A Diffusion model-based intelligent optimization method of rural road environments

International Journal of Transportation Science and Technology 21 (2026) 208–227

Contents lists available at ScienceDirect

International Journal of Transportation Science and Technology

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

A diffusion model-based intelligent optimization method of rural road environments

Bo Yu [a], Zehong Zhu [a], Yuren Chen [a], Junhua Wang [a], Kun Gao [b], Xin Qian [a,*]

[a] Key Laboratory of Road and Traffic Engineering of the Ministry of Education, College of Transportation Engineering, Tongji University, Shanghai 201804, China
[b] Department of Architecture and Civil Engineering, Chalmers University of Technology, Gothenburg SE-412 96, Sweden

ARTICLE INFO

Article history:
Received 3 December 2024
Received in revised form 22 January 2025
Accepted 29 January 2025
Available online 3 February 2025

Keywords:
Diffusion model
Intelligent optimization
Image generation technology
Rural road environments
Explainable machine learning

ABSTRACT

Well-designed rural road environments can guide drivers to adopt reasonable driving behaviors, thereby significantly improving the driving experience and ensuring road safety. Existing methods for optimizing rural road environments mainly rely on expert knowledge, have low automation degrees, and are limited in efficiency and accuracy. Therefore, this study aims to propose an intelligent optimization method for rural road environments by using image generation technology. Using environment images from a naturalistic driving dataset, the area and location information of semantic components (e.g., lane markings, vegetation, guardrails, and traffic signs) in rural road environments are extracted, and their impacts on driving speed are analyzed based on explainable machine learning (extreme gradient boosting (XGBoost) and Shapley additive explanations (SHAP)). These impacts are then utilized to determine how to adjust and optimize the road environment components at appropriate locations (i.e., obtain the optimization scheme). Then, a novel image generation technique, Diffusion model, is employed to establish an intelligent optimization method, which can directly generate optimized images of rural road environments. Compared with traditional manual mapping and other popular image generation algorithms such as CycleGAN, the method proposed in this study has the advantages of high efficiency, labor saving, and superior image generation quality. This study can facilitate the design and optimization of rural road environments and enhance rural road safety in a more intelligent way.

1 Introduction

Even with the low traffic volume, accidents still occur frequently and are often serious on rural roads (Drosu et al., 2020; Li et al., 2021). In China, the urban per capita vehicle ownership and miles driven were much higher than those in rural areas, but the mortality rate of accidents on rural roads was about two to three times that on urban roads in 2023 (Zhang et al., 2024). In the European Union, 52% of road traffic fatalities occurred in rural areas, versus 39% in urban areas in 2022 (EU Road Safety Statistics, 2023). It is widely reported that speeding is the leading cause of rural road accidents. About 70% of injury accidents and 60% of fatal accidents on rural roads were related to speeding in New Zealand (Job and Brodie,

Peer review under the responsibility of Tongji University.
* Corresponding author.
E-mail address: xqian2@tongji.edu.cn (X. Qian).

https://doi.org/10.1016/j.ijtst.2025.01.014

2022). In the United States, of the 17 283 rural traffic fatalities, 4 805 (28%) were killed in speeding-related accidents in 2022 (NHTSA, 2024). Therefore, guiding drivers to drive within the appropriate speed range and avoid speeding is of great importance to ensure traffic safety on rural roads.

Previous studies have corroborated the role of well-designed rural road environments in speed guidance (Li et al., 2021; Martinelli et al., 2022). For example, road signs played a significant role in indicating the change of road alignment and improving driving safety, and the average speed reduction after fixating the signs was found to be −21.89 km/h (Vignali et al., 2019). The existence of guardrails could reduce driving speed and guide drivers to travel closer to the centerline of the road, and the average speed increased from 47.73 km/h with guardrails to 53.47 km/h without guardrails (Gilandeh et al., 2018). Tree coverage was negatively associated with crashes and driving speed. When the tree coverage decreased from 10% to 0%, the total crashes were projected to increase by 24.5% (from 50.1 per year to 62.3) (Wesley et al., 2018). Rural road environments mainly include road alignment, road facilities, and surrounding landscape. Due to the limitations of arable land and topography, there are numerous intersections and slopes in rural roads, leaving limited room for road alignment design, coupled with limited construction funds (Fan and Chan-Kang, 2008; Coakley et al., 2016; Khuzan and Al-Jumaili, 2023). Therefore, considering the challenge of changing the alignment design of rural roads, optimizing road facilities and surrounding landscape (i.e., traffic signs, guardrails, and vegetation) is a more feasible method to ensure rural road safety.

Current road environment optimization research has a similar process that designers propose and implement various measures to reduce driving speed and collect crashes or speed data from optimized road sections to verify the effects (Ren et al., 2024). The traditional road environment optimization methods face numerous challenges, including over-reliance on expert experience, limited automation, and difficulty in timely verification of optimization effects (Naveen et al., 2017). Through adding road markings to the modeling of the driving simulator, Babić and Brijs (2021) analyzed how road markings affected driver behavior (driving speed, lateral movement, and acceleration) in dangerous curves. After optimizing the road markings on the driving simulator, the speed of the vehicle was significantly reduced, and the lateral movement was closer to the edge line. However, this method required substantial human input. The three-dimensional (3D) geographic information system (GIS) technique was also applied in the modeling and optimization design of highway environments (Huang et al., 2020). By visualizing the highway design, the study of Huang et al. (2020) showed that drivers could experience better visual effects and traffic safety could be improved if traffic facilities were set within 200 m of the proposed line. A notable drawback of the 3D GIS technique is the heavy reliance on manual collection of contour data and other information.

With the rapid advancement of technology in computers and artificial intelligence, many computer vision algorithms and technologies are applied to the optimization of road environments to achieve intelligent and automated processes. The infrastructure building information model (I-BIM) was used to generate a 3D parametric model of a section of the highway in northern Italy, and visualized the road infrastructure, achieving the optimization and verification of road projects according to specifications before construction (Vignali et al., 2021). A template-based 3D road modeling method was proposed to design and generate road facilities using digital elevation model data and engineering design rules, which efficiently simulated large-scale road scenes, and supported interactive editing for detailed optimization (Zhang et al., 2019). Digital twin technology could establish the existing road model by using multi-source data, and assist in road widening optimization decisions in many aspects, such as digital image processing, alignment fitting, and cross-section assembly creation (Jiang et al., 2022). However, road environments contain multiple components, including traffic signs, guardrails, and vegetation. Existing research has not proposed accurate optimization methods suitable for various types of components, resulting in a limited applicability of optimization methods.

Emerging intelligent image generation algorithms are a potential solution to this challenge, such as normalization flows (NFs), variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion model. NFs generated point clouds by modeling them as a distribution and transforming a simple base distribution into a complex target distribution through a sequence of mappings, but the computation requirement can lead to high costs and slow generation speed (Yang et al., 2019). VAEs were used to generate drivers' perspective road environment images based on four parameters of visual curvature, slope, visibility, and curve direction, promoting road design and optimization by considering the drivers' perception (Wang et al., 2020). However, VAEs are difficult to generate high-quality and high-resolution images compared to GANs and diffusion model (Cao et al., 2024). GANs could translate road environment images from night to day, which increased background brightness and improved vehicle detection accuracy at night (Shao et al., 2020). Ren et al. (2024) employed CycleGAN to generate optimized visual images of the facility environment and verify the optimization effects, and the whole process took about an hour. Although GANs can generate optimized images automatically which saves labor, it still has the disadvantages of long time consuming and unstable image quality. Diffusion model overcomes the above disadvantages of GANs, and surpasses GANs in image generation quality and stability (Ho et al., 2020). Thus, diffusion model may be able to provide a more promising way for road environment optimization.

Given the above, traditional road environment optimization methods mainly rely on manual work, and current intelligent methods cannot achieve highly accurate optimization. In this study, a novel intelligent optimization method is proposed based on diffusion model, which can directly generate optimized and accurate road environment images. The area and location information of road environment components are automatically extracted by deep neural networks, and their impacts on driving speed are quantitatively analyzed using extreme gradient boosting (XGBoost) and Shapley additive explanations (SHAP) algorithms. Then, diffusion model is used to generate optimized road environment images. Integrating the driving speed prediction model, the effectiveness of optimization can be promptly validated. It is expected that this study could help

road designers to select specific types of components in specific locations for optimization according to the interpretable analysis by the SHAP algorithm, and to use diffusion model to directly generate the optimized images, thus enhancing the automation, efficiency, and image quality of rural road environment optimization.

2 Methodology

As shown in Fig. 1, this study extracts 27 independent variables from a newly proposed rural road environment model by using naturalistic driving data. Then, XGBoost and SHAP algorithms are applied to predict and analyze driving speed. Subsequently, a diffusion model-based intelligent optimization method for rural road environments is presented. This methodology section consists of three parts: (1) a road environment model is established, including the visual road alignment layer and the visual semantic layer; (2) employing XGBoost and SHAP algorithms, a quantitative analysis is conducted to ascertain the impacts of road environment components and their locations on driving speed; (3) a diffusion model-based intelligent optimization method is proposed, to achieve direct and accurate generation of optimized road environment images.

2.1 A road environment model

Considering the profound impact of the visual road environment on drivers' cognitive processing and behavioral responses (Yu et al., 2019b), this study establishes a road environment model that effectively quantifies the environment from the perspective of drivers' visual perceptions. This model consists of two layers, including the visual road alignment layer and the visual semantic layer. Distinguished from preceding models of visual road environment (Yu et al., 2018; Qin et al., 2020; Li et al., 2023), this model acquires both the area and location information of the road environment components, and realizes a more refined quantitative analysis of the road environment.

2.1.1 The visual road alignment layer

While driving, drivers can perceive the "road alignments", even when the lane markings are unclear or blocked. These perceived "road alignments" are referred to as the drivers' visual road alignments. Interestingly, there might be a difference between visual road alignments and actual alignments, and drivers always determine their actions based on visual perceptions (Yu et al., 2018). Therefore, extracting 12 independent variables from the visual road alignment layer can better quantify the drivers' perception of the road environment and improve the accuracy of the driving speed prediction model, which helps provide more reliable guidance for subsequent optimization. Our prior research (Yu et al., 2016; Yu et al., 2019a) suggests that the Catmull-Rom spline curve is superior to cubic and quadratic curves in precisely describing the visual road alignments. The Catmull-Rom spline curve is highly accurate in fitting the visual road alignments because it passes through all control points and guarantees smooth curves. Besides, the local control feature of the Catmull-Rom spline curve ensures that local changes in a control point do not affect other parts, which further enhances the applicability in controlling the curve shape. Therefore, the Catmull-Rom spline curve is employed to fit the visual road boundaries in this study.

As shown in Fig. 2, the visual road alignment layer builds a coordinate system that takes the bottom-left corner of the drivers' visual field as its origin. It utilizes two sets of four control points each, designated as $(P_{1L}, P_{2L}, P_{3L}, P_{4L})$ for the left

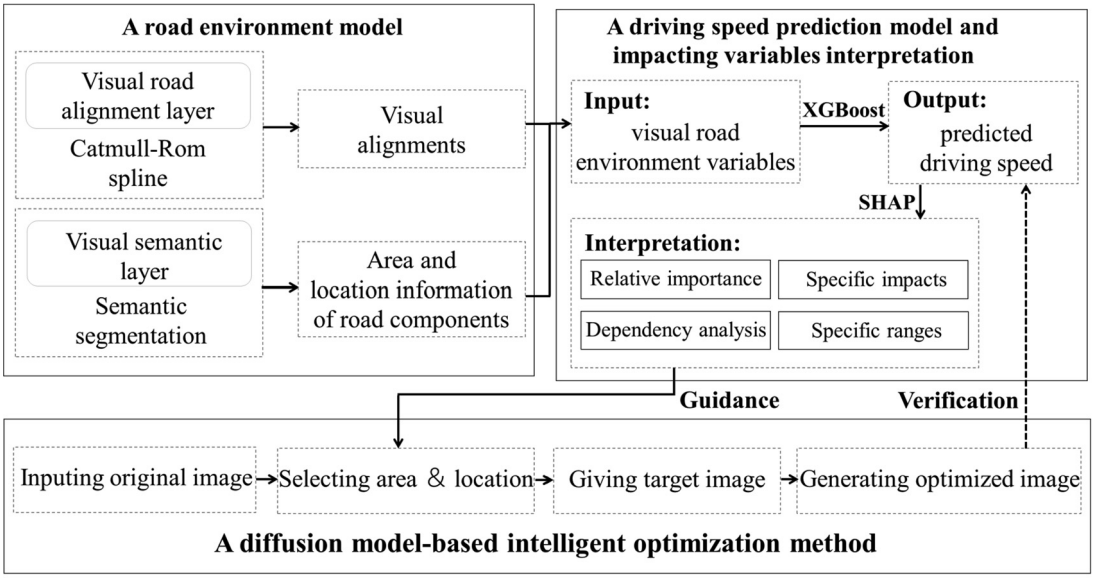

Fig. 1 The flow chart of the methodology.

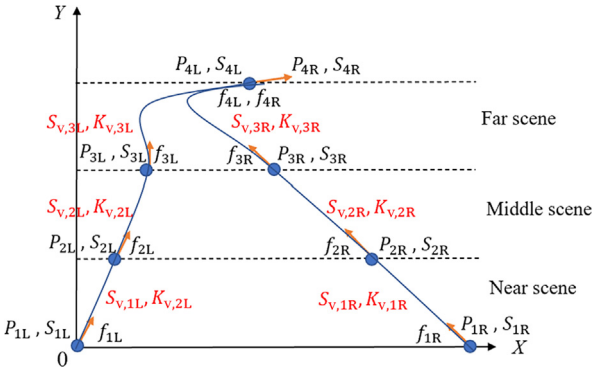

Fig. 2 The visual road alignment layer.

spline curve and $(P_{1R}, P_{2R}, P_{3R}, P_{4R})$ for the right spline curve, to fit the contour of the road boundaries. S_{iL} and S_{iR} signify the cumulative lengths (pixels) of the left and right visual road boundaries at the control points above. Besides, f_{iL} and f_{iR} are the tangent angles (radians) of the left and right visual road boundaries at the control points above. The horizontal connections of these control points on both sides of the road divide the road into three different regions, i.e., "near scene", "middle scene", and "far scene". These regions exhibit distinct visual characteristics, and aid in quantifying the visual road alignment. The parameters characterizing the shape of the three regions of the left visual road boundary are left visual curve length and left visual curve curvature, represented as $[S_{v,iL}, K_{v,iL}]$ (i = 1, 2, 3). Correspondingly, the parameters for the right visual road boundary are denoted as $[S_{v,iR}, K_{v,iR}]$ (i = 1, 2, 3). They can be calculated as follows:

$$S_{v,iL} = S_{(i+1)L} - S_{iL}, \tag{1}$$

$$K_{v,iL} = \frac{f_{(i+1)L} - f_{iL}}{S_{v,iL}}, \tag{2}$$

$$S_{v,iR} = S_{(i+1)R} - S_{iR}, \tag{3}$$

$$K_{v,iR} = \frac{f_{(i+1)R} - f_{iR}}{S_{v,iR}}, \tag{4}$$

where i = 1,2,3; S_{iL} is the cumulative length of the left road boundary at the control point P_{iL} (pixels), and $S_{v,iL}$ corresponds to the curve length of the left road boundary in the three different regions (pixels); f_{iL} is the tangent angle of the left road boundary at the control point P_{iL} (radians), and $K_{v,iL}$ corresponds to the curve curvature of the left road boundary in the three different regions (radians/ pixels); S_{iR} represents the cumulative length of the right road boundary at control point P_{iR} (pixels), and $S_{v,iR}$ corresponds to the curve length of the right road boundary in the three different regions (pixels); f_{iR} represents the tangent slope of the right road boundary at control point P_{iR} (radians), and $K_{v,iR}$ corresponds to the curve curvature of the right road boundary in the three different regions (radians/ pixels).

2.1.2 The visual semantic layer

The visual semantic layer includes the components of the road environment that drivers visually perceive, such as vegetation, roads, guardrails, and traffic signs. Through semantic segmentation technology, components with the same semantic category can be labeled with identical colors. As shown in Fig. 3, using a pre-trained semantic segmentation model (Qin et al., 2020), a semantic segmentation network is constructed to establish the visual semantic layer. The network consists of an encoder and a decoder. The encoder integrates feature extraction and feature fusion parts to extract road environment images by a residual network (ResNet) and to perform feature fusion via a feature pyramid network (FPN). The convolution layer enables the network to transform the ResNet's output from a non-spatial representation to a matrix (Ding et al., 2021), which facilitates the generation of feature maps that correspond to the input image, effectively mapping the spatial distribution of features. $1 \times 1\,024$ means its output is a one-dimensional matrix of length 1 024. The FPN is used to handle objects of varying sizes. The fractions such as $1/4\,096, 1/8\,192, 1/16\,384, 1/32\,768$ represent the downscaling ratios of the feature maps relative to the original input image (Li et al., 2022). Larger denominators indicate smaller feature maps. The feature maps obtained by the encoder are subsequently input into the decoder, where they are first uniformly restored to $1/4\,096$ and then restored to the original image size through convolution and upsampling, thereby completing the process of semantic segmentation.

The challenge of semantic segmentation is to determine both the class of the object and its precise outline. Class determination requires the extraction of higher-order semantic features, while determining the outlines of the objects requires

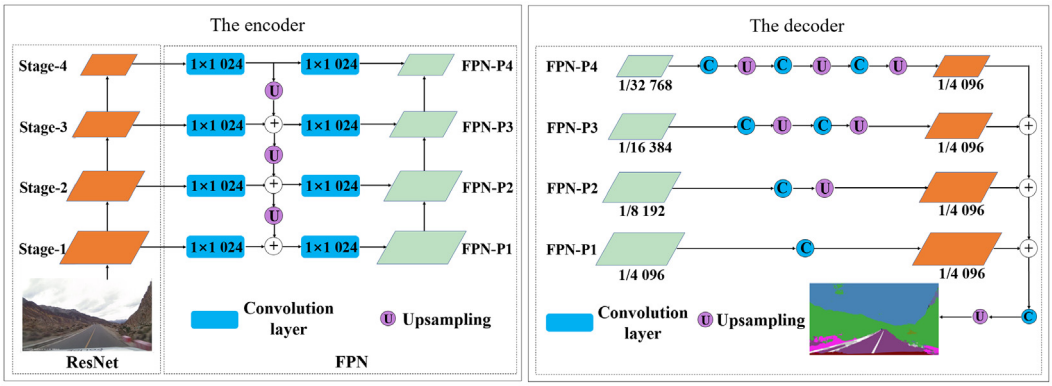

Fig. 3 The visual semantic layer.

lower-order detail features. ResNet addresses this challenge by introducing residual connections, which effectively mitigates the vanishing gradient problem (Koonce and Koonce, 2021). This enables the training of deeper networks, and helps to extract higher-order features. Concurrently, FPN enhances feature representation by integrating features from different scales, which helps to fuse lower-order features (Gong et al., 2021). By combining ResNet and FPN, the network is capable of robust feature extraction and fusion, which guarantees the accuracy of semantic segmentation.

This layer extracts seven kinds of semantic components in the visual road environment, including roads, shoulders, lane markings, vegetation, corrugated beam guardrails, concrete guardrails, and traffic signs. In the semantic segmentation process, each semantic category is assigned a corresponding RGB color. Subsequently, a Python program is employed to count the number of pixels for each RGB colour. By dividing these counts by the total number of pixels, the proportion of each semantic category can be calculated as variables extracted from this layer. To further investigate the impact of the location of traffic signs and guardrails on driving speed, the image is divided into three different regions, e.g., named "near scene", "middle scene", and "far scene" by two sets of four control points each, as shown in Fig. 2. The location information is then obtained by counting the number of pixels in each region and dividing it by the total number of pixels. These parameters are denoted by the area of roads (A.R), the area of shoulders (A.S), the area of lane markings (A.L), the area of vegetation (A.V), the area of corrugated beam guardrails (A.B), the area of concrete guardrails (A.C), the existence of traffic signs in the "near scene" (E.T.N), "middle scene" (E.T.M), or "far scene" (E.T.F), the area of corrugated beam guardrails in the "near scene" (A.B.N), "middle scene" (A.B.M), or "far scene" (A.B.F), and the area of concrete guardrails in the "near scene" (A.C.N), "middle scene" (A.C.M), or "far scene" (A.C.F).

2.2 Experiments

2.2.1 Data extraction

The naturalistic driving experiments were carried out in Xizang, China, with an accumulated travel distance of more than 10 000 km. Rural road sections consisting of mountain roads, plain roads, township roads, and other categories were used in this study. As counted through driving recorder videos, the average hourly traffic volume throughout the daytime on these roads was 118 pcu/h, and the peak hour volume was 194 pcu/h. According to China's Technical Standard of Highway Engineering (Ministry of Transport of the People's Republic of China, 2015), the design capacity of two-lane roads is typically 1 200 pcu/h. Thus, the traffic volume on these roads is significantly lower than the design capacity. Moreover, in the driving recorder videos, there are fewer vehicles and basically no traffic jams. Therefore, the driving behavior and speed are primarily determined by the visual road environment, rather than by interactions with other vehicles. A driving recorder (GARMIN GDR35) was mounted on the windshield at the driver's eye level to capture videos of the visual road environment, position, speed, acceleration, etc., with a sampling rate of 1 Hz. There were 30 drivers who took part in the driving experiments, comprising 23 males and 7 females. Their ages ranged from 23 years to 53 years (mean = 35.3 years and standard deviation (S.D.) = 7.4 years). All participants had a minimum of three years of driving experience (mean = 11.1 years, S.D. = 6.8 years, and range 3–30 years).

2.2.2 Input variables

There are 27 independent variables to predict driving speed, which are derived from the visual road alignment layer and visual semantic layer. The descriptions and statistical distributions of these variables are illustrated in Table 1.

Table 1
Distributions of input variables.

Input variables	Variable code	Description	Min	Max	Mean	S.D.
Left visual curve length (pixels)	vS_{1L}	Left visual curve length in the "near scene"	251.31	999.90	452.11	164.45
	vS_{2L}	Left visual curve length in the "middle scene"	46.59	347.03	129.43	63.83
	vS_{3L}	Left visual curve length in the "far scene"	14.08	205.18	98.65	45.97
Right visual curve length (pixels)	vS_{1R}	Right visual curve length in the "near scene"	152.51	873.12	401.87	142.23
	vS_{2R}	Right visual curve length in the "middle scene"	38.25	449.22	168.58	51.25
	vS_{3R}	Right visual curve length in the "far scene"	10.78	275.96	112.65	34.29
Left visual curve curvature (radians/pixels)	vK_{1L}	Left visual curve curvature in the "near scene"	1.2×10^{-4}	0.99	0.34	0.28
	vK_{2L}	Left visual curve curvature in the "middle scene"	1.0×10^{-5}	0.22	0.09	0.04
	vK_{3L}	Left visual curve curvature in the "far scene"	1.4×10^{-6}	0.01	0.003	0.002
Right visual curve curvature (radians/pixels)	vK_{1R}	Right visual curve curvature in the "near scene"	1.1×10^{-4}	0.60	0.27	0.21
	vK_{2R}	Right visual curve curvature in the "middle scene"	1.6×10^{-4}	0.21	0.08	0.03
	vK_{3R}	Right visual curve curvature in the "far scene"	7.2×10^{-4}	0.02	0.01	0.006
Area of roads	A.R	The proportion of area of roads	0.06	0.29	0.16	0.07
Area of shoulders	A.S	The proportion of the area of shoulders	0	0.01	0.004	0.002
Area of lane markings	A.L	The proportion of area of lane markings	0	0.02	0.009	0.003
Area of vegetation	A.V	The proportion of area of vegetation	0.21	0.64	0.43	0.29
Area of corrugated beam guardrails	A.B	The proportion of area of corrugated beam guardrails	0	0.06	0.02	0.02
Area of concrete guardrails	A.C	The proportion of area of concrete guardrails	0	0.04	0.02	0.01
The existence of traffic signs in the "near scene"	E.T.N	Traffic signs in the "near scene" exist or not	Sign (90.6%)	No-sign (9.4%)	/	/
The existence of traffic signs in the "middle scene"	E.T.M	Traffic signs in the "middle scene" exist or not	Sign (85.9%)	No-sign (14.1%)	/	/
The existence of traffic signs in the "far scene"	E.T.F	Traffic signs in the "far scene" exist or not	Sign (89.1%)	No-sign (10.9%)	/	/
Area of corrugated beam guardrails in the "near scene"	A.B.N	The proportion of area of corrugated beam guardrails in the "near scene"	0	0.056	0.023	0.01
Area of corrugated beam guardrails in the "middle scene"	A.B.M	The proportion of area of corrugated beam guardrails in the "middle scene"	0	0.006	0.003	0.002
Area of corrugated beam guardrails in the "far scene"	A.B.F	The proportion of area of corrugated beam guardrails in the "far scene"	0	0.002	0.001	0.001
Area of concrete guardrails in the "near scene"	A.C.N	The proportion of area of concrete guardrails in the "near scene"	0	0.042	0.017	0.009
Area of concrete guardrails in the "middle scene"	A.C.M	The proportion of area of concrete guardrails in the "middle scene"	0	0.009	0.004	0.003
Area of concrete guardrails in the "far scene"	A.C.F	The proportion of area of concrete guardrails in the "far scene"	0	0.003	0.001	0.001
Driving speed (km/h)	Sp	Driving speed at the current position	37	93	62	16

2.3 A prediction model of driving speed

2.3.1 XGBoost algorithm

This study uses XGBoost to establish a driving speed prediction model. XGBoost is an efficient and robust machine learning algorithm that employs the principle of gradient boosting and integrates regularization to control for model complexity and overfitting. The benefits of XGBoost include expected computational speed, superior prediction accuracy, and convenient management of model complexity when addressing data-related issues (Chen et al., 2015).

Given a dataset with n samples and m features, we denote the sample i as (x_i, y_i), where $x_i \in \mathbf{R}^m$ is the feature vector and $y_i \in \mathbf{R}$ is the target. The objective of XGBoost is to find a function F that minimizes the loss function L:

$$L = \sum_{i=1}^{N} \mathrm{Loss}(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(F_t), \tag{5}$$

$$\Omega(F_t) = \gamma J + \frac{1}{2}\lambda \|w\|^2, \tag{6}$$

where $\sum_{i=1}^{N} \mathrm{Loss}(y_i, \hat{y}_i)$ is a differentiable convex loss function; $\sum_{t=1}^{T} \Omega(F_t)$ is the regularization term that penalizes the complexity of the model; F_t represents the tree structure, and $\Omega(F_t)$ counts the number of leaves and the L_2 norm of the leaf weights; J is

the number of leaves; w is the weight vector of leaves; γ is the parameter for complexity control; λ is the parameter to control the L_2 norm of leaf weights.

In each iteration, the algorithm adds a new function h_k to the current model $F_{k-1}(x)$ to minimize the loss function L. Then, $F_k(x) = F_{k-1}(x) + h_k(x)$, iterating repeatedly to find the function F that minimizes the loss function L. The main advantage of XGBoost over traditional gradient boosting algorithms is the regularization term. While traditional gradient boosting algorithms only control the complexity by the number of trees, XGBoost reduces overfitting by the penalty term to control the number and the size of trees.

2.3.2 SHAP algorithm

To visually explain the driving speed prediction model, SHAP is applied in this study. As an interpretable algorithm, SHAP effectively solves the black-box problem of machine learning, which aids in the comprehension and analysis of the prediction model. SHAP also can be integrated with other machine learning algorithms like XGBoost to ensure both precision and interpretability (Parsa et al., 2020), which allows us to identify important variables that affect driving speed and further analyze the specific range of speed reduction by variables. Traditional algorithms such as logistic regression typically illustrate only the generalized linear relationships between dependent and independent variables. This limits its ability to capture the nonlinear relationships that often occur in real-world driving data. However, SHAP can handle both linear and nonlinear relationships by providing feature importance and interaction values. Moreover, SHAP can help us understand how different variables collectively affect driving speed prediction by analyzing interaction values (He et al., 2023).

SHAP bases on cooperative game theory, and assigns each variable a Shapley value by considering the permutation and combination of subsets of variables. For each variable i within subset S, SHAP computes its marginal contribution ($\Delta\Phi_i$) and shapley value (Φ_i) by Eqs. (7) and (8). The shapley value of each variable i represents its contribution to the model calculation, which can be used to interpret the results of the model.

$$\Delta\Phi_i(S) = \Phi_i\left(\mathrm{S}\bigcup\{\mathrm{i}\}\right) - \Phi_i(S), \tag{7}$$

$$\Phi_i = \sum_{S \subseteq Nleft\{i\}} \frac{|S|!(|N| - ||S| - 1|!)}{|N|!}\Delta\Phi_i(S), \tag{8}$$

where N is the set of N variables; S is a subset of N variables; $\Phi_i(S)$ represents the payoff of the set S in the cooperative game; $|S|$ denotes the size of set S; $|N|$ denotes the size of the set of N variables.

2.3.3 Evaluation indices of speed prediction

To compare and analyze the prediction effect, this study utilizes some accuracy indices to evaluate the performance of the prediction model, including the mean absolute error (MAE, e_{MAE}), mean squared error (MSE, e_{MSE}), mean forecast error (MFE, e_{MFE}), and mean absolute percentage error (MAPE, e_{MAPE}). MAE is the average of the absolute deviation of the predicted values from the actual value. MSE is calculated as the average squared deviation of the predicted values from the actual value. MFE is computed by averaging the deviation of the predicted values from the actual value. MAPE is the percentage of the sum of the ratios of the deviations of all predicted values and actual values to the actual value (Yu et al., 2024b). Their calculation equations are as follows:

$$e_{\mathrm{MAE}} = \frac{\sum |y_t - \widehat{y}_t|}{n}, \tag{9}$$

$$e_{\mathrm{MSE}} = \frac{\sum (y_t - \widehat{y}_t)^2}{n}, \tag{10}$$

$$e_{\mathrm{MFE}} = \frac{\sum (y_t - \widehat{y}_t)}{n}, \tag{11}$$

$$e_{\mathrm{MAPE}} = \left(\frac{100}{n}\right)\sum \left|\frac{y_t - \widehat{y}_t}{y_t}\right| 100\%, \tag{12}$$

where y_t represents the actual value of the sample t; $\widehat{y}_t$ represents the predicted value of the sample t; n is the total number of data samples.

2.4 A Diffusion model-based intelligent optimization method for rural road environments

2.4.1 An intelligent optimization method

The framework of the intelligent optimization method is proposed in Fig. 4. First, as shown in Fig. 4(a), road environments where driving speed exceeds the speed limit are screened out. Then, as illustrated in Fig. 4(b), the unsafe road environment is semantically segmented and interpretably analyzed through the road environment model, and the component for optimiza-

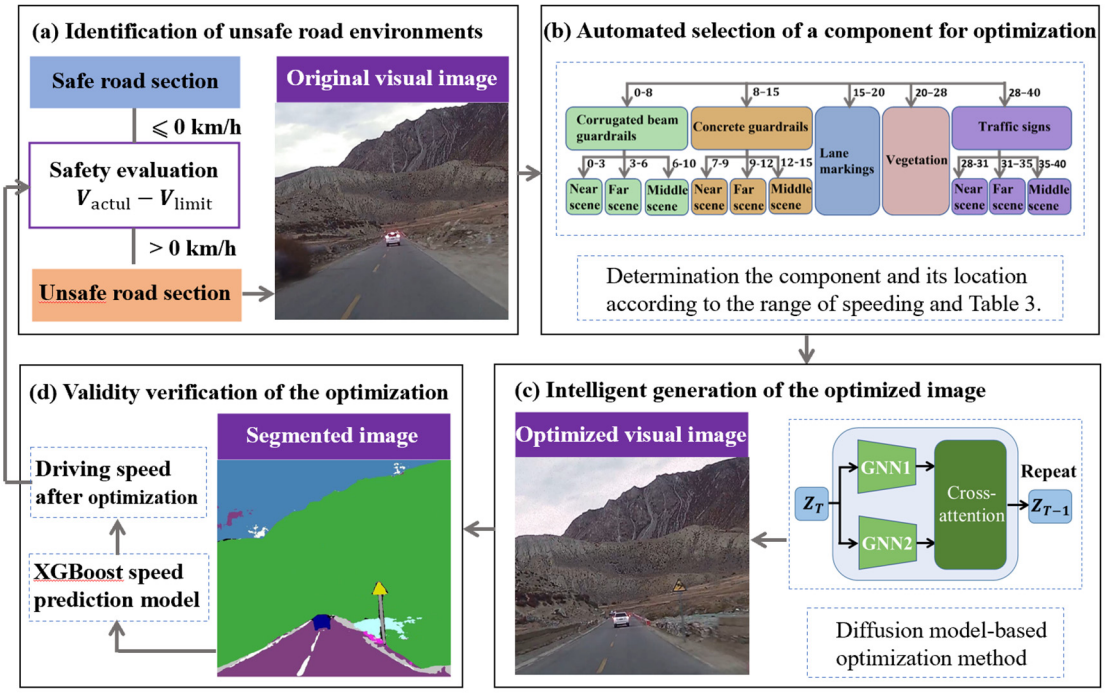

Fig. 4 The intelligent optimization method.

tion and its location are determined automatically through the Python program according to the range of speeding and the impacts of variables in Table 3. After that, the diffusion model-based optimization method is employed to automatically optimize the original environment and generate the optimized image (see Fig. 4(c)). At last, it is presented in Fig. 4(d) that the optimized environment is semantically segmented again, and the XGBoost speed prediction model is employed to verify the optimization effect. If the predicted driving speed of the optimized environment is below the speed limit, the optimization is complete. If not, the optimization process should be repeated until the predicted driving speed satisfies the speed limit's requirement.

2.4.2 Diffusion model

Diffusion model has emerged as a hot topic in computer vision research, especially in the area of image generation. It has numerous advantages, including superior image quality, flexible image control, and robust image generation (Yang et al., 2023). Diffusion model is used to adjust and optimize the rural road environment in this study, which can effectively guide drivers' behavior and enhance traffic safety.

The framework of diffusion model can be formulated as a dual Markov chain, consisting of a diffusion process and an inverse process called denoising process (Sohl-Dickstein et al., 2015). The diffusion process adds Gaussian noise iteratively to the original data, which increases the diversity of data. Conversely, the denoising process is to recover the data by removing noise, which generates the target data.

In diffusion process, at a time step $t = 0, 1, 2, \cdots, T$, the conditional distribution of the intermediate data state Z_t, given the previous state Z_{t-1}, is defined by the Gaussian distribution. Hence, this sequence is formulated as a Markov chain, which successively adds Gaussian noise into the original data through a variance schedule $\beta_1, \beta_2, \beta_3, \cdots, \beta_T$ ($\beta_t \epsilon (0, 1)$) by Eqs. (13) and (14). By setting $\bar{\alpha}_t = \prod_{s=1}^{t} 1 - \beta_s$, a delightful feature of the diffusion process is achieved that the data at any arbitrary time step t have a simple formula by Eq. (15) through a reparameterization trick. If the time step is sufficiently large, the final distribution becomes a standard Gaussian distribution.

$$q(Z_t|Z_{t-1}) = N\left(Z_t; \sqrt{1-\beta_t} Z_{t-1}, \beta_t I\right), \tag{13}$$

$$q(Z_T|Z_0) = \prod_{t=1}^{T} q(Z_t|Z_{t-1})), \tag{14}$$

$$q(Z_t|Z_0) = N\left(Z_t; \sqrt{\bar{\alpha}_t}Z_0, \left(1-\bar{\alpha}_t\right)I\right), \tag{15}$$

where I is an identity matrix; Z_0 represents the original data state in diffusion process; Z_{t-1} represents the data state at time step $t-1$ in diffusion process; Z_t represents the data state at time step t in diffusion process; Z_T represents the final Gaussian distribution data state in diffusion process; Z_{t-1} is mixed with Gaussian noise to yield Z_t, while β_t controls the degree of mixture; parameter $\bar{\alpha}_t$ controls how much signal is retained.

The denoising process is to recover the data from the data state Z_T generated by the diffusion process. This inverse process is also a Markov chain with parameters to be learned, formulated by Eqs. (16), (17), and (18) (Ho et al., 2020). A neural network is trained to perform the maximum likelihood estimation (denoted by ϵ_θ). First, the neural network samples from the Gaussian noise Z_T $N(0,I)$, then iteratively samples Z_{t-1} $p(Z_{t-1}|Z_t)$ for $t = T, T-1, T-2, \cdots, 1$, and finally samples Z_0 which is the data state at the time step $t = 0$.

$$\mu_\theta(Z_t, t) = \frac{1}{\sqrt{1-\beta_t}}\left(Z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_{\theta(Z_t,t)}\right), \tag{16}$$

$$p_\theta(Z_{t-1}|Z_t) = N(Z_{t-1}; \mu_\theta(Z_t, t), \sigma_t^2 I), \tag{17}$$

$$p_\theta(Z_0|Z_T) = \prod_{t=1}^{T} p_\theta(Z_{t-1}|Z_t), \tag{18}$$

where I is an identity matrix; Z_T represents the Gaussian distribution data state in denoising process; Z_t represents the data state at time step t in denoising process; Z_{t-1} represents the data state at time step $t-1$ in denoising process; Z_0 represents the final generated data state in denoising process; ϵ_θ is the parameter estimated by the neural network; μ_θ denotes the mean approximated by the neural network; σ_t^2 denotes the variance given by users.

Fig. 5 presents the components and processes of the diffusion model used in this study. The diffusion process gradually adds Gaussian noise to improve the variety of image generation. The denoising process contains two graph neural networks, i.e., GNN1 and GNN2. GNN1 and GNN2 can obtain geometric and appearance consistency information, respectively. Geometric consistency ensures that the generated object has a reasonable size and shape, while appearance consistency ensures that the generated object is coordinated with the boundary, light, and other parameters of the original road environment (Niu et al., 2021). The cross-attention layer is based on the transformer. The key component of the transformer is the self-attention mechanism, which enables the model to weigh the importance of different parts of the input sequence dynamically (Vaswani et al., 2017). This mechanism relies on three types of vectors: query (q), key (k), and value (v). The cross-attention layer combines two sequences of geometric and appearance consistency information. One sequence serves as the query vector, while the other sequence provides the key vector and value vector. Then, the layer computes a weighted sum of the value vector based on the similarity between the query vector and the key vector, which serves as the output. By incorporating both geometric and appearance consistency information, the cross-attention layer helps generate more realistic images. This step is repeated until the noise is removed to generate the target image.

2.4.3 Evaluation index of image generation

This study employs the structural similarity index (SSIM, M_{SSI}) to evaluate the image generation quality. The SSIM and the peak signal-to-noise ratio (PSNR) are two widely utilized measurement indices in the field of image processing. The PSNR is based on the MSE, which quantifies the average squared difference between the corresponding pixels of the generated and target images (Huynh-Thu and Ghanbari, 2012). However, the PSNR does not always correlate well with human visual perception, because it fails to account for the structural similarity between the generated and target images, which is important for human visual quality assessment (Setiadi, 2021). In contrast, the SSIM is an image measurement index based on three factors, i.e., luminance distortion, contrast distortion, and structure comparison, which is more aligned with the human visual system and provides a more accurate assessment of image generation quality. The range of the SSIM value is between 0 and 1. The closer it is to 1, the smaller the difference between the generated and target images, indicating better image generation quality. Given two images X and Y, the SSIM value can be calculated by

$$M_{SSI} = \frac{(2\mu_X\mu_Y + c_1)\cdot(2\sigma_X\sigma_Y + c_2)\cdot(\sigma_{XY} + c_3)}{(\mu_X^2 + \mu_Y^2 + c_1)\cdot(\sigma_X^2 + \sigma_Y^2 + c_2)\cdot(\sigma_X\sigma_Y + c_3)}, \tag{19}$$

where μ_X and μ_Y are the mean luminance of images X and Y, respectively; σ_X^2 and σ_Y^2 are the variance in the luminance of images X and Y, respectively; σ_{XY} denotes the covariance between images X and Y; c_1, c_2, and c_3 are constants used to avoid a null denominator.

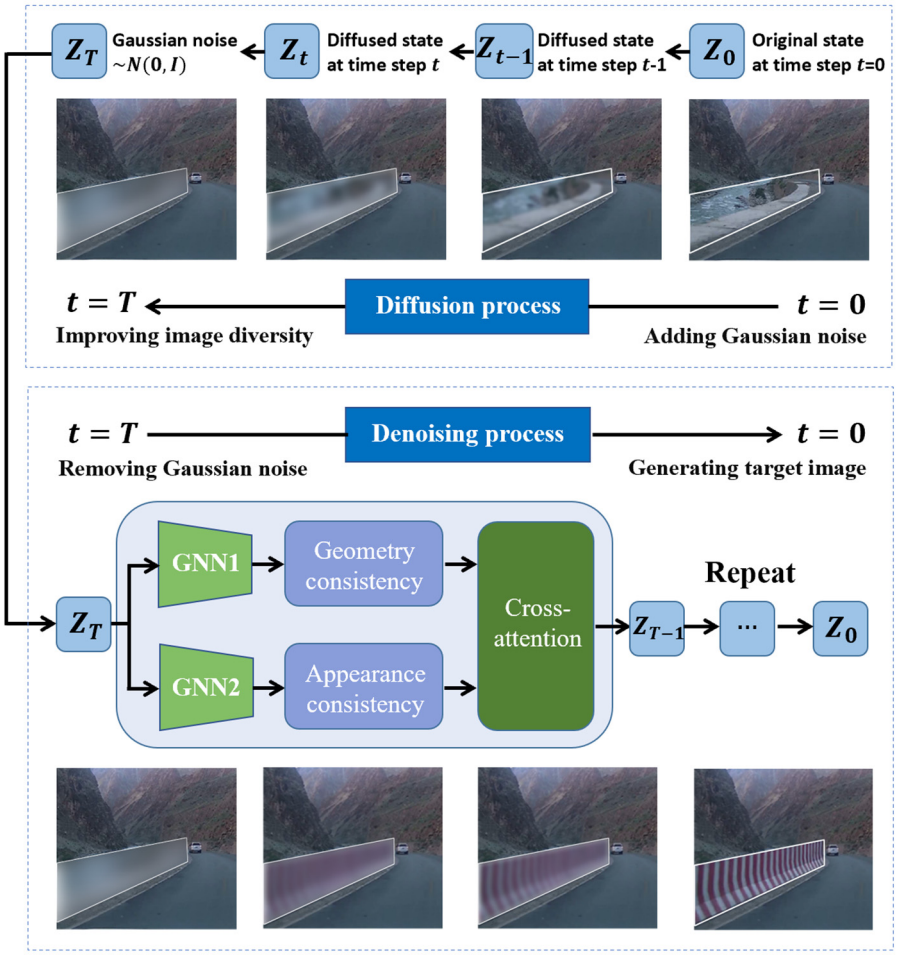

Fig. 5 The components and processes of the diffusion model.

3 Results

3.1 A prediction model of driving speed

In this study, 1 210 valid samples are extracted. The extracted samples cover the terrains encountered during driving, including mountainous, plain, and township regions, which is necessary for training driving speed prediction model suitable for different environmental conditions (Xiu et al., 2025). To ascertain the impacts of road environment components on driving speed, the images containing vegetation, guardrails, lane markings, and traffic signs in the drivers' field of view are selected. Moreover, clear and unobstructed images are preferentially extracted to ensure that the details of road environment components are clearly visible, which is essential for semantic segmentation and the training of diffusion model. The training and testing groups are allocated according to the ratio of 0.8: 0.2. With 27 input variables, XGBoost is utilized to establish a prediction model of driving speed. Based on the optimal results of the grid search and the five-fold cross-validation, the parameters of XGBoost are set as follows: the number of trees is 200, the learning rate is 0.5, and the max-

Table 2
Driving speed prediction model accuracy results.

Algorithm	MFE/(km · h^{-1})	MAE/(km · h^{-1})	MSE/(km^2 · h^{-2})	MAPE/%
Linear regression (LR)	2.34	10.87	106.70	14.34
Random forest (RF)	1.67	6.65	35.15	9.35
Support vector regression (SVR)	1.80	7.90	43.26	10.98
XGBoost	1.29	5.13	20.43	5.65

imum depth of each tree is 5. As illustrated in Table 2, XGBoost performs well in prediction accuracy, with an MFE of 1.29 km/h, an MAE of 5.13 km/h, an MSE of 20.43 km^2/h^2, and an MAPE of 5.65%.

For comparison, this study also employs three additional algorithms to develop the driving speed prediction model, including LR, RF, and SVM. LR is a fundamental and widely used algorithm for regression tasks. Its simplicity and interpretability make it a valuable benchmark for evaluating the more complicated algorithms (Su et al., 2012). RF is an ensemble learning algorithm that integrates multiple decision trees to improve prediction accuracy. It performs well in various regression tasks, thus serving as a robust comparison model (Gu et al., 2023). SVM excels in high-dimensional spaces and is known for its ability to handle complex data structures and high accuracy (Li et al., 2018). By comparing XGBoost to LR, RF, and SVM, we establish a comprehensive benchmark that demonstrates the superior performance of XGBoost in predicting driving speed. As shown in Table 2, the MFE, MAE, MSE, and MAPE of XGBoost are less than those of LR, RF, and SVM. These results indicate that XGBoost can predict driving speed more accurately than other algorithms.

3.2 Impacting variables interpretation by SHAP

Despite XGBoost's high prediction accuracy, it cannot provide a detailed explanation of the model outputs. To obtain a better understanding of the prediction by XGBoost, SHAP is used to interpret the model results in this study. Impacting factors on driving speed are analyzed interpretably from four aspects, containing relative importance, specific impacts, variable dependency, and numerical ranges.

3.2.1 Relative importance of variables

Fig. 6 demonstrates the relative importance of variables, ranked by their absolute SHAP values. A variable with a higher absolute SHAP value has a larger effect on driving speed. The left and right visual curve lengths in the "middle scene" (denoted by vS_{2L} and vS_{2R}) emerge as the primary variables, with absolute SHAP values of 2.36 and 2.02, which are much higher than other variables. Two other variables showing a marked effect on driving speed are the existence of traffic signs in the "middle scene" (E.T.M) and the existence of traffic signs in the "far scene" (E.T.F), with absolute SHAP values of 1.67 and 0.89. The following is the existence of traffic signs in the "near scene" (E.T.N), ranking seventh with an absolute SHAP value of 0.64. The variables extracted from the visual semantic layer of the road environment model also have strong impacts. Among them, the relative importance of the variables decreases in a sequence of the area of vegetation (A.V), area of lane markings (A.L), area of roads (A.R), area of concrete guardrails (A.C), and area of corrugated beam guardrails (A.B). Their corresponding absolute SHAP values are 0.59, 0.48, 0.41, 0.32, and 0.11, respectively. Notably, the impact of the area of shoulders (A.S) on driving speed is slight, with an absolute SHAP value of merely 0.02.

The impact of variables' location on driving speed is characterized by the absolute SHAP value of 1.67 for the existence of traffic signs in the "middle scene" (E.T.M), 0.89 for the existence of traffic signs in the "far scene" (E.T.F), and 0.64 for the existence of traffic signs in the "near scene" (E.T.N). The influence diminishes as traffic signs exist from the "middle scene" to the "far scene" and subsequently to the "near scene". This may be because drivers pay more attention to the middle scene than the far scene, and traffic signs in the "near scene" are about to exit the driver's field of vision. The impact of the area of concrete guardrails and corrugated beam guardrails has a similar decreasing trend from "middle scene" to "far scene" and then to "near scene", with absolute SHAP values of 0.36, 0.27, 0.19 and 0.24, 0.05, 0.04, respectively. It might be due to the same reasons as those of traffic signs.

3.2.2 Specific impacts of variables

The SHAP summary plot can reflect the specific impacts of input variables on driving speed. As shown in Fig. 7, the color of each sample (i.e., a dot) ranges from blue to red, indicating a gradual increase in the feature value. The SHAP value for each sample is the abscissa value of the dot, which denotes the influence of the variable. A variable has a positive impact on driving speed when high feature value dots have positive SHAP values (i.e., their abscissa values are greater than 0). Conversely, a negative impact on driving speed can be inferred when high feature value dots have negative SHAP values. The specific impacts of the variables on driving speed are interpreted below according to Fig. 7.

In the visual road alignment layer, longer left and right visual curve lengths in the "middle scene" (denoted by vS_{2L} and vS_{2R}) increase the driving speed since the SHAP values for vS_{2L} and vS_{2R} with high feature values (i.e., red dots) are distributed above 0. This suggests an increased likelihood for drivers to accelerate as the road length increases. As for the visual semantic layer variables, red dots for the existence of traffic signs in the "near scene" (E.T.N), "middle scene" (E.T.M), and "far scene" (E.T.F) scatter on the negative axis of SHAP values, while blue dots are distributed on the positive axis. It suggests that E.T.N, E.T.M, and E.T.F can cause a reduction in driving speed. This may be because drivers are more vigilant when traffic signs exist, which reduces driving speed (Guan et al., 2014). In addition, the area of vegetation (A.V), lane markings (A.L), concrete guardrails (A.C), and corrugated beam guardrails (A.B) are both negatively associated with driving speed, because their points with high feature values have negative SHAP values. When the areas of these components decrease, the road environment is more open. In such an open environment, drivers tend to relax and be less vigilant (Meng et al., 2020), generally leading to an increase in speed.

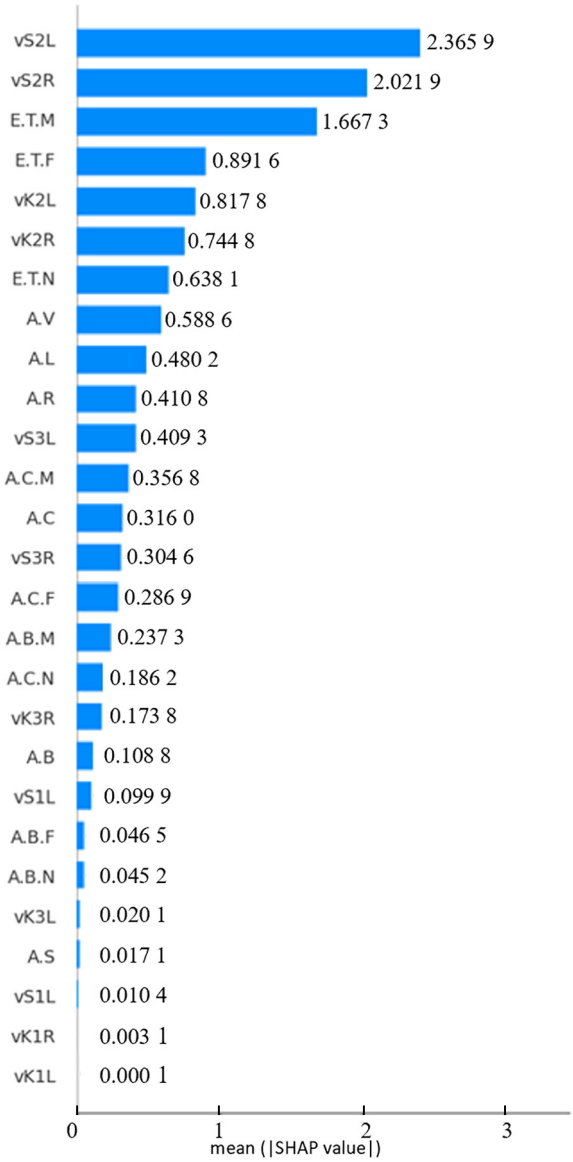

Fig. 6 Relative importance by SHAP.

3.2.3 Dependency analysis of variables

As illustrated in Fig. 8, each small graph that is not on the diagonal shows the interactions between the two variables from the abscissa axis and the ordinate axis. Each dot in a small graph represents a sample, and its color ranging from blue to red indicates a gradual increase in the feature value. The abscissa represents the SHAP interaction value. A significant interaction exists between the two variables of the abscissa axis and ordinate axis when the SHAP interaction value for dots with high feature values (i.e., red dots) is positive. In contrast, an abscissa value below zero for red dots indicating a negative SHAP interaction value, suggests the interaction between the variables from the abscissa axis and the ordinate axis is low. Taking the bottom-left graph in Fig. 8 as an example, the abscissa value for red points is positive, which indicates a notable interaction between the area of the concrete guardrails (A.C) and the existence of traffic signs in the "near scene" (E.T.N). Therefore, when optimizing the road environment, it is advisable to avoid altering these two variables concurrently to prevent mutual influence on each other.

3.2.4 Specific ranges of impacts of variables

According to the results in Fig. 7, the existence of traffic signs in the "near scene" (E.T.N), "middle scene" (E.T.M), and "far scene" (E.T.F), and the area of vegetation (A.V), lane marking (A.L), concrete guardrails (A.C), and corrugated beam guardrails (A.B) are negatively associated with driving speed. As shown in Table 3 and Fig. 9, the specific ranges of speed reduction by

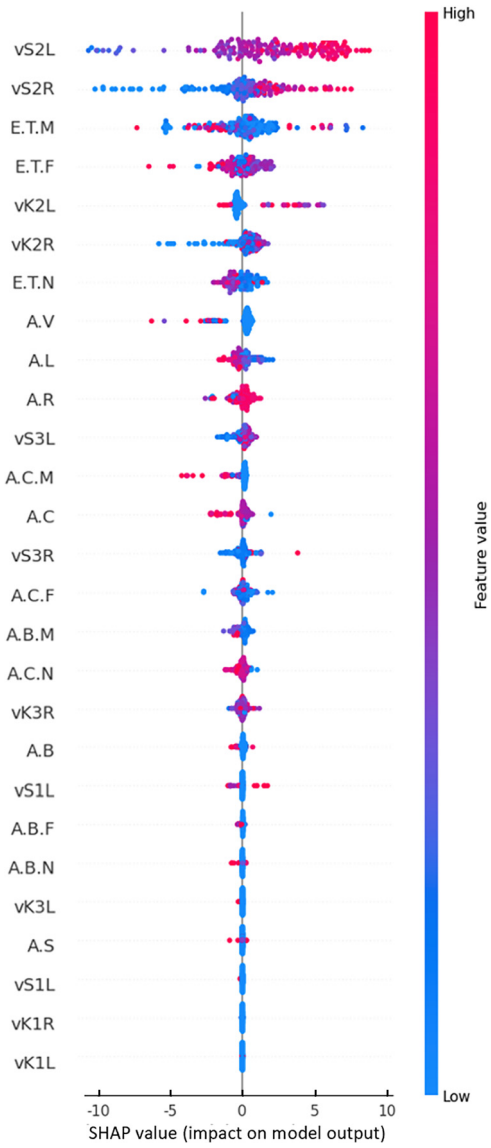

Fig. 7 Specific impacts of SHAP.

these variables can be obtained based on SHAP, which can be used as guidance for the intelligent optimization of road environments. The detailed selection process of the component and its location for road environment optimization is demonstrated in Fig. 10. After calculating how much driving speed is greater than the speed limit, and finding the influencing ranges of road environment variables on driving speed from Table 3, the component and its location in unsafe road environments for optimization can be determined.

3.3 Image generation through intelligent optimization

Using diffusion model, an intelligent optimization method is developed, enabling the selection of the component and its location for optimization and the generation of the optimized images. Fig. 11 shows the example images after optimization of concrete guardrails, corrugated beam guardrails, lane markings, vegetation, and traffic signs. The first, second, third, and fourth columns represent the original image, the selection of location (denoted by gray area), the target image, and the optimized image, respectively. Driving speeds on these road sections significantly exceed the speed limit, urgently requiring optimization, and all optimized speeds are below the speed limit. In this study, the parameters in diffusion model are set as follows: the learning rate is 1.0×10^{-5}, the linear learning rate starting value is 0.000 85, the linear learning rate ending value is 0.012, the timestep is 1 000, and the final generated image resolution is 512×512 pixels. To meet the large memory

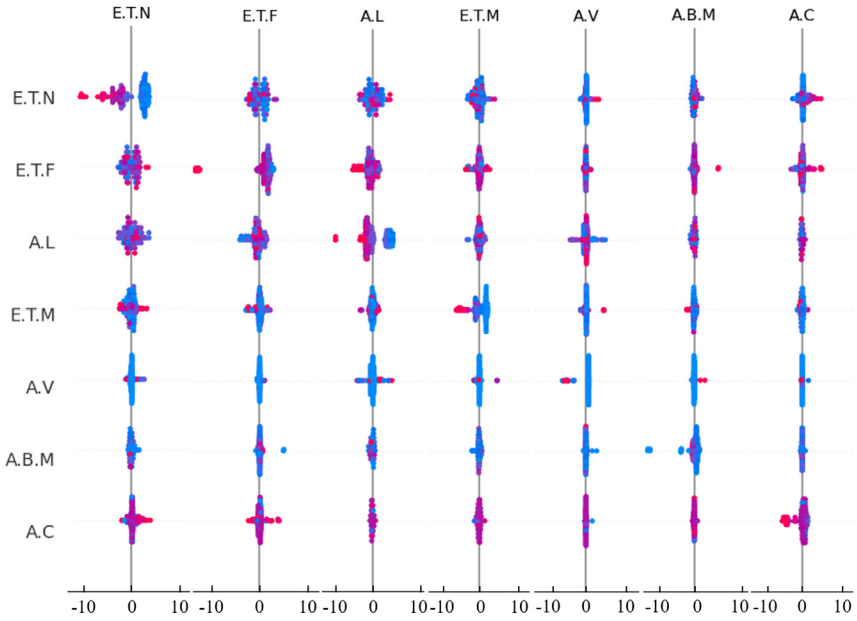

Fig. 8 Dependency analysis by SHAP.

Table 3
The specific ranges of speed reduction by variables.

Variable	Range/(km · h^{-1})	Mean/(km · h^{-1})	S.D./(km · h^{-1})
Traffic signs in the "near scene"	28–31	30.2	3.1
Traffic signs in the "middle scene"	35–40	37.4	3.6
Traffic signs in the "far scene"	31–35	33.8	4.2
Vegetation	20–28	24.3	10.8
Lane markings	15–20	16.9	5.3
Concrete guardrails in the "near scene"	7–9	7.9	2.2
Concrete guardrails in the "middle scene"	12–15	13.7	2.7
Concrete guardrails in the "far scene"	9–12	10.5	2.3
Corrugated beam guardrails in the "near scene"	0–3	2.0	1.6
Corrugated beam guardrails in the "middle scene"	6–10	8.3	3.9
Corrugated beam guardrails in the "far scene"	3–6	5.1	1.9

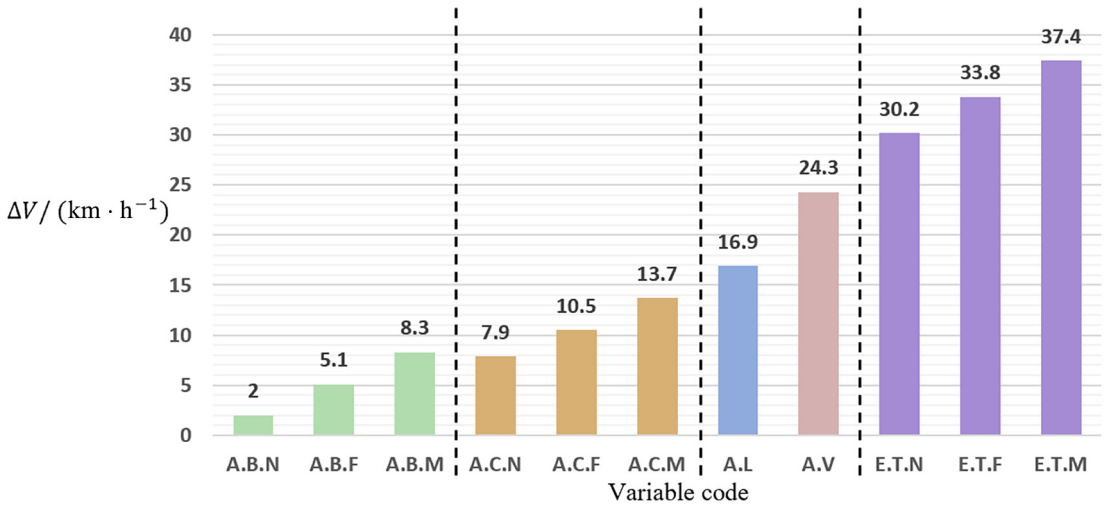

Fig. 9 The specific ranges of speed reduction.

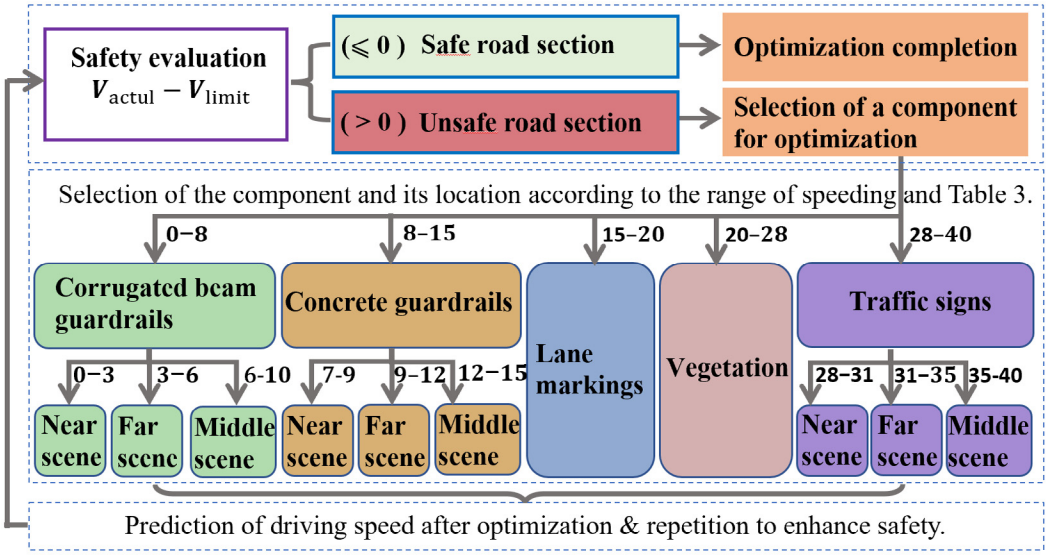

Fig. 10 Selection of the component and its location for optimization.

requirements and handle the complex computations in the denoising process of diffusion model, four NVIDIA RTX 3090 GPUs, each equipped with 24 GB of graphics memory, are used for training.

The last row of Fig. 11, with an original speed of 89 km/h, is taken as an example to describe the optimization process. First, the visual environment is quantitatively quantified with the road environment model proposed in Section 2.1. The speed limit on this rural road is 60 km/h and the current driving speed is 29 km/h above the limit. According to Table 3, traffic signs in the "near scene" can reduce the driving speed by 28 km/h to 31 km/h. Then, the area and location of the components for optimization are selected as shown in the gray area, and the ideal traffic sign image, serving as the target image, is input into the model. After that, the trained diffusion model is used to generate the optimized image. After semantically segmenting and information extraction, the driving speed of the optimized road environment predicted by the trained XGBoost model is 54 km/h. The optimized driving speed is less than the speed limit, and the optimization is complete.

To evaluate the advantages of diffusion model in the quality and stability of image generation, the optimized images generated by diffusion model and CycleGAN are compared. CycleGAN is an image-to-image translation algorithm first proposed by Zhu et al. (2017), and has demonstrated high performance in various image translation and generation tasks. Fig. 12 demonstrates the optimized images generated by Diffusion model and CycleGAN. The first, second, third, and fourth columns represent the original image, the target image, the optimized image of diffusion model, and the optimized image of CycleGAN, respectively. In this study, as illustrated in Table 4, the SSIM between the target image and generated image is calculated. The average SSIM between the optimized images generated by diffusion model and the target images is 0.753, which is significantly higher than the 0.501 achieved by CycleGAN. In addition, the standard deviation of SSIM for diffusion model is 0.103, which is less than that of CycleGAN (0.170). This indicates that the images generated by diffusion model are closer to the target images, exhibiting better image generation quality and stability than to CycleGAN.

4 Discussion and conclusion

This study proposes an intelligent optimization method for rural road environments based on diffusion model. Firstly, a road environment model consisting of the visual road alignment layer and the visual semantic layer is established, and XGBoost and SHAP algorithms are used to quantitatively analyze the impacts of the road environment component and its location on driving speed. Then, a diffusion model-based intelligent optimization method is developed. The optimized road environment image can be directly generated, which enhances the automation, efficiency, and image quality of road environment optimization.

The SSIM results in this study show that the quality of the images generated by diffusion model is better and more stable than those of CycleGAN. This may be because diffusion model is a parameterized Markov chain trained using variational inference, which is capable of generating high-quality images matching the data. GAN has expanded into numerous integrated applications. For example, a scarce scene transformation generative adversarial network (SST-GAN) achieves the generation of rare driving scenes (Wang et al., 2022), and a GAN-based translation algorithm enables the driving environment image transform from adverse weather to regular weather, enhancing the performance of vehicle detection and related tasks (Lee et al., 2022). However, its transformation is limited to the entire image, and cannot optimize specific regions accurately

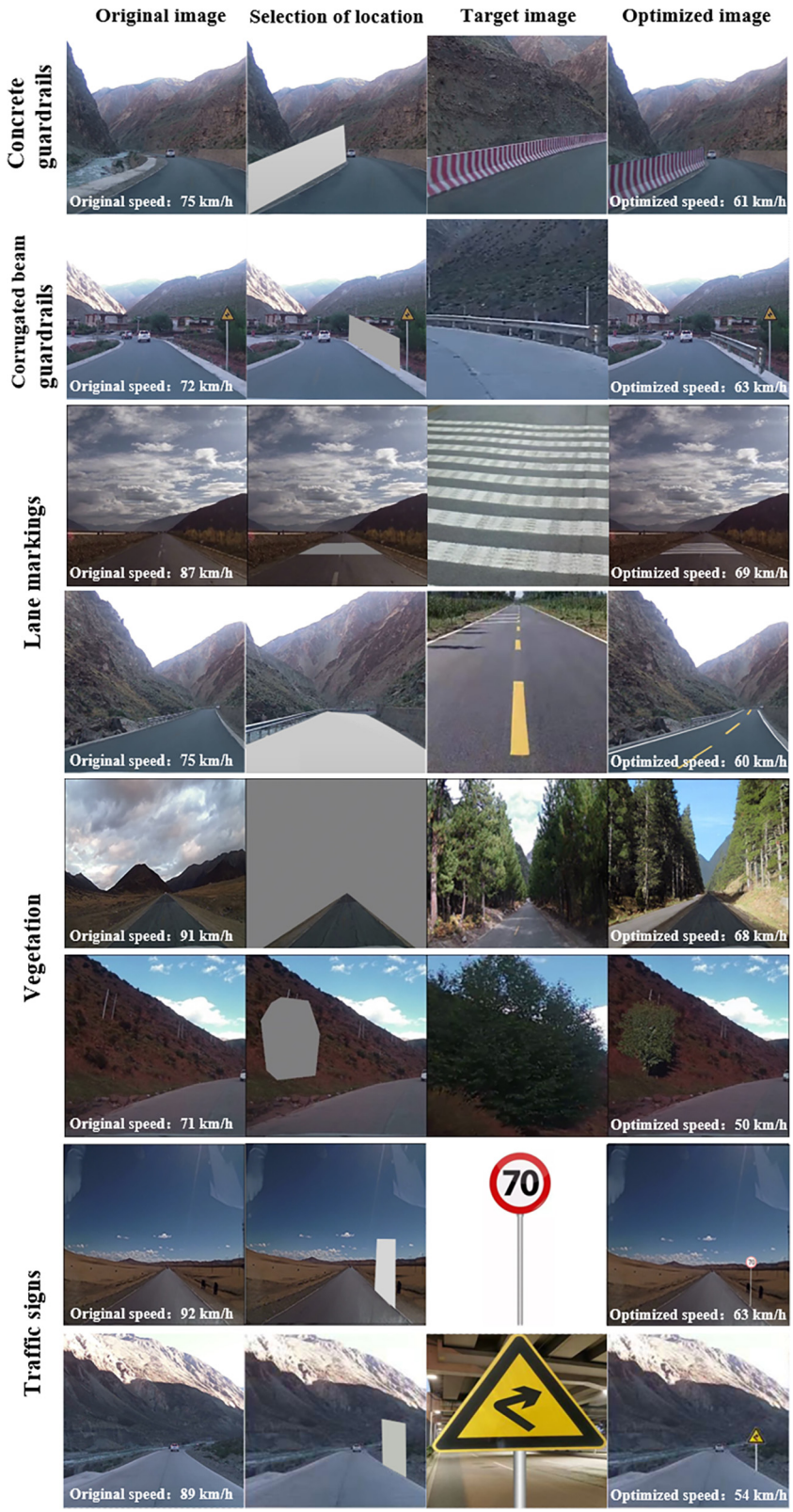

Fig. 11 Examples of optimized images by diffusion model.

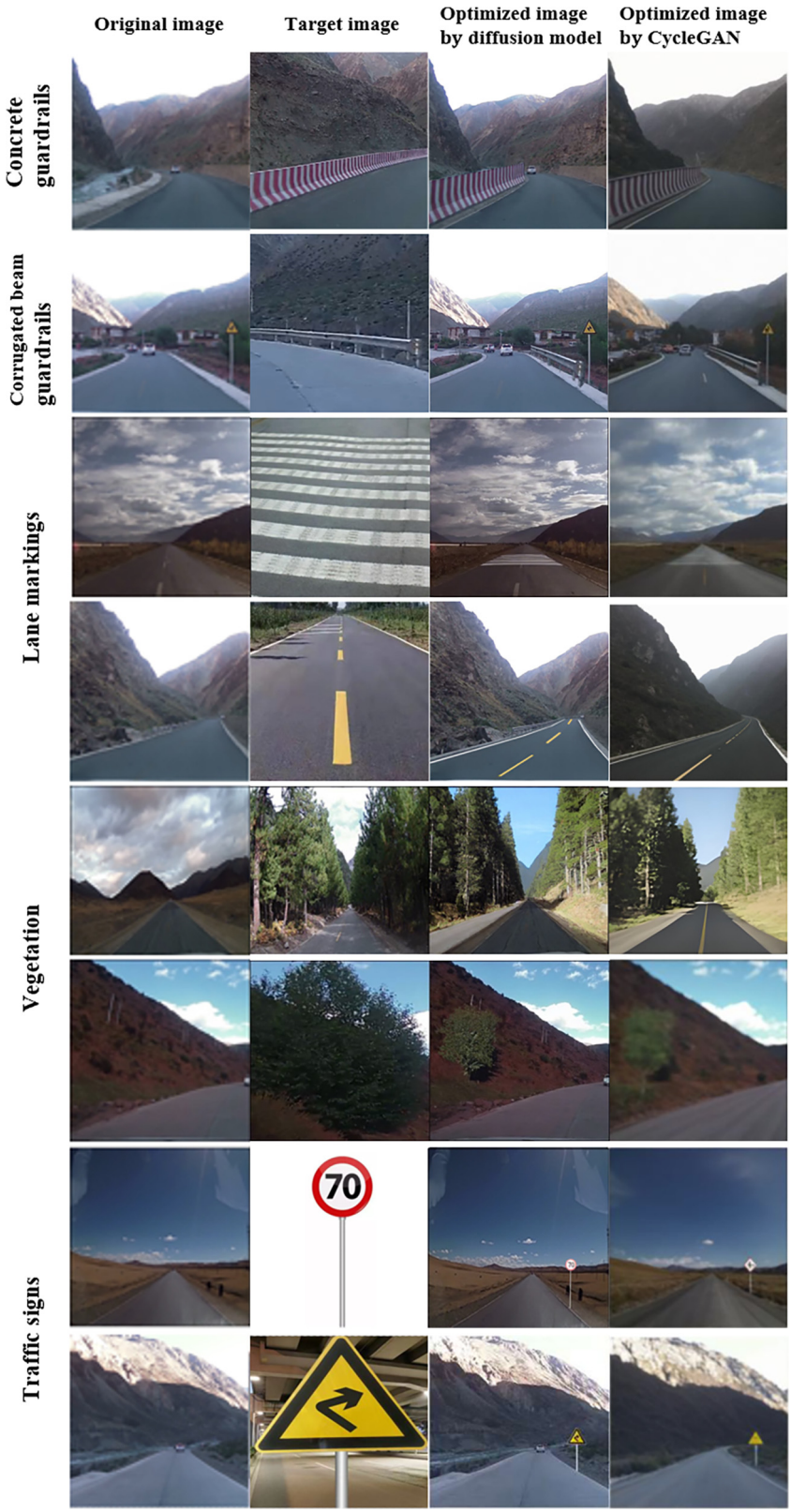

Fig. 12 Comparison of images generated by diffusion model and CycleGAN.

Table 4
Comparison of SSIM.

Algorithm	Min	Max	Mean	S.D.
Diffusion model	0.565	0.897	0.753	0.103
CycleGAN	0.287	0.710	0.501	0.170

within the driving environment, resulting in non-ideal image generation quality. In the future, more integrated applications of diffusion model need to be developed to enhance road environment image generation quality and ensure traffic safety.

This study can contribute to the following aspects.

(1) Drivers' visual perception plays an increasingly important role in road environment design (Wang et al., 2020; Gao et al., 2024b). However, drivers' perception of the road environment is often different from the actual situation, and drivers' behaviors are mostly based on visual perception (Yu et al., 2018). Existing technologies cannot obtain images from driver's perspectives, and thus the impacts of the optimized facility on the drivers' perception and driving behavior cannot be accurately evaluated (Bobermin et al., 2021; Jiang et al., 2022). Our intelligent optimization method automatically generates images from the driver's perspective and then the optimized driving speed can be predicted, which aligns more closely with the actual perceptual process of drivers, ensuring the effectiveness and applicability of the optimization.

(2) Utilizing diffusion model, various road environment components can be integrated into the existing driving environment, enabling the direct and intelligent generation of synthesized images. It is feasible to extend the application of our intelligent optimization method to the field of autonomous driving. Autonomous driving systems demand stringent safety standards, which require comprehensive testing of diverse scenarios (Feng et al., 2021; Yao et al., 2025). Autonomous driving systems need to handle a variety of complex scenarios and edge situations (Liu et al., 2019; Yu et al., 2024a). The large amounts of data are crucial to solve the long-tail problem, because the insufficient amount of data to train the model can cause the model to fail to properly handle these edge scenarios. However, the commonly used datasets have shortcomings such as limited region and limited scene type. The Mapillary Vistas (Neuhold et al., 2017) and BDD100K (Yu et al., 2020) are crowdsourced datasets, which are limited by inconsistent collecting devices. The KITTI (Geiger et al., 2012) was collected in the urban areas of Karlsruhe, Germany, and is not representative of scenes in other countries. WayveScenes101 contains only 101 scenes (Zürn et al., 2024). Our method can generate road environment images that integrate various components and can be precisely tailored to specific requirements, which contributes to solving this problem. This enhances the diversity of testing scenarios, and aids in conducting an overall assessment of the safety systems in autonomous vehicles, thereby improving their safety.

(3) With the planning and upgrading of roads, changes in the road network, such as the introduction of new highways or bypasses, lead to alterations in road types and speed limits. Consequently, road design needs to be optimized to visually signal changes in the speed limit to drivers. Driving simulation and testing are essential tools for the verification of road optimization design, but they demand substantial human and material resources (Bella, 2009; Gao et al., 2024a). Thus, this study proposes an intelligent method that can verify the optimization effect in real time. After utilizing diffusion model to generate an optimized road environment image, this image will be automatically segmented and quantitatively analyzed to predict the driver's driving speed on the upgraded road. The process will not end until the speed of the optimized road environment image reaches the safety standard.

There are several limitations of this study. First, this study only extracts variables from the road environment model to reflect the common characteristics of road environments, but the impacts of driver characteristics (e.g., age, gender, and psychology) and vehicle characteristics (e.g., type and vehicle dynamic characteristics) on driving speed are neglected. In the future, a more refined optimization method will be proposed by integrating and analyzing these factors. Another limitation is that this study is limited to the generation of a single image. In the driver's actual cognitive process, facilities in the road environment do not enter the driver's field of view at a certain moment, but enter the driver's field of view for a continuous period of time and then leave. Future studies can try to generate continuous driving scene videos from the driver's perspective and analyze the impacts of road environments on driving speed and driving behavior from two dimensions of space and time.

CRediT authorship contribution statement

Bo Yu: Writing – original draft, Methodology, Data curation, Conceptualization, Writing – review & editing. **Zehong Zhu:** Writing – original draft, Methodology, Data curation, Conceptualization. **Yuren Chen:** Writing – original draft, Funding acquisition. **Junhua Wang:** Writing – original draft, Methodology, Conceptualization. **Kun Gao:** Methodology, Conceptualization. **Xin Qian:** Writing – original draft, Methodology, Data curation, Writing – review & editing.

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.

Conflict of interest

Dr. Junhua Wang is an editorial board member/editor-in-chief for International Journal of Transportation Science and Technology and was not involved in the editorial review or the decision to publish this article. All authors declare that there are no competing interests.

Acknowledgment

This project is jointly supported by the National Key R&D Program of China (No. 2023YFE0202400), the National Natural Science Foundation of China (No. 52102416), and the Natural Science Foundation of Shanghai (No. 22ZR1466000).

References

Babić, D., Brijs, T., 2021. Low-cost road marking measures for increasing safety in horizontal curves: a driving simulator study. Accid. Anal. Prev. 153, 106013.
Bella, F., 2009. Can driving simulators contribute to solving critical issues in geometric design? Transp. Res. Rec. 2138 (1), 120–126.
Bobermin, M.P., Silva, M.M., Ferreira, S., 2021. Driving simulators to evaluate road geometric design effects on driver behaviour: a systematic review. Accid. Anal. Prev. 150, 105923.
Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.A., Li, S.Z., 2024. A survey on generative diffusion models. IEEE Trans. Knowl. Data Eng. 36 (7), 2814–2830.
Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., 2015. XGBoost: Extreme Gradient Boosting, R Package, Version 0.4. R Package Version 0.4-2 1 (4), pp. 1–4.
Coakley, R., Storm, R., Neuman, T., 2016. Relationship between geometric design features and performance. Transp. Res. Rec. 2588 (1), 80–88.
Ding, L., Zheng, K., Lin, D., Chen, Y., Liu, B., Li, J., Bruzzone, L., 2021. MP-ResNet: multipath residual network for the semantic segmentation of high-resolution PolSAR images. IEEE Geosci. Remote Sens. Lett. 19, 1–5.
Drosu, A., Cofaru, C., Popescu, M.V., 2020. Influence of weather conditions on fatal road accidents on highways and urban and rural roads in romania. Int. J. Automot. Technol. 21, 309–317.
EU Road Safety Statistics, 2023. Road Safety in the EU: Fatalities Below Pre-pandemic Levels but Progress Remains too Slow. https://transport.ec.europa.eu/news-events/news/road-safety-eu-fatalities-below-pre-pandemic-levels-progress-remains-too-slow-2023-02-21_en.
Fan, S., Chan-Kang, C., 2008. Regional road development, rural and urban poverty: evidence from China. Transp. Policy 15 (5), 305–314.
Feng, S., Yan, X., Sun, H., Feng, Y., Liu, H.X., 2021. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. Nat. Commun. 12 (1), 748.
Gao, J., Yu, B., Chen, Y., Bao, S., Gao, K., Zhang, L., 2024a. An ADAS with better driver satisfaction under rear-end near-crash scenarios: a spatio-temporal graph transformer-based prediction framework of evasive behavior and collision risk. Transp. Res. Part C Emerging Technol. 159, 104491.
Gao, J., Yu, B., Chen, Y., Gao, K., Bao, S., 2024b. A multi-perspective fusion model for operating speed prediction on highways using knowledge-enhanced graph neural networks. Computer-aided Civ. Infrastruct. Eng. 40 (8), 1004–1027.
Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The kitti vision benchmark suite. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361.
Gilandeh, S.S., Hosseinlou, M.H., Anarkooli, A.J., 2018. Examining bus driver behavior as a function of roadway features under daytime and nighttime lighting conditions: driving simulator study. Saf. Sci. 110, 142–151.
Gong, Y., Yu, X., Ding, Y., Peng, X., Zhao, J., Han, Z., 2021. Effective fusion factor in FPN for tiny object detection. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1160–1168.
Gu, Y., Liu, D., Arvin, R., Khattak, A.J., Han, L.D., 2023. Predicting intersection crash frequency using connected vehicle data: a framework for geographical random forest. Accid. Anal. Prev. 179, 106880.
Guan, W., Zhao, X., Qin, Y., Rong, J., 2014. An explanation of how the placement of traffic signs affects drivers' deceleration on curves. Saf. Sci. 68, 243–249.
He, L., Yu, B., Chen, Y., Bao, S., Gao, K., Kong, Y., 2023. An interpretable prediction model of illegal running into the opposite lane on curve sections of two-lane rural roads from drivers' visual perceptions. Accid. Anal. Prev. 186, 107066.
Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Adv. Neural Inf. Proces. Syst. 33, 6840–6851.
Huang, C.C., Xu, G.P., Yang, H., Wei, M., Jia, L.Q., Chen, L., Hu, H., Wang, Z.M., 2020. Research of ecological landscape assessment system of highway based on 3D GIS. IOP Conference Series Earth and Environmental Science 446 (3), 032063.
Huynh-Thu, Q., Ghanbari, M., 2012. The accuracy of PSNR in predicting video quality for different video scenes and frame rates. Telecommun. Syst. 49, 35–48.
Jiang, F., Ma, L., Broyd, T., Chen, K., Luo, H., 2022. Underpass clearance checking in highway widening projects using digital twins. Autom. Constr. 141, 104406.
Job, R.S., Brodie, C., 2022. Road safety evidence review: understanding the role of speeding and speed in serious crash trauma: a case study of New Zealand. J. Road Saf. 33 (1), 5–25.
Khuzan, T.S., Al-Jumaili, M.A., 2023. A review of studying the relationship of rural road accidents with geometric design. AIP Conf. Proc. 2787, 090040.
Koonce, B., Koonce, B., 2021. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization, Apress Berkeley, CA, pp. 63–72.
Lee, J., Shiotsuka, D., Nishimori, T., Nakao, K., Kamijo, S., 2022. Gan-based lidar translation between sunny and adverse weather for autonomous driving and driving simulation. Sensors 22 (14), 5287.
Li, Y., Chen, M., Lu, X., Zhao, W., 2018. Research on optimized GA-SVM vehicle speed prediction model based on driver-vehicle-road-traffic system. Sci. China Technol. Sci. 61, 782–790.
Li, R., Wang, L., Zhang, C., Duan, C., Zheng, S., 2022. A2-FPN for semantic segmentation of fine-resolution remotely sensed images. Int. J. Remote Sens. 43 (3), 1131–1155.
Li, M., Xie, H., Shu, P., 2021. Study on the impact of traffic accidents in key areas of rural roads. Sustainability 13 (14), 7802.
Li, Z., Yu, B., Wang, Y., Chen, Y., Kong, Y., Xu, Y., 2023. A novel collision warning system based on the visual road environment schema: an examination from vehicle and driver characteristics. Accid. Anal. Prev. 190, 107154.
Liu, S., Liu, L., Tang, J., Yu, B., Wang, Y., Shi, W., 2019. Edge computing for autonomous driving: Opportunities and challenges. Proc. IEEE 107 (8), 1697–1716.

Martinelli, V., Ventura, R., Bonera, M., Barabino, B., Maternini, G., 2022. Effects of urban road environment on vehicular speed. Evidence from Brescia (Italy). Transp. Res. Procedia 60, 592–599.
Meng, Y., Chen, L., Liu, B., Chen, B., Pan, X., 2020. Calculation method of visual information for driver in mountainous highway. J. Transp. Syst. Eng. Inf. Technol. 20 (5), 45.
Ministry of Transport of The People's Republic of China, 2015. Technical Standard of Highway Engineering, JTGB01-2014.
Naveen, N., Rajesh, M., Srinivas, M., Fasioddin, M., 2017. Road safety audit of a rural road. Int. J. Civ. Eng. Technol. 8 (4), 752–761.
Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P., 2017. The mapillary vistas dataset for semantic understanding of street scenes. Proceedings of the IEEE International Conference on Computer Vision, pp. 4990–4999.
National Highway Traffic Safety Administration, 2024. Rural/Urban Traffic Fatalities, Report, No. DOT HS 813 599, National Highway Traffic Safety Administration.
Niu, L., Cong, W., Liu, L., Hong, Y., Zhang, B., Liang, J., Zhang, L., 2021. Making images real again: a comprehensive survey on deep image composition. arXiv preprint, arXiv:2106.14490.
Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A.K., 2020. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accid. Anal. Prev. 136, 105405.
Qin, Y., Chen, Y., Lin, K., 2020. Quantifying the effects of visual road information on drivers' speed choices to promote self-explaining roads. Int. J. Environ. Res. Public Health 17 (7), 2437.
Ren, W., Yu, B., Chen, Y., Gao, K., Bao, S., Wang, Z., Qin, Y., 2024. An intelligent optimization method for the facility environment on rural roads. Computer-aided Civ. Infrastruct. Eng. 39, 2559–2580.
Setiadi, D.R.I.M., 2021. PSNR vs SSIM: imperceptibility quality assessment for image steganography. Multimed. Tools Appl. 80 (6), 8423–8444.
Shao, X., Wei, C., Shen, Y., Wang, Z., 2020. Feature enhancement based on CycleGAN for nighttime vehicle detection. IEEE Access 9, 849–859.
Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. Proceedings of the 32nd International Conference on Machine Learning, Lille, France, pp. 2256–2265.
Su, X., Yan, X., Tsai, C.L., 2012. Linear regression. Wiley Interdiscip. Rev. Comput. Stat. 4 (3), 275–294.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I., 2017. Attention is all you need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
Vignali, V., Bichicchi, A., Simone, A., Lantieri, C., Dondi, G., Costa, M., 2019. Road sign vision and driver behaviour in work zones. Transp. Res. Part F: Traffic Psychol. Behav. 60, 474–484.
Vignali, V., Acerra, E.M., Lantieri, C., Di Vincenzo, F., Piacentini, G., Pancaldi, S., 2021. Building information modelling (BIM) application for an existing road infrastructure. Autom. Constr. 128, 103752.
Wang, F., Chen, Y., Wijnands, J.S., Guo, J., 2020. Modeling and interpreting road geometry from a driver's perspective using variational autoencoders. Computer-aided Civ. Infrastruct. Eng. 35 (10), 1148–1159.
Wang, J., Wang, Y., Tian, Y., Wang, X., Wang, F. Y., 2022. SST-GAN: Single sample-based realistic traffic image generation for parallel vision. 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), IEEE, pp. 1485–1490.
Wesley, P.E., Coppola, N., Golombek, Y., 2018. Urban clear zones, street trees, and road safety. Res. Transp. Bus. Manag. 29, 136–143.
Xiu, H., Liu, X., Kim, T., Kim, K.S., 2025. Advancing ALS applications with large-scale pre-training: dataset development and downstream assessment. arXiv preprint, arXiv:2501.05095.
Yang, G., Huang, X., Hao, Z., Liu, M. Y., Belongie, S., Hariharan, B., 2019. Pointflow: 3D point cloud generation with continuous normalizing flows. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4541–4550.
Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F., 2023. Paint by example: exemplar-based image editing with diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18381–18391.
Yao, S., Yu, B., Chen, Y., Gao, K., Bao, S., Shangguan, Q., 2025. Does road environment aesthetics influence risky driving behavior of autonomous vehicles? An evaluation on road readiness using explainable machine learning and random parameters multinomial logit with heterogeneity. Accid. Anal. Prev. 211, 107877.
Yu, B., Chen, Y., Wang, R., Dong, Y., 2016. Safety reliability evaluation when vehicles turn right from urban major roads onto minor ones based on driver's visual perception. Accid. Anal. & Prev. 95, 487–494.
Yu, B., Chen, Y., Bao, S., Xu, D., 2018. Quantifying drivers' visual perception to analyze accident-prone locations on two-lane mountain highways. Accid. Anal. Prev. 119, 122–130.
Yu, B., Bao, S., Chen, Y., Chen, Y., 2019a. Using 3D mobile mapping to evaluate intersection design through drivers' visual perception. IEEE Access 7, 19222–19231.
Yu, B., Chen, Y., Bao, S., 2019b. Quantifying visual road environment to establish a speeding prediction model: an examination using naturalistic driving data. Accid. Anal. Prev. 129, 289–298.
Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T., 2020. Bdd100k: a diverse driving dataset for heterogeneous multitask learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2636–2645.
Yu, B., Gao, K., Cheng, Z., Chen, Y., Yue, L., 2024a. A human-like visual perception system for autonomous vehicles using a neuron-triggered hybrid unsupervised deep learning method. IEEE Transactions on Intelligent Transportation Systems 25 (7), 8171–8180.
Yu, B., Feng, X., Kong, Y., Chen, Y., Cheng, Z., Bao, S., 2024b. Using meta-learning to establish a highly transferable driving speed prediction model from the visual road environment. Eng. Appl. Artif. Intel. 130, 107727.
Zhang, K., Kan, D., Chen, D., 2024. Study on predicting the severity of traffic accidents in different grades of rural highways. J. Saf. Environ. 24 (4), 1515–1522.
Zhang, X., Zhong, M., Liu, S., Zheng, L., Chen, Y., 2019. Template-based 3D road modeling for generating large-scale virtual road network environment. ISPRS Int. J. Geo Inf. 8 (9), 364.
Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232.
Zürn, J., Gladkov, P., Dudas, S., Cotter, F., Toteva, S., Shotton, J., Simaiaki, V., Mohan, N., 2024. WayveScenes101: a dataset and benchmark for novel view synthesis in autonomous driving. arXiv preprint, arXiv:2407.08280.

Further Reading

Cantisani, G., Panesso, J.D.C., Del Serrone, G., Di Mascio, P., Gentile, G., Loprencipe, G., Moretti, L., 2022. Re-design of a road node with 7D BIM: geometrical, environmental and microsimulation approaches to implement a benefit-cost analysis between alternatives. Autom. Constr. 135, 104133.
Charlton, S.G., Mackie, H.W., Baas, P.H., Hay, K., Menezes, M., Dixon, C., 2010. Using endemic road features to create self-explaining roads and reduce vehicle speeds. Accid. Anal. Prev. 42 (6), 1989–1998.
Theeuwes, J., 2021. Self-explaining roads: What does visual cognition tell us about designing safer roads? Cognit. Res.: Princ. Implic. 6 (1), 15.