Vision-Based Pedestrian-Vehicle Interaction Risk Assessment in Safety-Critical Scenarios

Yiran Luo, Siyuan Liu, Wei Lu, Di Miao, Qing Zhou, Bingbing Nie*

I. INTRODUCTION

In automated driving, pedestrian injury in vehicle conflicts is strongly influenced by how reliably hazardous pedestrian-vehicle interactions can be recognised before impact [1]. Existing safety studies have mainly used surrogate safety measures and kinematic indicators to assess pre-crash risk, and vision-based deep learning methods have also been introduced for traffic risk prediction and traffic accident anticipation [2][3]. However, multi-level pedestrian-vehicle interaction risk in real world automated driving scenes remains insufficiently studied. In these interactions, pedestrian posture is an important visual cue for judging danger and may provide information that is not fully captured by simplified risk variables. This study focused on pedestrian-vehicle interaction risk assessment using vision backbone networks with real world automated driving data. Pedestrian posture was further incorporated to strengthen risk-related visual cues. The aim was to support more reliable and interpretable risk assessment in automated driving, and ultimately contribute to pedestrian injury mitigation.

II. METHODS

To support pre-crash injury mitigation, this study developed a vision-based framework for assessing pedestrian-vehicle interaction risk in real-world autonomous driving scenarios. As shown in Fig. 1, front-view images were collected from naturalistic road tests of autonomous vehicles in urban traffic, with safety drivers providing the reference for risk annotation in critical interactions. We defined three hierarchical risk levels based on safety driver judgment. Level 3 represents urgent risk, corresponding to situations in which immediate takeover was required to avoid an imminent collision and was typically accompanied by strong evasive manoeuvres. Level 2 represents moderate risk, in which the pedestrian did not yet pose an immediate collision threat but the situation already required cautious observation or mild preventive adjustment. Level 1 represents non-urgent risk, where sufficient spatiotemporal margin remained for safe passage without specific avoidance action. The risk prediction framework is shown in Fig. 1, and the overall formulation is given in Eq. (1):

$$\hat{r} = f_{MLP}\left(f_{backbone}(X)\right) \tag{1}$$

$$X = (1-\gamma)\, I + \gamma(\alpha I + (1-\alpha)\, P) \tag{2}$$

Where I is the original image, P is the pedestrian pose representation, X is the final input to the network. $\gamma \in 0,1$ controls whether pose fusion is used. α is the blending coefficient. $f_{backbone}(\cdot)$ denotes an ImageNet-pretrained visual backbone, such as Swin-T or ViT, and $f_{MLP}(\cdot)$ denotes the prediction head that maps the extracted features to the final risk level.

Beyond the final risk prediction, we projected the backbone features into a low-dimensional space for visualization. The purpose was to examine whether the learned image features preserved the hierarchical structure of the three risk levels and showed clear discrimination across different risk conditions. This analysis helps reveal whether the proposed framework learns interpretable risk-related representations from visual input. The specific methodology for this projection is detailed in [4].

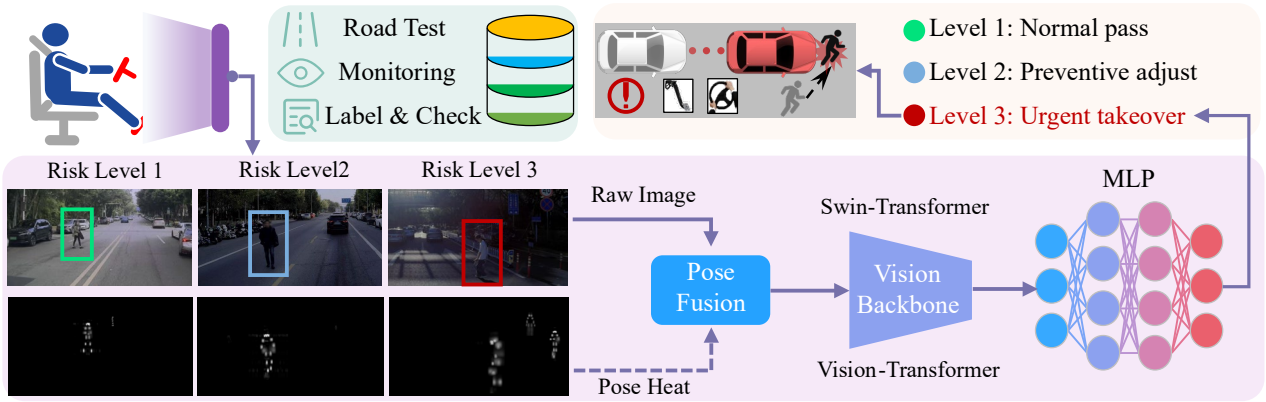

Fig. 1. Framework for hierarchical pedestrian-vehicle interaction risk assessment based on vision network.

B. Nie (e-mail: nbb@tsinghua.edu.cn; tel: +86-10-6278-8689) is an Associate Professor, Y. Luo and S. Liu are PhD students, and Q. Zhou is a Professor, all in the School of Vehicle and Mobility at Tsinghua University and State Key Laboratory of Intelligent Green Vehicle and Mobility, China. W. Lu and D. Miao are Engineers at Didi Global, China.

III. INITIAL FINDINGS

The experimental dataset comprised 423 Level 1, 423 Level 2, and 501 Level 3 cases from real-world autonomous driving road tests, with only cases confirmed by at least three safety drivers retained for analysis. Experiments were conducted with two visual backbones, Swin-T and ViT, to examine whether pedestrian risk levels can be inferred from image observations. Table 1 indicates that meaningful risk-related information can be recovered from visual observations, and that this ability becomes stronger when pedestrian pose is incorporated. Without pretraining from ImageNet, both backbones showed limited discrimination, suggesting that risk-related cues are not readily organized from scratch. After pretraining, the distinction between risk levels became much clearer, indicating that the visual backbone can form more reliable representations for recognizing interactions with different risk implications. Adding pose fusion further improved the results, indicating that pedestrian posture provides complementary information beyond scene appearance alone.

Figure 2 provides a feature-space view of the progressive separation among the three risk levels during training. Early in training, the three risk levels remain highly mixed, whereas later epochs show clearer grouping. This suggests that the visual backbone is able to capture cues that are relevant to human risk judgment.

TABLE I

EFFECT OF POSE FUSION AND BACKBONE PRETRAINING ON RISK PREDICTION PERFORMANCE

Backbone	Pose fusion	Precision	Recall	Specificity	F1-score	Geo-mean
Swin-T (Do not pretrain)	-	0.411	0.400	0.700	0.355	0.327
Swin-T	-	0.789	0.791	0.894	0.790	0.791
	√	0.868	0.870	0.932	0.869	0.869
ViT (Do not pretrain)	-	0.479	0.358	0.684	0.329	0.302
ViT	-	0.748	0.752	0.875	0.750	0.750
	√	0.844	0.839	0.916	0.841	0.838

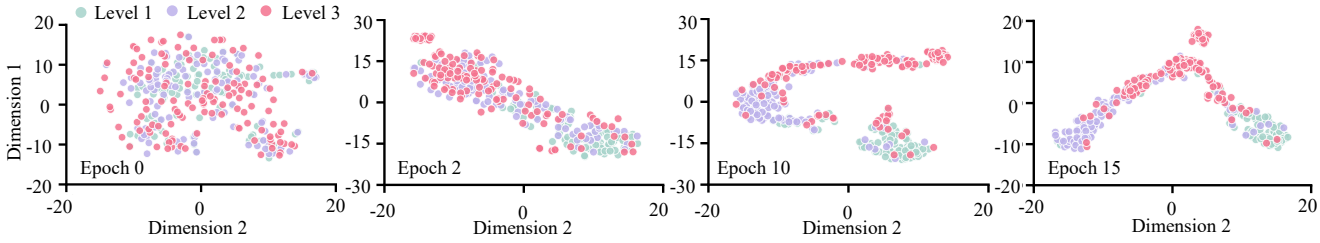

Fig. 2. t-SNE visualization of feature evolution on the validation set across training epochs for three risk levels.

IV. DISCUSSION

Pedestrian injury prevention in automated driving depends on timely recognition of hazardous pedestrian-vehicle interactions before impact. The present findings show that visual information from real-world driving scenes may support hierarchical risk assessment for autonomous driving safety and pre-crash injury mitigation. This suggests that injury-relevant interaction risk is not reflected only by simplified kinematic indicators, but can also be inferred from visual scene observations. Present results show that lightweight visual backbones can distinguish image features associated with different risk levels. This indicates that pretrained visual models can be adapted for pedestrian-vehicle risk assessment in safety-relevant events, while pose fusion adds complementary information beyond scene appearance alone. However, human risk judgment is inherently dynamic. Future work should therefore incorporate temporal modeling and other human-related cues to better reflect the development of pre-crash risk and strengthen the injury mitigation value of the framework.

V. ACKNOWLEDGEMENTS

This study is supported by the National Natural Science Foundation of China (52572413).

VI. REFERENCES

[1] JB Cicchino *et al., Accid Anal Prev*, 2022.
[2] GP Corcoran *et al.*, *CRV*, 2019.
[3] Bao W *et al., ACM MM*, 2020.
[4] L Van der Maaten *et al.*, *JMLR*, 2008.