Supplementary Materials for TransIFF: An Instance-Level Feature Fusion Framework for Vehicle-Infrastructure Cooperative 3D Detection with Transformers

Ziming Chen
Beihang University
chenzm@buaa.edu.cn

Yifeng Shi*
Baidu Inc.
shiyifeng@baidu.com

Jinrang Jia
Baidu Inc.
jjr5401@163.com

1. Loss Function

As mentioned in DETR [1], we utilize the Hungarian matching algorithm [2] to match predicted boxes with ground truth boxes and calculate the object detection loss function. Given the number of predictions, denoted as N, we search for a permutation σ^* of N elements that yields the lowest cost. This is defined as:

$$\begin{aligned}\sigma^* = \arg\min_{\sigma\in\mathcal{P}} \sum_{i=1}^{N} -1_{\{c_i\neq\varnothing\}}\hat{p}_{\sigma(i)}\left(c_i\right) \\ + 1_{\{c_i=\varnothing\}}\mathcal{L}_{\text{box}}\left(\boldsymbol{b}_i, \hat{\boldsymbol{b}}_{\sigma(i)}\right)\end{aligned} \tag{1}$$

where $\mathcal{P}$ denotes sets of permutations, and $\varnothing$ represents no object. The term $\hat{p}_{\sigma(i)}\left(c_i\right)$ represents the probability of class c_i for the prediction with index $\sigma^*(i)$.

We then calculate the total loss through classification loss and box loss as follows:

$$\mathcal{L} = \sum_{i=1}^{N} -\lambda_{\text{cls}} \log \hat{p}_{\sigma^*(i)}\left(c_i\right) + 1_{\{c_i=\varnothing\}}\mathcal{L}_{\text{box}}\left(\boldsymbol{b}_i, \hat{\boldsymbol{b}}_{\sigma^*(i)}\right) \tag{2}$$

The classification loss is cross-entropy loss, and the box loss $\mathcal{L}_{\text{box}}$ is defined below:

$$\mathcal{L}_{\text{box}} = \lambda_{\text{iou}}\mathcal{L}_{\text{iou}}\left(b_i, \hat{b}_{\sigma(i)}\right) + \lambda_{\text{L1}}\left\|b_i - \hat{b}_{\sigma(i)}\right\|_1 \tag{3}$$

In the above equation, we use L1 loss and generalized IoU loss [3]. $\lambda_{\text{cls}}, \lambda_{\text{iou}}, \lambda_{\text{L1}} \in \mathbb{R}$ are hyperparameters to weight the different components of the loss function.

2. Positional Encoding Visualization

To better understand the role of instance-level features and their matching relationship with obstacles in the scene, we visualize the positional encodings of these features, both before and after filtering, as shown in Fig.1. In addition, we also visualize the LiDAR point clouds and camera images of both the vehicle and infrastructure sides, with the former represented by the color red and the latter by blue. Our visualization results reveal that our filtering method effectively removes redundant instance-level features, leaving only those relevant to the obstacles in the scene. The remaining positional encodings of these features can then accurately correspond to the obstacles detected in the LiDAR point cloud and camera image. Furthermore, we observe that most of the positional encodings on both sides overlap in the shared view area, indicating that the spatial position of our fusion queries between both sides is accurate. Overall, the visualization of the positional encodings helps clarify the role of instance-level features in obstacle detection and demonstrates our method's effectiveness.

3. Cross-Domain Adaptation Visualization

To intuitively show the effect of the impact of Cross-Domain Adaptation (CDA), we have included visualizations of the query before and after undergoing CDA from both the vehicle and infrastructure sides, as depicted in Fig.2 (a)-(d). Comparing (a) and (c) (before CDA) with (b) and (d) (after CDA), it can be seen that the color distribution is closer in the latter, indicating that the CDA module reduces the domain gap between the two sides. Additionally, the vehicle-side query is enriched with information from the infrastructure side, as seen in (b) (vehicle side after CDA), where the high score distribution is complemented on the left side by (c) (infrastructure side before CDA).

4. More Qualitative Results

We provided more qualitative results of DAIR-V2X [4] in Fig.3. The left-hand side of the figure shows the no collaboration results obtained by a perception system that does not utilize infrastructure-side information. In contrast, the right-hand side shows the results obtained by our TransIFF model. As can be seen from the figure, the results obtained

*Corresponding author

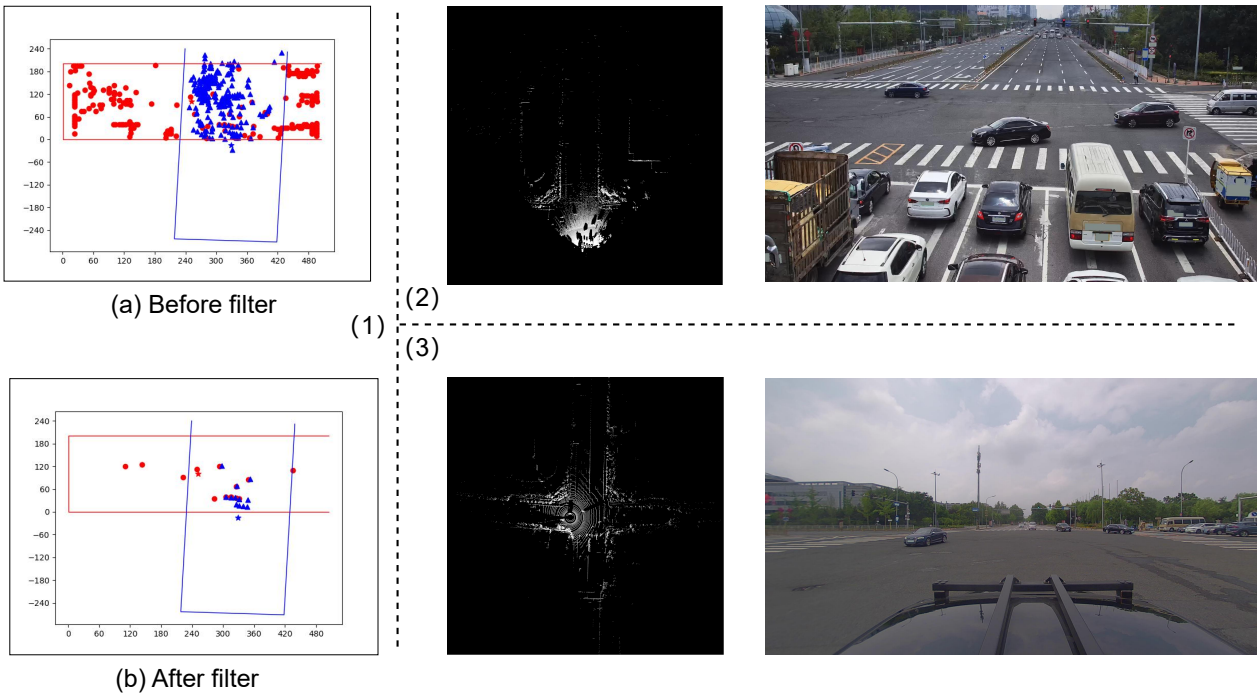

Figure 1. **Comparison of Position Encoding before and after Filtering.** In (1), we visualize the positional encoding (a) before the filtering module and (b) after the filtering module. In (2) and (3), we visualize the corresponding LiDAR point clouds and camera images at the vehicle and infrastructure sides. In (a) and (b), The pentagram represents the ego agent, red represents the vehicle, and blue represents the infrastructure. The red and blue pentagrams represent the vehicle and infrastructure agents, respectively. The orientation of both sensors is towards the opening of the rectangle. We choose the grid coordinate system of the vehicle as the coordinate system in figure (a) (b). The red circle and blue triangle represent the query of vehicle and infrastructure, respectively.

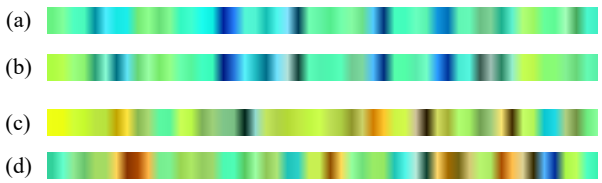

Figure 2. **Visualization of the query before and after Cross-Domain Adaptation.** (a) and (b) show the vehicle-side query before CDA and after CDA. (c) and (d) show the infrastructure-side query before CDA and after CDA.

by our model are better than the no collaboration results. The TransIFF method can effectively integrate vehicle and infrastructure-side information for cooperative perception, improving the accuracy and robustness of obstacle detection in complex driving scenarios.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 1

[2] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 1

[3] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 1

[4] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21361–21370, June 2022. 1

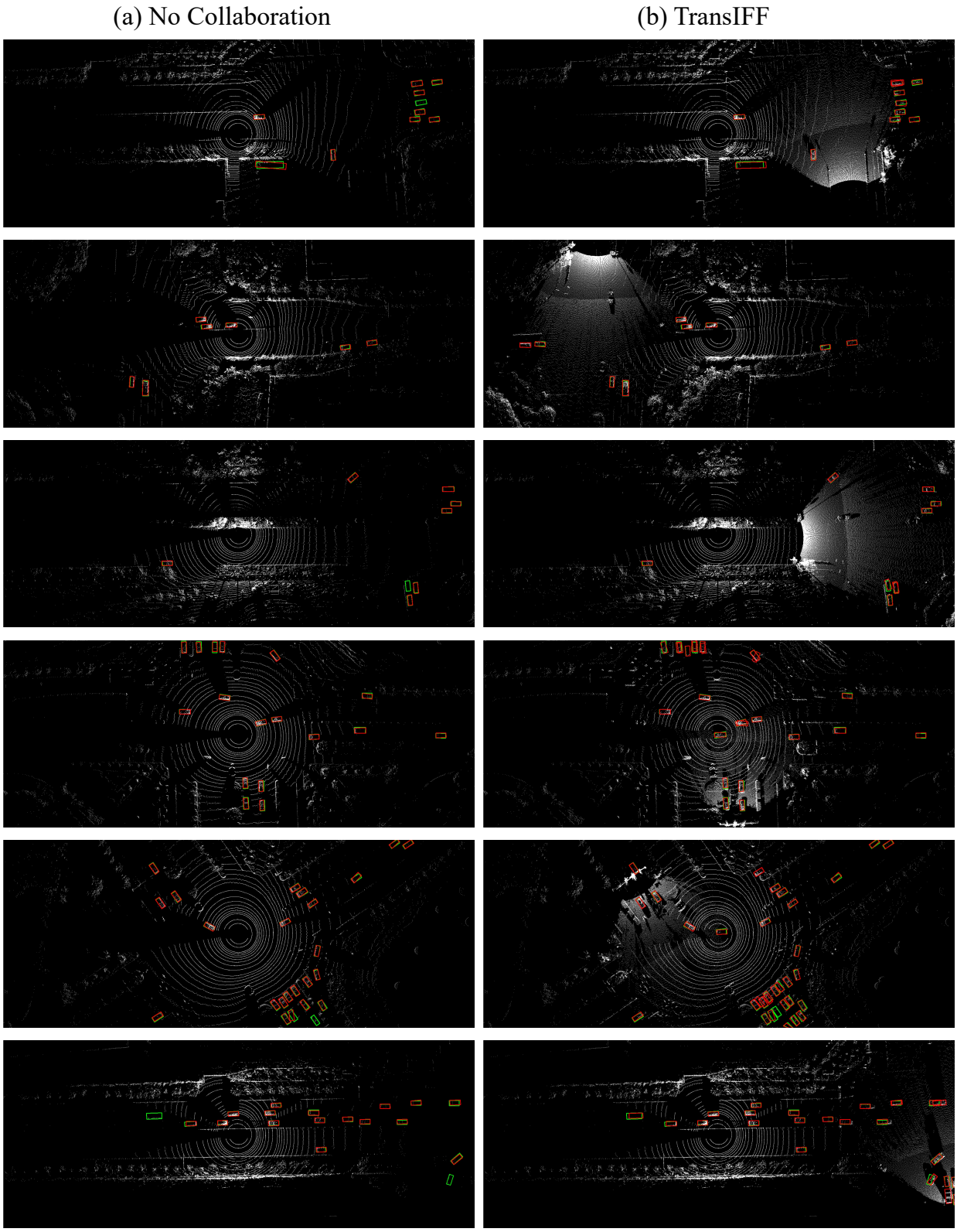

Figure 3. Qualitative results of No Collaboration and TransIFF in DAIR-V2X dataset.