LIFT: Learning 4D LiDAR Image Fusion Transformer for 3D Object Detection —Supplementary Material—

Yihan Zeng[1] Da Zhang[2] Chunwei Wang[1] Zhenwei Miao[2]
Ting Liu[2] Xin Zhan[2] Dayang Hao[2] Chao Ma[1*]
[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2] Alibaba DAMO Academy
{zengyihan, weiwei0224, chaoma}@sjtu.edu.cn, zhangda.zhang@alibaba-inc.com

In this supplementary material, we provide more details about sensor-time data augmentation and experimental results to complement the manuscript.

1. Sensor-Time Data Augmentation

Given a LiDAR virtual sample $O_{t'}$ from its original scene $S_{t'}$, we pick up the temporal consistent sample $O_{t'-1}$ from the previous scene $S_{t'-1}$, where $O_{t'}$ and $O_{t'-1}$ are respectively pasted into the sequential training scenes S_t and S_{t-1}. Besides, the corresponding image patches $I_{O_{t'}}$ and $I_{O_{t'-1}}$ are attached to the sequential camera images. Considering the cross-sensor consistency, we filter the point clouds that are occluded by pasted objects in each training scene. As illustrated in Figure 1, three objects are pasted into the sequential training scenes, including a static bicycle, a moving bicycle and a moving car. With our sensor-time projection operation, the pasted patterns maintain the temporal consistency across frames for both point clouds and camera images.

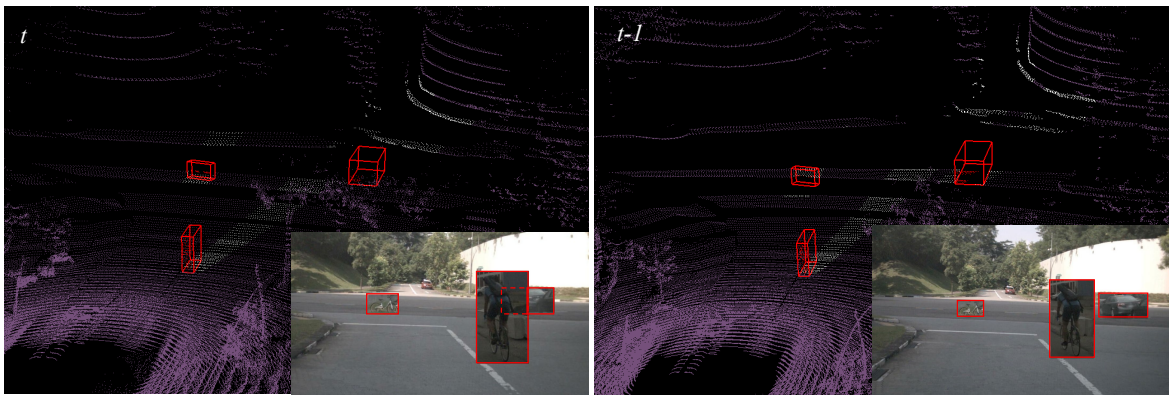

Figure 1. Visualization of sensor-time data augmentation. Red: Pasted virtual samples. Purple: Original point clouds of training scene. White: Filtered point clouds that are occluded by pasted objects. The pasted samples preserve both the cross- sensor and time consistency.

2. Additional Experimental Results

We show the qualitative results on the nuScenes dataset in Figure 2, which illustrates that LIFT improves the detection performance by greatly reducing the false positive predictions and increasing the detection consistency across frames. Specifically, the LiDAR-only detector PointPillar tends to produce false positives when detecting small or faraway objects with insufficient points, such as the pedestrians in the front camera view of Figure 2(a) and the car in the front camera view of Figure 2(b). While LIFT can effectively reduce those false predictions by incorporating the informative image features.

* Corresponding author.

Besides, due to the inconsistent sparse point clouds over time, the single-frame detector usually predicts discontinuous detection results as shown in Figure 2(a) and Figure 2(b). Our LIFT achieves much more consistent predictions by leveraging the temporal information from sequential input.

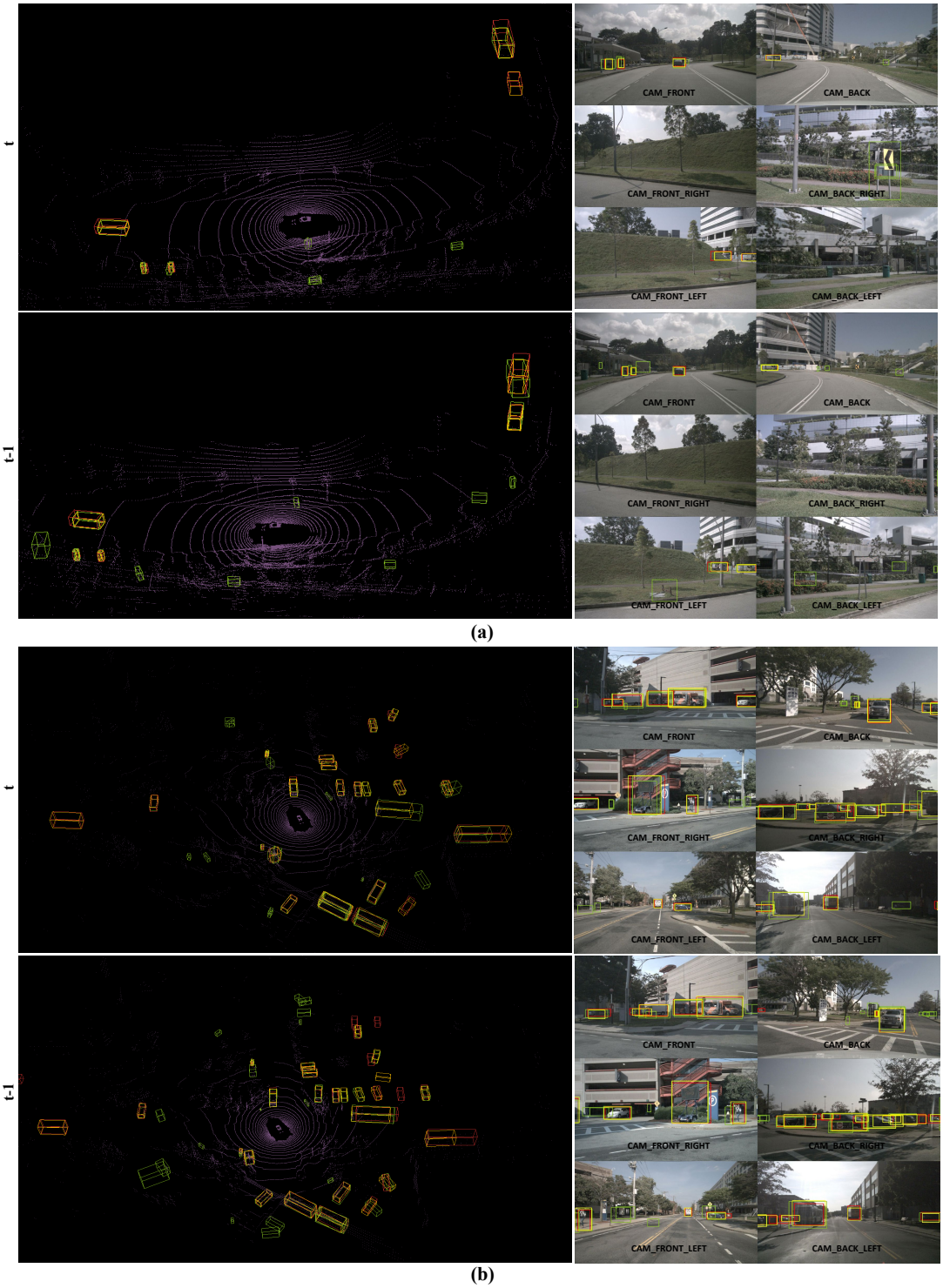

Figure 2. Qualitative results on the nuScenes dataset. Red: Ground truth. Green: PointPillar. Yellow: our LIFT. LIFT greatly reduces false positive predictions and increases the detection consistency across frames. Best viewed in color.