Accident Analysis and Prevention 194 (2024) 107360

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Improving model robustness of traffic crash risk evaluation via adversarial mix-up under traffic flow fundamental diagram

Rongjie Yu [a], Lei Han [b], Mohamed Abdel-Aty [b], Liqiang Wang [c], Zihang Zou [c,*]

[a] The Key Laboratory of Road and Traffic Engineering, Ministry of Education, 4800 Cao'an Road, 201804 Shanghai, China
[b] Department of Civil, Environmental & Construction Engineering, University of Central Florida, Orlando, FL 32816, USA
[c] Department of Computer Science, University of Central Florida, Orlando, FL 32816HEC 437E, USA

ARTICLE INFO

Keywords:
Crash risk evaluation model
Model robustness
Traffic flow fundamental diagram
Traffic flow adversarial example
Adversarial training

ABSTRACT

Recent state-of-art crash risk evaluation studies have exploited deep learning (DL) techniques to improve performance in identifying high-risk traffic operation statuses. However, it is doubtful if such DL-based models would remain robust to real-world traffic dynamics (e.g., random traffic fluctuations.) as DL models are sensitive to input changes, where small perturbations could lead to wrong predictions. This study raises the critical robustness issue for crash risk evaluation models and investigates countermeasures to enhance it. By mixing up crash and non-crash samples under the traffic flow fundamental diagram, traffic flow adversarial examples (TF-AEs) were generated to simulate real-world traffic fluctuations. With the developed TF-AEs, model accuracy decreased by 8% and sensitivity dropped by 18%, indicating weak robustness of the baseline model (a convolutional neural network, CNN-based crash risk evaluation model). Then, a coverage-oriented adversarial training method was proposed to improve model robustness in highly imbalanced crash and non-crash situations and various crash risk transition patterns. Experiments showed that the proposed method was effective to improve model robustness as it could prevent 76.5% accuracy drops and 98.9% sensitivity drops against TF-AEs. Finally, the evaluation model outputs' stability and limitations of the current study are discussed.

1. Introduction

Crash risk evaluation models are developed to identify high-risk traffic operation statuses based on real-time traffic flow data (Abdel-Aty et al., 2010). Their outputs are essential for triggering road crash warnings and implementing traffic management measures that proactively prevent crash occurrence (Hossain et al., 2019). In recent years, deep learning (DL) models (e.g., Convolutional Neural Networks, CNN in Yuan et al., 2019; Yu et al., 2020; Long Short-term Memory, LSTM in Li et al., 2020; Zhang and Abdel-Aty, 2022; Li and Abdel-Aty, 2022) have been widely applied in order to improve model performance. Compared with traditional statistical methods (Xu et al., 2013; Yu et al., 2013; Hossain et al., 2019), DL models are capable of fitting complicated nonlinear mapping functions with a modest number of parameters to extract high-dimension features among data, contributing to significant improvements of accuracy by 10–20 % (Yu et al., 2020; Li et al., 2020; Li and Abdel-Aty, 2022).

However, whether these improvements would remain consistent for real-world traffic dynamics is skeptical. DL models are sensitive to small perturbations, which could lead to misclassified results (Szegedy et al., 2013; Madry et al., 2017; Tsipras et al., 2018). This problem may be due to the fact that DL models excessively focus on the non-robust features within the data distributions, which could be obscured and corrupted when small perturbations are introduced (Ilyas et al., 2019). The fragile robustness of DL models has been proven in varying fields such as computer vision (Goodfellow et al., 2014; Silva and Najafirad, 2020), natural language processing (Jia and Liang, 2017; Wang et al., 2019a), and speech processing (Wang et al., 2020b; Omar et al., 2022). Similarly, in real-world traffic dynamics, traffic flow (characterized by the fundamental parameters such as speed, flow and density) would stochastically fluctuate under specific coupling relationships (Wang et al., 2013; Qu et al., 2017). Therefore, it is questionable if DL-based crash risk evaluation models would retain good model performance given the traffic flow fluctuations.

Before studying the robustness of crash risk evaluation models, one question that must be answered: what kind of perturbations would

* Corresponding author.
E-mail addresses: yurongjie@tongji.edu.cn (R. Yu), le966091@ucf.edu (L. Han), m.aty@ucf.edu (M. Abdel-Aty), lwang@cs.ucf.edu (L. Wang), zzz@knights.ucf.edu (Z. Zou).

https://doi.org/10.1016/j.aap.2023.107360
Received 31 March 2023; Received in revised form 2 October 2023; Accepted 16 October 2023
Available online 26 October 2023
0001-4575/© 2023 Published by Elsevier Ltd.

characterize traffic flow data in the real world? Unlike image processing where images with small perturbations would remain the same in human recognition (Szegedy et al., 2013), there is no guarantee that a small perturbation in traffic flow data would remain to be real. Since the parameterized traffic flow data (e.g., speed, occupancy, etc.) are in different scales or different units, finding a proper perturbation for each feature would be difficult. Furthermore, adding random perturbation around samples would no longer be applied as traffic flow data does follow certain principles (Greenshields et al., 1935; Greenberg, 1959; Qu et al., 2017). For example, traffic speed correlates with density in a negative relation (Greenshields et al., 1935). Therefore, it is necessary to consider classical traffic flow fundamental diagram as constraints during perturbation when studying model robustness in crash risk evaluation.

To analyze the model robustness issue for DL-based crash risk evaluation models, a preliminary experiment was first conducted to investigate the vulnerability of a formerly developed CNN-based crash risk evaluation model (Yu et al., 2020). As shown in Fig. 1, after adding small-scales (less than 0.05) of traffic flow fluctuations (e.g., changing average speed less than 2 km/h or changing average road occupancy less than 5 %), 13 % of the samples would reverse their prediction which lead to misclassification. The significant model accuracy drops indicating the necessity of improving the model robustness under the real-world traffic dynamics.

In summary, the model robustness issue in crash risk evaluation was investigated in this study. The problem of DL-based crash risk evaluation models' weak robustness was analyzed and a corresponding solution was proposed to improve their robustness. Main contributions of the study are as follows:

1) Proposed a traffic flow adversarial example (TF-AE) generation method via data mix-up under traffic flow fundamental diagram. With the developed TF-AEs under small scale of traffic fluctuations (0.1 of perturbation scale), model accuracy decreases by 8.0 % and sensitivity drops by 18.0 %, revealing the fragile robustness of the CNN-based crash risk model.
2) Developed a coverage-oriented adversarial training method to improve model robustness. Results showed that adversarial-trained model could avoid 76.5 % accuracy drops and 98.9 % sensitivity drops against TF-AEs.
3) Compared to conventional training method, the crash risk evaluation model with adversarial training provides more stable outputs to real-world traffic dynamic fluctuations, as reflected by its time-varying volatility measure reduction of 11.5 %.

The rest of the article was organized as follows. The second section reviews the previous studies on the model robustness in DL. The third section provides detailed description of our methodology. The fourth section is data preparation and the fifth section is modeling results. Finally, discussion and conclusion of the current work are presented.

2. Background

Model robustness is the ability of a model to maintain good model performance in the face of small deviations from the model inputs or assumptions (Huber, 2011). In the field of deep learning, although CNN and other DL models have become increasingly effective at many difficult tasks, researchers have found that they are vulnerable to data perturbations and have poor model robustness (Szegedy et al., 2013; Goodfellow et al., 2014). For example, an image with imperceptible perturbations can easily mislead a well-trained DL model to give a completely wrong result (Ilyas et al., 2019). The fragile robustness of DL models calls into question safety–critical applications deployed by DL models in varying fields (Jia and Liang, 2017; Silva and Najafirad, 2020; Omar et al., 2022).

Recent advances in deep learning (Carlini and Wagner, 2017; Madry et al., 2017; Xiao et al., 2018; Wang et al., 2019b; Liu et al., 2023; Wu et al., 2023) utilize adversarial examples as a key tool to study the robustness of neuron models. Adversarial examples are the perturbed samples that are almost identical to real examples but would lead to wrong prediction (Szegedy et al., 2013). These examples can commonly

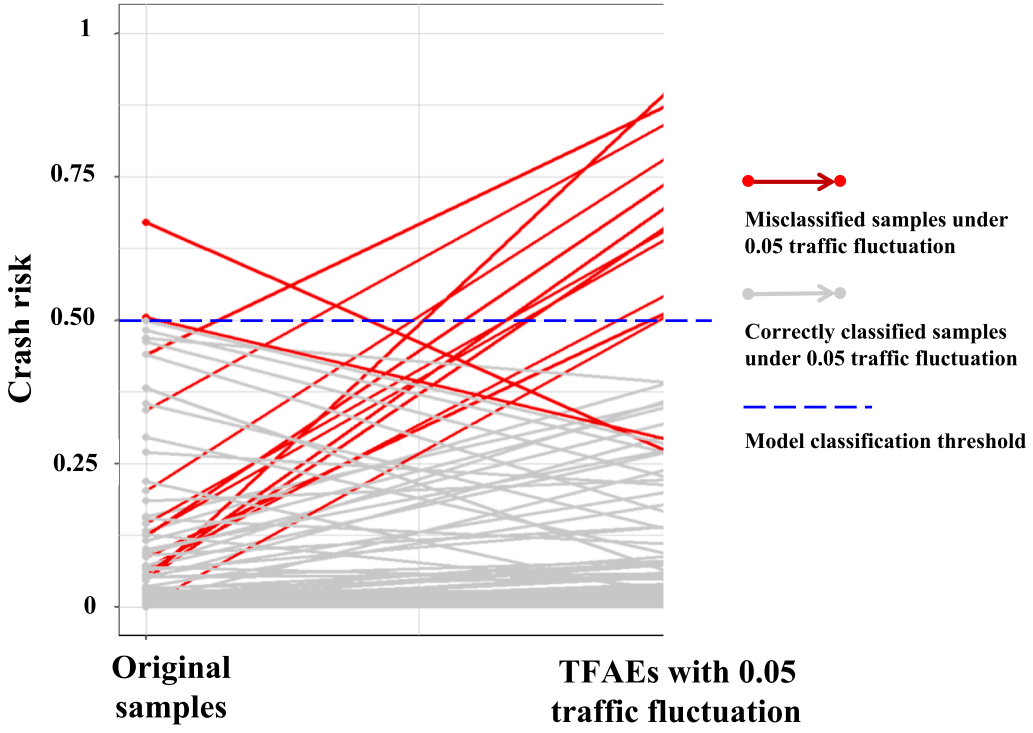


Fig. 1. A preliminary study conducted with CNN model on Shanghai urban traffic flow data. 13 out of 100 correctly classified samples wound reverse their prediction to be misclassified under a very small traffic fluctuation.

be found via gradient-based (Goodfellow et al., 2014; Madry et al., 2017), optimization-based (Papernot et al., 2016; Carlini and Wagner, 2017), and generative methods (Xiao et al., 2018; Wang et al., 2019b). Gradient-based methods (Madry et al., 2017; Liu et al., 2023) construct adversarial examples by finding perturbed example with gradient descent that can significantly increase model loss; While the optimization-based methods construct and minimize an objective function concerning the perturbation (Carlini and Wagner, 2017; Wu et al., 2023); For the generative methods, generative adversarial networks are being trained to generate perturbated samples almost identical to the original samples while being able to confuse the classifiers (Wang et al., 2019b; Hu and Tan, 2023). Among them, the gradient-based methods (e.g., fast gradient sign method in Goodfellow et al., 2014; projected gradient descent in Madry et al., 2017) have been widely utilized because of their high computational efficiency and flexibility.

To improve the fragile model robustness revealed by adversarial examples, adversarial training has been widely adopted as one of the most straightforward and efficient solution (Zhao et al., 2022). Instead of training original examples, adversarial training explores adversarial examples on the fly and use these examples for model training (Madry et al., 2017; Bai et al., 2021; Li and Spratling, 2023). An alternative adversarial training approach designs a specific penalty as regularization against adversarial examples, preventing misclassification and improve model robustness (Kannan et al., 2018; Huang et al., 2022).

However, the above methods are mainly in the image domain and domain knowledge is required to construct adversarial examples when they are applied to different fields (Silva and Najafirad, 2020). For example, in natural language processing, Jia and Liang (2017) first introduced word construction rules and semantic readability to traditional AE generation methods to construct word or sentence level adversarial examples; In speech processing researches, the generated perturbations are transferred into acoustic features like Mel Frequency Cepstral Coefficient and then reconstructed to speech waveform (Wang et al., 2020b; Omar et al., 2022). Similarly, there are coupled nonlinear relationships among traffic speed, flow, and density in traffic flow data (Greenshields et al., 1935; Underwood, 1961; Wang et al., 2013; Qu et al., 2017). On the purpose of evaluating and improving robustness of crash risk evaluation models, it is reasonable to construct adversarial examples that fit the traffic flow theory.

3. Methodology

3.1. Generating TF-AEs via adversarial mix-up under traffic flow fundamental diagram

Adversarial examples (Szegedy et al., 2013) are the key toward studying robustness of a model. By perturbing examples from real data and finding corresponding adversaries, the robustness of a model can be evaluated by comparing the change of accuracy on these perturbed examples. Following the same idea, a specific interpolation space for perturbing examples under traffic flow fundamental diagram is first studied. Then in this traffic flow interpolation space, adversarial examples are generated by gradient searching the neighborhood regions around data points.

(1) Traffic flow interpolation space via data mix-up.

Unlike image processing where adding a small perturbation would not change human recognition, random perturbation appended on traffic flow samples could create invalid perturbed examples. For example, a stochastic fluctuation on an urban highway (average speed is 90 km/h and traffic occupancy is 25 % in normal status) can be an increase of 5 km/h in traffic average speed along with 10 % increase in traffic occupancy, which is less likely to happen in reality.

Fortunately, the underline principles in traffic flow such as the negative relationship between traffic speed and density have been studied (Wang et al., 2013; Qu et al., 2017). As a result, it is tempting to adapt perturbation under these principles. However, it is impossible to obtain the proper scale of perturbation because of the varying scales of traffic flow data and complex interventions among parameters. To resolve this problem, an alternative data mix-up method (Zhang et al., 2017; Wang et al., 2020a; Wang et al., 2021) is proposed. Instead to add perturbation within empirical scales on either crash or non-crash data, perturbed traffic flow samples can be generated by taking a proper interpolation on a pair of crash and non-crash data.

To guarantee such interpolation space follows basic traffic flow fundamental diagram model, the flow-speed-density constraints (Greenshields et al., 1935; Greenberg, 1959; Qu et al., 2017) are considered. Formally, given parameterized traffic flow data, traffic flow x^F(vehs/h), speed x^S (km/h), and traffic occupancy x^O (%) follow:

$$x^F = f_{F-O,S}(x^O, x^S) = k^* x^O * x^S \tag{1}$$

where k is the linear estimation coefficient. Moreover, as previous research (Greenshields et al., 1935; Qu et al., 2017) show speed x^S negatively correlates with traffic occupancy x^O in various ways, speed x^S can also be denoted by a function of x^O,

$$x^S = f_{S-O}(x^O) \tag{2}$$

with its explicit formula shown in Table 1, the mean square error and relative error are compared to select the best speed-occupancy relations function among them.

From principles above which can be represented by functions of x^O, a perturbed traffic flow example x_λ can be generated by interpolating a pair of crash and non-crash sample $X = (x_{crash}, x_{non-crash})$ as,

$$x_\lambda = \begin{cases} x_\lambda^O = \lambda^* x_{crash}^O + (1-\lambda)^* x_{non-crash}^O \\ x_\lambda^S = f_{O-S}(x_\lambda^O) \\ x_\lambda^F = f_{F-O,S}(x_\lambda^O, x_\lambda^S) \\ x_\lambda^\alpha = \lambda^* x_{crash}^\alpha + (1-\lambda)^* x_{non-crash}^\alpha, \alpha \in [SO, SS, SF] \end{cases} \tag{3}$$

where λ is the interpolation rate, ranging from [0, 1]. Clearly, x_λ would become non-crash like example when λ is approaching 0 and vice versa. A linear interpolation is first applied on occupancy x^O between crash and non-crash data. Then speed and traffic flow are derived from traffic flow fundamental diagram in Equations (1) and (2). These standard deviation variables $x^{\alpha \in [SO,SS,SF]}$, which are found to be positively correlated with the crash risk (Roshandel et al., 2015; Hossain et al., 2019), are later linearly interpolated, to present the consecutive range of speed occupancy and volume between crash and non-crash examples from crash risk prediction. Through mixing up many pairs of crash and non-crash

Table 1
Several prominent speed-occupancy relations $f_{S-O}(x^O)$.

Models	Speed-occupancy function	Parameters
Greenshields et al. (1935)	$x^S = x^{S_f} * (1 - \frac{x^O}{x^{O_j}})$	x^{S_f}, x^{O_j}
Greenberg (1959)	$x^S = x^{S_c} * ln(\frac{x^{O_j}}{x^O})$	x^{S_c}, x^{O_j}
Underwood (1961)	$x^S = x^{S_f} * exp(-\frac{x^O}{x^{O_j}})$	x^{S_c}, x^{O_j}
Three-parameter logistics (3PL) model (Wang et al., 2011)	$x^S = \frac{x^{S_f}}{1 + exp(\frac{x^O - x^{O_c}}{\theta})}$	x^{O_j}, x^{S_f}, x^{O_c}, θ

Notation: x^{S_f} - free-flow speed; x^{O_j} - jam occupancy; x^{S_c} - at-capacity speed; x^{O_c} - at-capacity occupancy; θ - calibration coefficients.

data, a traffic flow interpolation space can be constructed, covering a large range of stochastic traffic flow fluctuations.

(2) Adversarial examples in traffic flow interpolation space.

Given the above interpolation space, it is reasonable to assume a robust crash risk evaluation model $f_\theta(x)$ would perform well under random traffic flow fluctuation. However, when the interpolation rate is around 0.5 (medium of crash and non-crash samples), it is difficult to justify if the perturbed example should still maintain its original prediction. Therefore, small fluctuations around crash and non-crash samples are considered because the model's predication should remain consistent within these regions since the ground-truth labels are most likely to be the same.

Consider a small scale of perturbation ε in interpolation space, a non-crash neighborhood can be denoted as $\boldsymbol{N}: \lambda \in [0, \varepsilon]$ and a crash neighborhood $\boldsymbol{C}: \lambda \in [1-\varepsilon, 1]$, respectively. To quantitatively justify if a model is robust in these regions, the traffic flow adversarial example (TF-AE) is introduced. Formally, the TF-AE is the perturbed example that achieves maximum loss within neighborhoods of traffic flow interpolation space,

$$x_{\lambda^*} = \underset{\lambda \in \boldsymbol{N} \cup \boldsymbol{C}}{\operatorname{argmax}} \mathscr{L}[f_\theta(x_\lambda), y] \tag{4}$$

where y is the ground-truth label for crash risk evaluation and the $\mathscr{L}$ is the loss function (e.g., the cross-entropy loss).

To optimize Equation (4), projected gradient descent (Madry et al., 2017) is commonly used to generate adversarial examples. As shown in Fig. 2, λ is updated iteratively by ascending its gradient,

$$\lambda_{t+1} = \prod_{\lambda_t \in \boldsymbol{N} \cup \boldsymbol{C}} \lambda_t + \alpha^* \operatorname{sgn}\{\nabla_\lambda \mathscr{L}[f_\theta(x_\lambda), y]\} \tag{5}$$

where α is the step size, t is the iteration number, and sgn is the symbolic function to return the sign (positive or negative) of an input. Additionally, if λ_t goes outside the bounds, a clipping operation will be performed to project λ_t back to range.

(3) Evaluating the model robustness via TF-AEs.

With the above methods, many pairs of TF-AEs can be generated with respect to two predefined neighborhoods and can be combined into an evaluation dataset. If the model's prediction accuracy drops significantly on this evaluation set, the model is said to be not robust to stochastic flow fluctuations, and vice versa.

3.2. Improving model robustness via coverage-oriented adversarial training

A formerly developed CNN-based crash risk evaluation model was utilized as base-line model in robustness improvement and more details about it can be seen in next section (*4. DATA PREPARATION)*. The TF-AEs and preliminary studies have shown the fragile robustness of it in adversarial scenario. It is of great importance to resolve this issue to the need of real-world applications. One common solution is adversarial training (Goodfellow et al., 2014; Madry et al., 2017; Zhao et al., 2022), which generates adversarial examples and uses them throughout training. Formally, a min–max objective function is being optimized for training in adversarial manner,

$$\min_\theta \max_\lambda E_{(x,y)\sim D} \mathscr{L}[f_\theta(x_\lambda), y)] \tag{6}$$

The process of adversarial training is shown in Fig. 3. Specifically, for each training iteration, the TF-AEs are generated on-the-fly and used for minimizing model loss.

Despite improving robustness via adversarial training, there are two challenges from reality that remain for crash risk evaluation. On one hand, the TF-AEs only depend on one pair of crash and non-crash samples while there could be various crash risk transition patterns from low-risk to high-risk traffic operation statuses. For example, a non-crash sample of free-flow traffic transits to a low-speed crash sample; A non-crash sample of congested traffic changes to a crash sample of high-speed traffic flow; It is also possible that a medium-speed non-crash sample transits to a crash sample of disorder traffic flow. On the other hand, the number of crashes is far less than non-crash observations, introducing data bias into the model and causing overfitting on crash samples in training set and less robust to unseen crash samples.

To overcome these shortages, a coverage-oriented adversarial training technique is proposed. As shown in Fig. 4, instead of randomly choosing one pair of crash and non-crash samples, one crash sample and multiple non-crash samples are chosen randomly. Thus, the coverage of non-crash-to-crash transition patterns can be enlarged, providing a larger interpolation space of traffic flow examples. Among interpolation space, the adversarial examples outside the decision boundary (from the non-robust area) would have much larger loss than others and will be optimized gradually until all adversaries lie in model decision boundary

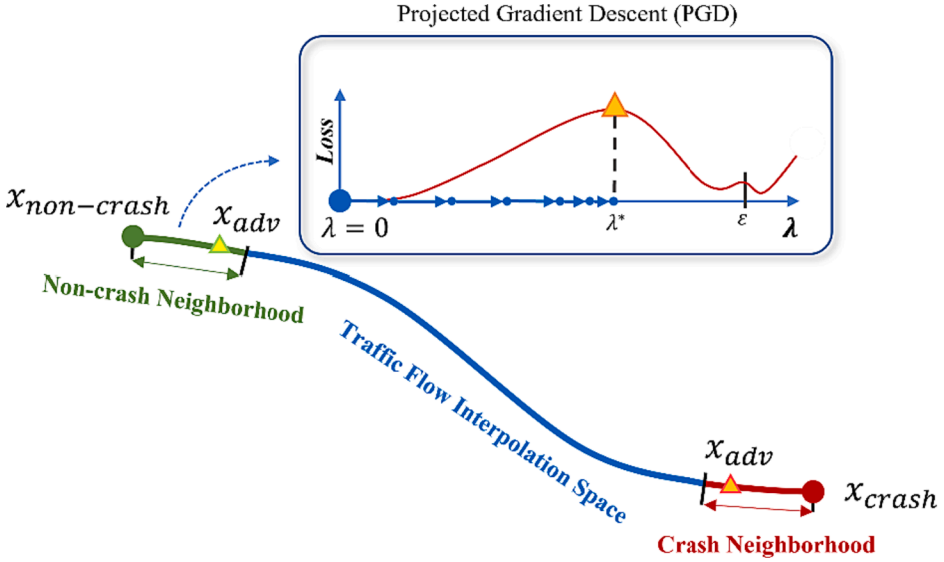


Fig. 2. Illustration of finding adversarial examples in traffic flow interpolation space.

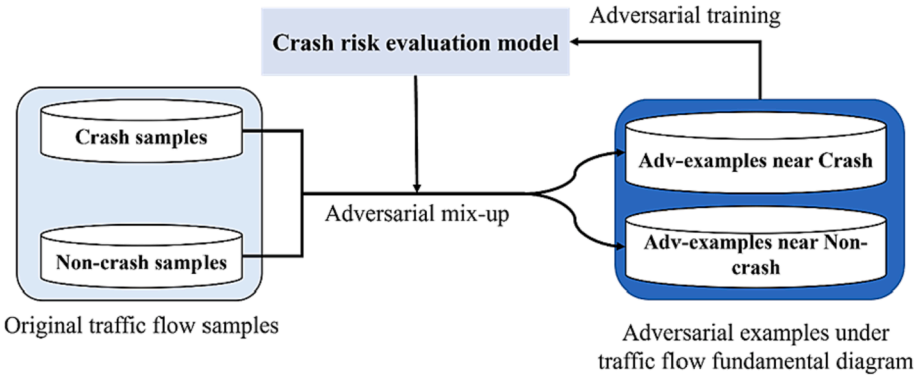


Fig. 3. Illustration of adversarial training process for model robustness improvement.

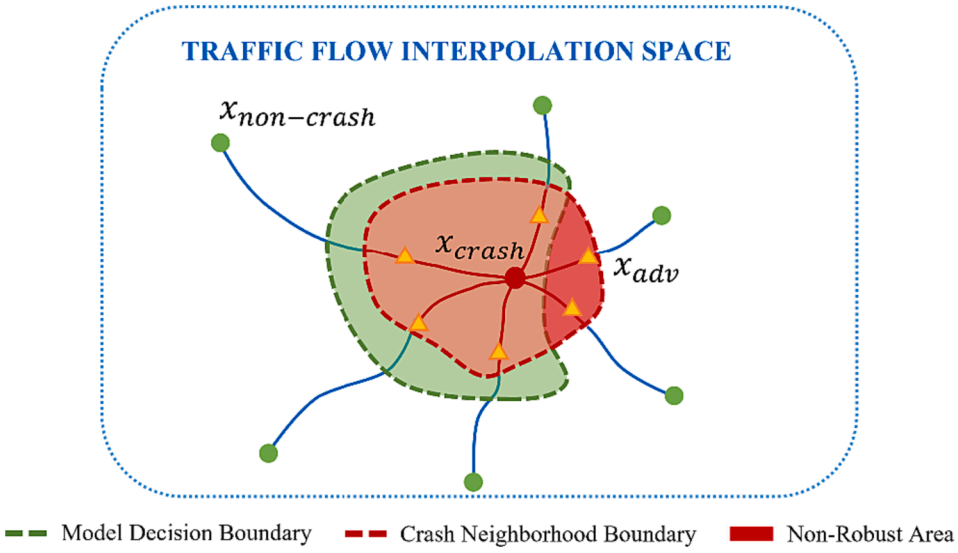


Fig. 4. Illustration of the TF-AEs generation in the coverage-oriented adversarial training.

with smaller loss. By generating TF-AEs on-the-fly and training iteratively, the robustness of the model can be improved. The detailed algorithm is shown in Algorithm 1 and Algorithm 2.

Algorithm 1: Coverage-oriented adversarial training with TF-AEs.

Input: x^i_{crash}, $x^j_{non-crash}$ are the original crash, non-crash samples, $f_\theta(X)$ is the crash risk evaluation model, $\mathscr{L}$ is the loss function, *TF-AE* (x, y) is the TF-AEs generation function.
Output: DL model parameter θ
Initialize θ_0
for each epoch $t = 0, 1, \ldots T$ **do**
  for each x^i_{crash} **do**
    $x^{i,t}_{adv-c} \leftarrow$ *TF-AE* (x^i_{crash}, $y^i = 1$) **end for**
  for each $x^j_{non-crash}$ **do**
    $x^{j,t}_{adv-nc} \leftarrow$ *TF-AE* ($x^j_{non-crash}$, $y^i = 0$) **end for**
  $X^t_{adv} \leftarrow \{x^{1,t}_{adv-c}, \ldots, x^{i,t}_{adv-c}, x^{1,t}_{adv-nc}, \ldots, x^{j,t}_{adv-nc}\}$ // the TF-AEs in epoch t
update θ_{t+1} using $\mathscr{L}[f_{\theta_t}(X^t_{adv}), Y]$
end for

Algorithm 2: TF-AEs generation function *TF-AE* (x, y).

Input: x is a crash/non-crash sample, y is the ground-truth label of x, K is the multi-pair interpolation parameter, ε is a perturbation scale, $f_\theta(X)$ is a crash risk evaluation model
Output: the TF-AE x_{adv}
if $y = 1$ **do** // x is a crash sample
  for $k = 0, 1, \ldots, K$ **do**
    $X \leftarrow (x, x_{non-crash})$// $x_{non-crash}$ is a randomly selected non-crash sample

(*continued on next column*)

(*continued*)

Algorithm 2: TF-AEs generation function *TF-AE* (x, y).

    x^k_{adv-c} is generated by **Equation (3)**, **(5)** using X, $f_\theta(X)$, y and ε
  end for
return $x_{adv-c} = x^l_{adv-c}$, where $l = \underset{1 \le l \le K}{argmax}\{L[f_{\theta_t}(x^l_{adv-c}), y^l_{adv}]\}$ // crash TF-AE
else do // x is a non-crash sample
  $X \leftarrow (x_{crash}, x)$ // x_{crash} is a randomly selected crash sample
  x_{adv-nc} is generated by **Equation (3)**, **(5)** using X, $f_\theta(X)$, y and ε
return x_{adv-nc} // non-crash TF-AE

Moreover, due to the imbalanced crash and non-crash samples, model's performance such as sensitivity can be affected by coverage-oriented adversarial training since most attention is on TF-AEs of the majority of non-crash observations. To reduce the imbalanced data training issue, focal loss (Lin et al., 2017) and sample augmentation (Man et al., 2022) are exploited to balance crash and non-crash predictions:

1) In focal loss, for p_i being the probability of crash for the i-th sample,

$$Loss = \sum_{i=1}^{N} \alpha(1-p_i)^\gamma y_i \log(p_i) + (1-\alpha)p_i^\gamma (1-y_i)\log(1-p_i) \quad (7)$$

where $\alpha \in (0, 1)$ is the class-wise weights and $\gamma \geq 0$ is the parameter of focal loss. With $\alpha > 0.5$ and suitable γ, the imbalance between crash and non-crash can be addressed as more weight would be put on crash TF-AEs during coverage-orient training.

2) In the sample augmentation method, the ratio of crash TFAEs and non-crash TFAEs is adjusted by generating more crash TFAEs (e.g., crash: non-crash = 2/3/4:4), guiding the model to learn more features of the minority crash TF-AEs.

4. Data preparation

In this study, empirical data from the Shanghai urban expressway system were utilized. Based on the data availability, historical crash data and traffic flow data from April 2014 were employed. A CNN-based crash risk evaluation model was constructed referring to our previous study (Yu et al., 2020).

(a) Traffic flow data.

The Shanghai expressway system was split into 206 roadway sections using on-ramps and off-ramps as dividing points. Then traffic flow data were obtained from dual loops detectors which cover all the urban expressway sections. The raw data at each station was originally aggregated to average speed, flow, and occupancy over a 20 s sampling period. It was further aggregated to 5 min. So, each roadway section could collect 288 data points ((24 h × 60 min)/5 min = 288) per day. Finally, 846,077 traffic flow data points were collected after data aggregation and abnormal data filtering.

(b) Crash data.

Crash data were obtained from the Shanghai Traffic Information Center. Only two-vehicle and multivehicle crashes were included in this study because single-vehicle crashes are more likely to be affected by vehicle mechanical failures or erroneous driving behaviors rather than the influence of traffic flow (Zhou and Chin, 2019). The one-month data contain 1,152 crash records; For each specific crash, crash data can be matched with traffic flow data at the section level based on section identify codes.

(c) Crash risk evaluation modeling.

Referring to our previous study (Yu et al., 2020), a 30-mintue traffic flow data prior to each crash in crash occurred section and its upstream, downstream sections were extracted. Traffic flow feature variables were calculated as shown in Fig. 5. For non-crash samples, a matched case-control data structure was adopted and the crash and non-crash ratios was set to 1:4 in keeping with the majority of studies (Abdel-Aty et al., 2004; Lord and Washington, 2018; Zhang and Abdel-Aty, 2022). After data processing, the final data set contains 1,152 crash cases and 4,608 non-crash cases. Each dataset was further split into training and testing data with the corresponding proportion of 3:1.

A 2-layer CNN model was applied to evaluate crash risk as shown in Fig. 6. The traffic flow feature variables were re-aligned into a $6 \times 3 \times 6$ tensor. Then, 2 convolutional layers with batch norm and RELU functions were utilized to extract high-level features among traffic flow data. Finally, a linear layer with sigmoid function was connected to obtain the evaluated crash risk. More details could be seen in previous work (Yu et al., 2020). In each experiment, the hyperparameters (e.g., batch size, and learning rate) in CNN model were adjusted based on the training data to achieve the best model performance.

5. Results

5.1. Evaluating crash risk evaluation model robustness via TF-AEs

First, the traffic flow-speed-density constraints in traffic flow fundamental diagram are modeled based on 1/200 of the full traffic flow dataset (4230 data points). The coefficient parameter of traffic flow-speed, occupancy model $x^F = f_{F-O,S}(x^O, x^S)$ is set to 9.14e-3 as shown in Fig. 7(a). Fig. 7(b) and Table 2 illustrate the model fitting results of four speed-occupancy models $x^S = f_{S-O}(x^O)$. Based on the mean square error and relative error, the 3PL model is best to quantify the speed-occupancy relationship and its model error is highly acceptable compared to relevant studies (Qu et al., 2015; Zhang et al., 2018).

To justify the correctness of adversarial examples, statistical comparisons are being conducted empirically. Following the same settings from existing studies (Cai et al., 2020; Islam et al., 2021), distributions for each traffic flow variables are being plotted and compared. Table 3 shows the statistical description and Fig. 8 visualizes their distributions. As the p-values of *t*-test and Kolmogrove Smirnov test (ks-test) are greater than 0.05 for all variables, the generated adversarial examples can be concluded to be statistically similar to real samples. Therefore, the adversarial examples could reflect the real neighborhood of real samples.

Within the established traffic flow interpolation space, TF-AEs under a specific perturbation scale (ε) can be found through the proposed projected gradient descent method. The maximum perturbation scale is set to 0.1 because the model's prediction should remain consistent within these small traffic fluctuation regions. The model's prediction accuracy drops on TF-AEs are used to evaluate model robustness. Fig. 9 shows the model performance metrics on TF-AEs with different ε. When ε is 0, it represents the original test samples. When $\varepsilon > 0$, it represents the TF-AEs generated using the test samples. The larger ε is, the greater the traffic flow fluctuation was imposed. As the perturbation scale of ε gradually increases from 0 to 0.1, the model accuracy decreases by 8.0 % from 0.94 to 0.86 and the sensitivity drops significantly by 18.0 % from 0.81 to 0.63. The above experimental results reflect the fragile robustness of the CNN-based crash risk evaluation model to the TF-AEs. Its model performance, especially for sensitivity, declines significantly in the face of TF-AEs.

5.2. Improving model robustness by coverage-oriented adversarial training

In the coverage-oriented adversarial training, K non-crash samples are randomly selected for each crash sample to enlarge the coverage of non-crash-to-crash transition patterns. K increases at a rate of 2 times to gradually increase the coverage of the traffic flow interpolation space. Table 4 compares the CNN-based crash risk evaluation model accuracy on TF-AEs with different K. The drops of model accuracy on TF-AEs increases as K value becomes larger. It indicates that as the coverage of traffic flow interpolation space becomes larger, more adversarial examples outside the decision boundary of this model can be found. However, if the value of K is too large, the computational efficiency of adversarial training will be seriously reduced. According to Fig. 10, the marginal effects of model accuracy drop become smaller (<0.002) when K exceeds 8. Therefore, in the subsequent experiments, K is set as 8 to balance the coverage of transition patterns and training efficiency.

Finally, coverage-oriented adversarial training is conducted to

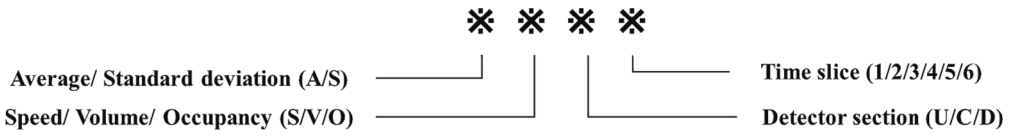


Fig. 5. Construction of traffic flow feature variables (Yu et al., 2020).

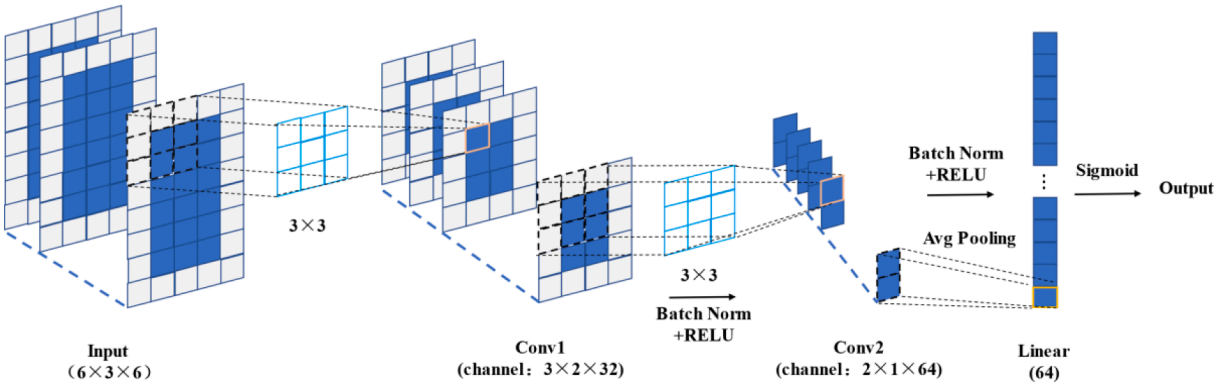


Fig. 6. Structure of CNN-based crash risk evaluation model (Yu et al., 2020).

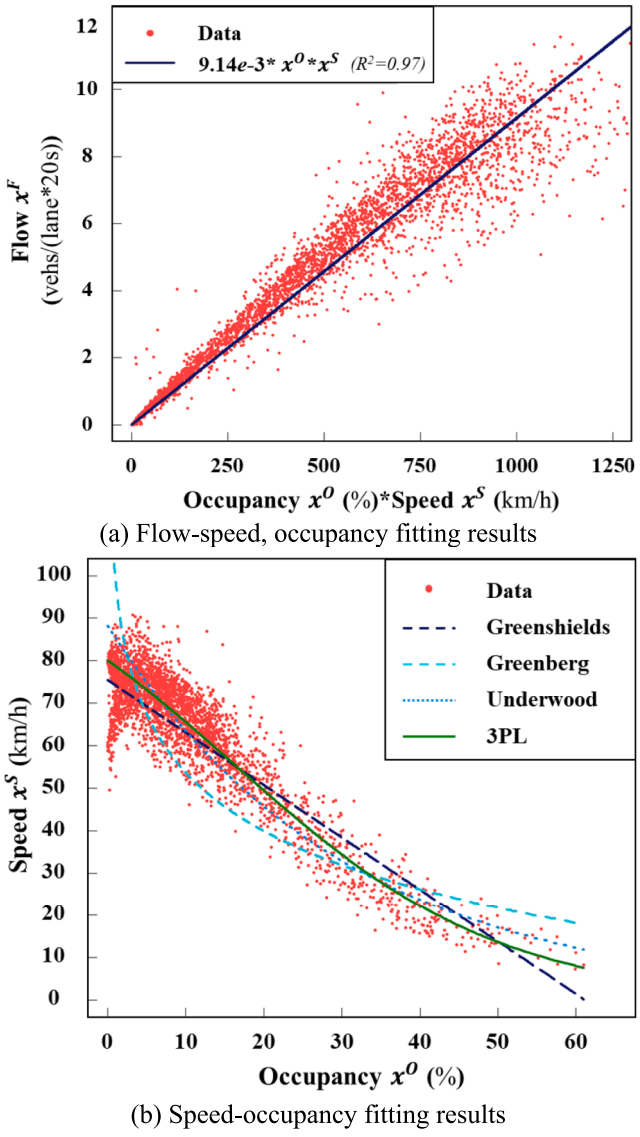


Fig. 7. Illustration of the traffic flow-speed-density constraints under Shanghai empirical data.

improve the crash risk evaluation model robustness. For the conventional training (Baseline), the CNN model was trained on the original real samples. For the adversarial training, the real samples were replaced by the generated TF-AEs under the specific perturbation scale

Table 2
Coefficient parameter results and validation indexes of four speed-occupancy models.

Models	Speed-occupancy function	R^2	Mean square error	Relative error
Greenshields	$x^S = 75.42*(1-\frac{x^O}{61.12})$	0.82	45.67	9.77 %
Greenberg	$x^S = 19.86*ln(\frac{149.35}{x^O})$	0.57	382.58	21.72 %
Underwood	$x^S = 88.24*exp(-\frac{x^O}{30.40})$	0.73	65.80	10.43 %
3PL model	$x^S = \frac{110.88}{1+exp(\frac{x^O-16.33}{17.12})}$	**0.84**	**38.43**	**8.12 %**

Notation: The significance of the calibration coefficients in the above models are less than 0.001.

(ε = 0.1). In each training iteration, new TF-AEs were generated based on the current model gradients and then used for model parameter update. Though the iteratively training on TF-AEs, the robustness of the crash risk evaluation model could be gradually improved. Moreover, focal loss and sample augmentation were utilized to address the imbalance features between crash and non-crash samples. All the models were finally evaluated on the TF-AEs generated using the test datasets. Table 5 illustrates the model performance metrics of the CNN-based crash risk evaluation model on TF-AEs using different training methods. There are three conclusions that can be drawn:

1) Compared to the baseline model, three adversarial training methods can improve the robustness of the CNN model, as reflected by the improvement in the model accuracy (from 0.857 to 0.906–0.919) and sensitivity (from 0.629 to 0.651–0.807). The result shows that putting the TF-AEs into the training process can help model learn more traffic flow fluctuating features and optimize their decision boundary, thus making the model no longer misclassify most TF-AEs.
2) Adversarial training with sample augmentation method works best among the three adversarial training methods. The 98.9 % of sensitivity drop and 76.5 % of accuracy drop can be avoided by adjusting the ratio of crash TF-AEs to non-crash TF-AEs to 2:4. It indicates that more crash TF-AEs during adversarial training can help to avoid the learning bias caused by non-equilibrium between crash to non-crash samples.
3) Meanwhile, the results show that the adversarial training with focal loss can also significantly avoid the accuracy by 71.6 % (from −0.180 to −0.054) and sensitivity drop by 70.0 % (from

Table 3
Statistical features and two test p-values of the TF-AEs and real samples.

Variables	Artificial samples (TF-AEs)				Real samples				Statistical Tests	
	Min	Max	Mean	STD	Min	Max	Mean	STD	*t*-test(p-value)	ks-test(p-value)
AOC2	1.24	72.2	29.3	12.9	2.04	78.7	30.9	13.6	0.81	0.96
SOC2	0	51.7	13.7	7.09	1.56	47.7	14.1	7.33	0.77	0.97
ASC2	9.55	84.4	40.6	16.4	7.59	80.8	38.1	17.3	0.79	0.72
SSC2	0	34.7	13.8	6.24	0	37.1	14.1	6.6	0.90	0.97
AFC2	0	12.1	6.94	1.69	0	12.9	7.18	1.77	0.33	0.12
SFC2	0	7.91	2.41	0.78	0	8.5	2.44	0.84	0.46	0.55

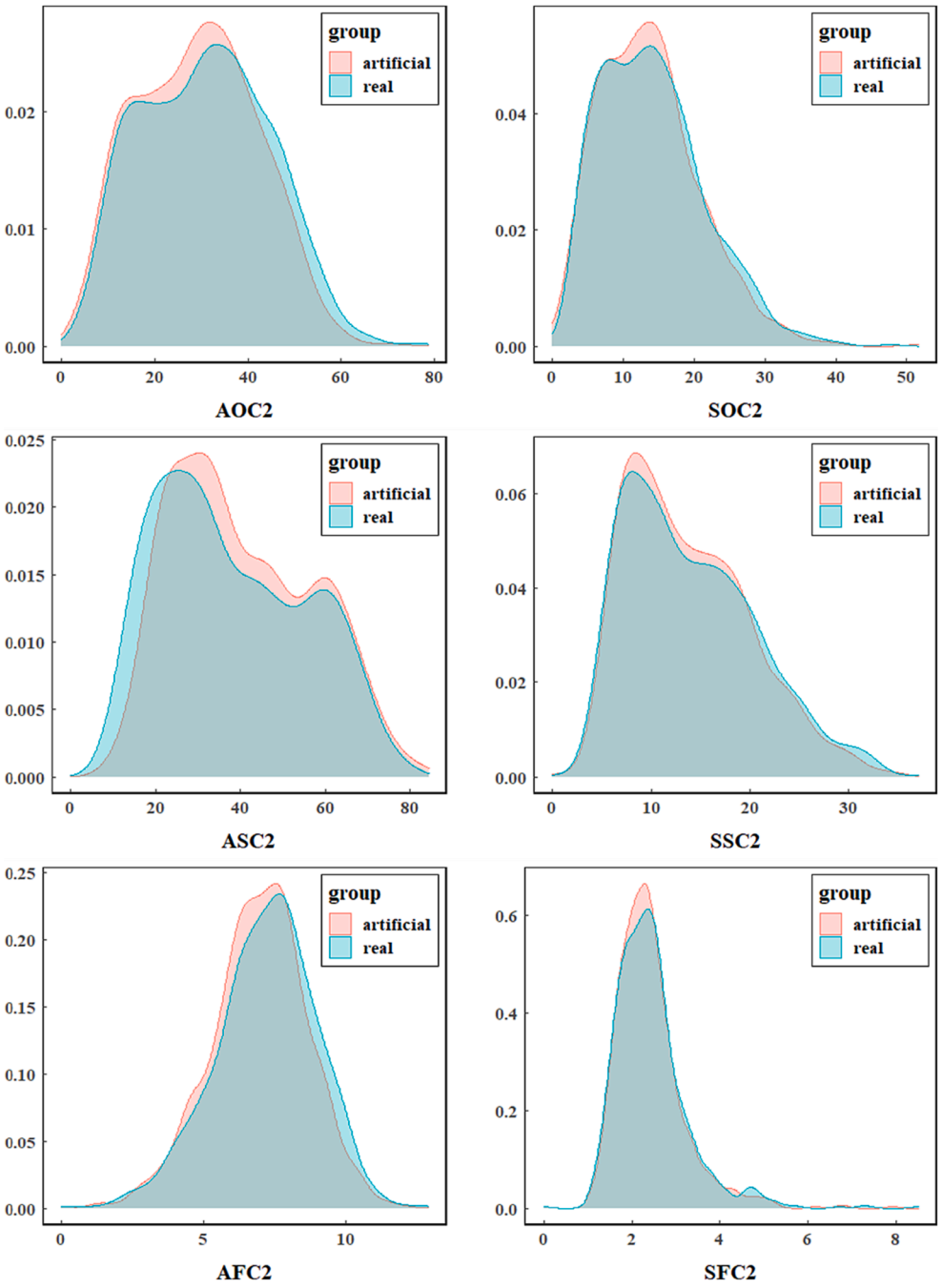


Fig. 8. Illustration of traffic flow variable distributions of the TF-AEs and real samples.

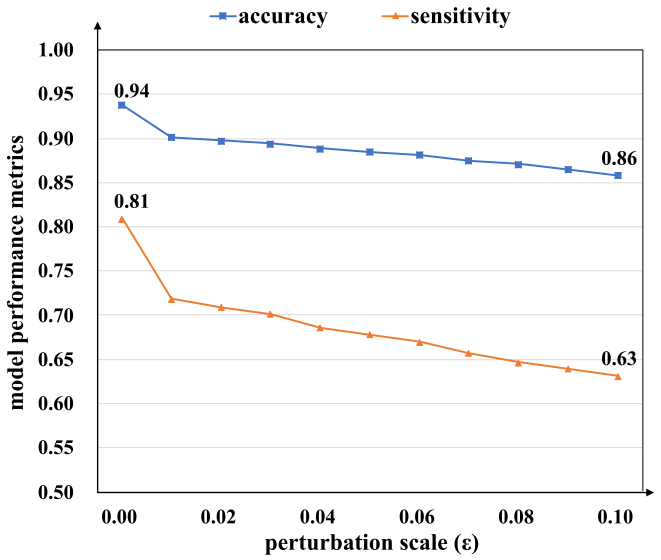


Fig. 9. Model performance metrics on the TF-AEs under different perturbation scales of ε.

Table 4
Model accuracy on TF-AEs with K-pair crash-noncrash samples.

	K	Accuracy	Accuracy drops
original samples	–	0.947	0
TF-AEs	1	0.893	−0.054
(ε = 0.1)	2	0.880	−0.067
	4	0.871	−0.076
	8	0.857	−0.090
	16	0.846	−0.101
	32	0.832	−0.115
	64	0.820	−0.127

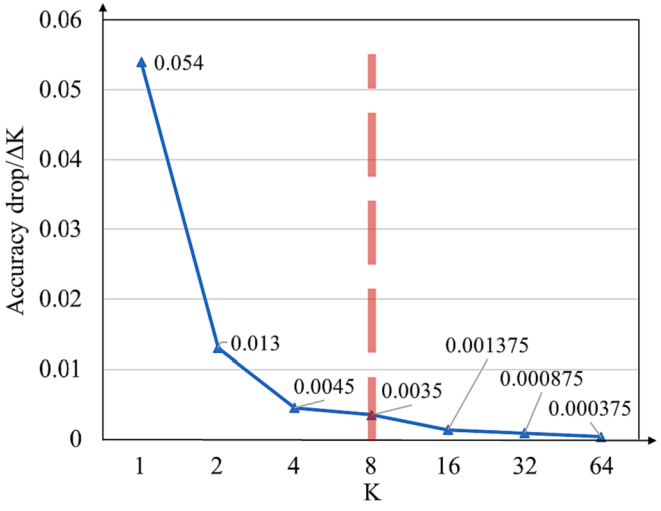


Fig. 10. Marginal effects analysis of TF-AEs with K-pair crash-noncrash samples.

Table 5
Model robustness comparison on TF-AEs using different training methods.

Models	Accuracy		Sensitivity	
	Value (drop)	Improvement for its drop	Value (drop)	Improvement for its drop
Conventional training (Baseline)	0.857 (−0.081)	–	0.629 (−0.180)	–
Coverage-oriented adversarial training	0.906 (−0.032)	+60.5 %	0.651 (−0.158)	+12.2 %
Coverage-oriented adversarial training with focal loss ($\alpha = 0.6, \gamma^1 = \gamma^2 = 5$)	0.915 (−0.023)	+71.6 %	0.755 (−0.054)	+70.0 %
Coverage-oriented adversarial training with sample augmentation (Crash: Noncrash = 2:4)	**0.919 (−0.019)**	**+76.5 %**	**0.807 (−0.002)**	**+98.9 %**

Notation: The focal loss parameters and crash-noncrash ratio are determined by Grid Search.

−0.081 to −0.023) by giving higher weights to the crash TF-AEs during model training.

6. Discussion

6.1. Model robustness with more, less and rare examples

To explore the performance of the proposed method in real world scenarios, we explore the following settings:

1) Firstly, more real-world data were included to compile large real datasets from 5,760 (×1) to 11,520 (×2), 17,280 (×3) and 23,040 (×4). As shown in Table 6, although model accuracies were very similar, the model sensitivities of adversarial training model had been improved significantly. This experiment reflects that by learning pairs of crash-non-crash examples, the proposed adversarial training can help model to capture the crash risk features under severe data imbalance, and thus become more robust in predicting unseen examples, especially the high-risk ones.
2) Secondly, experiments were conducted on smaller training data size (from 80 % to 5 %). As shown in Table 7, the proposed model outperforms baseline models in sensitivities. When there are very few training data (e.g., ×10 % and × 5 % contains less than 100 crashes), traditional method could easily collapse while the proposed adversarial training method remains consistent. This is because the proposed method places strong regularization on datapoints' neighborhood regions while the baseline model will be easily overfitting to training datapoints.
3) Finally, experiments were conducted to validate the model performance on rare examples. The top 10 % examples with largest loss were selected as "rare examples" using baseline model. As show in

Table 6
Model performance comparison on large real traffic flow samples.

Data size	Accuracy		Sensitivity	
	Ours	Baseline	Ours	Baseline
×1	0.922	**0.924**	**0.872**	0.797
×2	**0.895**	0.891	**0.666**	0.577
×3	**0.887**	0.881	**0.566**	0.498
×4	**0.880**	0.858	**0.507**	0.351

Table 7
Model performance comparison with limited training data.

Data size	Accuracy		Sensitivity	
	Ours	Baseline	Ours	Baseline
×80 %	0.920	**0.926**	**0.879**	0.784
×40 %	0.904	**0.906**	**0.824**	0.738
×20 %	0.892	**0.893**	**0.785**	0.714
×10 %	**0.889**	0.889 (Abdel-Aty et al., 2010)	**0.731**	0.689 (Abdel-Aty et al., 2010)
×5%	**0.852**	0.850 (Abdel-Aty et al., 2004)	**0.698**	0.667 (Abdel-Aty et al., 2004)

Notation: (Abdel-Aty et al., 2010) and (Abdel-Aty et al., 2004) are the early-stopping results before training procedure crashed.

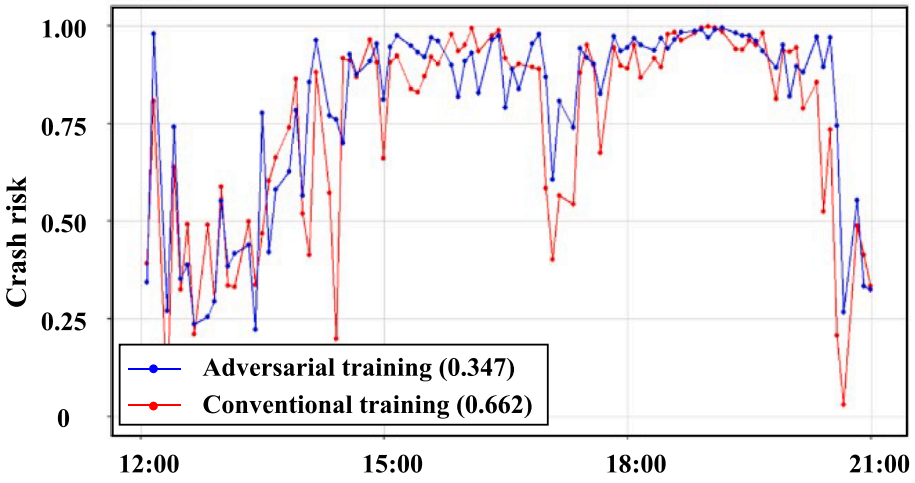
Table 8
Model performance comparison on rare examples.

Data size	Accuracy		Sensitivity	
	Ours	Baseline	Ours	Baseline
×80 %	**0.409**	0.323	**0.600**	0.350
×40 %	**0.236**	0.228	**0.267**	0.100
×20 %	0.244	**0.285**	**0.283**	0.167
×10 %	0.315	**0.323**	**0.367**	0.217
×5%	0.291	**0.354**	**0.267**	0.200

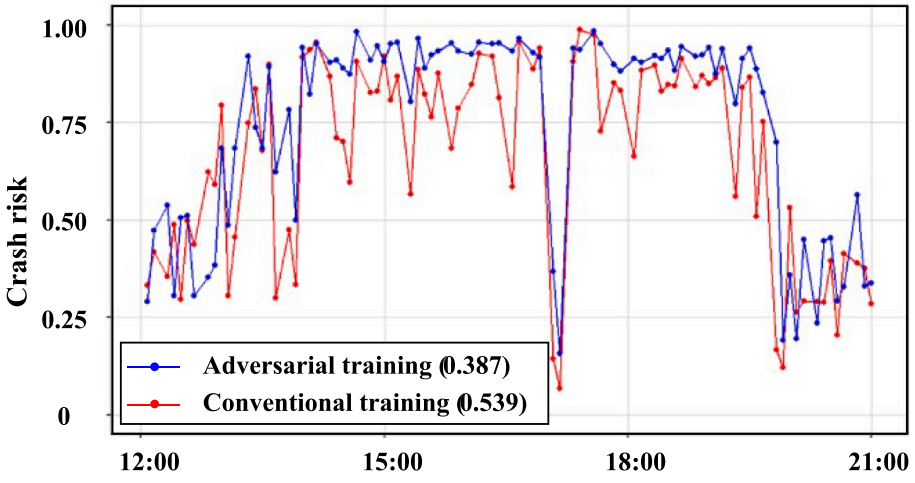
Table 8, the proposed method achieved good performance gain to "rare examples" on both accuracy and sensitivity by training with moderate amounts of training data (×80 %). The proposed method outperforms baseline model in sensitivity for all the experiments. It indicates that the proposed adversarial training method could improve robustness in a larger space, especially rare data points.

6.2. Model stability in real-world traffic dynamics

To better observe the model performance under real-world traffic dynamics, the stability of model results is compared. The time-varying volatility (Figlewski, 1994) is chosen to quantify the stability of the model results, which can be calculated by Equation (8). Fig. 11 shows the model results of conventional and adversarial training models in two



(a) Road ID=21103 in 2004-04-04



(b) Road ID=21112 in 2004-04-12

Fig. 11. Comparation of model performance in Shanghai empirical data.

Table 9
The time-varying volatility measure of the two models in the Shanghai expressways.

Expressway	Yan'an	North-South	Inner Ring	Middle Ring	Yixian	Humin	**Average**
Conventional training	2.01	2.02	1.72	1.91	**1.56**	1.63	**1.81**
Adversarial training	1.76	1.79	**1.43**	1.64	1.49	1.49	**1.61**
Degree of decline	−12.4 %	−11.4 %	**−16.9 %**	−14.1 %	−4.5 %	−8.6 %	**−11.5 %**

typical road sections. The crash risk output of the conventional model is unstable and lots of "jump points" can be seen in adjacent short-time slices. By contrast, the adversarial training model has stable model outputs and its crash risk results fluctuate little in most of the time. Table 9 shows the time-varying volatility measure of the two models in the six Shanghai expressways. In each expressway, adversarial training model has a less time-varying volatility value. The average time-varying volatility measure reduction is 11.4 %. Overall, the results show that compared with conventional training method, the adversarial training model has better robustness to traffic flow fluctuation, thus maintaining a stable crash risk evaluation result.

$$TV = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2},\ r_i = \ln\left(\frac{x_i}{x_{i-1}}\right) \tag{8}$$

where TV is the time-varying volatility measure of the time series data, x_i and x_{i-1} are the observations at time i and i−1, and ln is the natural logarithm.

7. Conclusion

Crash risk evaluation models need to identify the high-risk traffic operation statuses based on real-time traffic flow data. Given the complicated and stochastic traffic flow fluctuations in empirical applications, it is of importance for the crash risk evaluation models to keep good model accuracy, which is considered as model robustness. In recent years, DL models have been applied in most crash risk evaluation studies because of their excellent evaluation performances. However, DL models are shown to be sensitive to input perturbations and have poor model robustness. It is still questionable if the DL-based crash risk evaluation models would be affected by various traffic flow oscillations in real-world traffic dynamics to lose accuracy.

In this study, the model robustness issue in crash risk evaluation is introduced and a robustness improvement method for DL-based crash risk evaluation model is proposed. Specifically, TF-AEs are first generated via data mix-up under classical traffic flow fundamental diagram constraints to evaluate the model robustness. Then, coverage-oriented adversarial training is conducted to improve model robustness. By generating TF-AEs on-the-fly and training iteratively, the robustness of crash risk evaluation model can be effectively improved. In addition, focal loss and sample augmentation are integrated to solve the imbalanced data issue. The proposed method was validated based on the empirical data of Shanghai expressways. A CNN-based crash risk evaluation model (Yu et al., 2020) was established for robustness evaluation and improvement. The results showed that with the developed TF-AEs under a small-scale traffic fluctuation (0.1), the CNN-based crash risk evaluation model suffered 8 % decrease in accuracy and 18 % drops in sensitivity. After the proposed adversarial training, it could achieve avoiding 76.5 % loss in model accuracy and 98.9 % loss in model sensitivity to TF-AEs.

Based on the experimental results, the main findings of the study can be summarized as follows:

1) Although the DL model can improve the crash risk evaluation accuracies, they are not robust to real-world traffic flow fluctuations. Experiments have demonstrated that small scale traffic flow oscillations could make the CNN-based crash risk evaluation model accuracy and sensitivity significantly degraded.
2) The proposed coverage-oriented adversarial training can effectively improve the robustness of the DL-based crash risk evaluation models, as reflected by the significant improvement in model accuracy to TF-AEs. However, the improvement in model sensitivity is still unsatisfactory. While it is encouraging that both sample enhancement and focal loss can be well integrated into the proposed adversarial training processing to significantly improve model sensitivity.
3) Compared with conventional training methods, the crash risk evaluation model with adversarial training can provide more stable output to real-world traffic dynamics.

With the proposed model robustness analysis framework, this study can help the traffic managers to actively identify the high-risk traffic flow status and trigger safety countermeasures such as the variable speed limit. Early warnings can be issued in current traffic management system or even for the future cooperative vehicle infrastructure system. To avoid high-risk traffic flow scenarios, improvement suggestions can be implemented by analyzing adversarial perturbation generated on specific dimension of variables. However, there are still a few limitations of the current study. First, traffic flow variables were mainly considered following previous studies (Roshandel et al., 2015; Mannering et al., 2020; Cheng et al., 2022; Roy et al., 2022). However, other factors such as weather, lighting, traffic composition that might have significant impact on crash prediction are not included in current case study. In addition, to generate TF-AEs, a fixed-parameter traffic flow fundamental diagram model was utilized to construct traffic flow interpolation space. Traffic flow fundamental diagram models with mixed or random-distributions parameters (Qu et al., 2017) could be explored to better fit the stochastic feature in realistic traffic flow changing patterns.

CRediT authorship contribution statement

Rongjie Yu: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Lei Han:** Investigation, Methodology, Writing – original draft, Writing – review & editing, Visualization. **Mohamed Abdel-Aty:** Writing – review & editing, Visualization. **Liqiang Wang:** Conceptualization, Methodology, Supervision. **Zihang Zou:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This study was supported by the Chinese National Natural Science Foundation (NSFC 52172349), the Belt and Road Cooperation Program under the 2023 Shanghai Action Plan for Science, Technology and Innovation (No. 23210750500), and the Science and Technology Plan Project of Zhejiang Provincial Department of Transport.

References

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M.F., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transp. Res. Rec. 1897 (1), 88–95.

Abdel-Aty, M., Pande, A., Hsia, L., 2010. The concept of proactive traffic management for enhancing freeway safety and operation. ITE J. 80 (4), 34.

Bai, T., Luo, J., Zhao, J., Wen, B., & Wang, Q. 2021. Recent advances in adversarial training for adversarial robustness. arXiv preprint arXiv:2102.01356.

Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J., Wu, Y., 2020. Real-time crash prediction on expressways using deep generative models. Transp. Res. Part C: Emerg. Technol. 117, 102697.

Carlini, N., Wagner, D., 2017. May. Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (sp), pp. 39–57.

Cheng, Z., Yuan, J., Yu, B., Lu, J., Zhao, Y., 2022. Crash risks evaluation of urban expressways: A case study in Shanghai. IEEE transactions on intelligent transportation systems 23 (9), 15329–15339.

Figlewski, S. 1994. Forecasting volatility using historical data.

Goodfellow, I. J., Shlens, J., & Szegedy, C. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Greenberg, H., 1959. An analysis of traffic flow. Oper. Res. 7 (1), 79–85.

Greenshields, B.D., Bibbins, J.R., Channing, W.S., Miller, H.H., 1935. A Study of Traffic Capacity. in Highway Research Board Proceedings Vol. 14(1, 448–477.

Hossain, M., Abdel-Aty, M., Quddus, M.A., Muromachi, Y., Sadeek, S.N., 2019. Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. Accid. Anal. Prev. 124, 66–84.

Hu, W., Tan, Y., 2023. Generating adversarial malware examples for black-box attacks based on GAN. In: Data Mining and Big Data: 7th International Conference, DMBD 2022, Beijing, China, November 21–24, 2022, Proceedings, Part II. Singapore, Springer Nature Singapore, pp. 409–423.

Huang, P., Xu, M., Fang, F., & Zhao, D. 2022. Robust reinforcement learning as a Stackelberg game via adaptively-regularized adversarial training. arXiv preprint arXiv:2202.09514.

Huber, P.J., 2011. Robust statistics. In *International encyclopedia of statistical science*. In: Lovric, M. (Ed.), International Encyclopedia of Statistical Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1248–1251.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. 2019. Adversarial examples are not bugs, they are features. Adv. Neural Inform. Process. Syst., 32.

Islam, Z., Abdel-Aty, M., Cai, Q., Yuan, J., 2021. Crash data augmentation using variational autoencoder. Accid. Anal. Prev. 151, 105950.

Jia, R., & Liang, P. 2017. Adversarial examples for evaluating reading comprehension systems. arXiv preprint arXiv:1707.07328.

Kannan, H., Kurakin, A., & Goodfellow, I. 2018. Adversarial logit pairing. arXiv preprint arXiv:1803.06373.

Li, P., Abdel-Aty, M., 2022. Real-Time Crash Likelihood Prediction Using Temporal Attention-Based Deep Learning and Trajectory Fusion. J. Transp. Eng., Part A: Syst. 148 (7), 04022043.

Li, L., & Spratling, M. 2023. Data augmentation alone can improve adversarial training. arXiv preprint arXiv:2301.09879.

Li, P., Abdel-Aty, M., Yuan, J., 2020. Real-time crash risk prediction on arterials based on LSTM-CNN. Accid. Anal. Prev. 135, 105371.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988.

Liu, F., Zhang, C., Zhang, H., 2023. In: Towards Transferable Unrestricted Adversarial Examples with Minimum Changes. IEEE, pp. 327–338.

Lord, D., Washington, S., 2018. Safe mobility: Challenges, methodology and solutions. Emerald Group Publishing.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Man, C.K., Quddus, M., Theofilatos, A., Yu, R., Imprialou, M., 2022. Wasserstein Generative Adversarial Network to Address the Imbalanced Data Problem in Real-Time Crash Risk Prediction. IEEE Trans. Intell. Transp. Syst. 23 (12), 23002–23013.

Mannering, F., Bhat, C.R., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Analytic methods in accident research 25, 100113.

Omar, M., Choi, S., Nyang, D., & Mohaisen, D. 2022. Robust natural language processing: Recent advances, challenges, and future directions. arXiv preprint arXiv: 2201.00768.

Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A., 2016. In: Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. IEEE, pp. 582–597.

Qu, X., Wang, S., Zhang, J., 2015. On the fundamental diagram for freeway traffic: A novel calibration approach for single-regime models. Transp. Res. B Methodol. 73, 91–102.

Qu, X., Zhang, J., Wang, S., 2017. On the stochastic fundamental diagram for freeway traffic: model development, analytical properties, validation, and extensive applications. Transp. Res. B Methodol. 104, 256–271.

Roshandel, S., Zheng, Z., Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. Accid. Anal. Prev. 79, 198–211.

Roy, A., Hossain, M., Muromachi, Y, 2022. A deep reinforcement learning-based intelligent intervention framework for real-time proactive road safety management. Accident Analysis & Prevention 165, 106512.

Silva, S. H., & Najafirad, P. 2020. Opportunities and challenges in deep learning adversarial robustness: A survey. arXiv preprint arXiv:2007.00753.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. 2013. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. 2018. Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152.

Underwood, R.T., 1961. Speed, Volume, and Density Relationship. Quality and Theory of Traffic Flow, Yale Bur.

Wang, H., Li, J., Chen, Q.Y., Ni, D., 2011. Logistic modeling of the equilibrium speed–density relationship. Transp. Res. A Policy Pract. 45 (6), 554–566.

Wang, H., Ni, D., Chen, Q.Y., Li, J., 2013. Stochastic modeling of the equilibrium speed–density relationship. Journal of advanced transportation 47 (1), 126–150.

Wang, W., Wang, L., Tang, B., Wang, R., & Ye, A. 2019a. Towards a robust deep neural network in text domain a survey. arXiv preprint arXiv:1902.07285.

Wang, X., He, K., Song, C., Wang, L., & Hopcroft, J. E. 2019b. At-gan: An adversarial generator model for non-constrained adversarial examples. arXiv preprint arXiv: 1904.07793.

Wang, D., Wang, R., Dong, L., Yan, D., Zhang, X., Gong, Y., 2020. Adversarial examples attack and countermeasure for speech recognition system: A survey. In: International Conference on Security and Privacy in Digital Economy. Springer, Singapore, pp. 443–468.

Wang, D., Li, Y., Wang, L., & Gong, B. 2020a. Neural networks are more productive teachers than human raters: Active mixup for data-efficient knowledge distillation from a black-box model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1498-1507).

Wang, D., Zhang, S., & Wang, L. 2021. Deep epidemiological modeling by black-box knowledge distillation: An accurate deep learning model for covid-19. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 17, pp. 15424-15430).

Wu, B., Liu, L., Zhu, Z., Liu, Q., He, Z., & Lyu, S. 2023. Adversarial machine learning: A systematic survey of backdoor attack, weight attack and adversarial example. arXiv preprint arXiv:2302.09457.

Xiao, C., Li, B., Zhu, J. Y., He, W., Liu, M., & Song, D. 2018. Generating adversarial examples with adversarial networks. arXiv preprint arXiv:1801.02610.

Xu, C., Tarko, A.P., Wang, W., Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accid. Anal. Prev. 57, 30–39.

Yu, R., Abdel-Aty, M., Ahmed, M., 2013. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. Accid. Anal. Prev. 50, 371–376.

Yu, R., Wang, Y., Zou, Z., Wang, L., 2020. Convolutional neural networks with refined loss functions for the real-time crash risk analysis. Transp. Res. Part C: Emerg. Technol. 119, 102740.

Yuan, J., Abdel-Aty, M., Gong, Y., Cai, Q., 2019. Real-time crash risk prediction using long short-term memory recurrent neural network. Transp. Res. Rec. 2673 (4), 314–326.

Zhang, S., Abdel-Aty, M., 2022. Real-time crash potential prediction on freeways using connected vehicle data. Anal. Methods Accid. Res. 36, 100239.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

Zhang, J., Qu, X., Wang, S., 2018. Reproducible generation of experimental data sample for calibrating traffic flow fundamental diagram. Transp. Res. A Policy Pract. 111, 41–52.

Zhao, W., Alwidian, S., Mahmoud, Q.H., 2022. Adversarial Training Methods for Deep Learning: A Systematic Review. Algorithms 15 (8), 283.

Zhou, M., Chin, H.C., 2019. Factors affecting the injury severity of out-of-control single-vehicle crashes in Singapore. Accid. Anal. Prev. 124, 104–112.