

Evaluating Occluded Pedestrian Detection and Enhancing Distant Recognition via Virtual Telephoto Cameras in Autonomous Driving

Yitong Wang¹⁾ Jishu Miao¹⁾ Yukiya Hattori¹⁾ Tsubasa Hirakawa¹⁾ Takayoshi Yamashita¹⁾
Hironobu Fujiyoshi¹⁾

*1) Chubu University, Graduate School of Engineering
1200 Matsumoto-cho, Kasugai-shi, Aichi 487-8501, Japan (E-mail: tp25001-9894@sti.chubu.ac.jp)*

KEY WORDS: Electronics and Control, Autonomous driving, Image processing [E1]

Reliable autonomous driving requires early perception of occluded road users and small distant objects. We constructed an ataset of pedestrians occluded by oncoming vehicles with stationary passenger cars or large vehicles facing the ego vehicle and pedestrians at varying positions behind or near their rear sides. Some pedestrians were visible from a distance, whereas others became visible only as the ego vehicle changed viewpoint. Across 22 evaluation scenes, initial stable detection occurred at an average distance of 40.5 m. Late detection may provide insufficient time for risk recognition and vehicle control, motivating longer-range detection without adding physical sensors.

To supplement distant image features, we add a virtual telephoto view to BEVFormer. The central distant region of the front-camera image is cropped, enlarged to the original resolution, and used as an additional view with adjusted intrinsics and the front camera's extrinsics. The original front view is retained for broad context. Simulator input therefore increases from six to seven views and aiMotive input from four to five views. Spatial Cross-Attention receives features in which distant objects occupy more pixels, whereas the remaining detector is unchanged. This strengthens forward long-range representation without modifying the physical sensor arrangement.

We compared the baseline BEVFormer and the proposed configuration on a simulator dataset generated from a single driving sequence and the public aiMotive Dataset. In the simulator, the proposed method achieved higher NDS and mAP in the near-, mid-, and long-range intervals. Because training and evaluation data came from the same sequence on one road, this is a preliminary within-condition comparison, not evidence of generalization. The aiMotive evaluation included highway, urban, night-time, and rainy scenes. The largest improvement occurred within 51.2 m, while gains at longer ranges were limited. Qualitative results nevertheless showed plausible detections of distant adjacent-lane objects lacking ground-truth annotations. The virtual telephoto view shows potential, but real-world long-range effectiveness requires complete annotations and independent data.

Overall, Fig. 1(a) shows the seven-view simulator result, and Fig. 1(b) the five-view aiMotive result; predictions (PRED) are above ground truth (GT). In simulation, the proposed method improved NDS and mAP across all distance intervals, although the shared source sequence limits the result to a preliminary within-condition comparison. On aiMotive, the clearest gain occurred within 51.2 m, with only small improvements at longer ranges. Some plausible distant detections lacked GT annotations and could be counted as false positives. Future work will add long-range annotations, use independent datasets, and optimize the crop position, enlargement factor, and input resolution.



(a) Simulated dataset: six cameras plus the virtual telephoto view



(b) aiMotive Dataset: four cameras plus the virtual telephoto view

Fig. 1 Qualitative detections (PRED) and ground truth (GT)