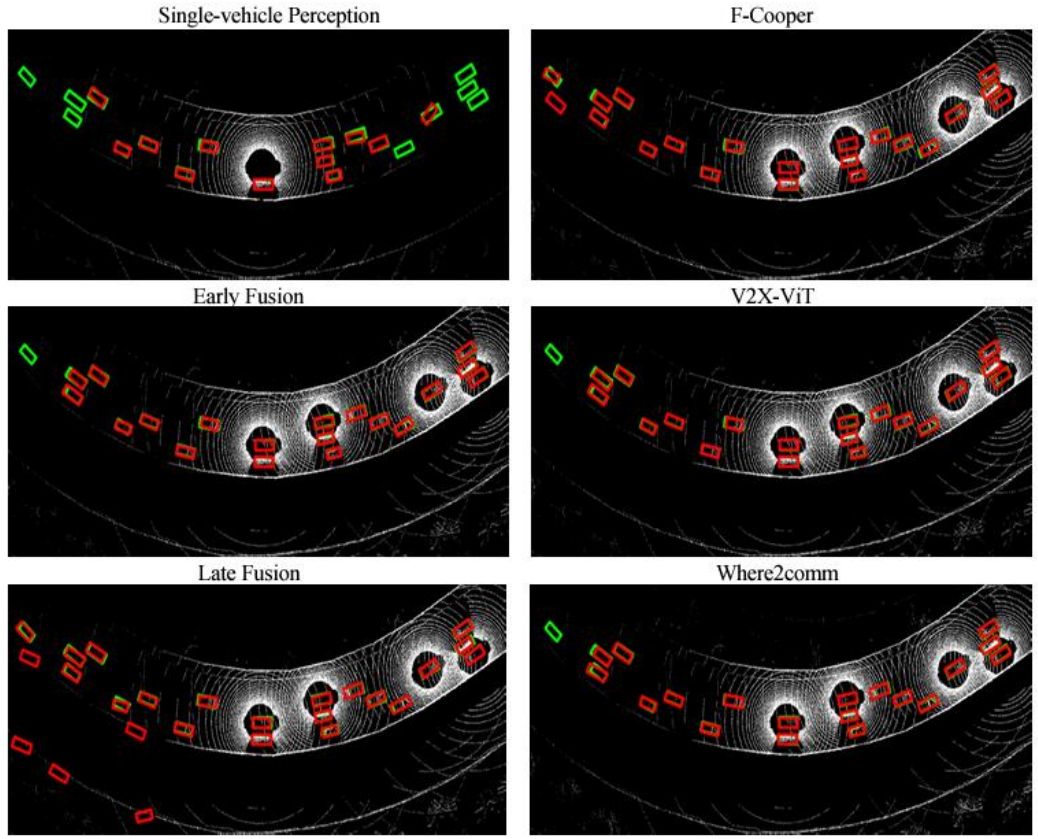


Fig. 9. Visualizations on V2XSet.

Collaborative Perception – Requirements, Stack

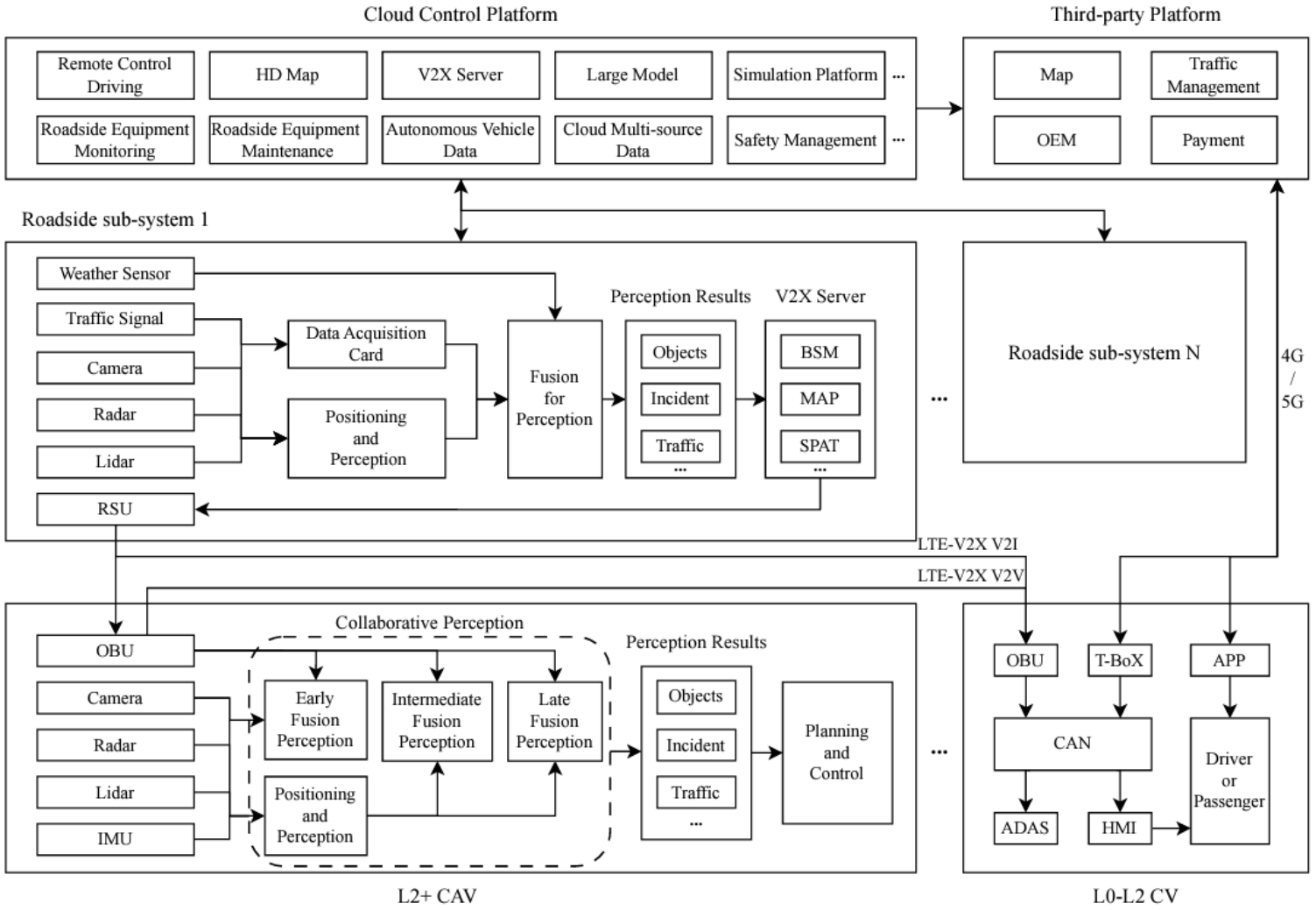


Fig. 3. The illustration of a typical V2X system architecture in practical applications.

Collaborative Perception – History

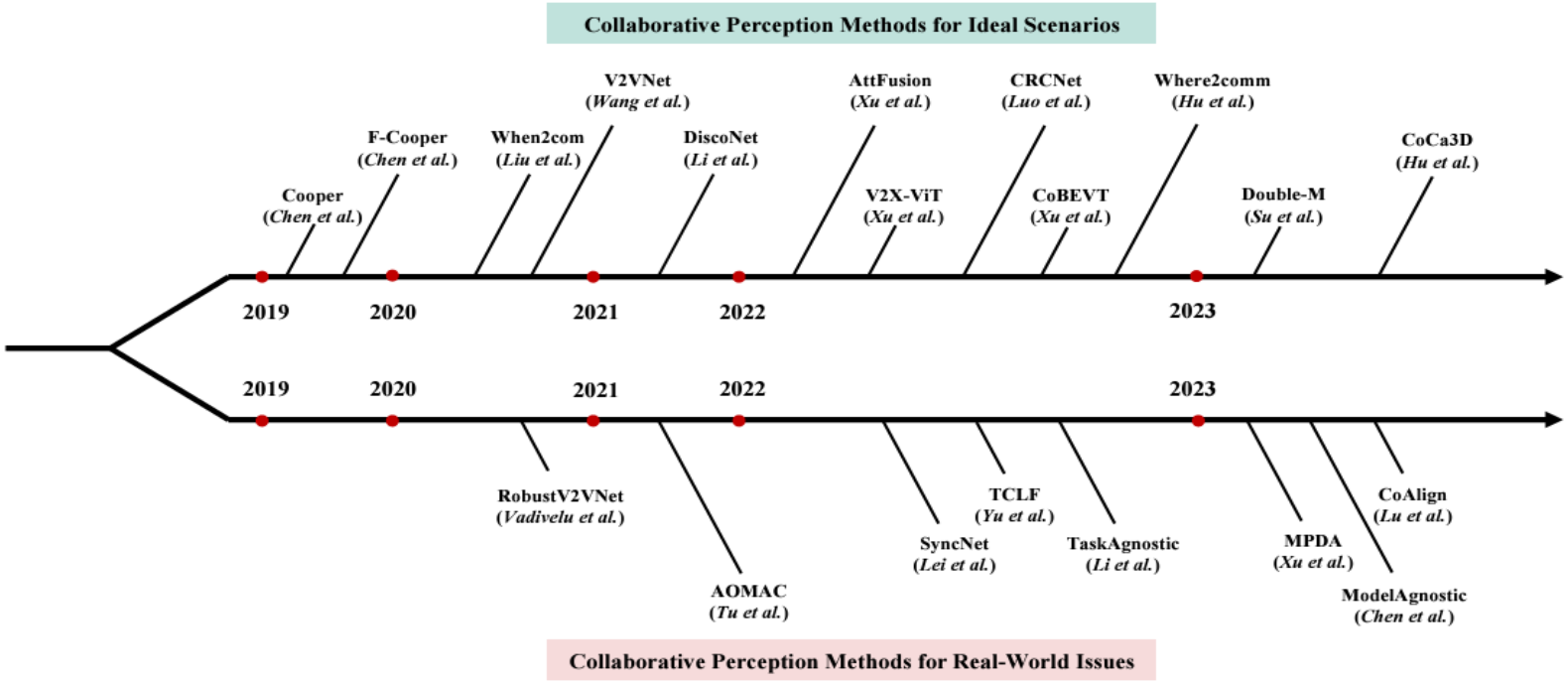


Fig. 2. Typical collaborative perception methods in autonomous driving are classified from two perspectives: 1) how to design common collaboration modules in ideal scenarios, which focus on collaboration efficiency and performance, and 2) how to address issues in real applications, which focus on robustness and safety. We categorized methods based on their most prominent contribution. Citation: 1) Cooper [11], F-Cooper [10], When2com [41], V2VNet [68], DiscoNet [38], AttFusion [79], V2X-ViT [78], CRCNet [44], CoBEVT [76], Where2comm [26], Double-M [59], CoCa3D [27], 2) RobustV2VNet [62], AOMAC [61], SyncNet [31], TCLF [82], TaskAgnostic [39], MPDA [75], ModelAgnostic [12], CoAlign [43].

UCF

Collaborative Perception – V2X

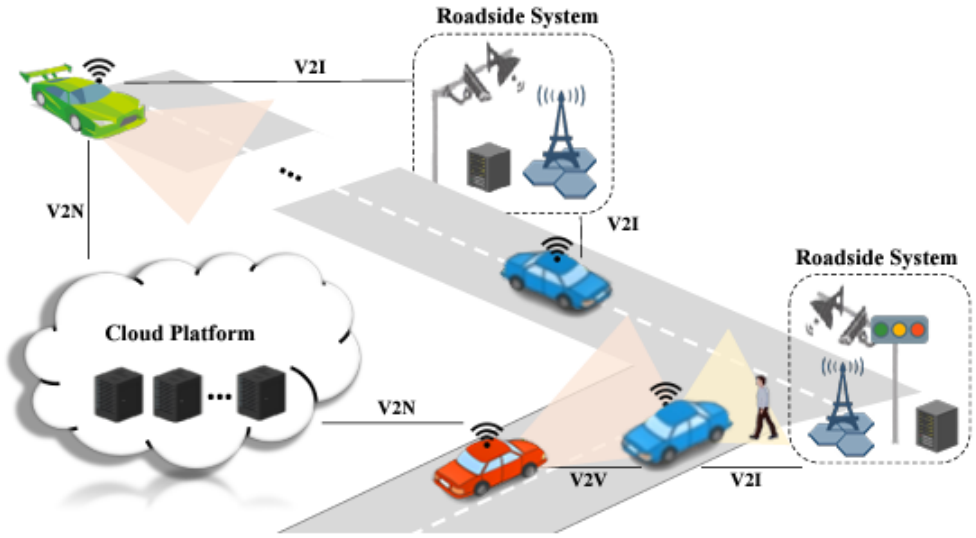


Fig. 1. A diagram illustrating V2X scenarios. The red car faces the occlusion issue, and the green car faces the long-range perception issue. By obtaining extra perceptual information from other vehicles (V2V) or infrastructure (V2I), these vehicles can achieve a holistic perception of their surroundings, improving traffic safety.

- Vehicle-to-Everything (V2X) communication can be used to communicate basic and/or advanced safety information.
- This form of low-latency communication can allow close and far range vehicles to obtain info without relying on the perfect functionality of sensors at all times.
- Two main communication technologies:
  - Dedicated Short-Range Communication (DSRC) – adapted from WiFi
  - Cellular-V2X (C-V2X) - LTE, 5G
  - mmWave - 6G (still to mature as a product)

Datasets

TABLE 1
A summary of existing datasets for collaborative perception in V2X scenarios. Most datasets are built upon various traffic simulators, and some datasets collect data from the real world. See §3 for more details.

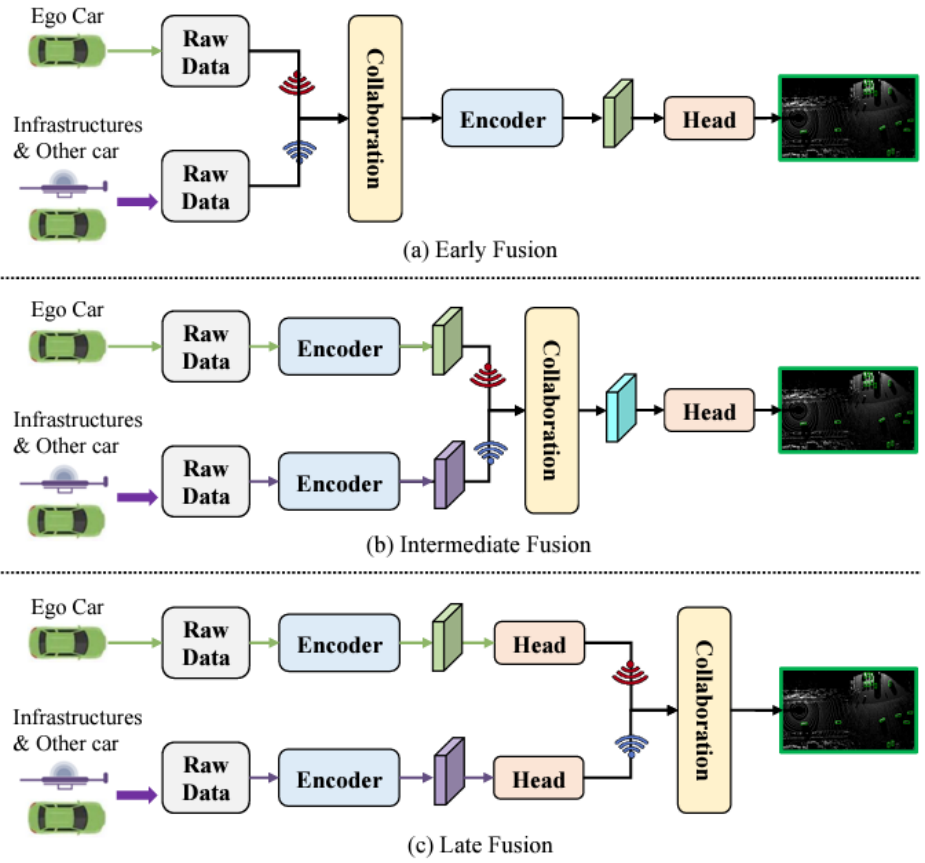
Dataset	Publication	Source	Scenario	Sensor			Tasks			Frames	Viewpoints	Link
				RGB	Depth	LiDAR	Detection	Tracking	Segmentation			
VANETs [30]	GLOBECOM 2017	KITTI [31]	V2V	✓		✓	✓			-	-	-
MFSL [32]	ICMEW 2018	KITTI [31]	V2V	✓					✓	-	-	-
T&J [23]	ICDCS 2019	Real-World	V2V			✓	✓			100	2	Link
V2V-Sim [20]	ECCV 2020	LiDARsim [33]	V2V			✓	✓	✓		51,200	-	-
CoopInf [34]	TITS 2020	CARLA [16]	V2I	✓	✓		✓	✓		10,000	6,8	Link
WIBAM [35]	BMVC 2021	Real-World	V2I	✓			✓	✓		33,092	2-4	Link
CODD [36]	RAL 2021	CARLA [16]	V2V			✓	✓	✓		8,783	10 (avg.)	Link
V2X-Sim [14]	RAL 2022	CARLA [16] & SUMO [37]	V2V,V2I	✓	✓	✓	✓	✓	✓	10,000	2-5	Link
COMAP [38]	RAL 2022	CARLA [16] & SUMO [37]	V2V			✓	✓			7,788	2-10	Link
OPV2V [15]	ICRA 2022	CARLA & OpenCDA [17]	V2V	✓		✓	✓	✓	✓	11,464	2-7	Link
AUTOCASTSIM [39]	CVPR 2022	CARLA [16]	V2V	✓		✓	✓			-	-	Link
DAIR-V2X-C [12]	CVPR 2022	Real-World	V2I	✓		✓	✓			38,845	2	Link
V2XSet [21]	ECCV 2022	CARLA [16] & OpenCDA [17]	V2V,V2I	✓		✓	✓			11,447	2-5	Link
DOLPHINS [40]	ACCV 2022	CARLA [16]	V2V,V2I	✓		✓	✓	✓		42,376	3	Link
CARTI [41]	ITSC 2022	CARLA [16]	V2I			✓	✓			11,000	2	-
V2V4Real [13]	CVPR 2023	Real-World	V2V	✓		✓	✓	✓		20,000	2	Link
V2X-Seq (SPD) [42]	CVPR 2023	Real-World	V2V,V2I	✓		✓		✓		15,000	2	Link
DeepAccident [43]	-	CARLA [16]	V2V,V2I	✓		✓	✓	✓	✓	57,000	5	Link

UCF

Collaborative Perception – Areas of Research

- Fusion Stage: Early, Intermediate, Late
- Performance-bandwidth tradeoff: encode features (feature compression, confidence maps)
- Communication latency (time delay between CAVs)
- Lossy communication: Feature partly damaged
- Domain Gaps: Different kinds of sensors, agents and configs
- Localization and Pose Errors

Areas of Research – Fusion Stage



SUMMARY OF DIFFERENT FUSION SCHEMES FOR COOPERATIVE PERCEPTION.

Fusion Scheme	Methodology	Pros. and Cons.	Highlighted Features	Author
Early Fusion	Deep Learning	Pros: Raw data is shared and gathered to form a holistic view.	Raw point cloud data is compressed to fit the limited bandwidth.	Chen et al. [17]
		Cons: Low tolerance to the noise and delay of the transmitted data; potentially constrained by the communication bandwidth.		
Deep Fusion	Deep Learning	Pros: High tolerance to the noise, delay, and difference between different nodes and sensor models.	Deep neural features are extracted and fused based on spatial correspondence.	Bai et al. [14]
		Cons: Require training data and hard to find a systematic way for model design.		
Late Fusion	Traditional	Pros: Easy to design and deploy in real-world system.	A late-fusion is proposed based on joint re-scoring and non-maximum suppression.	Zhang et al. [77]
		Cons: Significantly limited by the wrong perception results or the difference between sources.		

Thus, intermediate (deep) fusion is preferred in the works from the literature

Areas of Research – Performance Bandwidth Tradeoff

- Given the limitation of bandwidth, major portion of literature has focused on improving the performance-bandwidth tradeoff, such that the performance can still be improved while efficiently utilizing the bandwidth by encoding features
- Encoding features can be achieved by:
  - Feature compression
  - Confidence maps (consider reduced set of links)
- In addition to pure compression, recent studies have been focusing on only sharing the most important features in terms of confidence maps in the spatial or spatio-temporal domain

Areas of Research – Communication Latency

- Delay btw agent I obtaining a point cloud, processing, sending, and it being received at agent J. This can cause feature/pose alignment issues
- The feature misalignment can cause severe fusion problems, impacting the overall detection/segmentation performance.

Areas of Research – Lossy Communication

- Communication is prone to loss where features are partially/fully corrupted due to packet collision.
- Research community active in this domain in looking into trainable models that can repair the damaged parts of the feature, mainly using historical context.
  - Scene Reconstruction

Areas of Research – Domain Gaps

- Different agents use
  - Different sensors/modalities
    - Lidar -> Point cloud
    - Camera -> RGB image
    - IR -> Thermogram image
  - Sensors from different vendors, i.e
    - models with different inference capabilities
    - models with different sensing range
- Between vehicles and road-side units, additional discrepancies can be:
  - Sensor height -> pose estimation differences
- Sim-2-real
  - Simulation does not provide very diverse training data

Challenges specific to "co-operation"

- Training might be expensive, eg when:
  - Agent with a new modality is introduced
- How to manage confidence on the inference
- identifying correspondence of same neighbor in different feature sets

Study 1: Cooperative LiDAR Object Detection via Feature Sharing in Deep Networks

Problems:

- Early-stage feature fusion is efficient, but bandwidth-hungry
- Fully-processed feature sharing is bandwidth-efficient, but does not carry the expected benefits from cooperation

Major Contribution:

- Proposes intermediate feature sharing as an efficient fusion method
  - Sharing partially-processed data strikes a balance between performance and comm. cost.
- Proposes a training pipeline for cooperative perception.

Study 1: Cooperative LiDAR Object Detection via Feature Sharing in Deep Netw

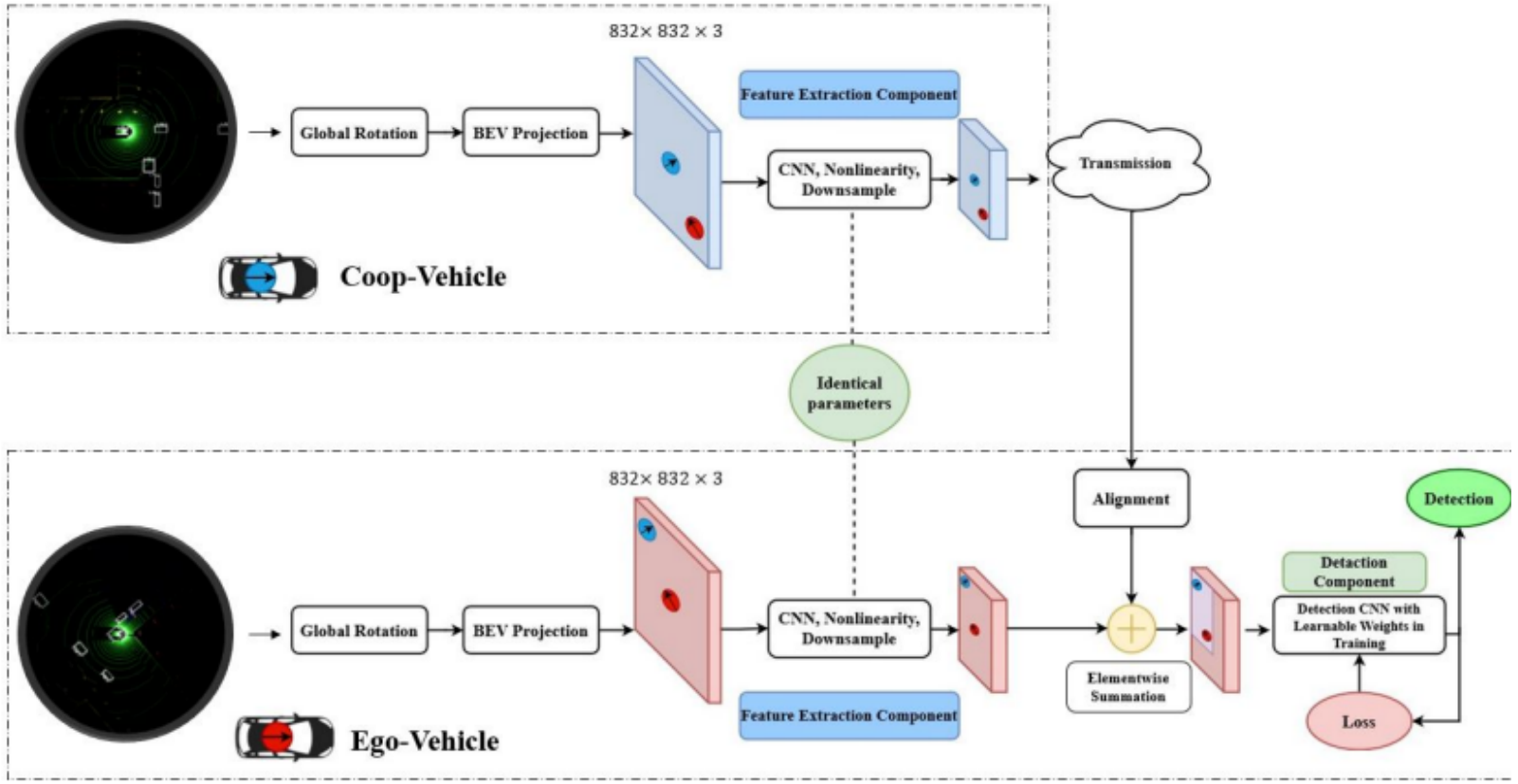
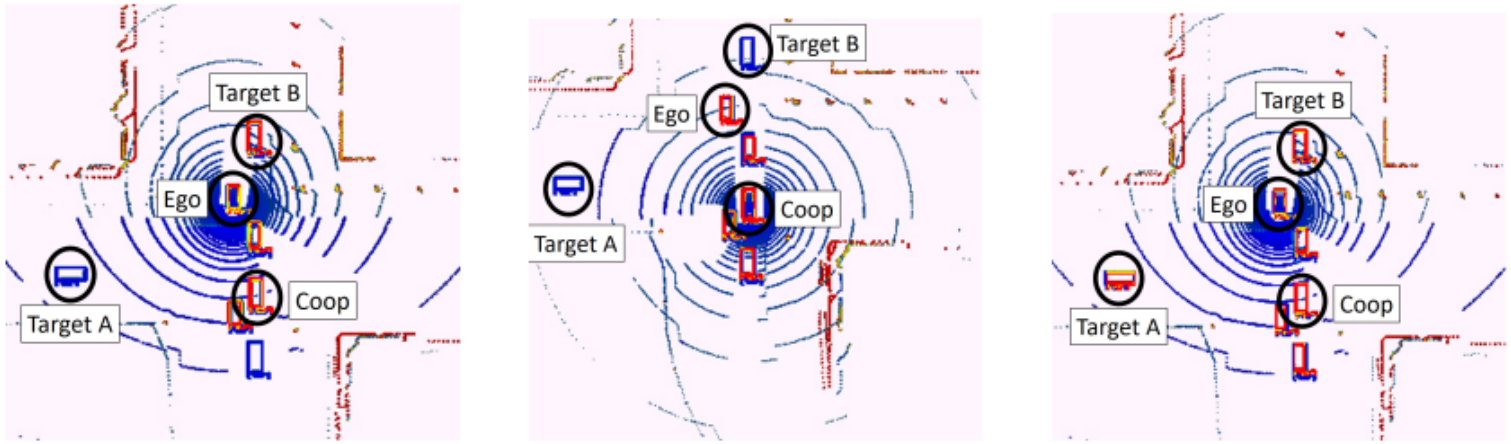


Fig. 1: The overview of feature sharing procedure. The cooperative vehicle transfers the LIDAR point-cloud to feat domain after an initial rotation alignment. After performing a translation transformation on the received feature-maps, aligned received feature-maps are accumulated with the feature-map produced by receiver vehicle and fed through the object detection module

TABLE I: The architecture of Proposed networks

10.4 ppm		4.16 ppm	
Baseline	**FS-COD**	**Baseline**	**FS-COD**
Input 832x832x3			
Feature Extraction Component			
3x3x24 Convolution Batch-Norm Leaky ReLU(0.1)			
Maxpool/2			
3x3x48 Convolution Batch-Norm Leaky ReLU(0.1)			
Maxpool/2			
3x3x64 Convolution Batch-Norm Leaky ReLU(0.1)			
3x3x32 Convolution Batch-Norm Leaky ReLU(0.1)			
3x3x64 Convolution Batch-Norm Leaky ReLU(0.1)			
Maxpool/2			
3x3x128 Convolution Batch-Norm Leaky ReLU(0.1)			
3x3x64 Convolution Batch-Norm Leaky ReLU(0.1)			
3x3x128 Convolution Batch-Norm Leaky ReLU(0.1)			
Maxpool/2		-	
3x3x128 Convolution Batch-Norm Leaky ReLU(0.1)			
1x1x64	1x1xC_t	1x1x64	1x1xC_t
Object Detection Component			
1x1x128 Convolution Batch-Norm Leaky ReLU(0.1)			
3x3x256 Convolution Batch-Norm Leaky ReLU(0.1)			
1x1x512 Convolution Batch-Norm Leaky ReLU(0.1)			
1x1x1024 Convolution Batch-Norm Leaky ReLU(0.1)			
3x3x2048 Convolution Batch-Norm Leaky ReLU(0.1)			
1x1x1024 Convolution Batch-Norm Leaky ReLU(0.1)			
1x1x2048 Convolution Batch-Norm Leaky ReLU(0.1)			
3x3x1024 Convolution Batch-Norm Leaky ReLU(0.1)			
1x1x20 Convolution			
Output 52x52x20		Output 104x104x20	

Study 1: Cooperative LiDAR Object Detection via Feature Sharing in Deep Networks



(a) (b) (c)

Fig. 3: Comparison between performance of single vehicle object detection and feature sharing cooperative object detection in an arbitrary scenario. blue and red bounding boxes represent ground truth and output of the object detection; (a)The single vehicle object detection at ego-vehicle, (b) The single vehicle object detection at coop-vehicle and (c) FS-COD at ego-vehicle

- target A is not detectable by either vehicles and there is a lack of consensus on target B between cooperative vehicles if they rely solely on their own sensory and inference units.
- However, target A is detectable if FS-COD is applied and the lack of consensus on target B is solved.

Study 2: Feature Sharing and Integration for Cooperative Cognition and Perception with Volumetric Sensors

Main challenge to address:

- Minimizing localization error in cooperative setting.

- Proposes Deep Feature Sharing;
  - robust to GPS-related localization error.
  - balanced in compute cost and information-richness.

Study 2: Feature Sharing and Integration for Cooperative Cognition and Perception with Volumetric Sensors

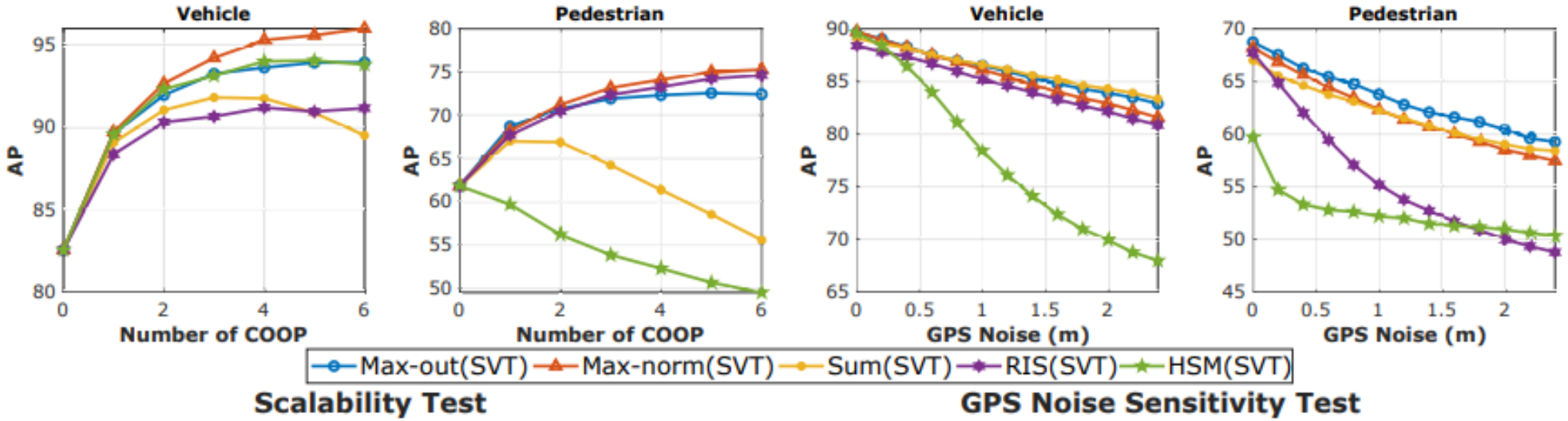


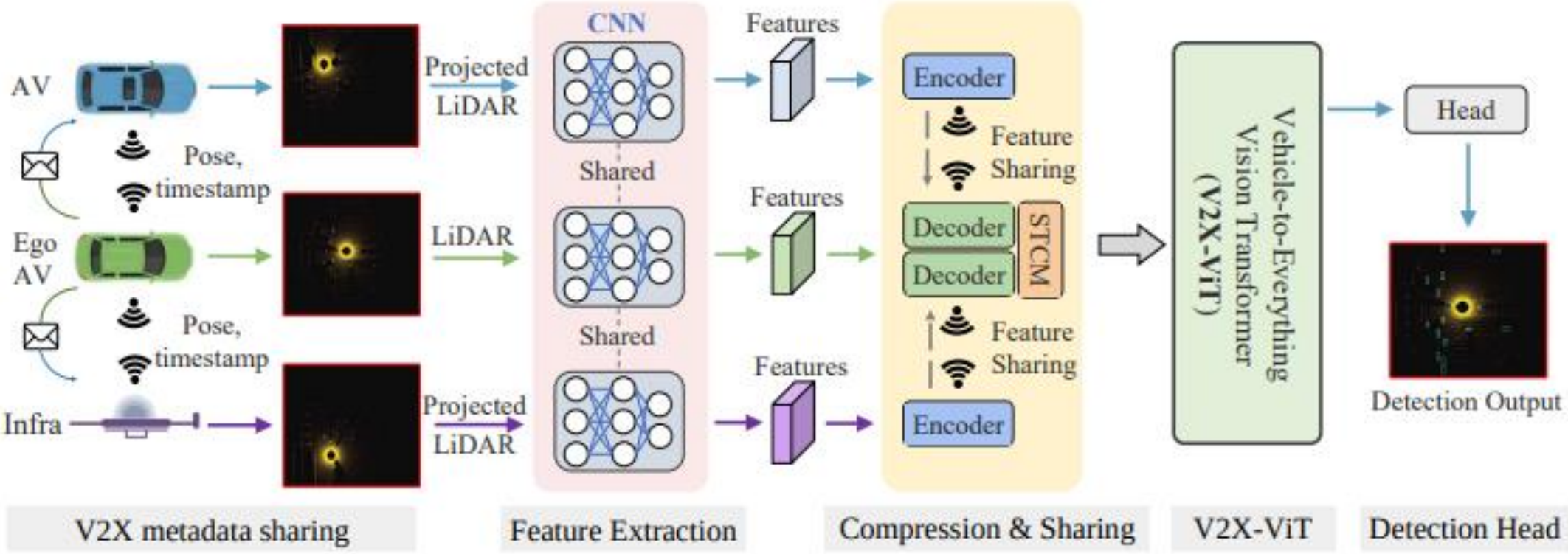
Fig. 7: The effect of different information aggregation functions with identical network paremeters. The network is trained based on SVT strategy.

Study 3: V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer

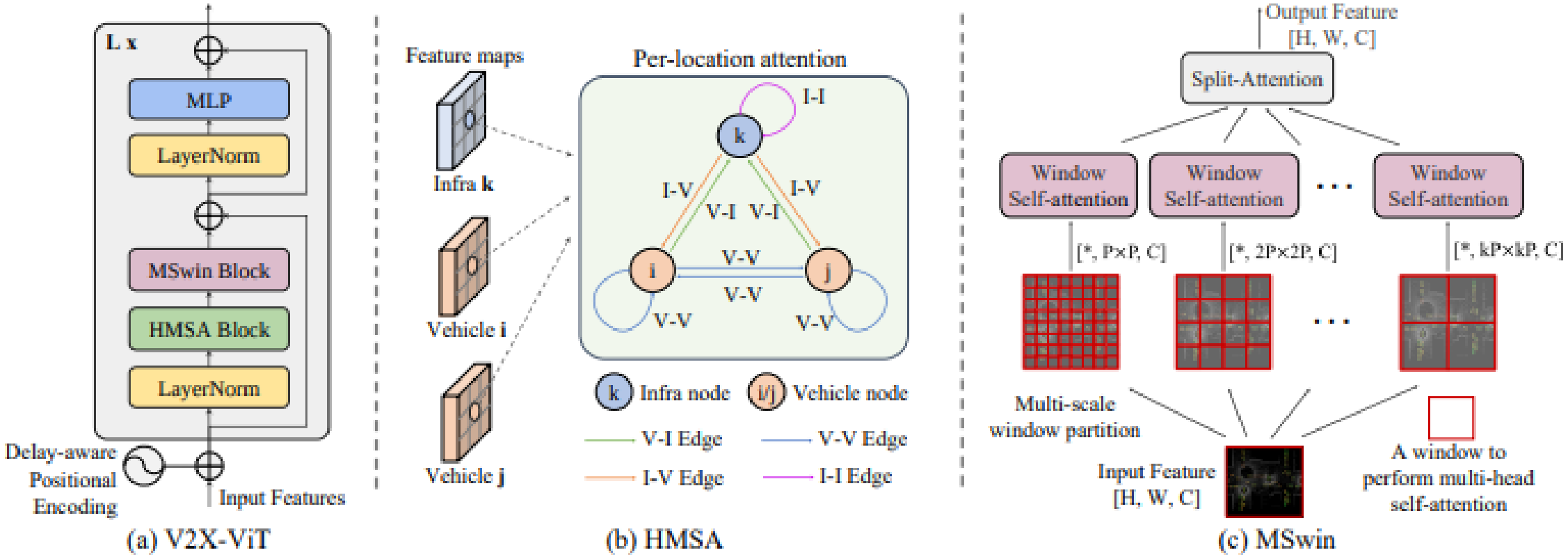
- Problems with SOTA:
  - Heterogeneous agents (infrastructure, agents) and configuration discrepancies such as noise levels, installation heights, and sensor modality
  - GPS localization noises and asynchronous sensor measurements of AVs and infrastructure can cause inaccurate coordinate transformation and lagged sensing info.
- Proposal:
  - Customized heterogeneous multi-agent self-attention module that explicitly considers agent types and their connections when performing attentive fusion
  - A multi-scale window attention module that can handle localization errors using multi-resolution windows in parallel
  - Integrate a delay-aware positional encoding to handle time delay uncertainty

UCF

Study 3: V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer



Study 3: V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer



Study 3: V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer

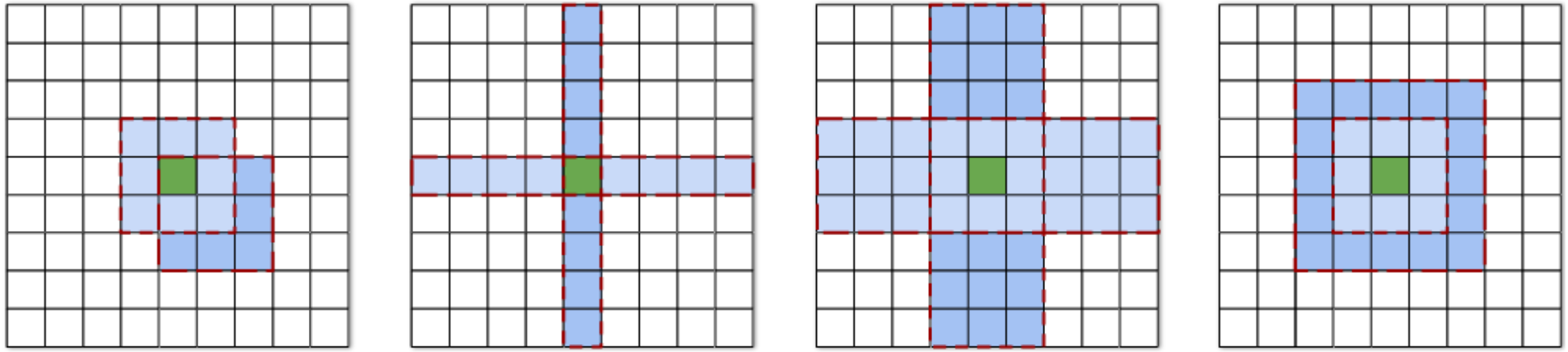


Fig. 8: Visualizations of approximated receptive fields (blue shaded pixels) for the green pixel for (a) Swin [30] (b) Axial [44], (c) CSwin [8] and (d) MSwin attention. MSwin obtains multi-scale long-range interactions at linear complexity.

UCF

Study 4: Where2comm: Communication-efficient Collaborative Perception via Spatial Confidence Maps

- Problems with SOTA:
  - Some previous works make an assumption that once agents collaborate, they are obligated to share perceptual info of all spatial areas equally. This can waste bandwidth as large proportion of spatial areas may contain irrelevant info.
  - Some previous works consider fully-connected communication graphs. This is excessive because agents that have similar global features do not necessarily need info from each other.
  - ad
- Proposals:
  - Includes a spatial confidence generator, which produces a spatial confidence map to indicate perceptually critical areas
  - Spatial confidence-aware communication module which leverages the spatial confidence map to decide where to communicate via novel message packing, and who to communicate via novel communication graph construction
  - Spatial confidence-aware message fusion, which uses novel confidence aware multi-head attention to fuse all message received from other agents, upgrading the feature map for each agent

UCF

Study 4: Where2comm: Communication-efficient Collaborative Perception via Spatial Confidence Maps

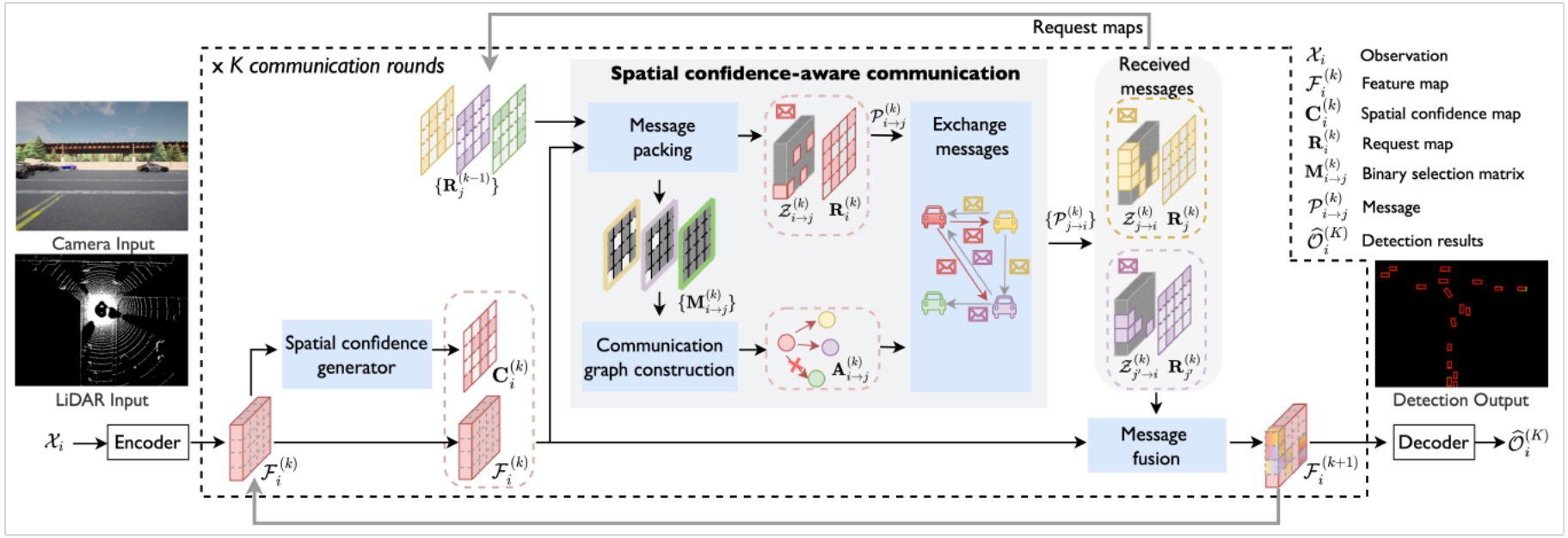
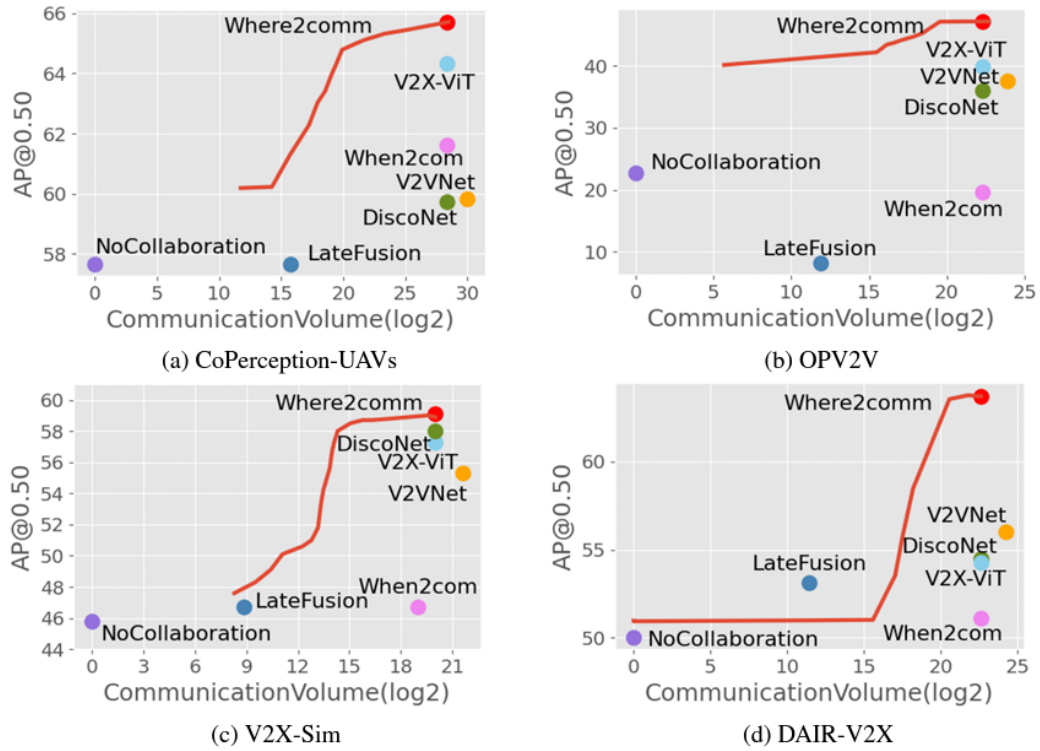


Figure 2: System overview. In `Where2comm`, spatial confidence generator enables the awareness of spatial heterogeneous of perceptual information, spatial confidence-aware communication enables efficient communication, and spatial confidence-aware message fusion boosts the performance.

Performance Comparison – Performance Bandwidth Trade-off

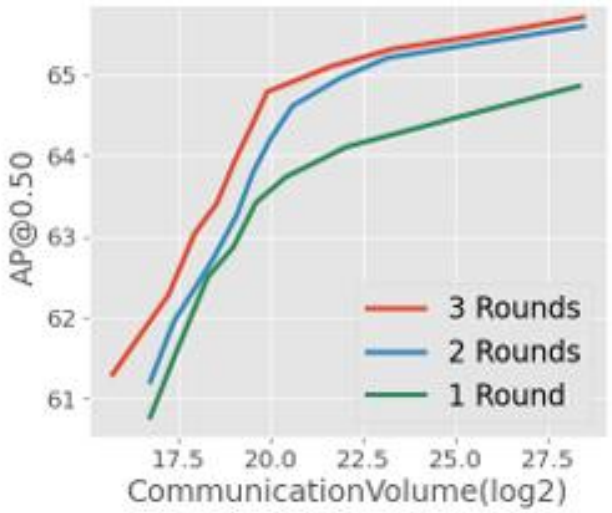


(a) CoPerception-UAVs (b) OPV2V

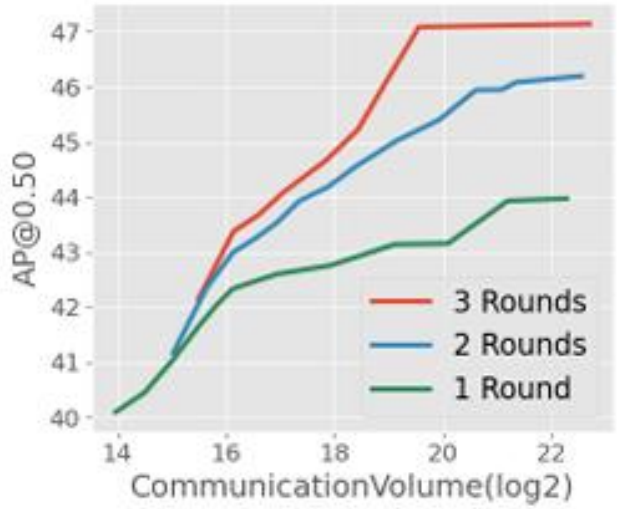
(c) V2X-Sim (d) DAIR-V2X

Figure 3: `Where2comm` achieves consistently superior performance-bandwidth trade-off on all the three collaborative perception datasets, e.g, `Where2comm` achieves *5,000* times less communication volume and still outperforms When2com on CoPerception-UAVs dataset. The entire red curve comes from a single `Where2comm` model evaluated at varying bandwidths.

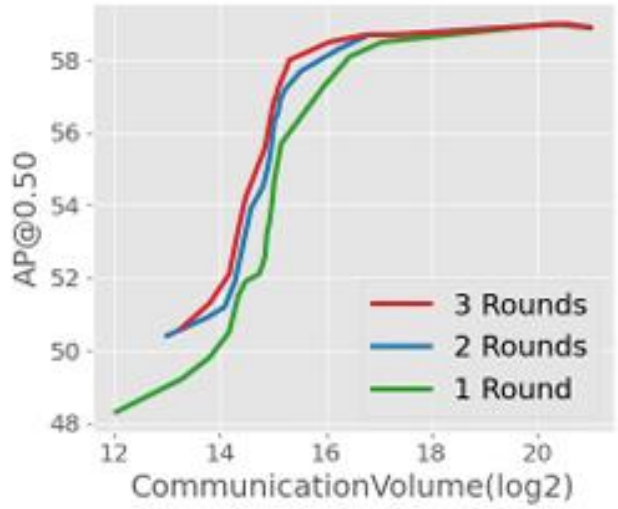
Performance Comparison – Robustness to Communication Latency



(a) CoPerception-UAVs



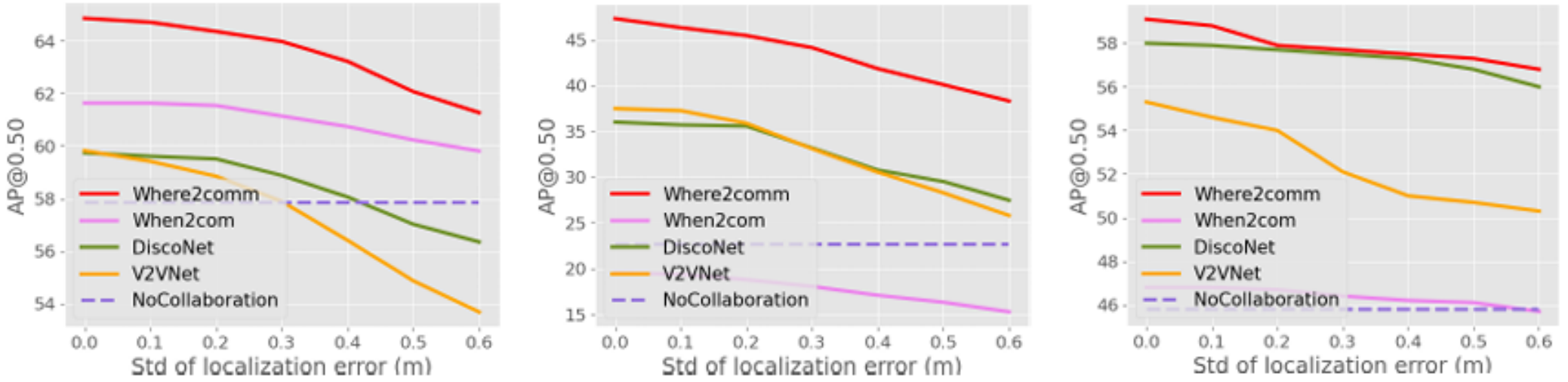
(b) OPV2V



(c) V2X-Sim

Figure 4: More communication rounds continuously improve performance-bandwidth trade-off.

Performance Comparison: Robustness to Localization Errors



(a) CoPerception-UAVs (b) OPV2V (c) V2X-Sim

Figure 5: Robustness to localization error. Gaussian noise with zero mean and varying std is introduced. *Where2comm* consistently outperforms previous SOTAs and No Collaboration.

Future Directions:

- Performance-bandwidth tradeoff:
  - To further improve the tradeoff, research community is considering spatial+temporal domain simultaneously so that only the most important features can be shared
- Pragmatic Communication:
  - Most studies consider a very abstract module of communication, with perfect reception and limited number of collaborative agents (<=5)
  - We are currently looking into integrating a pragmatic communication module using state-of-the-art simulator like NS-3 to consider all the eventualities of V2X, especially in congested scenarios
- Datasets:
  - Current datasets are limited to 5 collaborative agents; however this needs to be increased to replicate realistic scenarios
- Transmission Scheme:
  - Currently, most studies utilize either a broadcast or unicast approach where same/similar information is sent to the channel numerous times. Instead, we can focus on using infrastructure that can relay the info once to the related vehicles. This can greatly reduce the burden on the bandwidth.
- Domain Adaptation (explained earlier)
- Security and Privacy

Thank you

Ghayoor.shah@ucf.edu
Mahdi.zaman@ucf.edu
HEC 338