Traitement du Signal
Vol. 41, No. 6, December, 2024, pp. 3143-3151
Journal homepage: http://iieta.org/journals/ts

Improved Intelligent Learning Filter in Deep Learning Systems and Its Application in Traffic Object Detection

Xiaoyi Zheng[1], Xiaomin Fang[1], Kun Lan[2], Guofei Chai[3*]

[1] College of Information Engineering, Quzhou College of Technology, Quzhou 324000, China
[2] College of Mechanical Engineering, Quzhou University, Quzhou 324000, China
[3] College of Electrical and Information Engineering, Quzhou University, Quzhou 324000, China

Corresponding Author Email: chaig@qzc.edu.cn

https://doi.org/10.18280/ts.410630

Received: 19 May 2024
Revised: 28 October 2024
Accepted: 12 November 2024
Available online: 31 December 2024

Keywords:
intelligent learning filtering, deep learning, traffic object detection, Kalman Filter (KF)-Transformer, hybrid network, Kalman filtering

ABSTRACT

With the continuous development of intelligent transportation systems, traffic object detection technology has been widely applied in fields such as autonomous driving, traffic monitoring, and public safety. However, existing traffic object detection methods still face numerous challenges in complex traffic environments, such as occlusion, dynamic changes, and uneven lighting, which lead to a decrease in detection accuracy. Traditional deep learning methods, although performing well in static scenarios, often fail to maintain stable performance in dynamic, complex traffic scenes. Therefore, improving the robustness and accuracy of object detection has become a pressing issue in the field of intelligent transportation. To address these challenges, this paper proposes an intelligent learning filtering-based improvement to the deep learning training mechanism and applies it to traffic object detection. First, the training data is optimized using intelligent learning filtering techniques to eliminate noise and irrelevant information, improving data quality and enhancing the learning effectiveness of deep learning models. Next, a hybrid Kalman Filter (KF)-Transformer network for traffic object detection is constructed, combining the advantages of Kalman filtering and Transformer models to strengthen the model's ability to capture dynamic information and long-term dependencies. Experimental results show that the proposed model achieves higher accuracy and stability in traffic object detection tasks, especially in handling high-speed motion, partial occlusion, and complex backgrounds, demonstrating significant advantages. This study provides a novel solution to improve the accuracy and robustness of traffic object detection systems, with important theoretical and practical value.

1. INTRODUCTION

With the continuous development of traffic management and intelligent transportation systems, the application of traffic object detection technology has become increasingly widespread in fields such as autonomous driving, traffic monitoring, and road safety [1-4]. Traditional traffic object detection methods often rely on manually designed features and classical image processing techniques. However, in complex traffic environments, these methods face issues such as target occlusion, complex backgrounds, and lighting changes, which make it difficult for their performance to meet practical application requirements [5, 6]. In recent years, the rise of deep learning, especially convolutional neural networks (CNNs) and Transformer models, has greatly advanced the technology of traffic object detection. Deep learning, through automatic feature extraction and end-to-end training, has significantly improved the accuracy and efficiency of object detection [7-10]. However, existing deep learning systems still face problems such as data imbalance and noise interference, which limit their application effectiveness in real traffic environments.

Although a large number of studies have explored the application of deep learning in traffic object detection in recent years, there are still some notable flaws and shortcomings [11-14]. First, most existing research focuses on improving model structures, while less attention has been paid to improving data quality and optimizing the training mechanism [15-19]. Secondly, existing methods often exhibit instability when handling complex traffic scenes, especially in dynamic environments. Models are prone to interference from factors such as noise, occlusion, and lighting changes, thus affecting detection accuracy [20-24]. Moreover, many methods have failed to fully explore the potential of multi-source information, leading to insufficient robustness and real-time performance in object detection. Therefore, a new approach is urgently needed to address these issues and improve the performance of traffic object detection systems.

The main research content of this paper includes two aspects: on one hand, we propose a deep learning system training mechanism based on intelligent learning filtering, which optimizes the training data through filtering algorithms, improving the learning efficiency and robustness of deep learning models. On the other hand, we design and construct a

traffic object detection network based on the KF-Transformer hybrid network. This network combines the dynamic information estimation capability of Kalman filtering and the advantage of Transformer models in modeling long-term dependencies, which can effectively improve the accuracy and stability of object detection. Through these two aspects of research, this paper provides a new solution to improve the performance of traffic object detection systems, with significant theoretical and practical value.

2. DEEP LEARNING SYSTEM IMPROVED BY INTELLIGENT LEARNING FILTERING

The deep learning system training mechanism improved by intelligent learning filtering refers to the introduction of intelligent learning filtering algorithms during the training process of deep learning models, dynamically adjusting and optimizing input data and training parameters to enhance the model's learning efficiency and accuracy [25, 26]. This mechanism uses intelligent algorithms to analyze and filter training data in real time, identifying and eliminating noise and redundant information while retaining key information, thereby improving the model's generalization ability and convergence speed. With this method, the training process becomes more efficient, resource utilization becomes more reasonable, and the model's final performance is accordingly improved.

Traffic object detection tasks face complex and changing environments, as well as large amounts of real-time data, which contain a great deal of noise and redundant information. Traditional deep learning methods are difficult to process and learn efficiently in a short amount of time. At the same time, traffic object detection requires high precision and reliability, especially in autonomous driving and intelligent transportation systems, where any detection errors or missed, detections can lead to serious consequences. Therefore, this paper proposes applying the intelligent learning filtering improvement mechanism to traffic object detection. The intelligent learning filtering mechanism can dynamically optimize training data, enhance the model's adaptability in complex environments, and effectively improve the model's accuracy and robustness, further ensuring the safety and reliability of traffic object detection systems in real applications.

This paper proposes an optimized training strategy combining the KF algorithm and the deep learning network Transformer. This mechanism integrates the recursive update capability of the KF with the self-attention mechanism of Transformer. When processing video sequences in complex traffic object detection, the filter precisely estimates the state of the object, reduces the noise impact, and improves the global search ability during the training process by optimizing the learning rate, allowing the model to better converge to the global optimal solution in complex environments.

2.1 KF

In the Transformer-based traffic object detection model optimized by Kalman filtering proposed in this paper, the main function of the KF is to estimate and predict the dynamic state of traffic objects, thereby providing accurate and smooth input data for the Transformer model. In traffic surveillance videos, such as cars, pedestrians, etc., the dynamic changes in the motion trajectory of objects are often present, and due to the complexity of the environment, the position and speed of the objects are often affected by noise. The KF estimates and predicts the state using the historical motion information and current observations of the target, such as position and speed, which can effectively reduce the impact of these noise sources. Specifically, the working principle of the KF in traffic object detection is based on a recursive process of Bayesian estimation. At each moment, the KF predicts the target's position and speed at the next moment based on the current observation data and previous state estimates, combined with the system's motion model. Then, the filter fuses the current observation with the predicted value and updates the target's state estimate through a weighted average. This process includes two main steps: the prediction step and the update step. In the prediction step, the KF predicts the future state of the target based on the motion model; in the update step, the filter optimizes the state estimate by calculating the Kalman gain, combining the actual observations and predicted results. This recursive process not only allows real-time estimation of the target's motion trajectory but also effectively reduces detection errors caused by factors such as poor image quality and partial occlusion of the target. The state vector and observation vector of the system at time j are represented by A_j and B_j, the input variables at time j-1 are represented by i_{j-1}, and the system matrix, control matrix, observation matrix, and feed-forward matrix are represented by X, Y, Z, and F, respectively. The system noise and observation noise are represented by $q(j)$ and $n(j)$. The state transition equation and observation equation of the KF are given by the following formulas:

$$A_j = XA_{j-1} + Yi_{j-1} + q(j) \tag{1}$$

$$B_j = ZA_{j-1} + Fi_{j-1} + n(j) \tag{2}$$

2.2 Transformer

The main function of the Transformer model is to process the temporal information and spatial features in traffic surveillance video images through the self-attention mechanism, thereby effectively capturing and recognizing traffic objects. Although traditional CNNs can extract spatial features, they often have limitations when faced with dynamic scenes that involve long-term dependencies [27, 28]. The motion trajectory of traffic objects is usually continuous and changes over time. The Transformer model, through its self-attention mechanism, can establish relationships between targets across different time steps and capture dynamic features across time frames [29, 30]. For example, in traffic object detection, the Transformer can capture the motion trajectory, position changes, and interactions between the target and background by propagating information across different frames, thus achieving efficient object recognition and tracking. In this way, the Transformer model can effectively handle the spatial and temporal variations of objects in traffic surveillance videos and provide high-accuracy object detection results. Let the time step be represented by s, the feature dimension by u, the length of the input sequence by l, and the position encoding matrix by OR. The position encoding calculation method for the Transformer network is given by the following formulas, used to represent the position information of each element in the sequence:

$$OR(s,2u)=SIN\left(\frac{s}{10000}^{\frac{2u}{l}}\right) \tag{3}$$

$$OR(s,2u+1)=COS\left(\frac{s}{10000}^{\frac{2u}{l}}\right) \tag{4}$$

The Transformer network consists of two sub-layers: the multi-head self-attention mechanism layer and the feed-forward layer, which extract key features from each image frame and model and predict the target using global contextual information. In the traffic object detection scenario, the relationships between targets rely not only on the features of a single frame but also require cross-frame information to infer the motion state of the targets. The Transformer, by calculating the similarity between each target and other targets, generates attention weights, ensuring that the network can focus on important target features at each time step. Let the mapping weights be represented by QW^{Q}_{i}, KW^{K}_{i}, VW^{V}_{i} and the multi-head attention weights by $Q°$, with the self-attention mechanism parameters represented by W, J, and N. The multi-head self-attention mechanism can be represented by the following equation:

$$\begin{aligned}&\mathrm{HEAD}_u=\mathrm{ATT}\left(WQ_u^{W},JQu_i^{J},NQ_u^{N}\right),u=1,2...v\\&\mathrm{MULTIHEAD}(W,J,N)=\mathrm{CONCACT}(\mathrm{HEAD}_1,\mathrm{HEAD}_2...\mathrm{HEAD}_v)Q^{\circ}\end{aligned} \tag{5}$$

Let $f_g=f/g$, the key dimension be represented by f_g, and the transpose parameters by S. The attention weight matrix expression is:

$$\mathrm{ATT}(W,J,N)=\mathrm{SOFTMAX}\left(\frac{WJ^{S}}{\sqrt{f_g}}\right)N \tag{6}$$

Let the weights be represented by Q_1 and Q_2, and the biases by y_1 and y_2, with the input data represented by a. The calculation for the feed-forward layer is:

$$\mathrm{FNN}(a)=\mathrm{RELU}(aQ_1+y_1)Q_2+y_2 \tag{7}$$

3. KF-TRANSFORMER HYBRID NETWORK FOR TRAFFIC OBJECT DETECTION

3.1 Traffic object detection dataset

The datasets used in this paper are standard datasets specifically for traffic object detection and tracking tasks. These datasets contain a large amount of traffic surveillance video images used to validate the model's performance in different traffic scenarios. The selected traffic object detection datasets include KITTI, Cityscapes, which provide high-quality traffic scene images that cover various environmental conditions such as different lighting, weather, time, and speed. The KITTI dataset, as a widely used standard dataset in the autonomous driving and object detection fields, provides the localization and labeling of traffic scene targets such as vehicles, pedestrians, and cyclists. The dataset contains rich dynamic image sequences, making it suitable for validating the performance of traffic object detection systems in dynamic environments, including object detection, classification, and tracking tasks. The Cityscapes dataset provides more fine-grained urban traffic image data, suitable for object detection and segmentation tasks in street scenes. This dataset includes various traffic participants from urban environments, including pedestrians, cyclists, vehicles, etc., and covers different occlusions, background complexities, and various city road scenes, making it quite challenging. The dataset not only has accurate target position annotations but also provides high-resolution images and detailed pixel-level annotations, which help train and validate the model in a variety of complex traffic scenarios.

3.2 Traffic object detection task description

The traffic object detection task refers to the process of recognizing, locating, tracking, and classifying various objects in traffic scenes, such as vehicles, pedestrians, traffic signs, etc., from real-time or historical video images using computer vision and deep learning technologies, within traffic surveillance systems or autonomous driving scenarios. The core objective of this task is to accurately extract useful information from complex traffic environments to enable fast response and effective management of traffic objects. Traffic object detection is not limited to identifying the presence of objects but also includes tracking and predicting the spatial location and temporal changes of the objects. Therefore, traffic object detection is a comprehensive task involving object recognition, temporal modeling, target tracking, and dynamic prediction.

In traffic object detection, the challenges mainly arise from the dynamic and complex nature of traffic scenes. The appearance of objects in images is influenced by factors such as lighting, weather, occlusion, and viewing angle, making a single object recognition method insufficient for all situations. Furthermore, objects in traffic environments often exhibit strong dynamic features, such as acceleration, deceleration, and turning behaviors of vehicles, requiring the detection system not only to recognize static objects but also to predict and track the motion trajectories of objects. Therefore, the successful implementation of traffic object detection tasks requires high precision and robustness, especially in complex urban traffic environments. One of the major technical challenges in this task is how to handle rapidly changing target information and reduce false positives and missed detections.

3.3 Data preprocessing

Data preprocessing for traffic object detection based on the intelligent learning filter improved deep learning system involves effectively handling noise, redundant information, and dynamic changes in traffic surveillance video images, so that clean and high-quality input data can be provided for the subsequent object detection system. In traffic surveillance video images, due to camera movement, lighting changes, weather factors, and occlusion of objects, the images often contain a large amount of noise and unnecessary information. To address this, this paper performs image denoising as a preprocessing step to remove noise and improve the quality of the images. In this way, irregular fluctuations and irrelevant background information in the images are effectively suppressed, thereby improving the recognizability of traffic objects in the images. The preprocessed image data becomes more stable, providing more accurate input for the subsequent

KF and Transformer models, ensuring that the training and inference processes of the detection system are more efficient and robust.

3.4 Autoencoder

Traffic surveillance video images often contain considerable noise, such as lighting changes, weather effects, camera shake, etc. This noise can seriously interfere with the performance of the object detection system, especially in complex urban traffic environments, where object recognition and tracking become more challenging. Therefore, this paper introduces an autoencoder in the proposed traffic object detection model to effectively denoise and extract features from the initial traffic surveillance video images, further improving the accuracy and robustness of the subsequent object detection. The autoencoder, through its encoding-decoding structure, can compress the original image into a lower-dimensional latent space representation, capturing the most useful features of the data while suppressing unnecessary noise. Specifically, the autoencoder compresses the input image into a low-dimensional latent space representation, retaining the key features of the image while removing redundant information and noise. Let the output be represented by C, the activation function of the autoencoder by $d(a)$, the transpose of the autoencoder's weight matrix by Q^S, the image sequence data with added Gaussian noise by $\tilde{a}_s$, and the bias parameters obtained through gradient descent training by y. The output of the data processed by the autoencoder is:

$$c = d\left(Q^S \tilde{a}_s + y\right) \tag{8}$$

3.5 Network framework

Instead of placing the KF at the front or back end of the Transformer network, this paper chooses to insert the KF design between the encoder and decoder to achieve more refined dynamic information learning and prediction. Adding the KF at the front end can perform initial noise filtering in the input data stage but cannot fully utilize the self-attention mechanism of the Transformer for long-range dependencies in time series, nor can it handle dynamic changes in the model during the prediction stage. Adding the KF at the back end can improve prediction accuracy, but it may miss the effective capture of dynamic information during the encoding phase. By embedding the KF between the encoder and decoder, the model can, based on the high-quality input features, use the KF's dynamic updating mechanism to further accurately estimate the movement trajectory and state of the target, ensuring temporal consistency and dynamic adaptability during the object detection and tracking process. Especially in complex urban traffic scenarios, this strategy can better handle changes in the movement trajectories of different targets, occlusion, and background interference. Figure 1 shows the structure diagram of the KF-Transformer network designed in this paper.

The KF-Transformer hybrid network, by combining the advantages of the KF and Transformer network, is designed for traffic object detection with the aim of improving detection accuracy. The first step in the experimental process is to normalize the image data. Images in traffic surveillance videos often contain a lot of noise and redundant information, so normalization standardizes the image data, making it more consistent when input into the network, thus improving the training efficiency and stability of the network. Through this process, the network can focus on key information in the image content, such as vehicle behavior, pedestrian movement, etc., while also uncovering potential real-time traffic situations behind the images. Furthermore, in order to enhance the network's ability to learn the relationships between elements in the input image, the normalized image data is added to the position encoding matrix to generate an embedding matrix with positional information. Position encoding helps the network understand the relative positions and semantic relationships of different regions in the image, which is crucial for analyzing the temporal and spatial variations of object responses and traffic event developments in traffic surveillance video images. This enhances the model's generalization ability, allowing it to effectively handle a variety of real-time traffic situations.

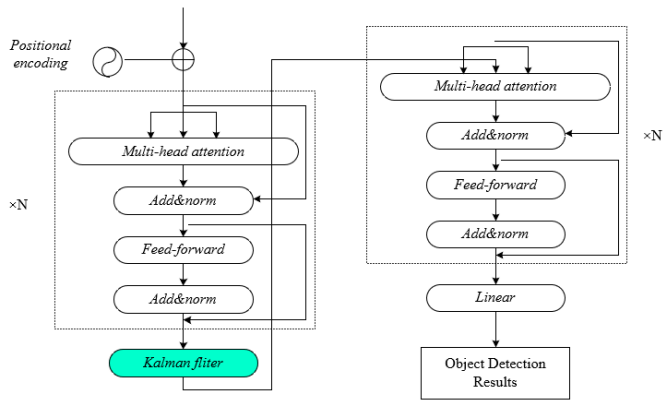


Figure 1. KF-Transformer network structure designed in this paper

Next, the image data, after position encoding, is input into the KF-Transformer hybrid network. The core of the network lies in the introduction of the KF, which is positioned between the Transformer encoder and decoder, primarily used for smoothing and noise removal of the temporal features of the image data. Since traffic objects usually exhibit sudden and fluctuating behaviors, the KF helps the network accurately track the dynamic changes of these image data and remove disturbances caused by noise, especially in fast-changing public opinion environments. Finally, the image features processed by the filter are sent to the decoder part of the Transformer network, where the complex relationships between the images are further explored through the self-attention mechanism. The decoded information is then passed to the linear layer, where the output signal is mapped, and the final detection result is obtained.

3.6 Network training principle

The loss function used in this paper is mean square error, and its formula is:

$$M = \frac{1}{v}\sum_{s=1}^{v} 1\left(\tilde{a}_f - \hat{a}_s\right)^2 \tag{9}$$

Let the network objective function be represented by M', the parameters that help control denoising and the object detection task be represented by β, the regularization parameter be represented by η, the regularization equation be represented by $E(a)$, and the model's learning parameters be represented by

Φ. The objective function formula for training the network is:

$$M' = x\sum_{u=1}^{v} l(\tilde{a}_u - \hat{a}_u) + \sum_{s=1}^{v} l(\tilde{a}_u - \hat{a}_u)^2 + \eta E(\Phi) \quad (10)$$

Training deep learning models often faces the issue of getting stuck in local optima, especially in complex traffic object detection tasks where the dynamic changes of objects and data noise may cause the model to fall into undesirable local solutions. In this paper, the cosine annealing method is introduced into the model training process to optimize the learning rate adjustment strategy. The cosine annealing method gradually reduces the learning rate, allowing the model to converge quickly in the early stages and gradually approach the optimal solution. Later, a smaller learning rate helps the model avoid excessive updates during the convergence process, thereby preventing oscillation or overfitting. Let the learning rate at the current epoch be represented by λs, the number of restarts be represented by c, the maximum and minimum learning rates be represented by λ^u_{MAX} and λ^u_{MIN}, the current number of periods executed be represented by S_{CU}, and the total number of epochs for the c-th restart be represented by S_u. The process of cosine annealing is represented by the following formula:

$$\lambda_s = \lambda^u_{MIN} + \frac{1}{2}\left(\lambda^u_{MAX} - \lambda^u_{MIN}\right)\left[1 + COS\left(\frac{S_{CU}}{S_u}\Pi\right)\right] \quad (11)$$

Figure 2 shows the traffic object detection network structure built in this paper.

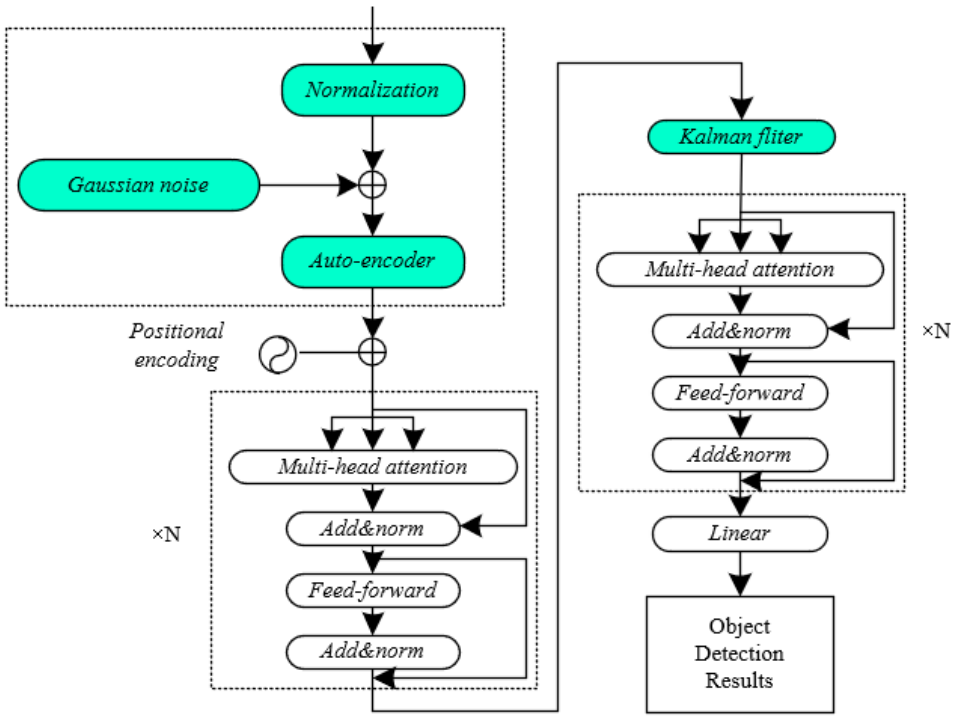


Figure 2. Traffic object detection network structure diagram

4. EXPERIMENTAL RESULTS AND ANALYSIS

From the experimental results shown in Figure 3, it can be seen that the deep learning training mechanism based on intelligent learning filtering proposed in this paper shows significant improvements across multiple metrics. Specifically, the losses for Box, Objectness, and Classification all show a decreasing trend as the training progresses, with significant drops observed at the 400th epoch. For example, the Box loss decreased from an initial value of 0.12 to 0.032, the Objectness loss decreased from 0.1 to 0.001, and the Classification loss decreased from 0.28 to 0. This indicates that the training data optimized by the intelligent learning filtering mechanism significantly improves the model's performance in object detection, with noise effectively reduced, resulting in more stable training and faster convergence. Moreover, the changes in Precision and Recall also show a gradual optimization of the model in the object detection task. Precision increased from 0 to 0.86, and Recall increased from 0 to 0.86, indicating significant improvements in both accuracy and recall. More importantly, the changes in mAP@0.5 and mAP@0.5:0.95 also show significant improvements. mAP@0.5 increased from an initial value of 0 to 0.8109, and mAP@0.5:0.95 increased from 0 to 0.6109, indicating that the model demonstrates good adaptability and stability in detection accuracy and multi-scale object detection.

Table 1. Parameter comparison of different traffic object detection models

Model	Pretrained Weight (MB)	Parameters (Million)	*F1 Score*	*mAP*(%)	*FPS*
Faster R-CNN	354	27.256	0.562	52.3	41
SSD	21.3	6.125	0.648	61.2	162
*YOLO*9000	3.6	1.784	0.632	65.8	123
YOLOv5n	12.9	7.125	0.723	74.5	112
DETR	15.6	8.562	0.758	76.2	87
Swin Transformer	21.4	12.235	0.762	76.2	83
PVT	16.5	7.562	0.745	75.1	92
The proposed model	13.8	8.326	0.823	82.6	78

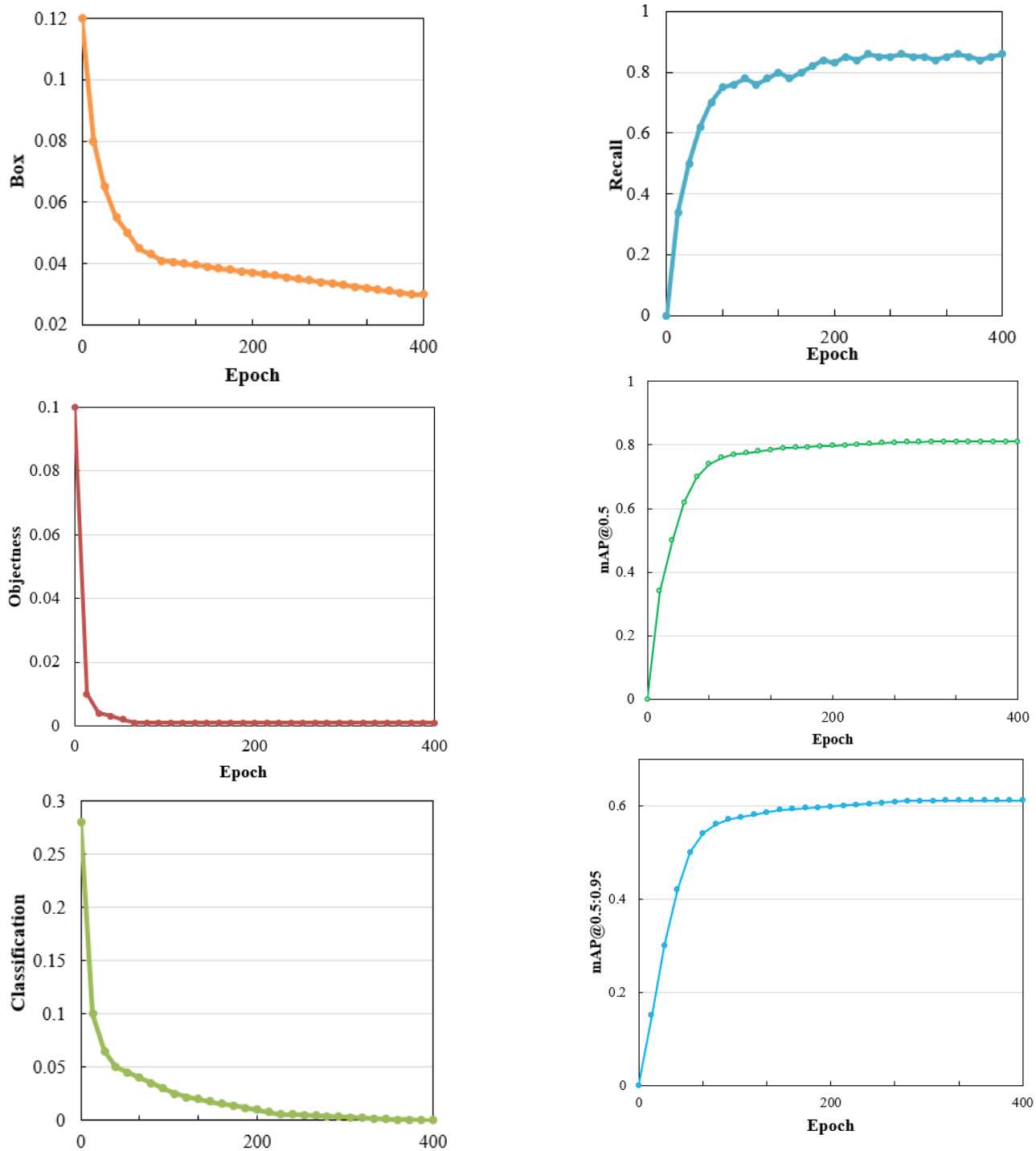


Figure 3. Experimental results of the model in this paper

From the data in Table 1, the KF-Transformer hybrid network-based traffic object detection model proposed in this paper shows superior performance across multiple key metrics. First, in the core indicators of F1 score and mAP, the proposed model achieves 0.823 and 82.6%, respectively. Compared to existing models, the proposed model performs excellently in these two metrics, far surpassing traditional models like Faster R-CNN (F1 score: 0.562, mAP: 52.3%) and SSD (F1 score: 0.648, mAP: 61.2%). Additionally, advanced models like YOLOv5n, DETR, and Swin Transformer also perform well in mAP and F1 score but still fall short of the proposed model's performance. Specifically, although YOLOv5n has a higher computational efficiency (FPS: 112), its mAP is only 74.5% and F1 score is 0.723, which is significantly lower than that of the proposed model. Although the FPS of the proposed model

(78) is lower than some efficient models like SSD (FPS: 162), its advantage in accuracy makes it more promising for practical applications.

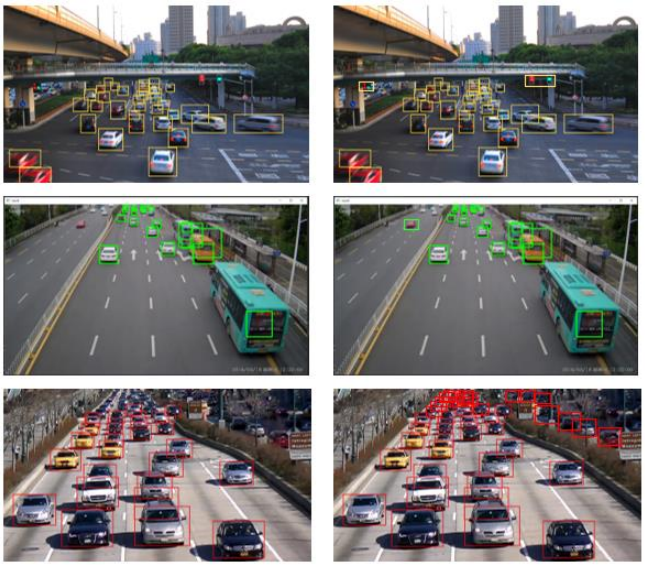
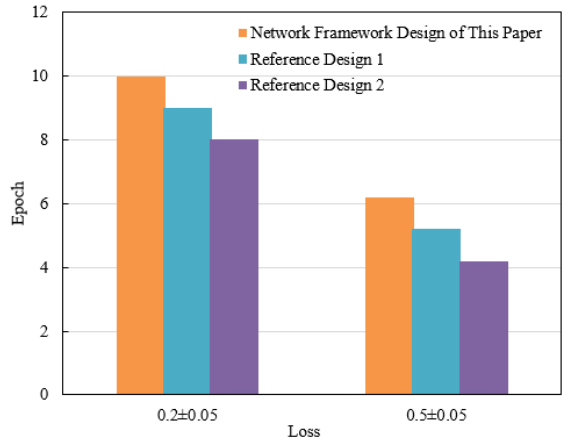


Figure 4. Comparison of traffic object detection results before and after introducing the autoencoder

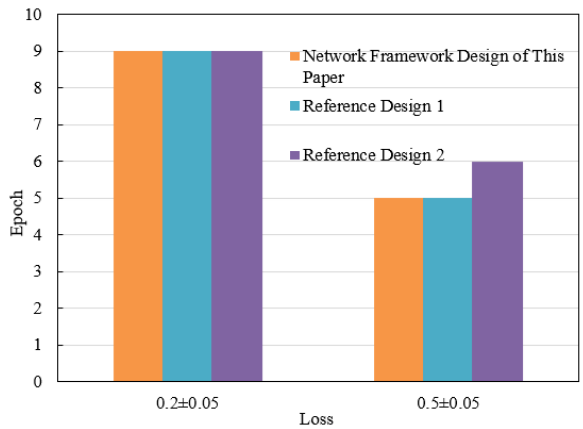
Figure 5. Confusion matrix

Figure 4 shows a comparison of traffic object detection results before and after introducing the autoencoder. After introducing the autoencoder, the performance of the traffic object detection model significantly improved. By applying the autoencoder for denoising and feature compression of the input data, the model became more robust when dealing with noise interference in complex scenarios. Experimental results indicate that the detection results after processing by the autoencoder are more accurate compared to the unprocessed data, especially in terms of object localization and classification.

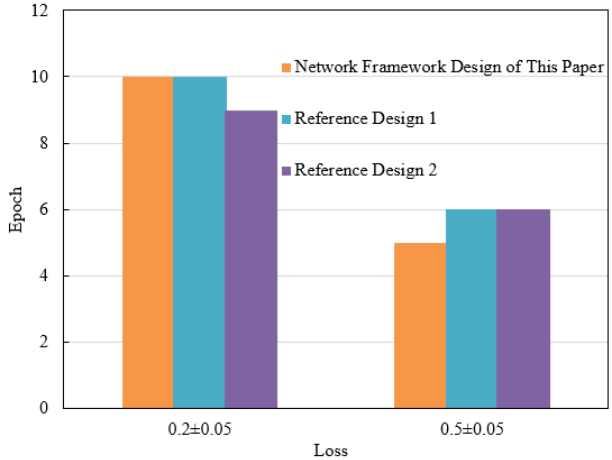
From the confusion matrix shown in Figure 5, the KF-Transformer hybrid network-based traffic object detection model demonstrates high accuracy and stability in identifying various types of targets. Particularly in detecting high-frequency targets such as Car, Bus, and Person, the model shows a high number of True Positives (TP), indicating that it can accurately identify these common targets, with low False Positives (FP) and False Negatives (FN). For example, for the "Car" target, both detection precision and recall are high, with the diagonal values (TP) in the confusion matrix significantly greater than other categories, demonstrating the model's strong recognition accuracy for this target. For relatively smaller targets like "bike" and "motor," although the recognition performance is slightly lower, the misdetection and missed detection rates are still significantly reduced compared to traditional detection models. Additionally, the model shows high accuracy in identifying the background (backg) class, indicating that it can effectively distinguish the background from actual traffic objects, reducing background noise interference and further proving its robustness and generalization ability.



(1) KITTI



(2) Cityscapes



(3) COCO

Figure 6. The impact of epochs on loss across different datasets

In Figure 6, Reference Designs 1 and 2 correspond to placing the KF at the front-end and back-end of the Transformer network, respectively, while the network framework in this paper places the KF between the encoder and decoder. Based on the experimental data provided in Figure 6, the proposed network framework outperforms Reference Designs 1 and 2 on different datasets (KITTI, Cityscapes, COCO), particularly in terms of the convergence speed and stability of the Loss value. On the KITTI dataset, the Loss value of the proposed framework reached 0.2±0.05, significantly lower than the Loss values of Reference Design 1 (0.5±0.05) and Reference Design 2 (0.5±0.05), indicating that the proposed framework can converge more quickly and stably to a lower loss value in the object detection task. Similarly, the proposed design shows similar advantages on the Cityscapes and COCO datasets. Particularly on the Cityscapes dataset, the Loss value of the proposed framework is 0.2±0.05, while the Loss values of Reference Design 1 and 2 are both 0.5±0.05, demonstrating that the proposed framework can maintain a low loss level in complex environments, improving the model's learning efficiency and stability. Furthermore, the performance on the COCO dataset also shows that the proposed design effectively controls Loss fluctuations while improving accuracy, allowing the model to achieve relatively ideal convergence in diverse object detection tasks.

5. CONCLUSION

This paper proposed a deep learning system training mechanism based on intelligent learning filtering and a KF-Transformer hybrid network for traffic object detection. By combining the dynamic information estimation capability of the KF and the advantages of the Transformer model in modeling long-term dependencies, the proposed framework significantly improved the precision, stability, and robustness of traffic object detection tasks. Through the filtering algorithm to optimize training data, the model's learning efficiency and robustness were enhanced, enabling it to efficiently recognize various targets in complex traffic scenes, including cars, pedestrians, buses, motorcycles, and more. Furthermore, the hybrid network design that combines Kalman filtering and Transformer demonstrated strong capabilities in processing dynamic information of targets, effectively eliminating noise, reducing false detections and missed detections, and improving practical value in various traffic environments. Experimental results validated the effectiveness of the proposed method, especially in terms of its performance on several public datasets such as KITTI, Cityscapes, and COCO. The results show that the network framework proposed in this research outperformed traditional designs and existing advanced models in various object detection tasks, especially in terms of Loss value convergence speed, accuracy, and robustness. Notably, in dynamic target detection and complex scenarios, this model demonstrated high stability and can achieve low Loss values within a short training period, with strong adaptability and generalization ability.

However, there are some limitations to this research. First, while the proposed method excels in object detection accuracy and stability, the computational resource consumption is relatively high when processing large-scale datasets, which may affect applications that require real-time performance. Second, although the combination of Kalman filtering and Transformer performs well in dynamic object detection, the model may still face challenges in high-density scenes or scenarios with multi-object occlusion. Additionally, this study primarily focuses on traffic object detection, and its generalization ability to other fields still needs further validation. Future research can deepen and expand in two directions: optimizing the network's computational efficiency and inference speed to reduce the demand for computational resources, allowing it to be better applied in real-time monitoring and intelligent transportation systems. Secondly, advanced multimodal perception technologies can be combined to further enhance object detection accuracy and robustness in complex environments such as multi-object occlusion and high-density scenes.

ACKNOWLEDGMENT

The work was supported by Science and Technology Research Project of Quzhou (Grant No.: 2023K257).

REFERENCES

[1] Kilic, I., Aydin, G. (2022). Traffic lights detection and recognition with new benchmark datasets using deep learning and TensorFlow object detection API. Traitement du Signal, 39(5): 1673-1683. https://doi.org/10.18280/ts.390525

[2] Ngo, T.Q., Toan, N.D., Le, L.H., Nguyen, T.D., Nguyen, H. (2023). An examination of advances in multistage object detection techniques utilizing deep learning. Mathematical Modelling of Engineering Problems, 10(5): 1587-1610. https://doi.org/10.18280/mmep.100510

[3] Kan, H.Y., Li, C., Wang, Z.Q. (2024). Enhancing urban traffic management through YOLOv5 and DeepSORT algorithms within digital twin frameworks. Mechatronics and Intelligent Transportation Systems, 3(1): 39-54. https://doi.org/10.56578/mits030104

[4] Saini, V., Kantipudi, M.V.V.P., Meduri, P. (2023). Enhanced SSD algorithm-based object detection and depth estimation for autonomous vehicle navigation. International Journal of Transport Development and Integration, 7(4): 341-351. https://doi.org/10.18280/ijtdi.070408

[5] Niu, C., Li, K. (2022). Traffic light detection and recognition method based on YOLOv5s and AlexNet. Applied Sciences, 12(21): 10808. https://doi.org/10.3390/app122110808

[6] Xu, Z.S., Liao, Z.M., Ahmad, S., Mat Diah, N. (2023). Exploration on vehicle target detection technology based on wireless networks and its application in intelligent traffic. Intelligent Decision Technologies-Netherlands, 17(4): 1233-1247. https://doi.org/10.3233/IDT-230243

[7] Liu, H.X., Lu, G.H., Li, M.X., Su, W.H., Liu, Z.Y., Dang, X., Zang, D.Y. (2024). High-precision real-time autonomous driving target detection based on YOLOv8. Journal of Real-Time Image Processing, 21(5): 174. https://doi.org/10.1007/s11554-024-01553-2

[8] Wang, F.H., Li, L.Y., Liu, Y.T., Tian, S., Wei, L. (2021). Road traffic accident scene detection and mapping system based on aerial photography. International Journal of Crashworthiness, 26(5): 537-548.

https://doi.org/10.1080/13588265.2020.1764719
[9] Lu, X., Mao, X.N., Liu, H.Q., Meng, X.L., Rai, L. (2021). Event camera point cloud feature analysis and shadow removal for road traffic sensing. IEEE Sensors Journal, 22(4): 3358-3369. https://doi.org/10.1109/JSEN.2021.3138736
[10] Wang, J.F., Chen, Y., Dong, Z.K., Gao, M.Y. (2023). Improved YOLOv5 network for real-time multi-scale traffic sign detection. Neural Computing and Applications, 35(10): 7853-7865. https://doi.org/10.1007/s00521-022-08077-5
[11] Lee, H.S., Kim, K. (2018). Simultaneous traffic sign detection and boundary estimation using convolutional neural network. IEEE Transactions on Intelligent Transportation Systems, 19(5): 1652-1663. https://doi.org/10.1109/TITS.2018.2801560
[12] Kotapati, G., Ali, M.A., Vatambeti, R. (2023). Deep learning-enhanced hybrid fruit fly optimization for intelligent traffic control in smart urban communities. Mechatronics and Intelligent Transportation Systems, 2(2): 89-101. https://doi.org/10.56578/mits020204
[13] Wang, F.H., Qiao, J., Li, L.Y., Liu, Y.T., Wei, L. (2022). Scene recognition of road traffic accident based on an improved faster R-CNN algorithm. International Journal of Crashworthiness, 27(5): 1428-1432. https://doi.org/10.1080/13588265.2021.1959156
[14] Ayachi, R., Afif, M., Said, Y., Atri, M., Ben Abdelali, A. (2023). Integrating recurrent neural networks with convolutional neural networks for enhanced traffic light detection and tracking. Traitement du Signal, 40(6): 2577-2586. https://doi.org/10.18280/ts.400620
[15] Guo, Y., Liang, R.L., Cui, Y.K., Zhao, X.M., Meng, Q. (2022). A domain-adaptive method with cycle perceptual consistency adversarial networks for vehicle target detection in foggy weather. IET Intelligent Transport Systems, 16(7): 971-981. https://doi.org/10.1049/itr2.12190
[16] Zou, S.Z., Chen, H., Feng, H., Xiao, G.Y., Qin, Z., Cai, W.W. (2022). Traffic flow video image recognition and analysis based on multi-target tracking algorithm and deep learning. IEEE Transactions on Intelligent Transportation Systems, 24(8): 8762-8775. https://doi.org/10.1109/TITS.2022.3222608
[17] Yu, L., Zhang, B.L., Li, R. (2020). Detection of unusual targets in traffic images based on one-class extreme machine learning. Traitement du Signal, 37(6): 1003-1008. https://doi.org/10.18280/ts.370612
[18] Zheng, H., Liu, J.F., Ren, X.G. (2022). Dim target detection method based on deep learning in complex traffic environment. Journal of Grid Computing, 20(1): 8. https://doi.org/10.1007/s10723-021-09594-8
[19] Hu, T., Gong, Z.W., Song, J. (2024). Research and implementation of an embedded traffic sign detection model using improved YOLOV5. International Journal of Automotive Technology, 25(4): 881-892. https://doi.org/10.1007/s12239-024-00082-y
[20] Li, W., Chen, Y.L., Zhao, L.X., Luo, Y.Z., Liu, X. (2023). Research on malicious traffic detection based on image recognition. International Journal of Embedded Systems, 16(2): 134-142. https://doi.org/10.1504/IJES.2023.136387
[21] Tao, Z.M., Li, Y.B., Wang, P.C., Ji, L.Y. (2022). Traffic incident detection based on mmWave radar and improvement using fusion with camera. Journal of Advanced Transportation, 2022(1): 2286147. https://doi.org/10.1155/2022/2286147
[22] Zhanikeev, M., Tanaka, Y. (2009). A framework for detection of traffic anomalies based on IP aggregation. IEICE Transactions on Information and Systems, 92(1): 16-23. https://doi.org/10.1587/transinf.E92.D.16
[23] Liu, F., Qian, Y.R., Li, H., Wang, Y.Q., Zhang, H. (2021). CAFFNet: Channel attention and feature fusion network for multi-target traffic sign detection. International Journal of Pattern Recognition and Artificial Intelligence, 35(7): 2152008. https://doi.org/10.1142/S021800142152008X
[24] Iftikhar, S., Asim, M., Zhang, Z., Muthanna, A., Chen, J., et al. (2023). Target detection and recognition for traffic congestion in smart cities using deep learning-enabled UAVs: A review and analysis. Applied Sciences, 13(6): 3995. https://doi.org/10.3390/app13063995
[25] Liu, H., Wang, L., Chen, C., Bian, A. (2023). On research of dispersion characteristics of multi-component surface waves from traffic-induced seismic ambient noise. Journal of Applied Geophysics, 213: 105038. https://doi.org/10.1016/j.jappgeo.2023.105038
[26] Sun, Y., Wu, M., Liu, X., Zhou, L. (2022). High-precision dynamic traffic noise mapping based on road surveillance video. ISPRS International Journal of Geo-Information, 11(8): 441. https://doi.org/10.3390/ijgi11080441
[27] Corinto, F., Biey, M., Gilli, M. (2006). Non-linear coupled CNN models for multiscale image analysis. International Journal of Circuit Theory and Applications, 34(1): 77-88. https://doi.org/10.1002/cta.343
[28] Li, J., Wang, Y., Luo, C., Zhou, W., Dong, Z. (2023). CNN-LDNF: An image feature representation approach with multi-space mapping. International Journal of Machine Learning and Cybernetics, 14(3): 739-759. https://doi.org/10.1007/s13042-022-01660-1
[29] Liu, Y., Wang, X., Qu, B., Zhao, F. (2024). ATVITSC: A novel encrypted traffic classification method based on deep learning. IEEE Transactions on Information Forensics and Security, 19: 9374-9389. https://doi.org/10.1109/TIFS.2024.3433446
[30] Guo, X., Zhu, Q., Wang, Y., Mo, Y. (2024). MG-GCT: A motion-guided graph convolutional transformer for traffic gesture recognition. IEEE Transactions on Intelligent Transportation Systems, 25(10): 14031-14039. https://doi.org/10.1109/TITS.2024.3394911