Journal of Southeast University (English Edition) Vol. 40, No. 1, pp. 1 - 12 Mar. 2024 ISSN 1003—7985

Identification of the spatiotemporal location of vehicle loads on highway bridges based on multi-view information fusion

Deng Lu[1,2] Deng Jiayu[1] Wang Wei[1] He Wei[1] Zhang Longwei[3]

([1]College of Civil Engineering, Hunan University, Changsha 410082, China)
([2]Hunan Provincial Key Laboratory for Damage Diagnosis of Engineering Structures, Hunan University, Changsha 410082, China)
([3]College of Civil Engineering, Hunan University of Science and Technology, Xiangtan 411201, China)

Abstract: To solve the problem that existing methods have difficulty in accurately obtaining the spatiotemporal distribution of vehicle loads on bridges in complicated traffic scenes, a spatiotemporal location identification method for vehicle loads based on multi-view information fusion is proposed. First, the vadYOLO-StrongSORT model is developed to detect and track vehicles simultaneously in a single view. Furthermore, based on image calibration and cross-view vehicle matching, an adaptive weighted least squares method is used for multi-view information fusion to correct the vehicle trajectory. Finally, the spatiotemporal distribution of axle loads is reconstructed by combining vehicle trajectories with axle configurations. The performance of the proposed method under typical traffic conditions is evaluated using model tests. The results show that the multi-view information fusion method significantly improves tracking stability, localization accuracy, and anti-occlusion performance compared with the single view-based vehicle location identification method. In the lane-changing scenes, the highest average localization error of the proposed method is less than 2.0 cm, which is significantly better than the 17.0 cm of the single-view method. In multivehicle occlusion scenes, the proposed method achieves a vehicle capture rate of up to 100%, compared with a maximum of only 72.5% for the single-view method. Meanwhile, vadYOLO-StrongSORT achieves the highest identification accuracy in the experiment compared with other detection and tracking models.

Key words: bridge engineering; vehicle loads; spatiotemporal location; multi-view information fusion; vehicle axle identification; bridge weigh-in-motion; bridge health monitoring

DOI: 10.3969/j.issn.1003 - 7985.2024.01.001

Vehicle loads are the major live loads that highway bridges encounter throughout their service life. The magnitude and spatiotemporal location of vehicle loads significantly impact the safety and durability of highway bridges[1-3]. Recently, the number of vehicles and the volume of vehicle loads have significantly increased. Therefore, the actual vehicle loads on bridges may differ from those expected during the design and construction periods. These significantly increased vehicle loads can pose a serious threat to bridge structural safety[4-5]. Therefore, it is necessary to accurately identify the spatiotemporal distribution of vehicle loads on highway bridges. Such identification is beneficial for comprehensively assessing in-service bridges and improving the design of new bridges.

The spatiotemporal distribution of vehicle loads, including information such as the magnitude and location of vehicle loads. The bridge weigh-in-motion (BWIM)[6] system is currently the mainstream method for obtaining this information, as it can effectively estimate the axle weight of vehicles directly based on the corresponding bridge responses. This estimation method is more flexible and has better durability and unbiased accuracy compared with traditional methods[7]. Currently, the BWIM system mainly uses strain sensors and other free-of-axle detector sensors installed under bridge decks to obtain vehicle information such as vehicle speed and axle configuration[8-9]. The vehicle is assumed to cross the bridge at a constant speed along a fixed lateral location of the bridge to obtain the axle weight of the vehicle. However, this assumption does not always conform to real situations, which can lead to erroneous calculation of the axle location and further produce errors in vehicle weight identification in real traffic scenes. To decrease identification errors due to speed variations, some scholars have proposed methods to correct vehicle speed and achieved better results in some simple traffic scenes[10-11]. However, these methods have similar limitations, i.e., they cannot accurately identify the axle weights in complicated traffic scenes, especially when the vehicle speed change, lane change, and multiple vehicles simultaneously cross over bridges. This limitation is that the specific contribution of each axle loading to the overall bridge response is difficult to determine. The reason for this is that the location of each axle at each moment during vehicle crossing over the

Received 2023-11-17, **Revised** 2024-01-02.

Biographies: Deng Lu (1984—), male, doctor, professor; He Wei (corresponding author), male, doctor, wei_he@hnu.edu.cn.

Foundation items: The National Natural Science Foundation of China Youth Program (No. 52108139), Hunan Provincial Natural Science Foundation Youth Program (No. 2023JJ40290).

Citation: Deng Lu, Deng Jiayu, Wang Wei, et al. Identification of the spatiotemporal location of vehicle loads on highway bridges based on multi-view information fusion[J]. Journal of Southeast University (English Edition), 2024, 40(1): 1 - 12. DOI: 10.3969/j.issn.1003 - 7985.2024.01.001.

bridge cannot be accurately and consecutively identified. In addition, these methods can only obtain the magnitude of the vehicle loads and not its location on the entire bridge. Therefore, it is necessary to conduct an in-depth study of this problem.

Previous studies[12-14] have proven that computer vision technology is efficient in identifying vehicle locations and axle configurations, which provide a new path for obtaining the spatiotemporal distribution of vehicle loads. Chen et al.[15] and Dan et al.[16] identified the spatiotemporal locations of vehicles from surveillance videos based on traditional vision methods such as background difference and template matching. However, these methods are sensitive to environmental conditions, resulting in poor performance in the spatiotemporal information acquisition of vehicles. In recent years, several deep learning-based computer vision methods have been proposed. Zhang et al.[17] used the faster region-based convolutional neural network (R-CNN) model to obtain vehicle trajectories and the number of axles. Xia et al.[18] used the YOLOv4 model to track vehicles on bridges and identify the gross vehicle weights and axle weights based on the single view-based method. On this basis, Zhao et al.[19] used binocular vision technology to track axle locations and obtained more accurate results. Yang et al.[20] initially realized the spatiotemporal distribution of vehicle loads on bridge decks using the YOLOv3 model combined with a pavement weigh-in-motion system. Xu et al.[21] and Dong et al.[22] also attempted to continuously track vehicles along the driving direction. However, the above methods may work well only in some simple traffic scenes because the identification accuracy of the spatiotemporal locations of the axles depends highly on the vehicle body tracking results and the completeness of the axle information. For more complicated traffic scenes, such as multivehicle crossing over bridges simultaneously, the tracking process is often unstable or even fails because of mutual occlusion of the vehicles. In addition, the tracking performance of vehicles from a single view is easily affected by the environment, which further affects the robustness of the entire identification system.

To solve these problems of tracking failure due to mutual occlusion of vehicles and low accuracy of vehicle location identification, a novel framework and system based on multi-view information fusion is proposed. First, a newly proposed vadYOLO-StrongSORT model is established to obtain vehicle trajectories and axle configurations for a single view. Then, a multi-view information fusion method based on the adaptive weighted least squares method is developed to further update the vehicle trajectories based on image calibration and cross-view vehicle matching. Finally, the spatiotemporal distribution of axle loads is reconstructed by combining vehicle trajectories with axle configurations. The performance of the proposed method is verified using model tests.

1 Theory and Method

1.1 Detection and tracking model for the vehicle and axle

1.1.1 Detection model for the vehicle and axle

YOLOv7[23] is an emerging object detection model with superior detection speed and accuracy compared with other single-stage object detection models. The detection process is shown in Fig. 1. First, the backbone network extracts feature information from the input image through a series of key modules. Among them, the Conv + BN + SiLU (CBS) module enhances the number of features learned by the backbone extraction network by stacking and integrating feature layers and the concat operation. The efficient layer aggregation network (ELAN) module integrates the results of the CBS layers to enhance the various levels of feature learning. In addition, max pooling 1 (MP1) downsamples the input feature maps to support the extraction of higher-level features in subsequent network layers. Effective fusion of low-resolution features with high-resolution features is realized by the upsample (UP) operation on the neck network. The spatial pyramid pooling and channel spatial pyramid convolution (SPPCSPC) module performs feature extraction through a path aggregation feature pyramid network and a parallel Maxpool layer to avoid image distortion. Subsequently, the target bounding boxes and categories of each grid are predicted simultaneously in the head network. Finally, the prediction result containing the location and the probability of the object is the output. However, when applied to complicated traffic scenes, this model is not efficient because unstable detection of vehicles and missing detection of axles often occur. Herein, an improved vadYOLO object detection model is proposed, and its architecture is shown in Fig. 1.

First, to improve the ability of the model to identify vehicles and axles in complicated traffic scenes, the squeeze-and-excitation (SE)[24] module is embedded between the backbone and neck networks of the YOLOv7 model. The SE module is an attentional mechanism that improves the performance of convolutional neural networks (CNNs) by adaptively learning the importance of feature mappings. The acting process of the SE module is illustrated in Fig. 1. Before applying the SE module (indicated by the blue cuboid), each channel of the feature map has the same weight. After passing through the SE module (marked by a multicolored cuboid with each color representing a different weight), the weight of each feature channel becomes different. This step enables the neural network to focus on the vehicles and axles that should be detected.

Generally, the imaging sizes of axles are smaller than those of vehicle bodies, whereas the downsampling factor

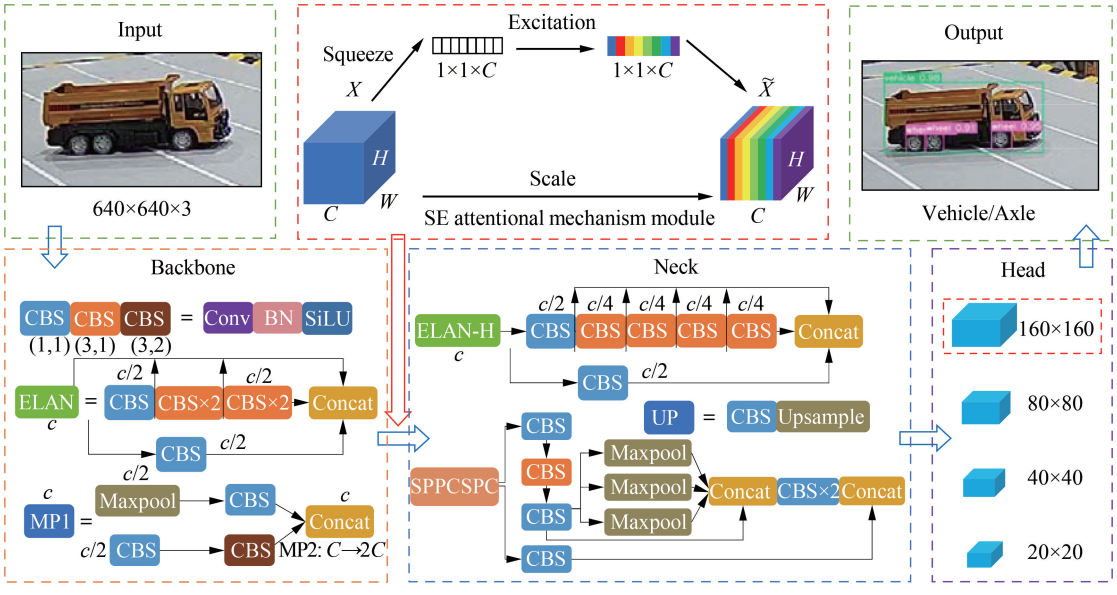


Fig. 1 Network structure of the proposed vadYOLO-based detection model

of the YOLOv7 model is larger. This difference makes it challenging to learn the feature information of the axles for the deep feature map. Therefore, a four-fold downsampling layer is added to the YOLOv7 model to enhance the receptive fields, which makes the YOLOv7 model more sensitive to axles.

Furthermore, the imaging sizes of vehicles increase as they move closer to the camera and decrease as they move away from the camera. The YOLOv7 model considers the complete intersection over union (CIoU[25]) as the loss function of location regression, in which the aspect ratio describes the relative value. However, a certain degree of ambiguity exists in this process. When vehicles are away from the camera, the differences in aspect ratios between these vehicles can become too unapparent to distinguish these objects for the CIoU loss function. Note that the efficient intersection over union (EIoU[26]) loss function considers these issues. The number of high-quality anchor boxes (vehicles) with small regression errors is less than that of low-quality anchor boxes (axles) with large regression errors in a single image. Low-quality anchor boxes usually produce excessive gradients that affect the training performance; thus, the EIoU loss function cannot directly work well. Therefore, the CIoU loss function of YOLOv7 is replaced by the Focal-EIoU[26] loss function.

1.1.2 Tracking model for the vehicle

StrongSORT[27] is a tracking model with outstanding performance. In this model, the method used to track the object is based on its appearance features, and Kalman filtering is used to predict the motion of the object to reduce confusion and false associations due to the similarity of the features, thus tracking the object more clearly. This model was selected for tracking vehicles in this study, and the specific tracking process is described in Ref. [27]. First, the original frames of the video are obtained, and the vehicles are detected through the vadYOLO network. Then, after using the enhanced correlation coefficient method for camera motion compensation, the noise scale adaptive Kalman filter is used to obtain the motion features of vehicles. A bottleneck feature extractor that uses ResNetSt50 as the main network is used to obtain the appearance features of the vehicle. Additionally, an exponential moving average feature update strategy is used to enhance the feature-matching performance. After calculating the cost and gate matrices, the pixel trajectories of the vehicles are obtained using Vanilla global linear assignment matching.

1.2 Coordinate transformation from image space to object space

To transfer pixel trajectories to physical trajectories, a coordinate mapping model of the image space and object space (see Fig. 2) must be established. Fig. 2(a) shows the camera imaging scene, where u-o-v is the pixel coordinate system, x_1-O_1-y_1 is the image coordinate system, $x_2y_2z_2$-O_2 is the camera coordinate system, and XYZ-O_3 is the world coordinate system. Set the point (X_1, Y_1, Z_1) in the world coordinate system to be mapped to the point (u, v) in the pixel coordinate system. Based on the pinhole camera model (see Fig. 2(b)), the perspective projection can be obtained by

$$Z_c\begin{bmatrix} u & v & 1 \end{bmatrix}^{\mathrm{T}} = \boldsymbol{KP}\begin{bmatrix} X_1 & Y_1 & Z_1 & 1 \end{bmatrix}^{\mathrm{T}} \quad (1)$$

where Z_c is the projection of point T in the world coordinate system in the z_2 direction under the camera coordinate system $x_2y_2z_2$-O_2; $\boldsymbol{K} = \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$ is the internal reference of the camera; $\boldsymbol{P} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{T} \\ \boldsymbol{O}_{1\times 3} & 1 \end{bmatrix}$ is the external

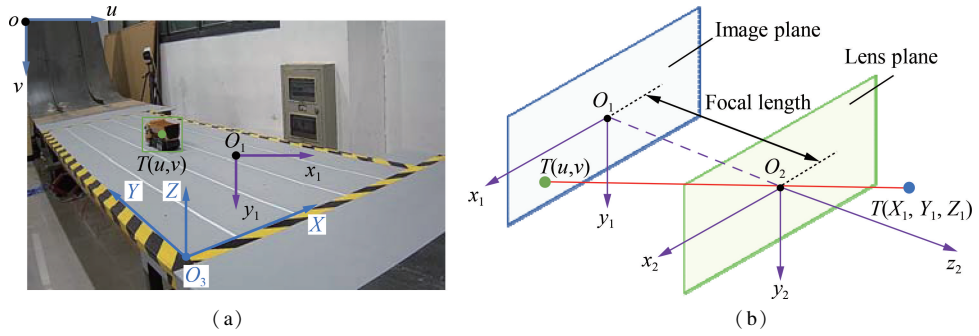


Fig. 2 Schematic of the coordinate transformation from image space to object space. (a) Camera image scene; (b) Pinhole camera model

reference of the camera, $\boldsymbol{R}$ denotes the rotation matrix, which is used to describe the rotational transformations of an object in space, and $\boldsymbol{T}$ denotes the translation matrix, which is used to describe the translational transformations of an object in space.

If the bridge deck is a plane and the world coordinate system is set on this plane, the coordinate of the point (X_1, Y_1) on the bridge deck can be found by

$$[X_1 \quad Y_1 \quad 1]^{\mathrm{T}} = \boldsymbol{H}_{3\times3}^{-1} Z_c [u \quad v \quad 1]^{\mathrm{T}} \tag{2}$$

where $\boldsymbol{H}_{3\times3}$ is the mapping matrix from the perspective transformation of the object space to the image space. The specific solution can be found in Refs. [18,28].

1.3 Cross-view vehicle matching

By matching vehicles in multiple views, the system can simultaneously track the vehicles throughout the scene rather than independently in each view. As a result, for the multi-view information fusion method, matching vehicles in the cross-view is required. This means matching the pixel trajectories of the same vehicle in multiple views as it crosses over the bridge. The specific matching steps are as follows:

1) Matching preparation. First, temporal synchronization of the multi-view system is performed, and mapping models from pixel coordinate systems in various views to the uniform world coordinate system are established. Then, the pixel coordinates of the center points and images are extracted for each frame of the vehicle tracking boxes in each camera. Finally, the pixel coordinates are transformed to the world coordinates, and the images are transformed to the appearance feature matrices. The world coordinates correspond to the appearance feature matrices in the order of the video frames.

2) Matching database establishment. If vehicles are tracked in the overlapping area (bridge deck) of N cameras, each camera assigns an identity (ID) number to the tracked vehicle. The vehicle of the i-th ID number in the n-th camera is denoted as $O_{C_n\mathrm{ID}i}$. The sets of world coordinates and appearance feature matrices of each frame for the tracked vehicles in each camera were constructed. For example, the set $C_1 = \{O_{\mathrm{C1ID1}}, O_{\mathrm{C1ID2}}, \cdots, O_{\mathrm{C1ID}M}\}$ denotes all vehicles tracked using camera 1, and M is the total number of vehicles tracked using camera 1.

3) Overall feature similarity calculation. The vehicle O_{C1ID1} in camera 1 is selected as the object to be matched. The Euclidean distance of the world coordinates between it and the vehicle O_{C2ID1} in the corresponding video frame in camera 2 is computed, and a threshold is set. Additionally, the similarity between the vehicle appearance feature matrices of the corresponding video frames is calculated, and a threshold is set. The calculated results of dividing the number of video frames within the threshold by the total number of frames of the corresponding video frames are counted as the spatiotemporal feature similarity and the appearance feature similarity, respectively. The overall feature similarity can be calculated by $aS_1 + (1-a)S_2$, in which a is the weighting factor, S_1 is the appearance feature similarity, S_2 is the spatiotemporal feature similarity. The above steps are repeated to calculate the overall feature similarity between the vehicle O_{C1ID1} and all vehicles tracked using camera 2.

4) Matching vehicles. Set an overall feature similarity threshold and eliminate the ID number exceeding the threshold in the set of C_2, and the set of C_2 is queried. If only one vehicle is left in the set, the matching of the vehicle under camera 2 is completed. If more than one vehicle is left in the set, the vehicle with the largest similarity is considered the matching object. If there is no vehicle left in the set, the vehicle that should be matched in camera 2 is occluded during the entire tracking process. The above steps are repeated to sequentially match the vehicle in other cameras.

1.4 Multi-view information fusion

Although single view-based methods can track and locate vehicles, relying on only one camera in a real bridge scene is more susceptible to environmental factors such as light variations and camera vibrations, which can affect the accuracy and stability of localization. In contrast, the proposed method uses multiple cameras for fusion localization. Although one camera is affected by light variations, the other cameras can still provide relatively stable information. Therefore, the proposed method exhibits higher robustness in response to lighting variations. To obtain accurate and stable vehicle trajectories, a multi-view information fusion method based on the adaptive weighted least squares method is developed.

Note that spatiotemporal synchronization is a prerequisite for achieving multi-view information fusion. Spatiotemporal synchronization ensures that images captured by various cameras are consistent in space and time. In this study, the four cameras are connected by a videocassette recorder that contains dedicated hardware synchronization circuitry, which ensures that all connected cameras capture according to the same clock signal, thus reducing delays at the hardware level. Spatial synchronization can be achieved by converting the image coordinate systems from four views into a uniform world coordinate system.

After the multi-view system is spatiotemporally synchronized, for each camera, the relational equation of the transformation from image space to object space can be obtained as follows:

$$\begin{bmatrix} h_{11}-h_{31}u_i & h_{12}-h_{32}u_i \\ h_{21}-h_{31}v_i & h_{22}-h_{32}v_i \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}=\begin{bmatrix} h_{33}u_i-h_{13} \\ h_{33}v_i-h_{23} \end{bmatrix} \tag{3}$$

where $h_{ij}(i,j=1,2,3)$ is the parameter of the mapping matrix; (X_1,X_2) is the world coordinate of the vehicle for the i-th frame; (u_i,v_i) is the pixel coordinate of the vehicle for the i-th frame.

When multiple cameras track the same vehicle, to obtain the relational equation of the transformation for the multi-view system, Eq. (3) can be extended to $\boldsymbol{AX}=\boldsymbol{b}$, where

$$\boldsymbol{A}=\begin{bmatrix} h_{11}^{(1)}-h_{31}^{(1)}u_i^{(1)} & h_{12}^{(1)}-h_{32}^{(1)}u_i^{(1)} \\ h_{21}^{(1)}-h_{31}^{(1)}v_i^{(1)} & h_{22}^{(1)}-h_{32}^{(1)}v_i^{(1)} \\ \vdots & \vdots \\ h_{11}^{(N)}-h_{31}^{(N)}u_i^{(N)} & h_{12}^{(N)}-h_{32}^{(N)}u_i^{(N)} \\ h_{21}^{(N)}-h_{31}^{(N)}v_i^{(N)} & h_{22}^{(N)}-h_{32}^{(N)}v_i^{(N)} \end{bmatrix}$$

$$\boldsymbol{b}=\begin{bmatrix} h_{33}^{(1)}u_i^{(1)}-h_{13}^{(1)} \\ h_{33}^{(1)}v_i^{(1)}-h_{23}^{(1)} \\ \vdots \\ h_{33}^{(N)}u_i^{(N)}-h_{13}^{(N)} \\ h_{33}^{(N)}v_i^{(N)}-h_{23}^{(N)} \end{bmatrix},\quad \boldsymbol{X}=\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \tag{4}$$

where $h_{ij}^{(n)}(n=1,2,\cdots,N)$ is the parameter of the mapping matrix for the n-th camera; $(u_i^{(n)},v_i^{(n)})$ is the pixel coordinate of the i-th frame for the n-th camera.

$\boldsymbol{AX}=\boldsymbol{b}$ is an overdetermined equation, and an adaptive matrix of weights is developed for the weighted least squares solution.

$$\boldsymbol{X}_{\mathrm{WLS}}=(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_i\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_i\boldsymbol{b} \tag{5}$$

$$\boldsymbol{W}_i=\mathrm{diag}(\boldsymbol{w}_1^i,\boldsymbol{w}_2^i,\cdots,\boldsymbol{w}_N^i) \tag{6}$$

where $\boldsymbol{W}_i$ is the weight matrix of the i-th frame in the multi-view system; $\boldsymbol{w}_n^i=\mathrm{diag}(w_{n,1}^i,w_{n,2}^i)$ is the weight matrix of the i-th frame for the n-th camera; $w_{n,j}^i(j=1,2)$ is calculated by

$$w_{n,j}^i=\begin{cases} 1 & |R_{n,j}^i|<R_0 \quad (7a) \\ \dfrac{R_0}{|R_{n,j}^i|} & |R_{n,j}^i|\geqslant R_0 \quad (7b) \end{cases}$$

where R_0 is a reasonable value of the residual set, which is determined as the average value of the residual values $R_{n,j}^i$ computed by selecting several frames of the tracked video. If the measurement point exceeds this value, the weight can be calculated according to Eq. (7b); $R_{n,j}^i$ is the value of the j-th element in the residual matrix $\boldsymbol{R}_n^i$ of the i-th frame for the n-th camera, and $\boldsymbol{R}_n^i$ is calculated by

$$\boldsymbol{R}_n^i=\boldsymbol{A}_n\boldsymbol{X}_{\mathrm{OLS}}-\boldsymbol{b}_n \tag{8}$$

where $\boldsymbol{A}_n$ and $\boldsymbol{b}_n$ are the mapping matrices for the n-th camera; $\boldsymbol{X}_{\mathrm{OLS}}$ is the solution of ordinary least squares.

2 Traffic Model Test

Traffic model tests were conducted in an indoor laboratory to verify the accuracy and stability of the proposed method. The entire indoor test scene is shown in Fig. 3. The monitoring area (bridge deck) is 2.7 m long and 0.9 m wide and comprises five traffic lanes. Four cameras with a 1 920 × 1 080 pixel resolution and a frame rate of 25 frame/s were used in the tests. Three types of vehicle models were considered, including a 2-axle car, two 3-axle trucks, and a 5-axle trailer. Thirty reference control points were selected on the bridge deck, as indicated by the red points in Fig. 3, to help solve the transformation matrix (Eq. (2)). Note that these cameras were placed in randomized erection positions and capture angles. Just ensure that the cameras can fully cover the area to be monitored. For actual scenes, camera positions and capture angles may need to be adjusted according to the actual situation.

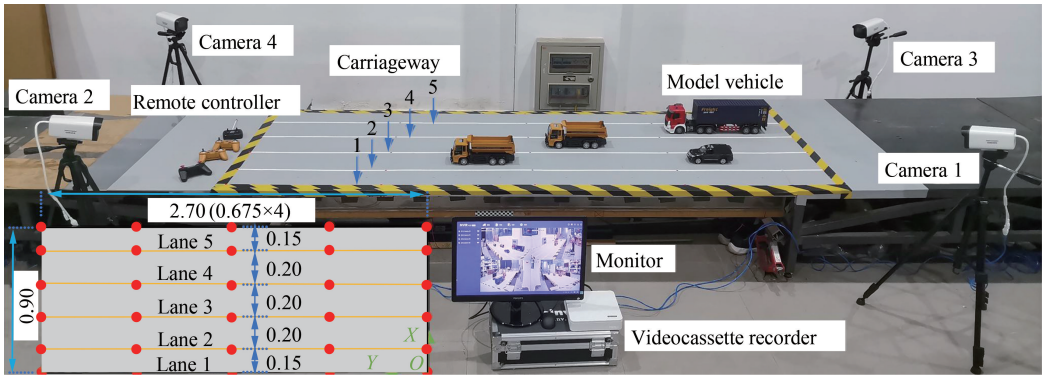


Fig. 3 Indoor laboratory test scene (unit: m)

The entire test was divided into two parts: one conducted in simulated random traffic scenes and the other in artificially set traffic scenes. The detection, tracking, and anti-occlusion performance of the proposed method, as well as its accuracy in identifying axle configurations, were verified through tests conducted in a simulated random traffic scene using the adapted remote manipulation. The localization accuracy of the proposed method was verified using artificial placement to set the precise vehicle location.

3 Result Analysis

3.1 Verification of object detection and tracking performance

Video frames of traffic flows were extracted from four cameras to establish an image dataset. In this test, 697 images of the original dataset were expanded to 3 881 images by performing a series of data enhancement operations such as random scaling, rotation, horizontal flipping, cropping, and brightness adjustment to form a diverse training sample to improve the generalization performance of the model. Finally, we divided the training, validation, and test sets in a ratio of 7∶2∶1 to ensure that the model fully learns and adapts to the input images under various changing conditions. The input image size in the training network was 640 × 640 pixels, and the batch size was eight. As shown in Fig. 4, after 1 000 epochs, the model converged, the loss was reduced to 0.015, and mAP@0.5 reached 98.2%.

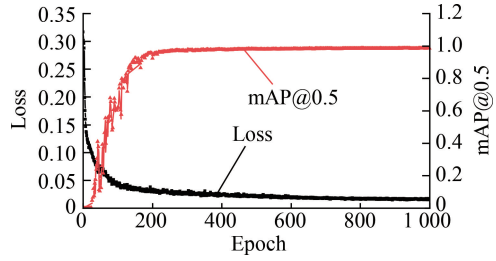
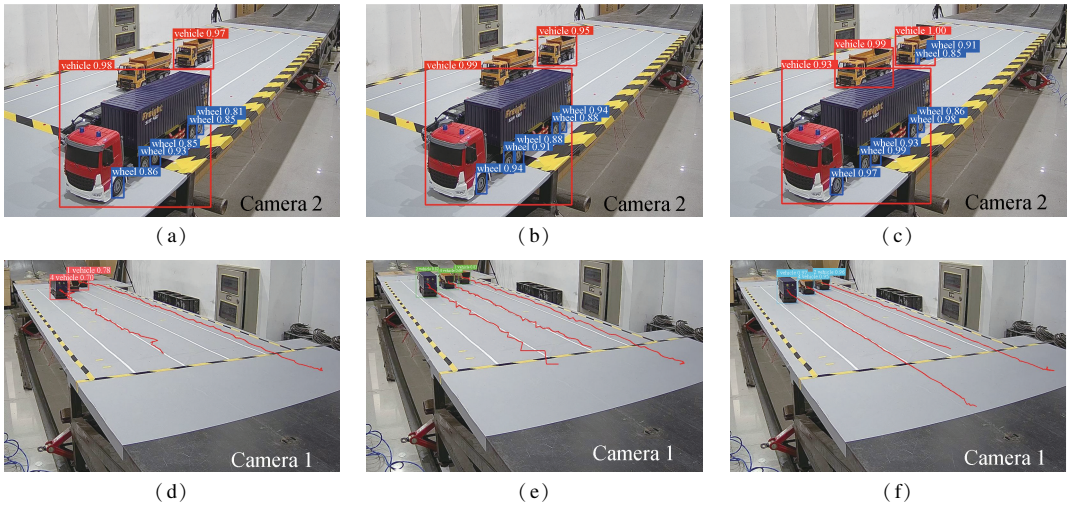


Fig. 4 Training process

To verify the detection and tracking performance of the proposed model (combined with the StrongSORT algorithm), a comparative validation was performed with YOLOv5 and YOLOv7 based on the traffic flow videos. The detection results are presented in Figs. 5(a)-(c). The original YOLO model failed to detect the axles farther away from the camera as well as the partially occluded vehicles, whereas the proposed vadYOLO model can not only accurately detect the vehicle information but also has the highest confidence level of the overall identification results among the three models. Figs. 5(d)-(f) show the tracking results of the three models in the same scene. The YOLOv5-StrongSORT model experienced the problem of missing track when tracking the vehicle. Although the YOLOv7-StrongSORT model tracked the vehicle better, the trajectories showed some fluctuations. In comparison, the proposed method successfully tracked all trajectories with stable performance under all working conditions. Two trucks with the same appearance were successfully tracked, indicating that the proposed model is effective in tracking vehicles with similar appearance features.



(a) (b) (c)

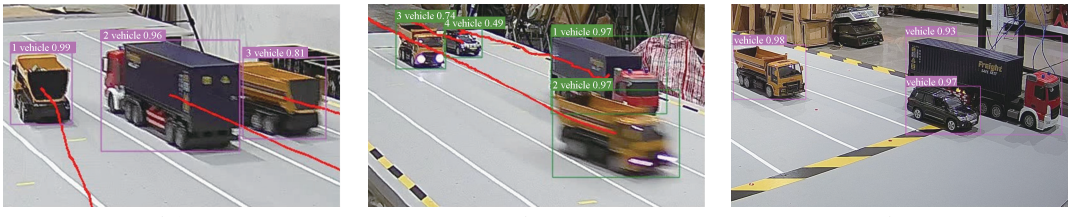
(d) (e) (f)

Fig. 5 Various model detection and tracking results. (a) YOLOv5 detection results; (b) YOLOv7 detection results; (c) vadYOLO detection results; (d) YOLOv5-StrongSORT tracking results; (e) YOLOv7-StrongSORT tracking results; (f) Proposed model tracking results

To further verify the performance of the proposed method for vehicle detection and tracking, Fig. 6 shows the identification results in some special scenes. Fig. 6(a) shows the identification result in the presence of the vehicle shadows at various angles. Fig. 6(b) shows the identification result when the vehicles were blurred because of a higher speed. Fig. 6(c) shows the identification result with the light variation caused by the shadow of the large-volume vehicle. The results show that the vehicles can be successfully detected and tracked using the proposed method in the special scenes of this test.

Comparison results regarding the quantitative perform-

ance of the three models are presented in Table 1. The detection accuracy is characterized by model average precision. Tracking accuracy is characterized by the ratio of the number of correct frames (frames other than false positives, missing tracking, and abnormal identity-switching situations) to the total number of video frames. Efficiency is quantified by the time consumed for per-frame detection or tracking. As shown in Table 1, when the adopted model was changed from YOLOv5 to YOLOv7 and the proposed vadYOLO, both the detection and tracking accuracy of the objects and the time consumed increased. Among the three models, YOLOv5 had the best efficiency but was also the least accurate model. In comparison, the proposed vadYOLO model was tested with the highest detection and tracking accuracy, especially in the detection of axles, although the efficiency slightly decreased. Specifically, the proposed method achieves the average accuracy of 98.5%, 97.9%, 97.1% for vehicle body detection, small-size vehicle axle detection and vehicle tracking with the 1.2%, 6.2%, 2.6% improvement, respectively.



(a) (b) (c)

Fig. 6 Identification results in special scenes. (a) Vehicle shadow; (b) Vehicle blur; (c) Light changes

Table 1 Comparison of the detection and tracking performance of various models

Model	Vehicle detection accuracy/%	Axle detection accuracy/%	Detection efficiency/(ms · frame^{-1})	Tracking accuracy/%	Tracking efficiency/(ms · frame^{-1})
YOLOv5-StrongSORT	96.8	87.7	15.9	93.7	20.5
YOLOv7-StrongSORT	97.3	91.7	16.2	94.5	21.3
vadYOLO-StrongSORT	98.5	97.9	16.9	97.1	21.6

3.2 Verification of the localization accuracy in occlusion-free scenes

To verify the localization accuracy of the proposed method in occlusion-free scenes, single-vehicle tests were conducted under both the straight-line driving and lane-changing driving scenes. The vehicle was artificially placed in a specified location so that its real location could be obtained to verify the localization results. Under the straight-line driving scenes, 10 detection points were set in Lanes 1, 3, and 5. Under lane-changing driving scenes, 31 detection points are set in a sinusoidal path to further verify the localization performance of the proposed method in the case of vehicle turning (change of vehicle direction).

Fig. 7 shows the typical identification results of the vehicle trajectories under straight-line driving conditions, including the tracking results of the vehicle trajectories obtained by the single-view method and the multi-view information fusion method, as well as the artificial setting trajectory. Localization based on the single view-based method is prone to an overall offset error. In contrast, the trajectory identification result based on the proposed method matches well with the set locations. For better comparison, the statistical results of the localization errors of vehicle trajectories before and after fusing the multi-view information under the same scene are displayed in Fig. 8(a), where the bars represent the mean error of all frames, the error bars represent the standard deviation. The offset error is significantly reduced by fusing the multi-view information.

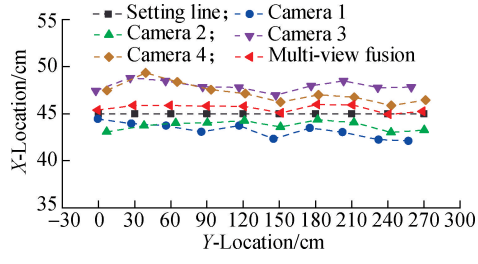


Fig. 7 Comparison of the vehicle trajectories obtained by the single-view and the proposed methods under straight-line driving conditions

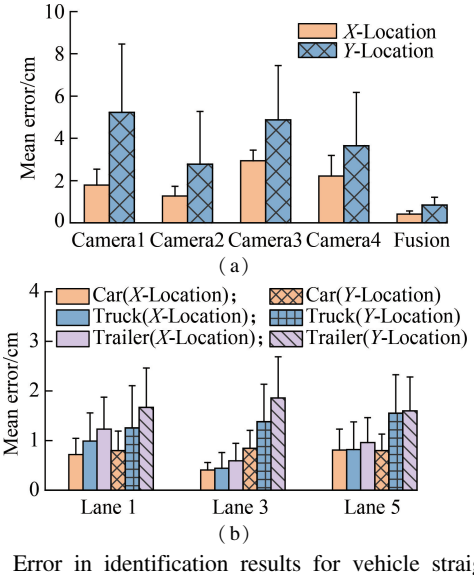


Fig. 8 Error in identification results for vehicle straight-line driving. (a) Localization errors before and after fusion; (b) Localization error for various vehicle types and driving lanes

Fig. 8(b) shows the statistical results of the localization errors after fusing the multi-view information for all straight-line driving conditions considered in the study. Various vehicle types (car, truck, and trailer), driving lanes (Lanes 1, 3, and 5), and two directions (*X*-Location and *Y*-Location) were considered. The localization error under all conditions is less than 2.0 cm. This indicates that the proposed method can lead to high localization accuracy and better tracking stability for various vehicle types, lanes, and directions. In addition, the localization errors (after fusion) of various vehicles driving in the *X*-direction fluctuate less, and the error values are also lesser than those in the *Y*-direction. This phenomenon arises because the object-space location change in the *Y*-direction in the camera field corresponds to a smaller projected distance. It was also found from Fig. 8(b) that a longer vehicle generally corresponds to a larger localization error. This may be because the center of the outer rectangle of the vehicle (the detection box) is the vehicle location in the tracking algorithm. Thus, a larger aspect ratio of the vehicle in the camera field usually results in larger localization errors.

Fig. 9 shows the typical identification results of the vehicle trajectories under lane-changing driving scenes. An overall offset similar to that in straight-line driving scenes is also observed. The proposed multi-view information fusion method can improve localization accuracy. Furthermore, this overall offset in lane-changing driving conditions is smaller in the peaks and troughs of the sinusoidal curve path than at other locations. This is because the vehicle direction in this position is parallel to the lane, and the center of the vehicle detection box is closer to the setting location.

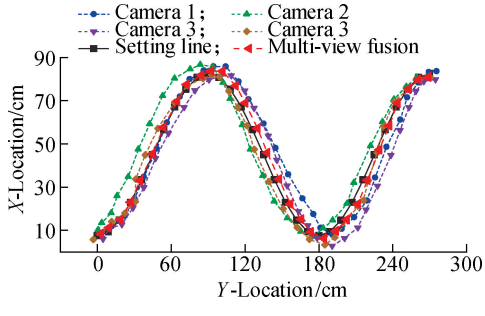


Fig. 9 Comparison of the vehicle trajectories obtained by the single-view and the proposed methods under lane-changing driving conditions

For further comparison, Fig. 10 shows the statistical results of the localization errors before and after fusing the multi-view information for various vehicle types under lane-changing driving conditions. The bars represent the mean error of all frames, and the error bars represent the standard deviation. As shown in Fig. 10, for various vehicle types and localization directions, the multi-view information fusion method can improve localization accuracy. In particular, for larger trailers, the maximum of the mean localization error in the *X*-direction was reduced from 6.0 to 1.5 cm, achieving a reduction of over 75%; the maximum of the mean localization error in the *Y*-direction was reduced from 17.0 to 2.0 cm, realizing a reduction of almost 90%.

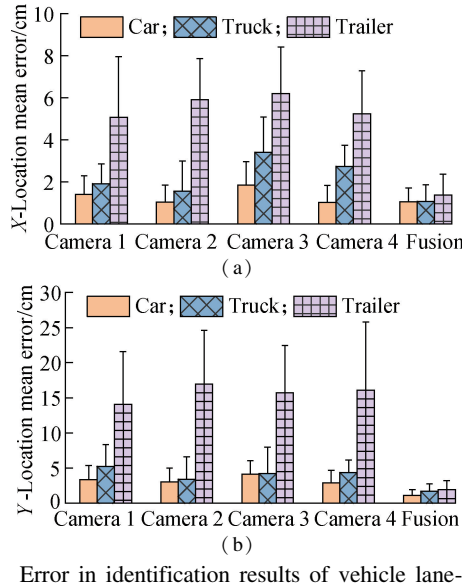


Fig. 10 Error in identification results of vehicle lane-changing driving before and after fusion. (a) *X*-Location errors based on single-view and multi-view fusion; (b) *Y*-Location errors based on single-view and multi-view fusion

To more comprehensively assess the effectiveness of the proposed method compared with the traditional single-view method, Table 2 demonstrates the relative error values of the localization results of the three-vehicle types with relation to the bridge width before and after the fusion of localization based on the four views under two working conditions. The overall relative error in the lane-changing driving scene was larger than that in the straight-line driving scene, and the error increased accordingly as the vehicle size increased. Further observation of the table shows that the relative error of localization based on the single-view approach reached 18.88%. In contrast, with the proposed multi-view information fusion localization method, the relative errors of the three-vehicle types under the test conditions remained below 2.15%.

3.3 Verification of the anti-occlusion performance in multivehicle occlusion scenes

To verify the anti-occlusion performance of the proposed method, the common scenes of multiple vehicles simultaneously driving were considered, as shown in Fig. 11. The test vehicle was operated by a remote control, and the vehicle was not controlled to travel in a straight line due to possible control errors. Fig. 12 shows the tracking results of the vehicle trajectories for cameras 1 and 4. When using a single camera for vehicle tracking, the phenomenon of vehicles occluding each other may lead to the miss of tracking objects, making it impossible to track the complete trajectories of vehicles.

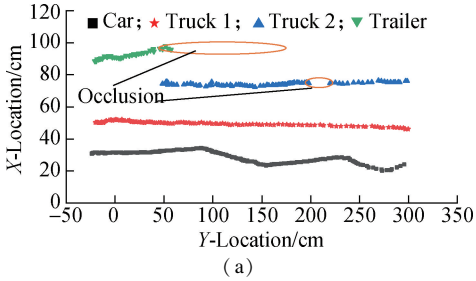
Table 2 Relative error of localization in occlusion-free scenes %

Working condition	Vehicle type	Camera 1		Camera 2		Camera 3		Camera 4		Fusion	
		X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
Straight-line driving	Car	1.99	5.81	1.41	3.09	3.26	5.42	2.46	4.05	0.66	0.94
	Truck	2.02	5.32	1.51	3.96	3.36	4.52	2.86	4.58	0.83	1.55
	Trailer	5.44	9.42	4.36	10.39	5.69	10.04	4.69	10.46	1.03	1.89
Lane-changing driving	Car	1.56	3.74	1.15	3.41	2.06	4.60	1.14	3.24	1.18	1.26
	Truck	2.13	5.83	1.73	3.80	3.79	4.67	3.03	4.88	1.19	1.87
	Trailer	5.64	15.66	6.57	18.88	6.90	17.45	5.82	17.87	1.53	2.14

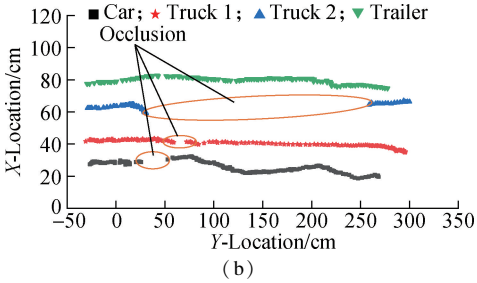
(a)

(b)

Fig. 11 Multivehicle occlusion scenes. (a) Camera 1 scenes; (b) Camera 3 scenes



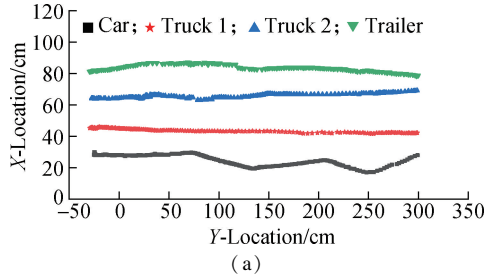
(a)



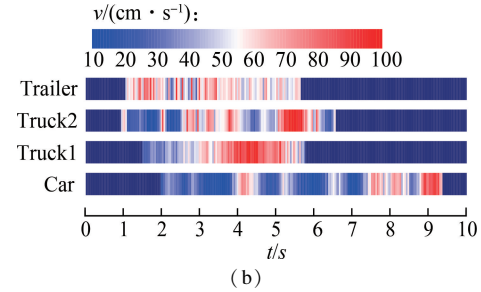
(b)

Fig. 12 Identification results of vehicle trajectories based on the single view-based method. (a) Camera 1 tracking trajectories; (b) Camera 4 tracking trajectories

After fusing the four-view information, the complete vehicle trajectories can be identified, as shown in Fig. 13 (a). Fig. 13 (b) shows the results of the vehicle speed distribution when the first vehicle entered the monitoring area and the last vehicle left the area. The vehicle speeds were mainly distributed between 10 and 100 cm/s. This shows that the proposed method can track vehicles relatively well for vehicles driving at various speeds under the test conditions. Note that there are some sudden changes in vehicle speed. As the vehicles were controlled by manually manipulating the remote control, the manipulator performed intermittent control to keep them moving along the set lanes.



(a)



(b)

Fig. 13 Identification results of vehicle trajectories in multivehicle occlusion scenes. (a) Vehicle trajectories for fusing four-view information; (b) Vehicle speeds for fusing four-view information

To further quantitatively analyze the anti-occlusion performance of the proposed method, various occlusion scenes were also simulated, including the two-vehicle occlusion, three-vehicle occlusion, and four-vehicle occlusion scenes. Ten randomized trials were conducted for each scene. Table 3 shows the statistical results for the object capture rate of the proposed method. Table 3 shows that the more complicated the occlusion situation, the lower the single-view object capture rate. Even for simpler two-vehicle occlusion scenes, the vehicle capture rate is only 97.6%. In complicated traffic scenes with four vehicles, the capture rate decreases dramatically to 62.5%, with the highest capture rate being only 72.5%. In contrast, the proposed method maintains a target capture rate of 100% for all types of occlusion scenes. Hence, the proposed method can effectively improve the anti-occlusion performance of vehicle tracking in complicated traffic scenes.

Table 3 Object capture rate in various occlusion scenes %

Number of vehicles	Camera 1	Camera 2	Camera 3	Camera 4	Fusion
2	95.3	93.8	91.5	97.6	100
3	85.7	77.5	89.5	90.3	100
4	62.5	70.0	72.5	67.5	100

3.4 Results of axle identification and axle spatiotemporal distribution

Table 4 summarizes the relative errors in axle spacing identification under various working conditions. Among them, the identification results of the number of axles for various types of vehicles under each working condition are correct, and the identification accuracy reaches 100%. The relative error of axle spacing identification is kept within 5.0% overall. As the number of vehicles on the bridge increases, the relative errors also increase. Even so, the maximum error is still controlled within 5.23%, which achieves an accuracy close to that of direct measurement by road sensors and can fulfill the accuracy requirements of techniques such as BWIM[29].

Table 4 Relative errors in axle spacing identification under various traffic conditions %

Number of vehicles	Car	Truck		Trailer			
	S_1 = 12 cm	S_1 = 14 cm	S_2 = 4 cm	S_1 = 11 cm	S_2 = 5 cm	S_3 = 19.5 cm	S_4 = 5 cm
1	2.08	0.75	1.00	3.50	1.00	0.50	0.60
2	1.92	2.75	1.88	1.36	2.22	2.46	1.80
3	1.00	3.17	1.63	3.36	2.80	−2.92	2.40
4	0.88	3.85	2.63	5.23	4.10	3.53	3.50

Note: S_1 represents the distance from the first axle to the second axle of the vehicle; the real value of the axle spacing is shown in parentheses.

Fig. 14 shows the spatiotemporal distribution of axles on the bridge deck at the 60th, 90th, and 120th frames based on multi-view information fusion. As shown in Fig. 14, the axles of the four types of vehicles on the bridge are continuously tracked and localized, and the axle spacings and the number of axles for the vehicles can be clearly represented. The results show that the proposed method performs well in identifying vehicle trajectories and axle configurations. The localization method for axles based on matching the vehicle trajectory and the axle configurations can effectively compensate for the lack of location information when the axles are occluded and obtain the spatiotemporal distribution of all the axles at each moment.

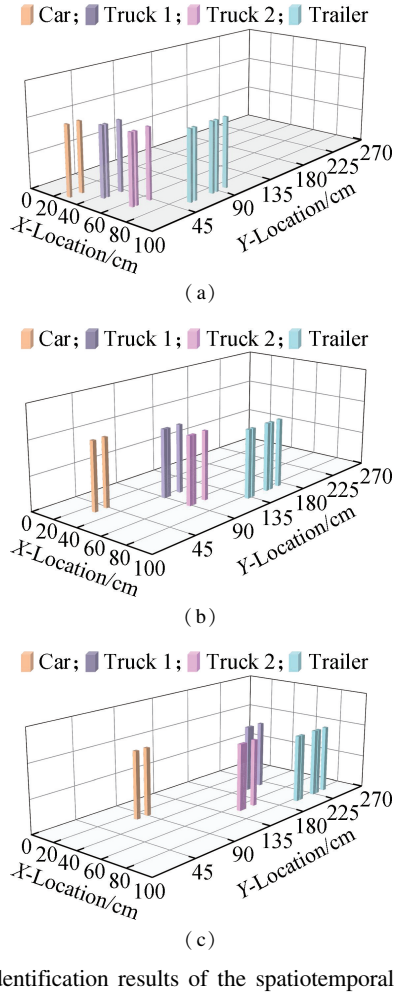


Fig. 14 Identification results of the spatiotemporal distribution for axle loads. (a) Frame 60 result; (b) Frame 90 result; (c) Frame 120 result

4 Conclusions

1) An effective object detection and tracking model vadYOLO-StrongSORT was proposed to simultaneously detect and track both vehicle bodies and axles, which showed better stability and higher accuracy than traditional object tracking models under complicated traffic scenes. Specifically, compared with traditional models, the average accuracy of vehicle body detection, small-size vehicle axle detection, and vehicle tracking of the proposed model increased by 1.2%, 6.2%, and 2.6%, reaching 98.5%, 97.9%, and 97.1%, respectively.

2) The proposed multi-view information fusion method for locating vehicles can significantly reduce the overall offset error compared with single view-based methods. Laboratory tests show that for various vehicle types under straight-line and lane-changing driving scenes, the highest average localization error of the proposed method can be kept within 2.0 cm, whereas that of single view-based methods can reach up to 17.0 cm. Furthermore, the relative error of vehicle localization based on multi-view fusion was reduced from 18.88% in the single-view method to less than 2.15%.

3) The proposed multi-view information fusion method can greatly improve anti-occlusion performance in vehicle

tracking. Random traffic tests show that for the single view-based methods, the capture rate of the vehicle is only 97.6%, even for a simple traffic scene in the presence of two vehicles, and the capture rate drops sharply to 62.5% under a complicated traffic scene in the presence of four vehicles. In contrast, the object capture rate of the proposed method under various occlusion conditions can be maintained at 100%.

4) The proposed vadYOLO method can effectively identify the number and spacing of vehicle axles. The test results show that the proposed method achieves an identification accuracy comparable to that of direct measurement through road sensors. The relative identification errors of vehicle axle spacings were maintained within 5.0% overall, with a maximum error is 5.23%. These results fulfill the requirements of the commercial BWIM system.

References

[1] Zhao H W, Ding Y L, Li A Q, et al. Digital modeling of vehicle load-bridge effect and system condition monitoring[J]. *Journal of Southeast University* (*Natural Science Edition*), 2022, **52**(2): 203 - 211. DOI: 10.3969/j. issn. 1001-0505. 2022. 02. 001. (in Chinese)

[2] Zheng J Y, Tang J Y, Zhou Z X, et al. Intelligent cognition of traffic loads on road bridges: From measurement to simulation—A review[J]. *Measurement*, 2022, **200**: 111636. DOI: 10. 1016/j. measurement. 2022. 111636.

[3] Zhao K L, Zong H, Zhu Q X, et al. Analysis on vehicle loads on Nanjing Qixiashan Yangtze River Bridge based on long-term field measurement[J]. *Journal of Southeast University* (*Natural Science Edition*), 2021, **51**(6): 979 - 985. DOI: 10. 3969/j. issn. 1001-0505. 2021. 06. 009. (in Chinese)

[4] Zhang L, Feng D M, Wu G. Dynamic vehicle load identification method based on LSTM network[J]. *Journal of Southeast University* (*Natural Science Edition*), 2023, **53**(2): 187 - 192. DOI: 10. 3969/j. issn. 1001-0505. 2023. 02. 001. (in Chinese)

[5] Yan J Y, Deng L, He W. Evaluation of existing prestressed concrete bridges considering the randomness of live load distribution factor due to random vehicle loading position[J]. *Advances in Structural Engineering*, 2017, **20**(5): 737 - 746. DOI: 10. 1177/1369433216664350.

[6] Moses F. Weigh-in-motion system using instrumented bridges[J]. *Transportation Engineering Journal of ASCE*, 1979, **105**(3): 233 - 249. DOI: 10. 1061/tpejan. 0000783.

[7] Ren W X, Zuo X H, Wang N B, et al. Review of non-pavement bridge weigh-in-motion[J]. *China Journal of Highway and Transport*, 2014, **27**(7): 45 - 53. DOI: 10. 19721/j. cnki. 1001-7372. 2014. 07. 006. (in Chinese)

[8] Žnidaric A, Lavric I, Kalin J. Free-of-axle detector bridge WIM measurements on short slab bridges in proceedings of the 3rd international WIM conference[C]// *Proceedings of the 3rd International WIM Conference*. Florida, USA, 2002: 231 - 239.

[9] Deng L, Shi H, He W, et al. Vehicles' BWIM based on virtual simply-supported beam method[J]. *Journal of Vibration and Shock*, 2018, **37**(15): 209 - 215. DOI: 10. 13465/j. cnki. jvs. 2018. 15. 029. (in Chinese)

[10] Lansdell A, Song W, Dixon B. Development and testing of a bridge weigh-in-motion method considering nonconstant vehicle speed[J]. *Engineering Structures*, 2017, **152**: 709 - 726. DOI: 10. 1016/j. engstruct. 2017. 09. 044.

[11] Zhuo Y, An J H, Zhang B, et al. Application of non-constant speed algorithm in bridge weigh-in-motion[J]. *Highway Engineering*, 2020, **45**(5): 100 - 107, 128. DOI: 10. 19782/j. cnki. 1674-0610. 2020. 05. 017. (in Chinese)

[12] Buch N, Velastin S A, Orwell J. A review of computer vision techniques for the analysis of urban traffic[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2011, **12**(3): 920 - 939. DOI: 10. 1109/TITS. 2011. 2119372.

[13] Peng B, Cai X Y, Zhang Y J, et al. Automatic vehicle detection from UAV videos based on symmetrical frame difference and background block modeling[J]. *Journal of Southeast University* (*Natural Science Edition*), 2017, **47**(4): 685 - 690. DOI: 10. 3969/j. issn. 1001-0505. 2017. 04. 010. (in Chinese)

[14] Peng B, Cai X Y, Tang J, et al. Automatic vehicle detection with UAV videos based on modified faster R-CNN[J]. *Journal of Southeast University* (*Natural Science Edition*), 2019, **49**(6): 1199 - 1204. DOI: 10. 3969/j. issn. 1001-0505. 2019. 06. 025. (in Chinese)

[15] Chen Z C, Li H, Bao Y Q, et al. Identification of spatio-temporal distribution of vehicle loads on long-span bridges using computer vision technology[J]. *Structural Control and Health Monitoring*, 2016, **23**(3): 517 - 534. DOI: 10. 1002/stc. 1780.

[16] Dan D H, Ge L F, Yan X F. Identification of moving loads based on the information fusion of weigh-in-motion system and multiple camera machine vision[J]. *Measurement*, 2019, **144**: 155 - 166. DOI: 10. 1016/j. measurement. 2019. 05. 042.

[17] Zhang B, Zhou L M, Zhang J. A methodology for obtaining spatiotemporal information of the vehicles on bridges based on computer vision[J]. *Computer-Aided Civil and Infrastructure Engineering*, 2019, **34**(6): 471 - 487. DOI: 10. 1111/mice. 12434.

[18] Xia Y, Jian X D, Deng L, et al. Research on traffic-video-aided bridge weigh-in-motion approach[J]. *China Journal of Highway and Transport*, 2021, **34**(12): 104 - 114. DOI: 10. 19721/j. cnki. 1001-7372. 2021. 12. 009. (in Chinese)

[19] Zhao D D, He W, Deng L, et al. Trajectory tracking and load monitoring for moving vehicles on bridge based on axle position and dual camera vision[J]. *Remote Sensing*, 2021, **13**(23): 4868. DOI: 10. 3390/rs13234868.

[20] Yang G, Wang P, Han W S, et al. Automatic generation of fine-grained traffic load spectrum via fusion of weigh-in-motion and vehicle spatial-temporal information[J]. *Computer-Aided Civil and Infrastructure Engineering*, 2022, **37**(4): 485 - 499. DOI: 10. 1111/mice. 12746.

[21] Xu Z F, Wei B, Zhang J. Reproduction of spatial-tempo-

ral distribution of traffic loads on freeway bridges via fusion of camera video and ETC data[J]. *Structures*, 2023, **53**: 1476 - 1488. DOI: 10.1016/j.istruc.2023.05.023.

[22] Dong Y Q, Wang D L, Pan Y, et al. Large field monitoring system of vehicle load on long-span bridge based on the fusion of multiple vision and WIM data[J]. *Automation in Construction*, 2023, **154**: 104985. DOI: 10.1016/j.autcon.2023.104985.

[23] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Vancouver, BC, Canada, 2023: 7464 - 7475.

[24] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA, 2018: 7132 - 7141.

[25] Zheng Z H, Wang P, Ren D W, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. *IEEE Transactions on Cybernetics*, 2022, **52**(8): 8574 - 8586. DOI: 10.1109/TCYB.2021.3095305.

[26] Zhang Y F, Ren W Q, Zhang Z, et al. Focal and efficient IOU loss for accurate bounding box regression[J]. *Neurocomputing*, 2022, **506**: 146 - 157. DOI: 10.1016/j.neucom.2022.07.042.

[27] Du Y H, Zhao Z C, Song Y, et al. StrongSORT: Make DeepSORT great again[J]. *IEEE Transactions on Multimedia*, 2023, **25**: 8725 - 8737. DOI: 10.1109/TMM.2023.3240881.

[28] Zhang X, Hao X Y, Li J S, et al. Fusion and visualization method of dynamic targets in surveillance video with geospatial information[J]. *Acta Geodaetica et Cartographica Sinica*, 2019, **48**(11): 1415 - 1423. (in Chinese)

[29] He W, Deng L, Shi H, et al. Novel virtual simply supported beam method for detecting the speed and axles of moving vehicles on bridges[J]. *Journal of Bridge Engineering*, 2017, **22**(4):04016141. DOI:10.1061/(asce)be.1943-5592.0001019.

基于多视角信息融合的公路桥梁车辆荷载时空位置识别

邓 露[1,2] 邓佳宇[1] 王 维[1] 何 维[1] 张龙威[3]

([1] 湖南大学土木工程学院, 长沙 410082)

([2] 湖南大学工程结构损伤诊断湖南省重点实验室, 长沙 410082)

([3] 湖南科技大学土木工程学院, 湘潭 411201)

摘要:针对现有方法在复杂交通场景下难以准确获取桥上车辆荷载时空分布的问题,提出了一种基于多视角信息融合的车辆荷载时空位置识别方法. 首先,开发了 vadYOLO-StrongSORT 模型,可在单视角下同时检测和跟踪车辆;然后,在图像标定和跨视角车辆匹配基础上,采用自适应加权最小二乘法进行多视角信息融合以修正车辆轨迹;最后,结合车辆轨迹和车轴配置,重构车轴荷载的时空位置分布. 通过模型试验评估了所提方法在典型交通场景下的性能. 结果表明:相较于基于单视角的车辆位置识别方法,多视角信息融合方法在跟踪稳定性、定位精度和抗遮挡性能上有显著提升;变道场景下,所提方法的最高平均定位误差低于 2.0 cm,明显优于单视角方法的 17.0 cm;多车遮挡场景下,所提方法的车辆捕获率可达 100%,而单视角方法最高仅为 72.5%;同时,与其他检测跟踪模型相比,vadYOLO-StrongSORT 在试验中取得了最高的识别精度.

关键词:桥梁工程;车辆荷载;时空位置;多视角信息融合;车轴识别;桥梁动态称重;桥梁健康监测

中图分类号:U441.3