IEEE IROS'25 Hangzhou, China.

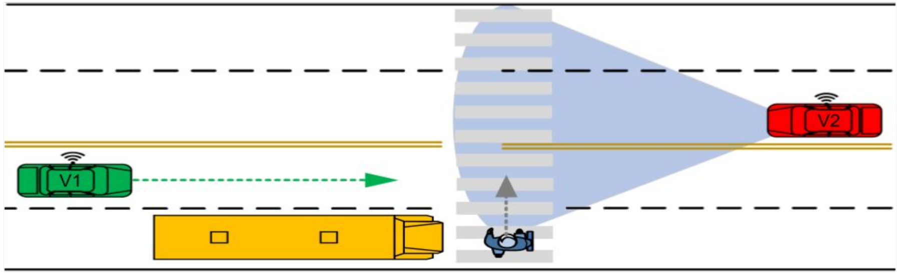
SCORPION: Robust Spatial-Temporal Collaborative Perception Model on Lossy Network

Ruiyang Zhu, Minkyoung Cho, Shuqing Zeng[†], Fan Bai[†], Z. Morley Mao

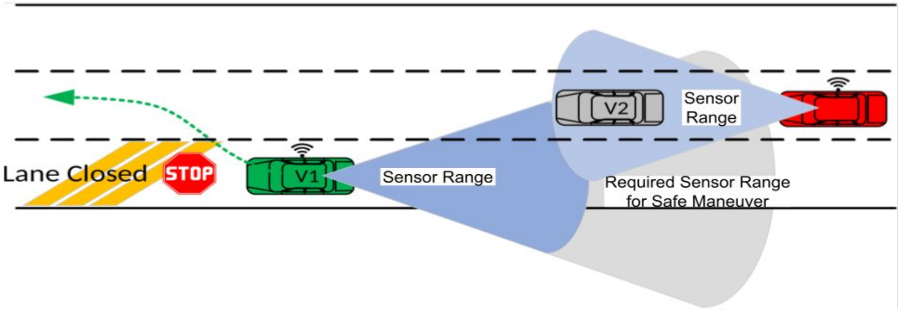
University of Michigan   [†]General Motors Research & Development

Background - Collaborative Perception

- Limited sensing on occluded or far-away objects



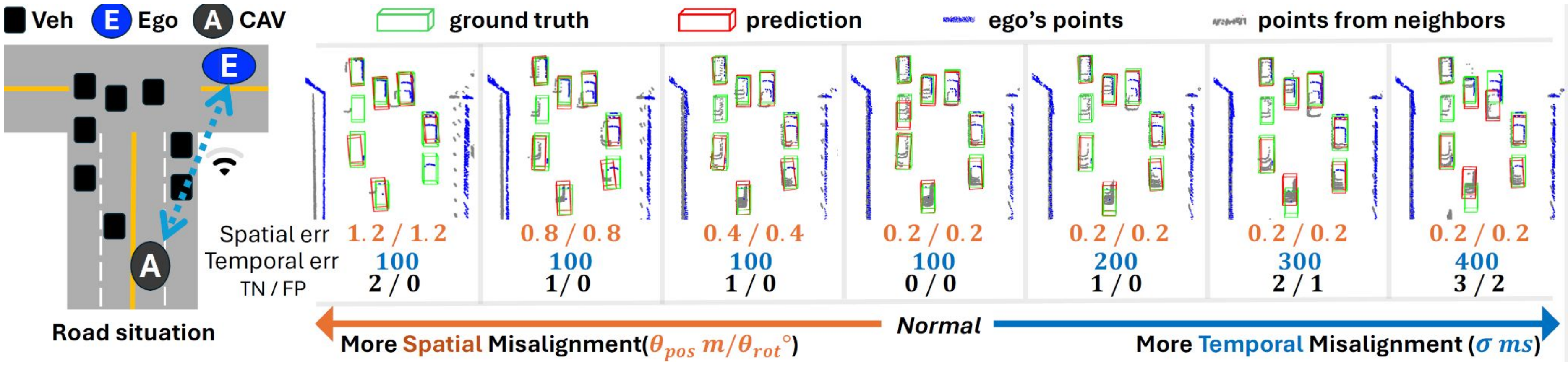
Occluded pedestrian



Far-away obstacles

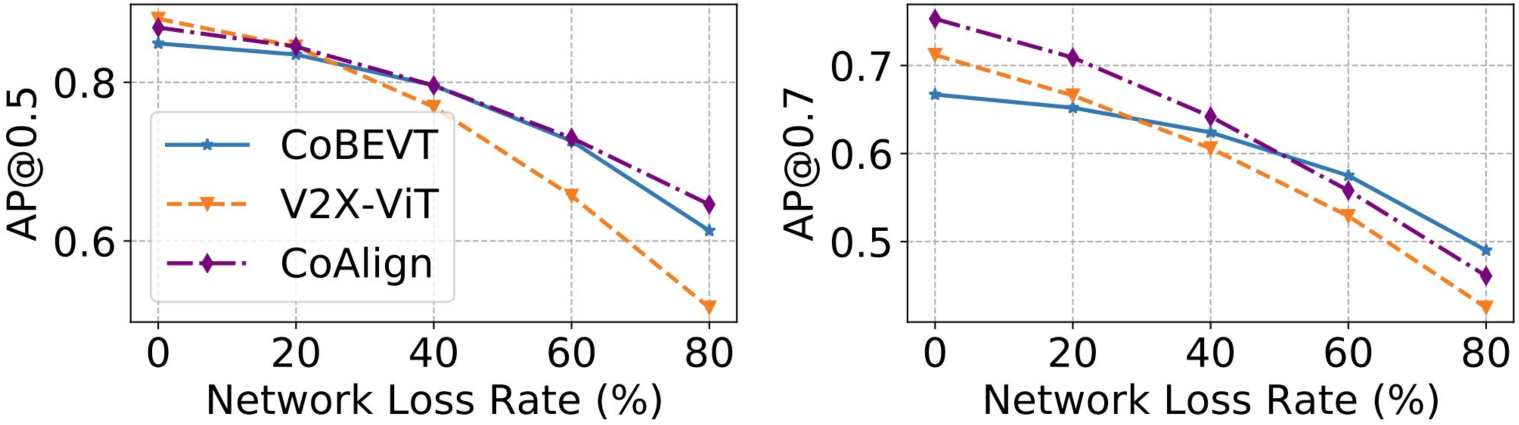
Motivation - Practical Challenges in Collaborative Perception

- Imperfections in underlying system layers
  - *Spatial misalignments occur due to <u>sensing errors</u> or dropped network packets*
  - *Temporal misalignments arise from <u>sensor asynchronization</u> and network delays*



Challenge: Lossy V2X Network Transmission

- Performance of existing collaborative perception methods drops significantly on V2V/V2X network packet loss



[1] Toward understanding characteristics of dedicated short range communications (dsrc) from a perspective of vehicular network engineers. MobiCom 2010.
[2] CoBEVT: Cooperative Bird's Eye View Semantic Segmentation with Sparse Transformers, CoRL 22
[3] V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer, ICCV 22
[4] Co-Align: Robust Collaborative 3D Object Detection in Presence of Pose Errors, ICRA 23

Related Work

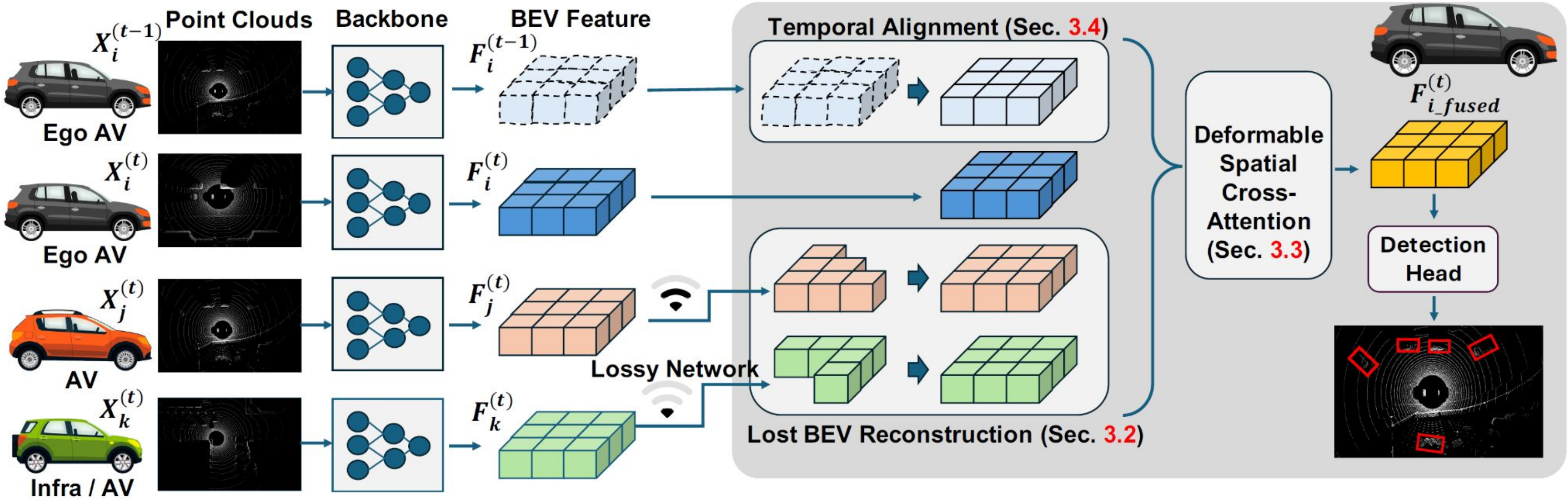
- Existing cooperative perception overlooks the synergy between different types of real-world dynamics
  - *None of the existing work tackles <u>all 3 challenges at the same time</u>*

Work	Sensing Errors	Sensor Asynchronization	Lossy V2X Network	Fusion Method
OPV2V [1]	x	x	x	Intermediate
Where2comm [2]	x	x	x	Intermediate
CoBEVT [3]	x	x	x	Intermediate
V2X-ViT [4]	✓	✓	x	Intermediate
RAO [5]	x	✓	x	Early
Co-Align [6]	✓	x	x	Intermediate
LCRN [7]	x	x	✓	Intermediate

[1] OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication, ICRA 21
[2] Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps, Neurips 2022
[3] CoBEVT: Cooperative Bird's Eye View Semantic Segmentation with Sparse Transformers, CoRL 22
[4] V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer, ICCV 22
[5] Robust Real-time Multi-vehicle Collaboration on Asynchronous Sensors, MobiCom 23
[6] Co-Align: Robust Collaborative 3D Object Detection in Presence of Pose Errors, ICRA 23
[7] Learning for Vehicle-to-Vehicle Cooperative Perception under Lossy Communication, IEEE IV 23

Solution Framework

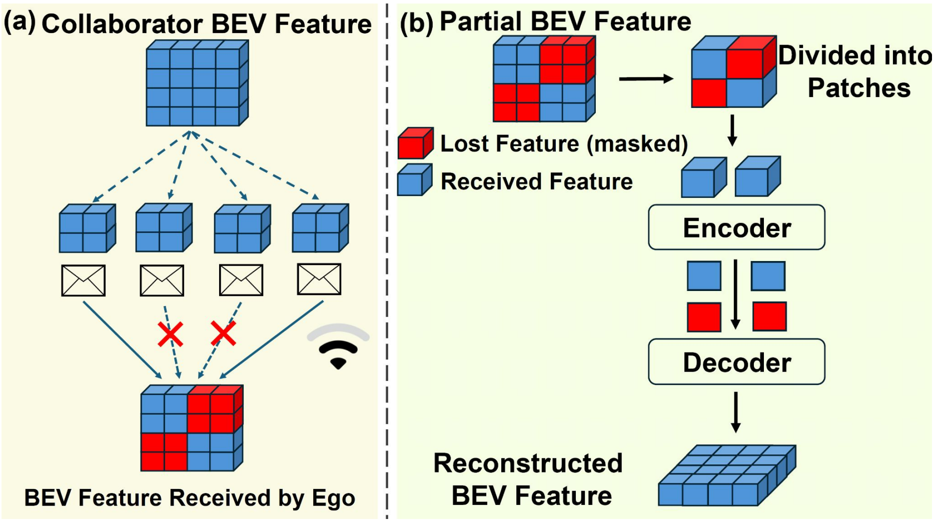
- **SCORPION:** <u>**S**</u>patial-temporal <u>**Co**</u>llabo<u>**r**</u>ative <u>**P**</u>ercept<u>**i**</u>on model on l<u>**o**</u>ssy <u>**N**</u>etwork
  - *An **end-to-end Intermediate-fusion model** to address and compensate for the imperfections in system layers*
  - *[Lossy V2X Network] Lost BEV Reconstruction (L-BEV-R)*
  - *[Spatial Alignment] Deformable Spatial Cross Attention (DSCA)*
  - *[Temporal Alignment] Historical BEV Temporal Alignment (TA)*

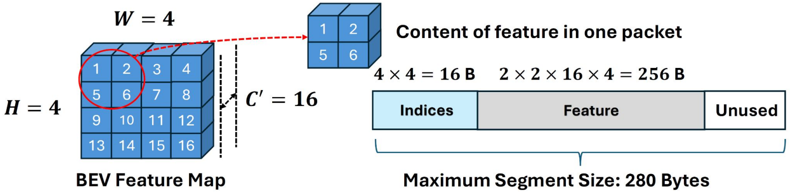


System Architecture of **SCORPION**

Lost BEV Feature Reconstruction (L-BEV-R)

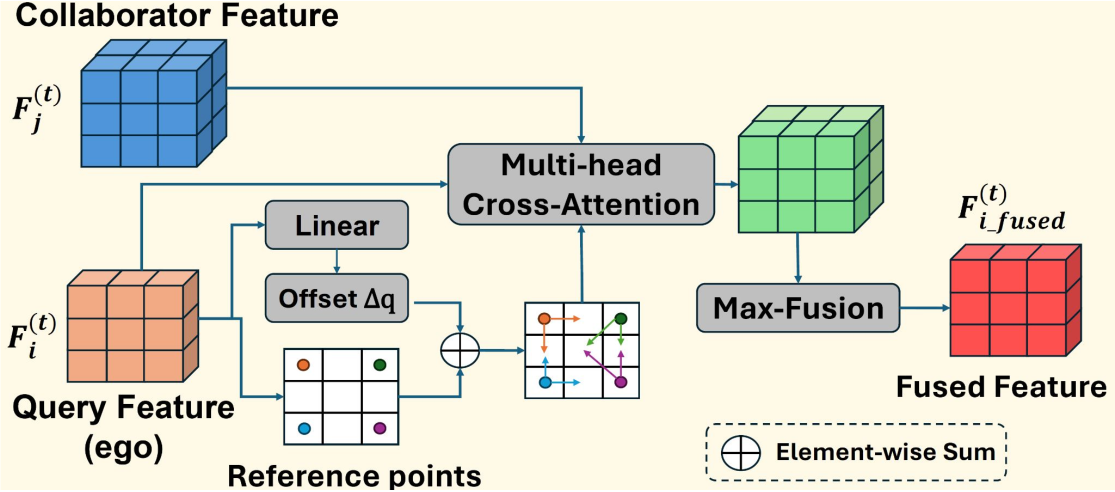
- *The received map has feature indices lost due to lossy V2X network*
- *The underlying MAE Encoder [1] processed the patches, and decoder recover the original BEV feature*

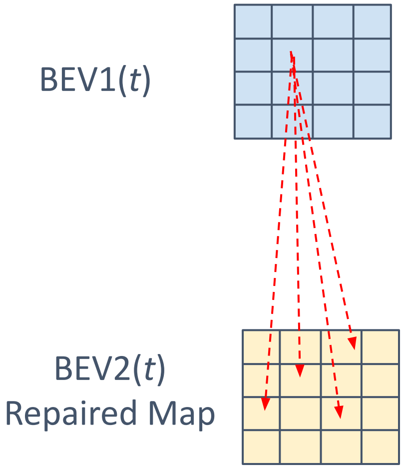




Deformable Spatial Cross Attention (DSCA)

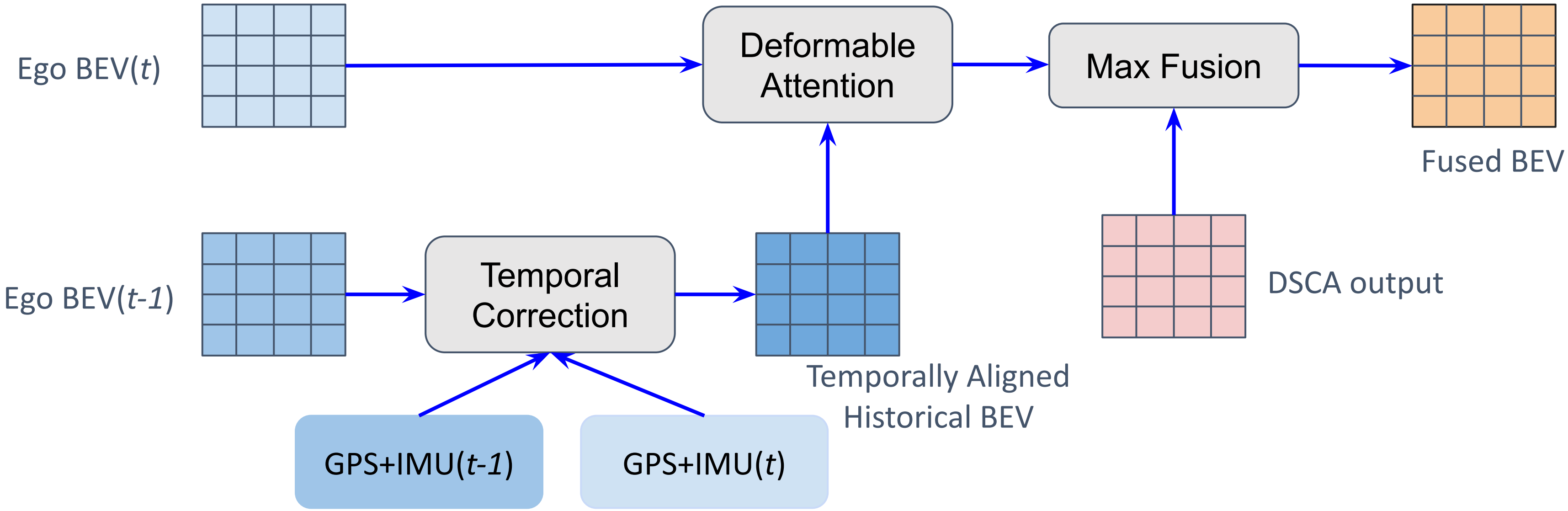
- Instead of a standard attention mechanism, DSCA interacts with a <u>learned set of offset points</u> across all vehicles' BEV maps, considering potential spatial misalignments
  - ***Benefits***: *DSCA allows the model to look for semantic information in areas that may be misaligned due to localization errors.*





Historical BEV Temporal Alignment (TA)

- The TA module incorporates historical BEV features to address temporal misalignment
- **Benefits**: By spatially wrapping the historical BEV map from the ego-vehicle using measured pose (GPS/IMU), the model can align temporal information.



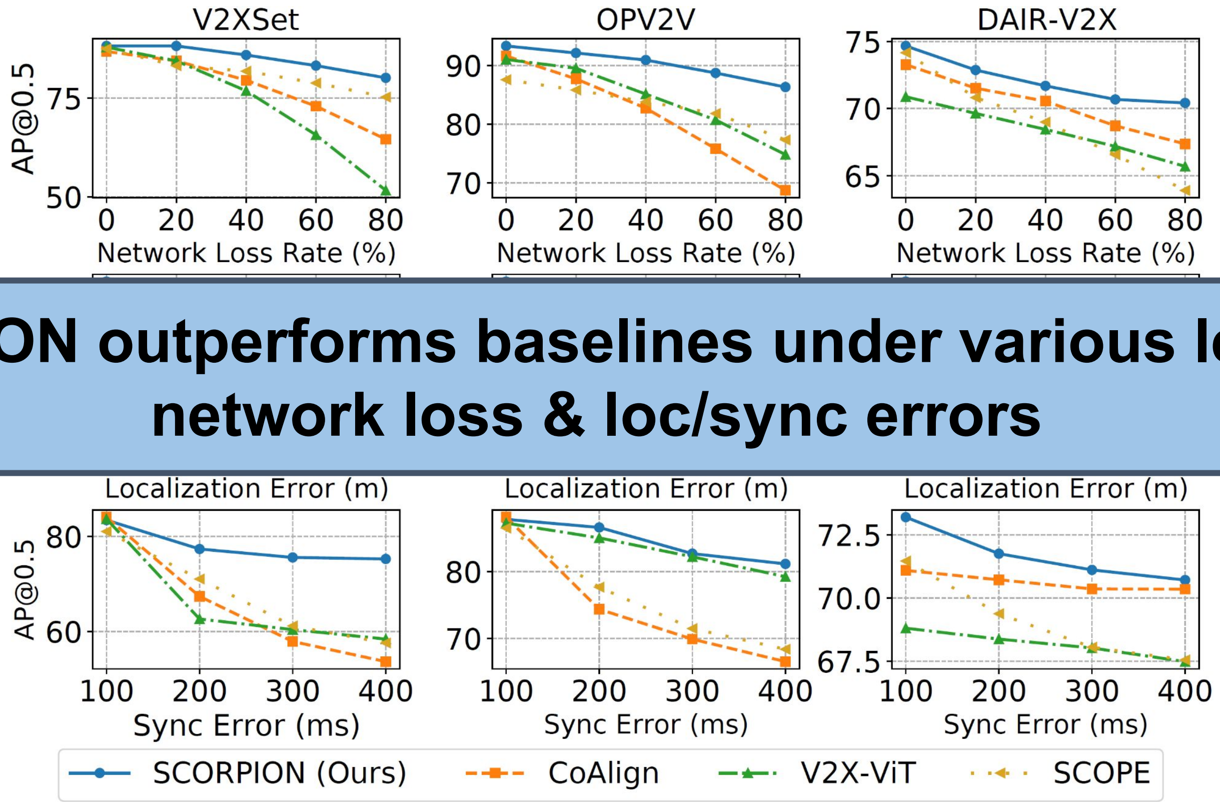
Evaluation

- Dataset: V2XSet [1], OPV2V [2] and DAIR-V2X [3]
- ***Perfect environment*** *setup*: no net loss, localization error or sync error
- SCORPION achieves SOTA performance

Model	V2XSet		OPV2V		DAIR-V2X	
	AP0.5	AP0.7	AP0.5	AP0.7	AP0.5	AP0.7
No Fusion	65.73	52.57	69.38	56.40	63.04	47.39
V2VNet [8]	87.82	74.28	86.76	73.38	65.09	48.18
F-Cooper [10]	82.82	69.38	89.22	79.66	70.54	52.21
AttFuse [7]	81.70	66.24	88.54	72.91	68.02	48.40
CoBEVT [1]	81.00	65.06	88.99	72.80	67.61	55.51
V2X-ViT [2]	82.32	71.21	86.74	75.70	70.87	54.35
CoAlign [5]	86.90	75.31	91.60	82.30	74.02	**56.81**
SCOPE [13]	87.55	75.67	89.60	80.71	74.15	56.52
SCORPION	**88.32**	**77.78**	**93.10**	**85.10**	**74.65**	56.76

[1] V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer, ECCV 22
[2] OPV2V: an open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication, ICRA 21
[3] DAIR-V2X and OpenDAIRV2X: Towards General and Real-World Cooperative Autonomous Driving, CVPR22

Performance under Noise Environment

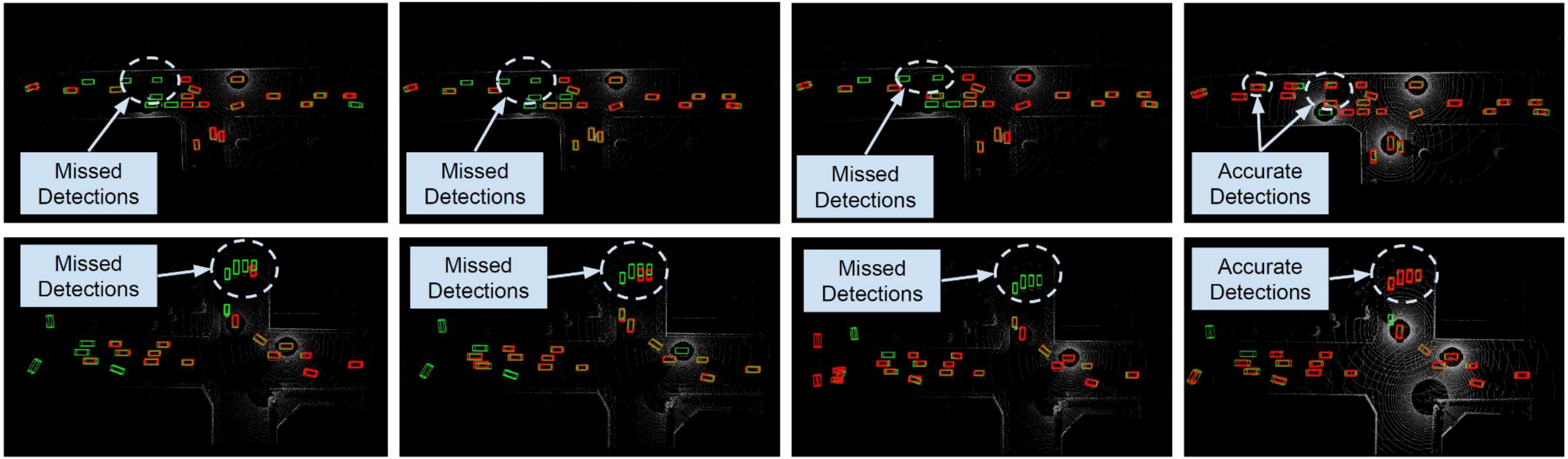


SCORPION outperforms baselines under various levels of network loss & loc/sync errors

Visualization of Detection Results

- Test on environment w/ coexistence of net loss, loc err and sync err

Green: Ground Truth   Red: Prediction



(a) V2X-ViT   (b) SCOPE   (c) CoAlign   (d) **SCORPION (Ours)**

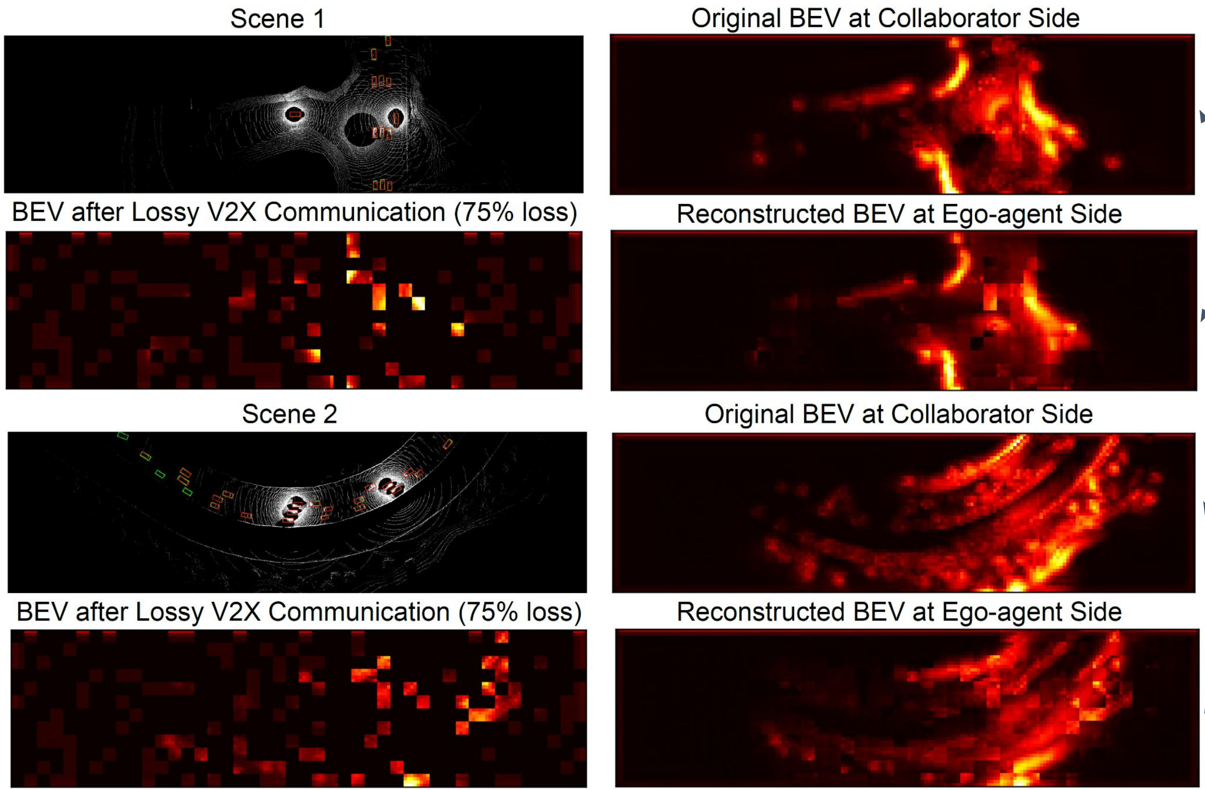
Thank You!

Our Team

Visualization of Reconstructed BEV map

Entrance Ramp

Curvy Road



Close to original

Close to original

SCORPION Demo Video