AFMCT: Adaptive Cross-modal Transformer Fusion for Stable 3D Pedestrian Perception Toward Robust Behavior Prediction

Yixin Wang [1], Bingli Zhang[1], Chengbiao Zhang[1], Xinyu Wang[1], Junzhao Jiang[1,*],Yifan Wang[1], Shuqing Zhao[1]

***Abstract*— effective Pedestrian Behavior Prediction (PBP) requires precise and continuous understanding of their 3D spatial states. However, existing multi-modal perception systems often suffer from suboptimal feature alignment, leading to noisy or flickering bounding boxes that severely degrade downstream trajectory forecasting. In this paper, we propose an Adaptive Fusion Module based on Cross-modal Transformer blocks (AFMCT) to construct a robust, unified multimodal spatial representation. By utilizing a bidirectional enhancing strategy, AFMCT deeply fuses the geometric cues from LiDAR with the rich semantic context from cameras. Evaluated on the widely-used KITTI 3D detection benchmark, our proposed AFMCT significantly improves 3D detection accuracy, outperforming existing advanced multi-sensor fusion modules by at least 0.56% in overall mean Average Precision (mAP), with improved detection of pedestrians and cyclists in safety-critical scenes. This high-fidelity 3D perception foundation proves essential for resolving the uncertainties in downstream pedestrian intent and behavior prediction.**

I. Introduction

In highly socialized environments such as urban intersections, effective Pedestrian Behavior Prediction (PBP) is a cornerstone for the safety of Autonomous Vehicles (AVs) and mobile robots [1]. A critical yet often overlooked challenge in current PBP pipelines is the heavy reliance on the quality of upstream perception. Predictors like LSTMs or Graph Neural Networks expect continuous, accurate tracklets of pedestrians [2]. However, in real-world scenarios, heterogeneous sensor data (such as cameras and LiDAR) are difficult to fuse thoroughly. Single-modal or poorly fused algorithms frequently produce false negatives or bounding box "flickering" under occlusion, cascading into failures in pedestrian intent prediction.

To address this context and representation challenge, we propose AFMCT, an Adaptive Fusion Module based on Cross-modal Transformer blocks. Instead of simple concatenation, AFMCT generates a discriminative multimodal spatial representation. High-quality 3D detection—capturing precise spatial locations from LiDAR and human-centric cues from cameras—is the prerequisite for handling pedestrian dynamics.

[1]School of Automotive and Transportation Engineering, Hefei University of Technology, Hefei, 230009, China

* Correspondence: chlgch.2006@hfut.edu.cn).

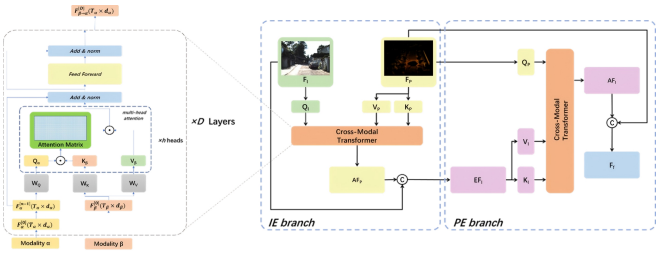

Fig. 1. Illustration of AFMCT. It consists of two components: IE branch utilizes cross-modal transformer to enhance image feature, and PE branch uses another cross-modal transformer to enhance point feature

II. The Proposed Approach

As shown in Fig. 1, AFMCT adopts a bidirectional structure consisting of an IE branch and a PE branch. The IE branch first enhances image feature F_I with point feature F_p to produce EF_I. Subsequently, EF_I and F_p are taken as inputs of PE branch to enhance F_p. We argue that our module can be seamlessly integrated into any multi-modal fusion network for 3D object detection.

A. IE branch

IE branch enhances image features based on the cross-modal transformer block. The image feature $F_I \in \mathbb{R}^{T_I \times d_I}$ and the point feature $F_P \in \mathbb{R}^{T_P \times d_P}$ are first normalized to the same feature channel d_n by fully-connected layers. The normalized image feature F_I' is projected into $Q_I \in \mathbb{R}^{T_I \times d_K}$, while the normalized point feature F_P' is projected into $K_P \in \mathbb{R}^{T_P \times d_K}$ and $V_P \in \mathbb{R}^{T_P \times d_I}$. Then Q_I, K_P, and V_P are fed into the cross-modal transformer block to generate an attention-based point feature $AF_P \in \mathbb{R}^{T_I \times d_n}$. Since AF_P is aligned with F_I and has the same size, IE branch concatenates F_I' and AF_P to enhance F_I' with geometric and depth information from Lidar. Another fully-connected layer transforms the feature channel into d_n. The output of this branch is an enhanced image feature $EF_I \in \mathbb{R}^{T_I \times d_n}$.

B. PE branch

Following the IE branch, the PE branch uses a similar strategy to enhance one modality with the other. The cross-modal transformer block takes the enhanced feature EF_I and the normalized point feature F_P' as input, where the former provides $Q_P \in \mathbb{R}^{T_P \times d_K}$, while the latter provides $K_I \in \mathbb{R}^{T_I \times d_K}$ and $V_I \in \mathbb{R}^{T_I \times d_P}$. The module then outputs an adaptively extracted image feature $AF_I \in \mathbb{R}^{T_P \times d_n}$. Next,

F'_P and AF_I are concatenated and fed into a fully-connected layer to obtain a channel number d_P, so that the fused feature has the same size as F_P. The final result of PE branch is denoted as F_f, an enhanced point feature that can be fed into any point-based detection network to generate specific predictions.

III. Experimental

A. Experimental Setup and Baselines

We evaluated our method on the challenging KITTI 3D Object Detection Benchmark, which serves as a representative proxy for assessing spatial representations critical to urban pedestrian behavior prediction. The evaluation focused on the accuracy and stability of 3D bounding boxes as essential inputs for downstream trajectory forecasting. To validate the effectiveness of our fusion strategy, we compared against several state-of-the-art modules, including LI-Fusion from EPNet [3], CB-Fusion from EPNet++ [4], LearnableAlign from DeepFusion [6], CMT from CAT-Det [5], and DFT from 3D Dual-Fusion [7]. Experimental results show that our adaptive design consistently outperforms these baselines.

B. Comparison with state-of-the-art fusion strategies

TABLE I

Comparison with state-of-the-art fusion strategies on the KITTI *val* set (Pedestrian category)

Fusion Module	Year	Pedestrian		
		Easy	Moderate	Hard
LI-Fusion[3]	2020	62.63	53.77	47.55
CB-Fusion[4]	2021	63.06	54.01	48.18
LearnableAlign[6]	2022	63.54	53.98	48.42
CMT[5]	2022	55.74	39.13	34.84
DFT[7]	2022	64.12	54.03	48.88
AFMCT(ours)	–	**64.74**	**54.32**	**49.65**

Table I reports results on the KITTI *val* set for the Pedestrian category. To further prove the effectiveness of our AFMCT, results on the KITTI *test* set for the Car category are additionally reported. Compared with state-of-the-art multi-modal detectors, the proposed method exhibits superior precision, and achieves a gain of 0.56% in mAP over the state-of-the-art method 3D Dual-Fusion [7].

C. Qualitative results and discussion

We evaluated AFMCT on the KITTI dataset to demonstrate its benefits. Compared with the LiDAR-only method, AFMCT shows three main advantages: (1) it improves class scores by increasing true positives and reducing false positives, with higher scores for the emphasized bounding boxes; (2) it detects objects missed by the LiDAR-only method, as shown in Fig. 2 for highlighted cars, pedestrians, and cyclists; and (3) it corrects some judgment errors of the LiDAR-only method, further improving precision. For example, as shown in Fig. 2, the LiDAR-only network incorrectly identifies a wall as an automobile, while our network corrects this error. Overall, AFMCT outperforms the baseline in classification performance. We attribute this improvement to the transformer in AFMCT, which precisely aligns multimodal data and enhances object detection performance.

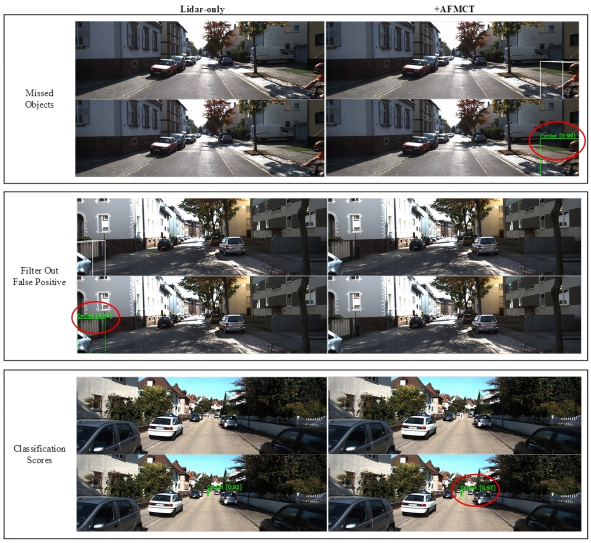

Fig. 2. Visualization of object detection results on the KITTI *val* set

IV. Conclusion

Robust pedestrian behavior prediction cannot be decoupled from robust perception. In this work, we present AFMCT, an adaptive cross-modal transformer fusion module that provides high-fidelity, unified 3D spatial representations. By significantly improving detection accuracy and stability, AFMCT lays a critical foundation for intent understanding. Future work will focus on closing this gap by integrating AFMCT's perception outputs with a stochastic trajectory prediction head, optimizing perception and forecasting jointly in dynamic social environments to directly validate the benefit for pedestrian behaviour prediction.

References

[1] A. Rasouli and J. K. Tsotsos, "Autonomous Vehicles That Interact With Pedestrians: A Survey of Theory and Practice," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 900–918, 2020, doi: 10.1109/TITS.2019.2901817.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human Trajectory Prediction in Crowded Spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[3] T. Huang, Z. Liu, X. Chen, and X. Bai, "EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 35–52.

[4] Z. Liu, T. Huang, B. Li, X. Chen, X. Wang, and X. Bai, "EPNet++: Cascade Bi-directional Fusion for Multi-Modal 3D Object Detection," arXiv preprint arXiv:2112.11088, 2022.

[5] Y. Zhang, J. Chen, and D. Huang, "CAT-Det: Contrastively Augmented Transformer for Multi-Modal 3D Object Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 908–917.

[6] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le, A. Yuille, and M. Tan, "DeepFusion: Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 17182–17191.

[7] Y. Kim, K. Park, M. Kim, D. Kum, and J. W. Choi, "3D Dual-Fusion: Dual-Domain Dual-Query Camera-LiDAR Fusion for 3D Object Detection," arXiv preprint arXiv:2211.13529, 2023.