CoCMT: Towards Communication-Efficient Cross-Modal Transformer for Collaborative Perception

Anonymous authors
Paper under double-blind review

A Appendix

A.1 Multi Range Label Selection

We adopted a multi-range label selection method and constructed corresponding ground truth for two stages: r_{single} and c_{single} for the single-agent independent prediction stage, and r_{co} and c_{co} for the cooperative fusion prediction stage. This strategy offers several advantages: not only does it expand the cooperative perception detection range under the V2V-C setting, but it also reduces the learning complexity during the cooperative fusion stage and effectively addresses challenges posed by differing detection ranges of heterogeneous sensors in the V2V-H setting. We configured the detection ranges and ground truth for the three cooperative perception settings: V2V-L, V2V-C, and V2V-H. Using the OPV2V dataset as an example, the selection results are shown in Table. 1.

Table 1: Specific Configuration Settings

Setting	Ego Detection and GT Range (m)	Collaborative Detection and GT Range (m)
V2V-L	$[-102.4, -102.4, +102.4, +102.4]$	$[-102.4, -102.4, +102.4, +102.4]$
V2V-C	$[-51.2, -51.2, +51.2, +51.2]$	$[-102.4, -102.4, +102.4, +102.4]$
V2V-H	L: $[-102.4, -102.4, +102.4, +102.4]$ C: $[-51.2, -51.2, +51.2, +51.2]$	$[-102.4, -102.4, +102.4, +102.4]$

For V2V-C Setting: Unlike most cooperative perception methods Xiang et al. (2023); Lu et al. (2024); Xu et al. (2022a) that use a detection range of only $51.2m$, we maintained the camera's detection range and ground truth of $51.2m$ in the single-agent independent prediction stage, while extending the detection range to $102.4m$ during the cooperative fusion stage. Through a cooperative deep supervision mechanism, the effective detection range for cooperative perception was successfully expanded.

For V2V-L Setting: Due to the larger detection range of the LiDAR, we used a $102.4m$ detection range for both the single-agent prediction and cooperative fusion stages. To improve individual vehicle detection performance, we introduced cooperative ground truth in the single-agent stage, increasing the number of prediction labels, thereby reducing the difficulty of subsequent cooperative fusion.

For V2V-H Setting: In the OPV2V Xu et al. (2022b) dataset, the camera's effective detection range is $51.2m$, while the LiDAR's is $102.4m$. Unlike HMViT Xiang et al. (2023), which simplifies heterogeneous feature fusion by unifying the detection range to $102.4m$, our framework flexibly handles differences in detection ranges between heterogeneous sensors. In the single-agent independent prediction stage, each sensor used its effective detection range and ground truth. During the cooperative fusion prediction stage, we unified the detection range to $102.4m$, leveraging the cooperative ground truth to further improve the accuracy of individual vehicle predictions.

A.2 Detection Visualization

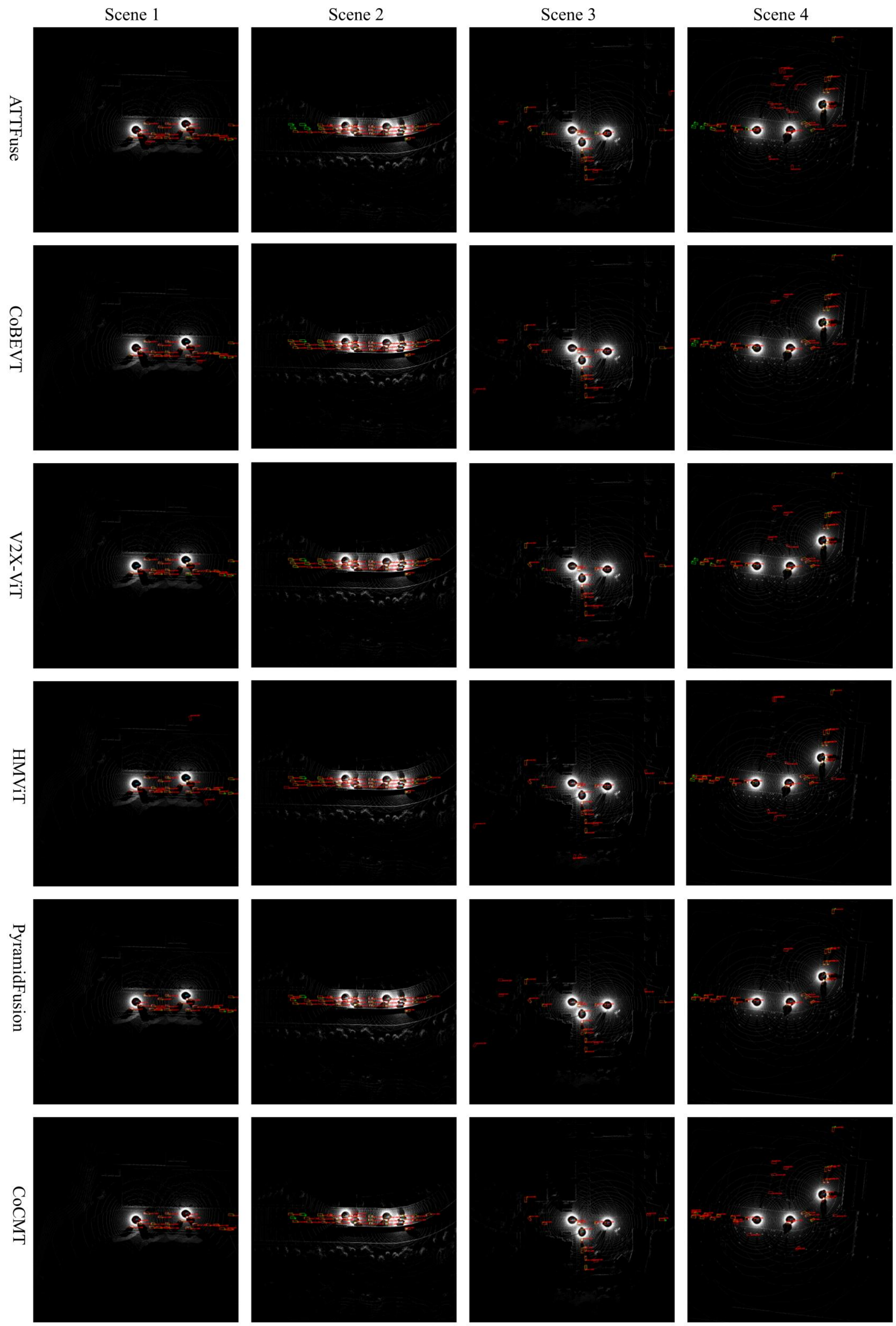

Figure 1: **Qualitative comparison on scenarios 1-4 under V2V-L setting in the OPV2V dataset.** The green and red bounding boxes represent the ground truth and prediction, respectively. Our method detected more dynamic objects.

REFERENCES

Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Siheng Chen, and Yanfeng Wang. An extensible framework for open heterogeneous collaborative perception. *arXiv preprint arXiv:2401.13964*,

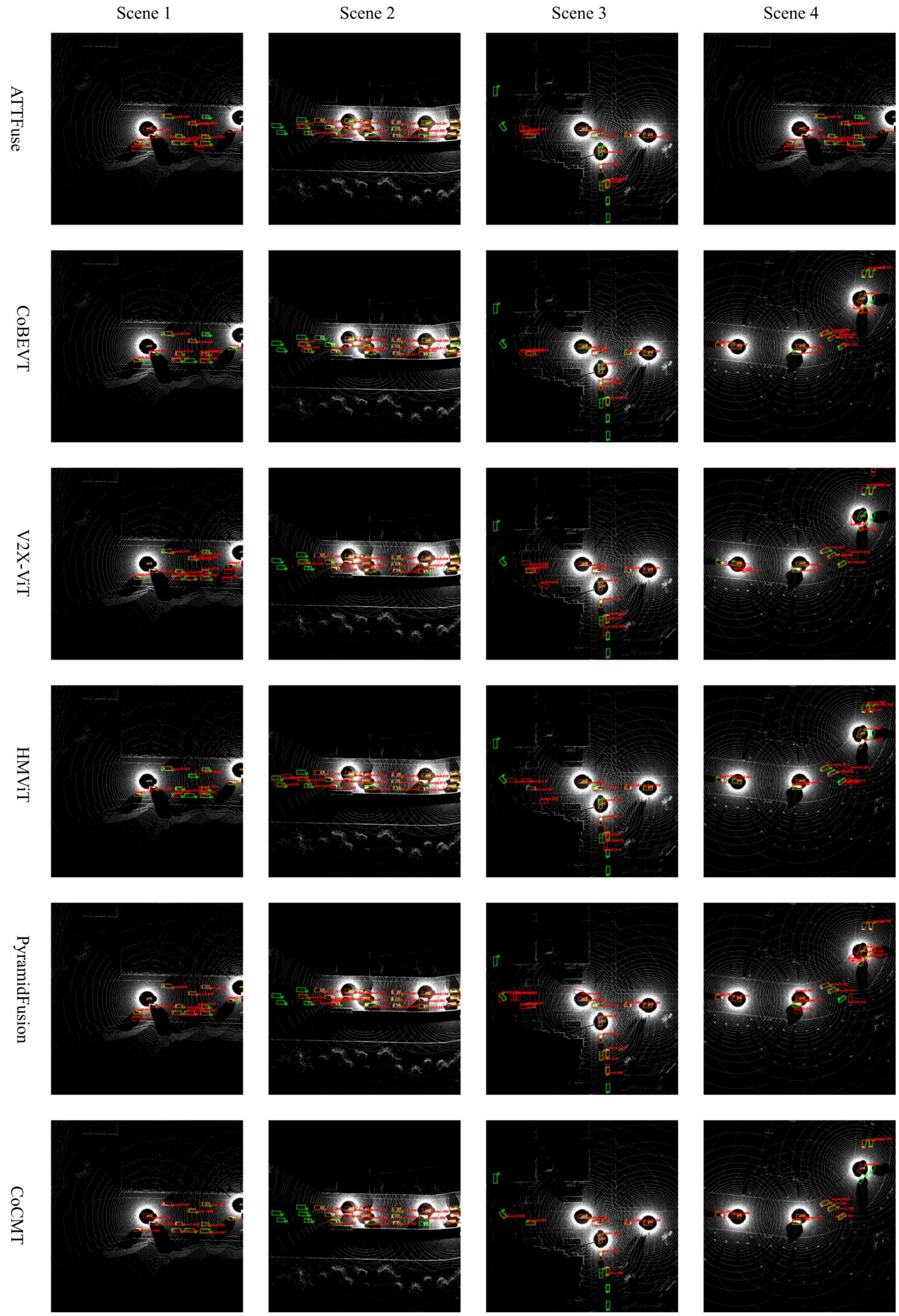

Figure 2: **Qualitative comparison on scenarios 1-4 under V2V-C setting in the OPV2V dataset.** The green and red bounding boxes represent the ground truth and prediction, respectively. Our method produced more accurate detection results.

2024.

Hao Xiang, Runsheng Xu, and Jiaqi Ma. Hm-vit: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 284–295, 2023.

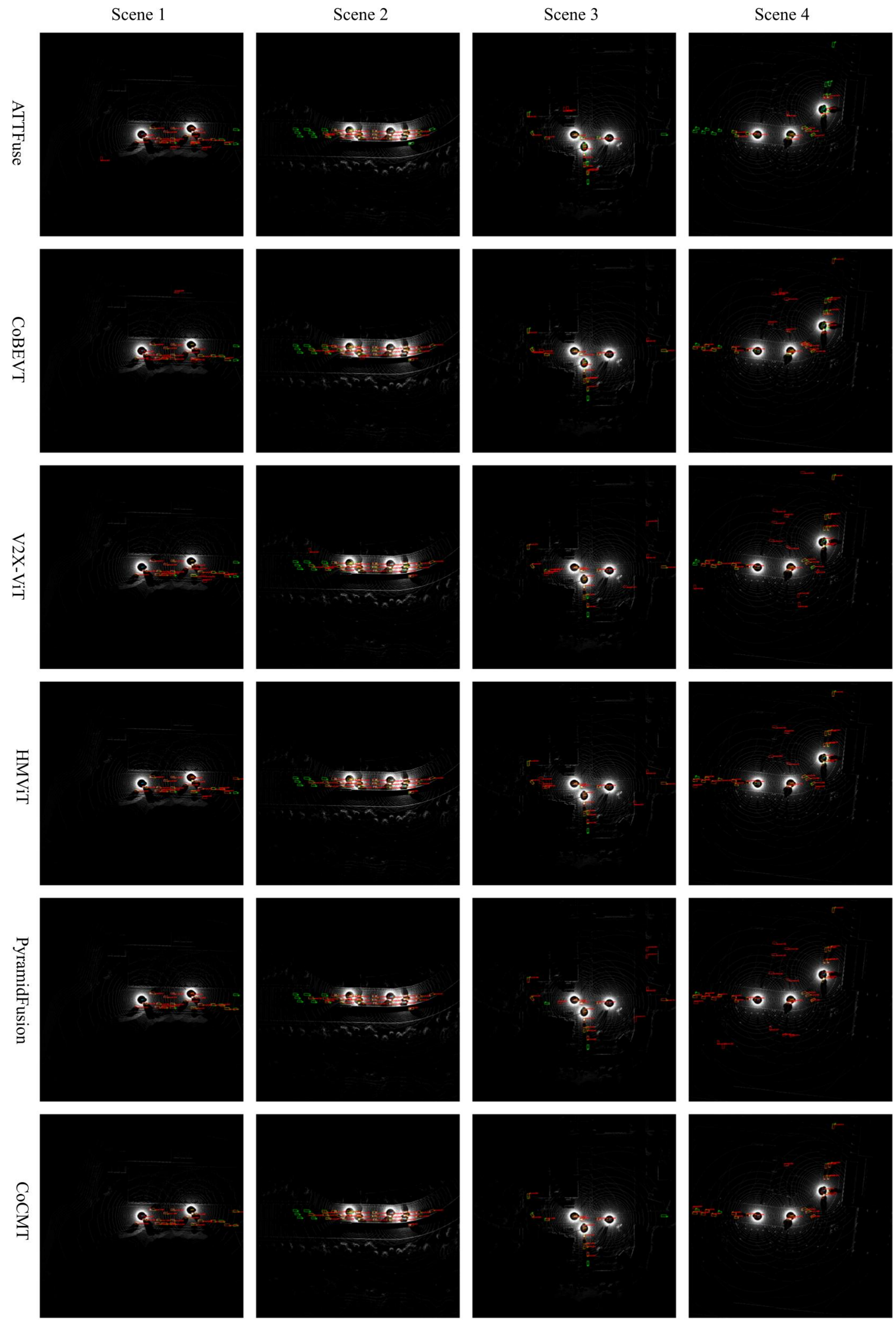

Figure 3: **Qualitative comparison on scenarios 1-4 under V2V-H setting in the OPV2V dataset.** The green and red bounding boxes represent the ground truth and predictions, respectively. Our method produced more accurate detection results and resulted in fewer false detection boxes.

Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*, 2022a.

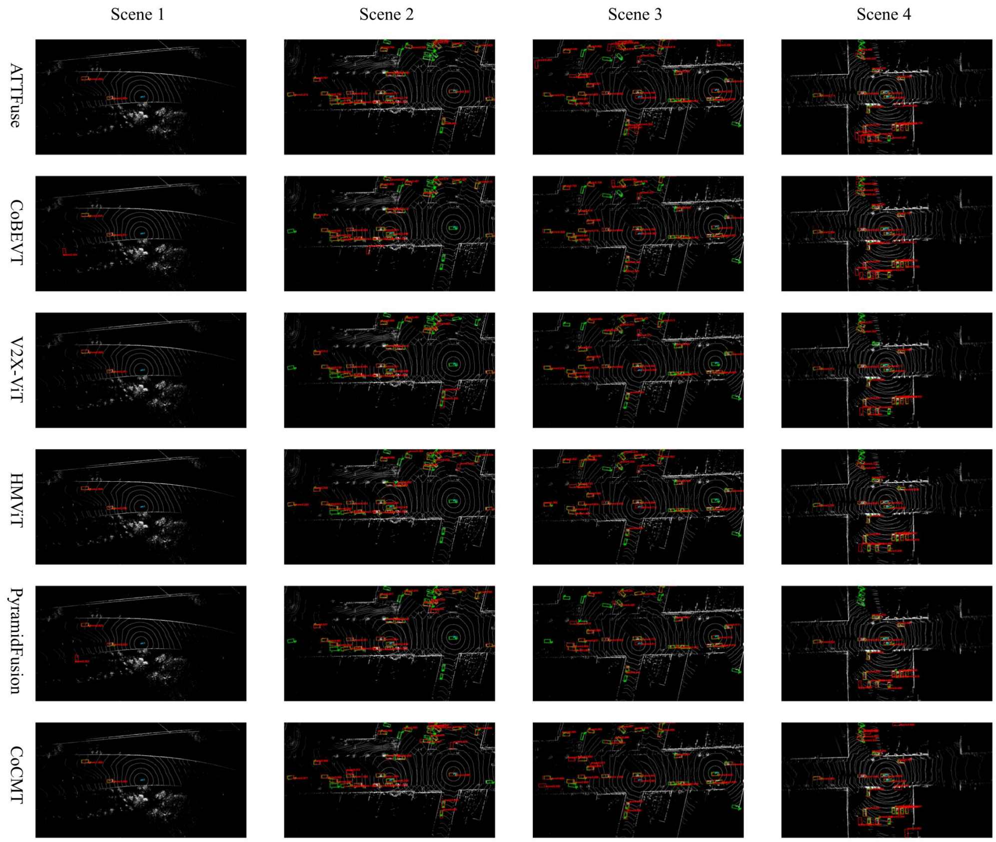

Figure 4: **Qualitative comparison on scenarios 1-4 in the V2V4Real dataset.** The green and red bounding boxes represent the ground truth and predictions, respectively. Our method produced more accurate detection results.

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 *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2583–2589. IEEE, 2022b.