

# Target image detection algorithm of complex road scene based on improved multi-scale adaptive feature fusion technology

Zhaosheng Xu<sup>1,2,\*</sup>, Zhongming Liao<sup>1,2</sup>, Xiaoyong Xiao<sup>2</sup>, Suzana Ahmad<sup>1</sup>, Norizan Mat Diah<sup>1</sup>, and Azlan Ismail<sup>1</sup>

<sup>1</sup> College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Shah Alam Branch, 40450 Shah Alam, Selangor Darul Ehsan, Malaysia

<sup>2</sup> College of Mathematics and Computer Science, Xinyu University, Xinyu 338004, Jiangxi, China

Received: 10 December 2024 / Accepted: 27 February 2025

**Abstract.** Understanding road scenes is crucial to the safe driving of autonomous vehicles, and object detection in road scenes is necessary to develop driving assistance systems. Current object detection algorithms are not very good at handling complex road scenes, and public datasets do not always adequately represent city traffic. Using Improved Multi-Scale Adaptive Feature Fusion Technology (IMSAFFT), this work suggests a real-time traffic information identification method to fix the issues of low detection accuracy of road scenes and high false detection rates in panoramic video images. In addition, a semantic recognition algorithm for a road scene based on image data is suggested. This study introduces computer vision-based approaches, including colour and texture recognition, object detection, and scene context understanding using Deep Neural Networks (DNN). An increasing number of deeper stacked layers allows the deep neural network to learn more complicated high-level semantic features, and the features' quality improves with time. A learning rate adaptive adjustment technique has been utilized to make training more efficient. After that, this improved detector is used to identify vehicles in original road environments. The suggested technique surpassed traditional detectors in the experiments with a high accuracy rate and processing speed. It worked well in real-world traffic situations for detecting overlapping, multiple, distant, and small objects. The simulation outcomes illustrate that the recommended IMSAFFT model increases the accuracy ratio of 98.4%, target image detection ratio of 97.4%, traffic prediction rate of 96.5%, processing speed rate of 10.4% and F1-score ratio of 95.4% compared to other existing models.

**Keywords:** Target image detection / complex road scenes / multi-scale adaptive feature fusion / real-time traffic information / object detection algorithms / urban traffic datasets / panoramic video images

## 1 Introduction

Image detection algorithms are of the utmost importance in improving smart transportation and advanced driver assistance systems (ADAS) [1]. These algorithms are crucial because they enable correct notions and interpretations of complex traffic congestion. To enhance driving safety and efficiency, locating and perceiving objects in real time is essential [2]. A type of impediment, including automobiles, pedestrians, traffic signs, and different categories of objects, is included in this class. Alternatively, object detection algorithms deal with great demanding situations while dealing with complex road sceneries [3]. This is especially unaffected in an urban environment characterized by an excessive density of devices and a diverse collection of goods. One of the primary drawbacks of the object detection algorithms that are now in use is

they cannot maintain high accuracy and occasionally false detection values in various circumstances [4]. Traditional algorithms typically confront limitations like fluctuating light conditions, occlusion, and the presence of small and long far-away objects. Because of those limits, the detection overall performance is reduced, making it difficult for these algorithms to be applied dependably in applications that might be used in the real world [5]. Another key aspect that has a giant effect on the effectiveness of object detection models is the quality of public datasets and the relevancy of those datasets [6]. Because an immoderate amount of public datasets employed to train these algorithms does now not appropriately constitute the situations of city site traffic, the performance of these algorithms is subpar when it comes to detecting and classifying objects. To discover solutions to these issues, it is vital to expand progressive processes to boost object identification algorithms' robustness and accuracy in complicated road scenarios [7]. An approach that can be deemed promising is using a

\* e-mail: [xuzhaosheng@xyc.edu.cn](mailto:xuzhaosheng@xyc.edu.cn)

technology that combines adaptive features on many scales. This method combines features received at adequate exclusive scales to obtain global and local information. Consequently, the detection performance is improved simultaneously during a wide spectrum of object sizes and distances [8]. This research aims to construct an actual-time traffic data identification method based on a more suitable multi-scale adaptive feature fusion technology. To summarize, the subsequent are the essential additives that constitute the proposed solution: the collection and getting ready of data, the introduction of an improved Single Shot MultiBox Detector (SSD) detector, the adaptive learning rate, and the experimental validation [9]. The testing results showed that the proposed technology became more advanced than traditional detectors. One of the most vital factors of the method that has been proposed is the concept of combining features, which can be adaptable on quite a few scales [10]. The technique can better cope with the range in item sizes and distances typically found in complicated road scenes because it integrates data from many scales. This allows the technique to handle the situation better. One of the most important aspects of improving the generalization capabilities of object identification algorithms is the utilization of data augmentation techniques. These techniques include single-data distortion techniques such as affine transformations and colour gamut modifications. Specifically, the significantly more desirable detection set of rules has extensive repercussions for packages within the real world, especially inside the context of smart transportation systems and ADAS. By imparting correct and up-to-date perception data about the road environment, the algorithm enhances ADAS's functionality to allow decisions based on accurate data. Consequently, there is an increase in both the efficiency of driving and the safety of driving on the road. Due to its accuracy and processing velocity, the solution [11] that has been produced is high-quality for implementation in numerous applications concerned with traffic management and monitoring. This is the case when considering that the solution becomes designed. The target image detection method that evolved, based on the advanced multi-scale adaptive feature fusion technology, gives a robust approach to the issues presently encountered using object identification systems in complicated situational circumstances that arise on roads [12]. Several potential areas of research could be taken up in subsequent research. These encompass incorporating additional sensor data, using the multi-scale adaptive characteristic fusion approach to various objects and situations, and investigating the usage of state-of-the-art machine mastering algorithms. Researchers and developers can continue pushing the boundaries of object detection technology if they address these potential future courses. This will ultimately lead to transportation systems that are safer and more efficient in terms of their operations. Object recognition in complicated road settings is much improved by the suggested multi-scale adaptive feature fusion technique compared to previous approaches. This is achieved by merging features gathered at different sizes. The method guarantees a more thorough

comprehension of spatial and semantic information for overlapping, tiny, and faraway objects, improving detection accuracy. The model can also generalize better across various traffic circumstances than traditional models since it uses sophisticated data augmentation techniques. The ability to process data in real-time is another way it differs from more conventional methods. The suggested detection algorithm can distinguish between various items seen in road scenes, such as people, road signs, lane markings, automobiles, buses, trucks, and scooters. The technique can capture objects of diverse sizes, shapes, and distances using deep neural networks and multi-scale adaptive feature fusion. Complex urban and traffic scenarios are ideal for identifying small, distant, overlapping, and obstructed objects. The multi-scale strategy trains the model to extract characteristics from many resolutions and scales to improve the detection algorithm's handling of objects of different sizes, shapes, and distances. The network uses multi-scale feature extraction and fusion to combine low-level spatial characteristics with high-level semantic information. This synergy ensures the accurate detection of local and distant objects and overlapping and veiled components, making the model reliable in complex traffic scenarios. Adaptive feature fusion highlights important data while reducing redundant data to increase detection accuracy across settings.

The adaptive feature fusion technique in the proposed IMSAFFT model allows for successful recognition in different and complicated traffic settings and enhances detection across many scales. It analyses traffic in real-time using a deep neural network and a novel data augmentation mechanism, setting it apart from traditional models. This synergy guarantees precise detection and classification, a huge leap forward for autonomous driving systems in theory and practice.

The main contributions of the article

1. An enhanced multi-scale adaptive feature fusion approach, IMSAFFT, manages overlapping, distant, and microscopic objects in difficult road situations with robust performance, boosting processing speed and object recognition accuracy to 98.4%.
2. An adaptive learning rate modification method optimizes training and enables real-time traffic analysis with accurate semantic recognition.
3. IMSAFFT outperforms traditional models in F1 scores, accuracy, and prediction rates for intelligent transportation systems and autonomous automobiles in real-world traffic scenarios.

The rest of the article is planned: [Section 2](#) discourses the related work, [Section 3](#) recommends the MSAFFT model, [Section 4](#) reflects the simulation findings, and [Section 5](#) concludes the research article.

## 2 Related work

Ryo Hasegawa et al. [13] suggested Convolutional Neural Networks (CNN) for Road Sign Detection and Recognition in Complex Scenes. Variations in size and contrast

significantly impact the accuracy of traffic sign identification and recognition. This issue is addressed in this work by the author's use of a DL approach that is resilient against scale changes to learn road signs. The author experimented to compare the strategy to previously suggested deep learning algorithms. Individual Japanese traffic signs are used in this research to test the suggested approach. Compared to Faster R-CNN and SSD, the suggested solution outperforms both when detecting and identifying road signs. Zhenfeng Shao et al. [14] suggested the Multi-task Road-related Extraction Networks (MRENet) for Simultaneous Road Surface and Road Centerline Extraction in Complex Urban Scenes from Very High-Resolution Imagery. To increase data abundance, include multi-level characteristics, and broaden the network's receptive field, the author employs atrous convolution and pyramid scene parsing pooling modules (PSP pooling) in the configuration of the network. Compared to approaches based on visual interpretation and comparative classification precision, the suggested algorithm performs better in experiments.

Lingzhi Shen et al. [15] recommended the YOLOv3 (You Only Look Once v3) with feature map cropping for multi-scale road object recognition. The K-means-GIoU (Generalized Intersection Over Union) method aims to produce a priori boxes with forms closely resembling actual boxes. The training process becomes more straightforward, allowing for rapid convergence. Subsequently, an identification branch is integrated to identify a small target; feature map cropping modules are inserted into this branch to eliminate regions with a high likelihood of background and easily detectable target. Experiments conducted on the KITTI dataset demonstrate that the suggested approach outperforms YOLOv3-paralytcs regarding small-scale object recognition while keeping detection speeds high and increasing mAP (mean average precision) values by a maximum of 2.86. Tsz-Yeung Chow et al. [16] discussed the conditional generative adversarial networks (CGAN)-based dehazing model for identifying targets in road scene imagery. Several picture dehazing datasets were used for comparison. By at least 2 dB in PSNR, the suggested model surpassed competing deep learning-based and hand-crafted picture dehazing techniques. By comparing the degraded and improved images, the author revealed that the suggested dehazing model produced noticeably better results for object detection.

Fan Yang and Yutai Rao [17] deliberated the Vision-Based Intelligent Vehicle Road Recognition and Obstacle Recognition Technique. The focus of this study is on how intelligent vehicles perceive their surroundings. It investigates the current issues with road identification and obstacle identification algorithms, such as road picture segmentation, vanishing point detection, and binocular vision-based road sceneries. The research presented here includes technologies for both 3D reconstruction and obstacle detection. Retaj Yousri et al. [18] presented the DL-Based Benchmarking Framework for Lane Segmentation in Dynamic and Complex Road Scenes. An automated segmentation algorithm was first tested using conventional computer vision methods. Using the nuScenes dataset's complicated urban imagery, our algorithm accurately

segments the host lane's semantic zone and creates matching weak labels. Next, five cutting-edge FCN-based models are trained and benchmarked utilizing the produced data, which is then assessed qualitatively. Among the models tested in complicated road settings with dynamic situations and different lighting conditions, ResUNet++ demonstrated the most robust performance in the output results.

Tianmin Deng et al. [19] introduced the Multi-Scale Hybrid Attention Mechanisms (MSHAM) for Occluded Vehicle Recognition in Road Scenes. Grouping convolution of various-sized multi-scale feature extraction networks improves multi-scale features; combining spatial attention modules with parallel connection channels forms various scale hybrid domain attention modules, enhancing occluded vehicles' local feature data, enabling multi-scale feature reinforcement learning and suppressing occlusion interference information. The experimental findings demonstrate that this technique outperforms the baseline network YOLOv5 by 1.5% and 2.9%, respectively, with an average mean accuracy of 95.2% and 59.3%. Lindong Tang et al. [20] offered the Highly Robust YOLO Networks (HRYNet) for Complex Road Traffic Object Recognition. First, dual fusion gradual pyramid structures (DFGPN) are presented. This structure uses two-stage gradient fusion strategies to improve the production of more thorough multi-scale high-level semantic data, as well as to strengthen the interconnection and decrease the data gap between non-adjacent feature layers. The residual multi-head self-interest mechanisms (RMA) is a brand new module in HRYNet that extracts features while stopping interference. LHRYNet progressed 6.7%, 10.9%, and 2.5% on all three datasets in a mean average precision of 0.5, proving it's far advanced to YOLOv8s.

Zhen Zhu et al. [21] investigated the polarimetric object detection benchmarks (PODB) for road scenes in antagonistic climate conditions. By integrating polarimetric imagery, the PODB gives a complete framework for evaluating the quality of DL-primarily based item popularity algorithms in complex road situations. In addition, the author validated and evaluated the performance of powerful benchmark algorithms by engaging in good-sized object detection tests using the PODB. In addition, a model for a multi-scale picture fusion cascaded item identification NN is suggested primarily based on bodily standards. An adaptive study of the multi-selection object identity NN version was used with the polarized images to enhance the prediction accuracy of complex street scenes in poor climates by about 10%. Arnav Vaibhav Malawade et al. [22] suggested the roadscene2vec for extracting and embedding road scene graphs. By offering tools for creating scene graphs, graph learning models to produce spatio-temporal scene-graph embedding and tools for viewing and evaluating scene-graphs-based techniques, roadscene2vec aims to facilitate study into the uses and abilities of road scene graphs. The author exhibits the practicality of roadscene2vec in various applications by presenting experimental findings and qualitative assessments of graph learning and CNN-based models.

### 3 Multi-Scale Adaptive Feature Fusion Technology (MSAFFT)

The tracking and monitoring of traffic vehicles is gaining more and more attention from city management today due to the rising popularity of automobiles. Accurate vehicle detection aids in urban traffic management since the positioning and nature of vehicles substantially impact the overall traffic environment in traffic situations. Autonomous driving technology relies on dependable environmental awareness to rapidly gather data about the vehicle's environment and generate predictions based on this data, including detecting road traffic signs. Autonomous driving cars' security is affected by the ability to identify road traffic signs. Detecting them using traditional research methods that rely on traffic signs' geometric and colour features is a laborious and error-prone process vulnerable to environmental factors like weather changes and occlusions. Vehicle identification technology has come a long way thanks to deep convolutional neural networks (CNNs) and their powerful feature learning capabilities, which are extensively used in computer vision. However, some issues with practical implementations regarding real-world traffic situations still exist. Conventional approaches often focus on identifying a single vehicle at a time. Nevertheless, several vehicles might be targets in situations with complicated traffic or on the battlefield. In the real-life traffic image, the vehicle's size ranges from small to gigantic, and the distance between objects changes. Furthermore, researchers occasionally only partially know the vehicle's shape, contour, and colour due to vehicle shielding, combat camouflage, or bad weather. This is a significant obstacle when trying to identify vehicles. Hence, this study proposes the Improved Multi-Scale Adaptive Feature Fusion Technology (MSAFFT) for real-time traffic detection from complex road scene video images.

Several meticulous stages are involved in the pre-processing of incoming photos to improve the object detection process's accuracy and efficiency:

1. **Colour Space Conversion:** The input photos transform from the RGB colour space to the HSV colour system, which stands for Hue, Saturation, and Value. Thanks to this modification, their distinct colour may more easily identify road signs, cars, and other features, which helps separate colour attributes like hue.
2. **Reducing Noise:** Morphological filters effectively remove photo noise, including opening and closing. These actions preserve the forms of more oversized, significant items while eliminating small, unnecessary features.
3. The photos are segmented into "regions of interest" (ROIs) based on the likelihood of the appearance of essential characteristics, such as automobiles and road components. By narrowing the focus, this segmentation improves the model's performance and decreases the number of false positives.
4. The input photos are downsized to match the anticipated dimensions of the model. This ensures consistency in the training and assessment processes. Normalizing the pixel values to fit within a standard range reduces calculation time and speeds up model convergence during training.

Figure 1 shows the proposed IMSAFFT model. The first step is to change the colour space of each original frame from RGB to hue, saturation, and value (HSV). Pixels may be assigned a specific colour, such as red or blue, by mapping their values using the component colour wheel in HSV. After that, the noise is removed using morphological filters, which are the opening and closing processes. Signs indicate the remaining categories of positive detection. For each recognized grouping in the image, the labelling algorithm finds a square block that may represent a road sign. Certain areas are either too large or too small. Because of this, there may not be signs directing traffic. Therefore, it is necessary to determine the applicants' height and length to exclude those not qualified. The input frame must be pre-processed and converted to a new colour space before processing can begin. Image segmentation for ROI extraction is the second step. Colour separation in the HSV colour space is used to accomplish the segmentation. Given four distinct geometric shapes: circle, triangle, rectangle, and octagon, the ROI retrieves shape information. This study suggests a multi-scale adaptive fusion network to enhance the algorithm's vehicle recognition resilience at multiple sizes, increase the variety of road lane features, and enable detection efficiency while changing distance and perspective. To identify small vehicles using visual and spatiotemporal feature data, the deep Neural Network (DNN) approach employs a deep segmentation network to glean insights from the mutually beneficial interaction between roadways and vehicles. The suggested approach fully characterizes the congestion state from multi-dimensional feature spaces, encompassing speed, traffic density, flow, and occupancy, instead of characterizing traffic congestion regarding a single feature dimension, as most conventional techniques ensure. These dynamic and static multi-dimensional visual characteristics complement and interact with one another to provide a more accurate depiction of traffic congestion in this research. The target image is matched with reference images, and the reference with minimum judgment value is chosen. There could be an uneven distribution of target groups in traffic circumstances. To fix this, data augmentation may provide additional samples for less frequent categories or change the weights of the classes.

The detection approach uses a multi-scale feature extraction mechanism to handle the difficulty of different sizes in complicated road situations. The model can successfully detect objects of varying sizes since it analyses characteristics at varied resolutions. Incorporating adaptive learning rates facilitates training by allowing the system to zero in on the finer details of both nearby and faraway objects. The suggested technique improves object detection under challenging occlusion or poor visibility by integrating contextual information across many dimensions.

A multi-scale feature fusion technique is integrated into the suggested model to overcome illumination and occlusion issues. This mechanism efficiently recovers from occlusion by capturing features across multiple resolutions. Adaptive learning and contrast enhancement techniques ensure further effective object recognition in varied visibility circumstances to minimize the impacts of variable illumination conditions.

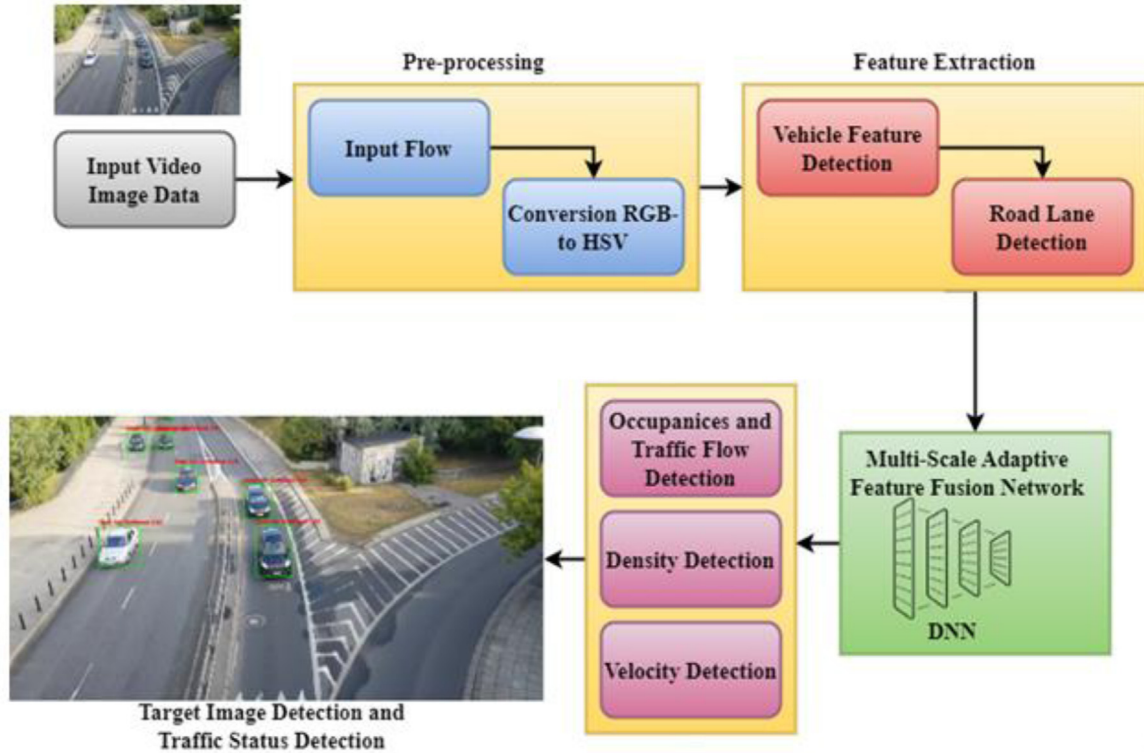


Fig. 1. Proposed IMSAFFT model.

The proposed approach utilizes a multi-task loss function that combines classification and regression loss to optimize efficiency. Using a softmax loss function to categorize things like cars and buses ensures accurate identification. The Smooth L1 Loss for bounding box regression decreases the difference between predicted and ground-truth bounding box coordinates, improving localization accuracy. Together, they balance training to recognize and place things effectively.

The model's many processes coordinate their efforts to lower the false positive and negative rates. First, we use the multi-scale feature extraction and fusion method, ensuring robust detection by combining characteristics at different resolutions to identify objects appropriately. Secondly, an adaptive learning rate improves the training process and helps the model comprehend the minute differences between objects and their backgrounds. Third, the model improves its object and non-relevance-differentiation capabilities by including contextual information across many dimensions. Data augmentation strategies enhance the model's resilience to environmental changes, such as lighting and occlusion, further reducing false detections. Regularisation techniques such as batch normalization and dropout are utilized to prevent overfitting.

### 3.1 Road occupancy detection

This research employs road extraction by monitoring from a starting location on the road since roads are linked as complete networks, a distinctive feature of the images. Automatically and accurately adjusting the tracker's

turning angle  $\varphi$  according to the current environment is the main challenge of tracking. Suppose the trackers are in the current position  $(j, i)$  with rotation angles  $\theta$ . A local picture patch  $J$ , centred at the position with the rotation angles, is then extracted as the present atmosphere of trackers. The critical issue of monitoring the target image is then articulated as neural networks  $f(y)$  that learn a likelihood distribution as demonstrated as

$$f(y) = q(\varphi|\Phi(J(j, i, \Delta z))) \quad (1)$$

As shown in equation (1), where  $\Delta z$  indicates dynamic window sizes, and  $\Phi$  denotes the function of pre-processing.

This research presents a DNN classifier based on the Alexnet structure to enhance the accuracy of foreground identification of road traffic targets. During DNN training, the input picture patterns are ranked using the trained neural network's scoring system for the input candidate foreground objects. To find out which category the item in the foreground belongs to, we compare the scores of each category to the highest score. Vehicles, scooters, people, and outliers are the usual categories into which the items in this research fall. The 2D imaging property states that in a captured frame, a scooter's attributes closely resemble a pedestrian's. Consequently, this research considers scooters and pedestrians in a similar group. Consequently, the overall object type  $D$  can be described by:

$$\begin{aligned} D_1 &= \{c|c : vehicles\} \\ D_2 &= \{c|c : pedestrians\} \\ D_3 &= \{D_1 \cap D_2\} \end{aligned} \quad (2)$$

The DNN built in this research comprises 11 core network layers, comprising five convolutional layers, three fully-connected layers and three pooling layers. It comprises SoftMax layers and input layers. After obtaining the foreground sets  $C$  in the DNN stage, the feature extraction progression for traffic flow and road occupancy can be determined concurrently. When analyzing the traffic condition recognition process, it is seen that when the road is congested, there is a considerable rise in the occupancy of vehicles and pedestrians. Therefore, this study considers traffic occupancy as a measure of traffic condition. This research defines road traffic occupancy as the ratio of vehicles using a particular road segment during a given time frame. There are two main ways to divide traffic occupancy. One relies on counting pixels, while the other depends on area estimates. In the first case, this study computes the overall number of pixels in the foreground object by the total number of pixels in the road in the background, excluding any foreground target. The number of foreground pixels indicates the space objects in the foreground that take up on the road relative to the background. To be more accurate, the likelihood of congestion increases as the number of foreground objects grows, which is why the road occupancy ratio rises.

Moreover, the interest objects (i.e., both  $D_1$  and  $D_2$  types) in road traffic can be preoccupied geometrically into a rectangle comprising scooters, pedestrians, and vehicles. Therefore, a region-based approach to road occupancy would be more practical than pixels. An area-based technique takes the related areas of each foreground object and finds their Minimum Enclosing Rectangles (MER) to determine the location of every foreground object. The region of the MER indicates the amount to which the foreground item fills the road while the road in the background is fixed. The road occupancy ratio, in particular, rises as the total MER areas increase.

Therefore, road traffic occupancies  $\sigma$  for processing speed is provided by the subsequent expression (3):

$$\sigma = \frac{\left(\sum_{j=1}^n W'(c_j)\right)}{W} \quad (3)$$

As inferred from equation (3), where  $W'(c_j)$  denotes the MER region of the  $j$ th foreground  $c_j$ ,  $W$  is the region of the identified road in images, and  $n$  specifies the number of foreground objects.

This research employs a weighted smoothing approach for dynamic video processing to enhance the present occupancy values. This is necessary since the road occupancy acquired by the DNN depends on recognition outcomes and the network samples' incoming frame once per second. Based on current  $\sigma_t$  and occupancies estimation from the prior 3 seconds, the enhanced occupancies can then be articulated as

$$\sigma_t = \omega_t \sigma_t + \omega_{t-1} \sigma_{t-1} + \omega_{t-2} \sigma_{t-2} + \omega_{t-3} \sigma_{t-3} \quad (4)$$

As discussed in equation (4), where  $\omega_t$ ,  $\omega_{t-1}$ ,  $\omega_{t-2}$ ,  $\omega_{t-3}$  represent the weight of the current occupancy  $\sigma_t$  and weight of the occupancy from the previous 3 seconds,

correspondingly, and the weight fulfils  $\omega_t + \omega_{t-1} + \omega_{t-2} + \omega_{t-3} = 1$ . Since road status is a time-based feature, the link between present features and their values is stronger for recently detected features and weaker for features discovered in the past. Because of this, the weights in our studies were 0.49, 0.33, 0.17, and 0.01, in that order, as determined by this research.

Figure 2 shows the Structure of the Target Image Detection Algorithm. For targeting, the detection and tracking fusion technique includes a fusion module, correlation filter tracking for video sequences, and deep learning detection for individual image frames. Initially, the research employed a DNN model to identify objects in the video. Then, it continually tracked the surface target using an SSD. Once the tracking process reaches a certain number of frames, the detection mechanism is reintroduced. The output of the target frame is then finalized using a machine learning-based feature fusion approach. Any new targets that come into view are initialized and tracked simultaneously. Improved bounding box placement effect and faster model convergence were the goals of this study's modifications to the bounding box loss, which allowed for more accurate regression of the bounding boxes.

### 3.2 Traffic flow detection

To enhance the accuracy of road congestion status, the visual feature of traffic flows may help recompense for the unique situation mentioned earlier and reduce the bias in road occupancy measurements, leading to erroneous congestion detection. The term "traffic flow" describes the quantity of traffic that moves across a particular stretch of road during a given time frame. This research obtains road traffic flow and occupancy using a DNN that samples incoming frames once every second. After the DNN stage is complete, the foreground object may be quantitatively examined. In other words, more engaging foreground items are likely to have a lot of traffic, while less interesting ones will have less, and vice versa.

By eliminating erroneous objects identified by the Gaussian mixture model (GMM), our model obtains the appropriate objects based on DNN object recognition in this work. Post-processing uses each final foreground morphological feature to get the MER of every foreground object to create a linked region. Next, this study utilizes a bounding box to signify the MERs of the linked foreground. The number of automobiles throughout the estimated time for traffic flows increases by one for each added bounding box in this model's observed frame. Based on the conversation above,  $l$  signifies the traffic flow features.

$$l = |c_j \in C| \quad (5)$$

As found in equation (5), where  $C$  and  $|\cdot|$  signify the last foreground object sets assimilated during the DNN stage and the number of set components correspondingly. This study uses a weighted smoothing technique to enhance current traffic flows further. Based on the present  $l_t$  and traffic flow estimation from the prior 3 seconds, the

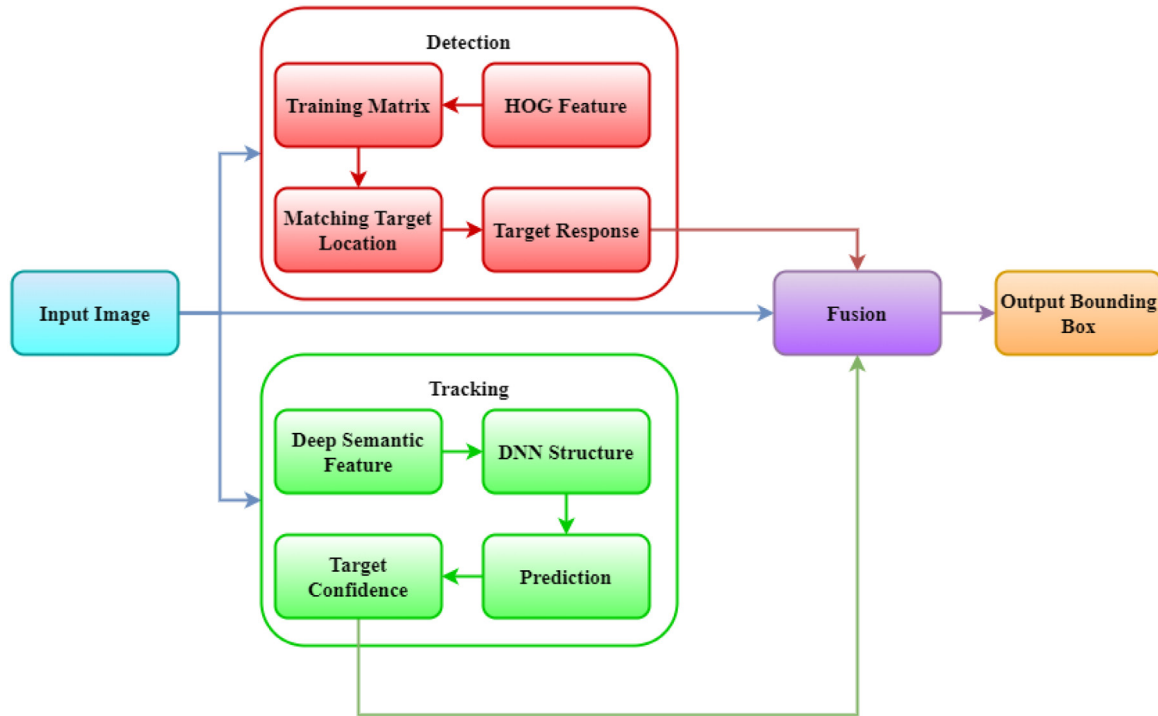


Fig. 2. Structure of target image detection algorithm.

improved traffic flow can then be expressed as

$$l_t = \omega_t l_t + \omega_{t-1} l_{t-1} + \omega_{t-2} l_{t-2} + \omega_{t-3} l_{t-3} \quad (6)$$

As deliberated in equation (6), where  $\omega_t$ ,  $\omega_{t-1}$ ,  $\omega_{t-2}$ ,  $\omega_{t-3}$  represent the weights of the current  $l_t$ , and the weight of the previous 3-second traffic flows can be set to 0.48, 0.32, 0.16, 0.01.

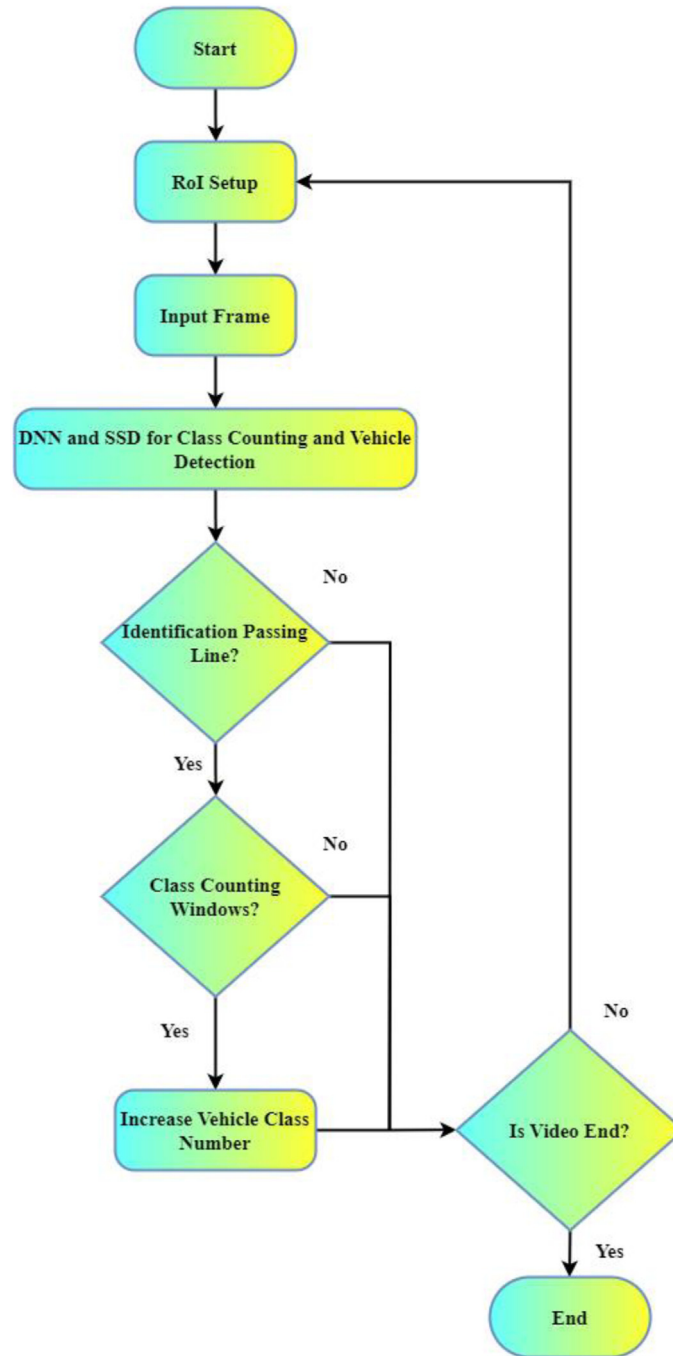
Figure 3 shows the flow chart of vehicle identification and class counting. Initially, using the road video as input, the ROI setup can find the forward-passing and backwards-passing objects via the object detection window. The second step is identifying the cars from the current frame utilizing the suggested Single Shot MultiBox Detectors (SSD) DNN model. Cars, buses, and trucks are the main categories into which these objects fall. Finally, vehicle classification is determined according to the configuration and operation of the ROI windows. The flow diagram shows that this procedure will be repeated for vehicle detection and class counting if the input video does not terminate. In this case, the steps will display the outcome once they finish. The passing detection lines monitor the vehicles' ability to pass through the detection lines. The responsibility for vehicle class counting is consistent with the class counting windows. In both instances, the class counting windows are positioned behind the passing detection line, and an object passes through them. Many parameters become incorrect and inefficient during the continuous convolution operations of deep neural networks as the weight values approach zero.

The suggested model's feature fusion process uses many well-thought-out methods to prevent the introduction of redundant information. The first step is to use a multi-scale

feature extraction method to ensure the model gets all the correct information at different levels, locally and globally, without duplicating any steps. Secondly, attention processes minimize redundancy by prioritizing important traits and suppressing less helpful ones. Third, features are fused across layers using pixel addition and convolutional procedures to remove redundancy and distribution inconsistencies instead of cascade fusion, which typically results in dimensional expansion. Lastly, during training, the model's capacity to concentrate on essential features is further improved by using batch normalization and regularisation approaches, including dropout layers, to prevent overfitting and provide feature relevance. Taken as a whole, these techniques guarantee a smooth and effective fusion procedure.

### 3.3 Deep neural network

DNNs are the subject of this research, which trains a model to identify road directions given just their input data. The DNNs become deeper (with several layers) and broader (with several maps) due to several beneficial features, such as receptive fields and max pooling. The structure of a DNN consists of an input layer with  $M \times M$  neurons, several max-pooling layers after each, convolution layers, and fully connected layers at the far end. In between the convolution layers are the max-pooling layers. The basic max pooling approach discovers the best-performing neurons in convolution layers by choosing the most active neurons from every sector of quadratic zones of local inhibition. The winning layer feeds into the subsequent layer up the hierarchy from



**Fig. 3.** Flow chart of vehicle identification and class counting.

lesser, down-sampled layers with lower resolutions. In convolution layers, feature vectors are reduced in dimension by pooling, a sub-sampling.

Figure 4 shows the multi-scale adaptive feature fusion model using DNN architecture, considering the target identification layer's dominance and other layer features. This article uses pixel addition to multi-scale feature fusions and the features of fused  $3 \times 3$  convolutional to decrease feature fusions after stack effects, as opposed to cascade fusion features, which quickly expand the dimension problem. Information on feature maps at all levels is

fused, and the discrepancy in feature distribution among them is erased. For deep learning to produce more accurate and resilient models, massive amounts of data are required for complete feature extraction and learning about the target. Insufficient knowledge of the network model makes convergence difficult, or an over-reliance on the available data, which makes the network rigid, occurs when the data is too little. Additionally, for the model to avoid performing poorly on particular objects while performing admirably on others, high generalization is required for target identification tasks. To ensure that the model properly learns for all

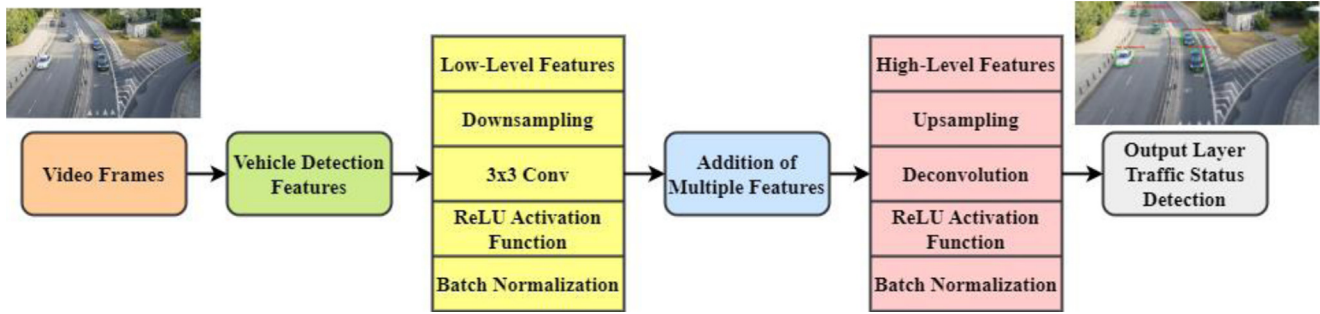


Fig. 4. Multi-scale adaptive feature fusion model using DNN.

object categories, it is recommended to have an evenly distributed amount of examples for each category. An approach to data enhancement that addresses this block's shortcomings is diversity in data generation, which improves data utilization. For instance, a deep learning neural network can learn the typical semantic data of target objects and object recognition colour information through random chromaticity adjustment. The suggested technique determines the input's confidence probability of being a traffic sign by using batch normalization, ReLU activation, and a softmax layer. Following the fully connected layer, batch normalization and a 0.5 dropout enhanced the model's stability and training speed.

More important for the deep architecture is the max-pooling layer. The max-pooling layer uses quadratic areas of local inhibition to reserve the most sensitive neurons with higher activation values. This approach is based on the hypothesis that such neural activation patterns are likely used by human brains, as supported by biological research. Because each layer's winners flow into the one below them, the feature input is independent of the layer transitions. Several well-known data sets have shown that DNN can successfully use this pooling method to extract deeper features, such as hand digit identification and face recognition.

This research maps all pixels to the input layer before training and pre-processes the picture patches using the abovementioned methodologies. According to equation (7), the input to a convolutional layer neuron in forward propagation is the weighted sum of a rectangular data segment from earlier layers. The convolution filter  $\delta$ 's initial value, derived from uniform random distributions in the ranges  $[-0.05, 0.05]$ , is shared by all rectangular cells in the same layer. In most cases, the activation function is not linear. The problem's size is reduced by adding sparseness using max-pooling layers. In our networks,  $k \times k$  neuron blocks are decreased to single values. The output of ultimately linked layers relates to the accurate class, i.e. the 5 pre-described way categories of the road.

$$y_{j,i}^k = \sum_{b=0}^n \sum_{a=0}^n \delta_{b,a} x_{(j+b)(i+a)}^{k-1} \quad (7)$$

As shown in equation (7), where  $y_{j,i}^k$  denotes the inputs for the  $(j, i)$  neurons of the  $k$ th layer,  $x_{j,i}^k$  symbolizes activation values linked with  $y_{j,i}^k$ , and  $\delta_{b,a}$  denotes the convolution filters of  $n \times n$  dimension for 1 layer.

In the backward propagations, error functions  $E$  are described amongst the outputs and the ground truths. The objective of the training is to minimize it by optimizing the weights matrices  $W$ . The weights are rationalized with the rules:

$$W = W - \frac{\eta \partial E}{\partial W} \quad (8)$$

After the networks have been trained, deeper features may be extracted from road environment picture patches using hierarchical learning, hidden layers with max-pooling, and convolution.

This study employs a multi-task loss  $P(y, d, k, h)$  on every training batch to jointly optimize manifold vehicle classification and offset regression from the suggested deep network as subsequent functions:

$$P(y, d, k, h) = \frac{1}{M} (P_{conf}(y, d) + \beta P_{localization}(y, k, h)) \quad (9)$$

Expression (9) shows where  $h$  symbolizes the ground truth boxes, and  $k$  specifies the predicted boxes.  $y$  denotes whether the matched box (between the ground truth boxes and default boxes) belongs to a category  $q$ , and  $d$  designates the confidence values if  $y$  is 1. If the default box is negative,  $y$  is 0. The confidence losses  $P_{conf}$  denotes the softmax losses over four classes (i.e., van, bus, car, and others). The SSD method has been fine-tuned to increase its detection speed, and it now uses neighbouring features to fuse with its target detection layer. The use of multi-scale feature correlation and level-specific feature complementarity is maximized. The road scene target detection data set was used to feature fuse the detection layers utilizing the transfer learning principle. An experiment comparing DNN and SSD was additionally carried out. The suggested IMSAFFT model increases the accuracy ratio, target image detection ratio, traffic prediction ratio, processing speed rate and F1-score ratio compared to traditional models.

The suggested detection technique uses pre-trained weights to extract features from DNNs using transfer learning effectively as part of its training process. The annotated dataset, which contains the location of target items in road scenes and their bounding boxes, is used to fine-tune the model. The training uses softmax for object classification and regression; it uses Smooth L1 Loss, making bounding box changes. This multi-task loss

function is then applied to the data. Accelerated convergence in the beginning and steady optimization in the end are achieved through an adaptive learning rate scheduler, which dynamically changes the learning rate. Adam (Adaptive Moment Estimation), a well-known optimization technique, guarantees practical training across different object sizes. Adam can manage sparse gradients and dynamic learning rates.

## 4 Simulation analysis

This study presents the Improved Multi-Scale Adaptive Feature Fusion Technology (IMSAFFT) for a real-time traffic data identification method to fix the issues of low detection accuracy of road scenes and high false detection rates in panoramic video images. This dataset [23] aims to identify vehicles in traffic scenes (images with dimensions  $1080 \times 1920 \times 3$ ). This dataset was created using video footage freely accessible on YouTube, particularly a video licensed under the Creative Commons Attribution license. The datasets used to train and test the model consist of images at a resolution of  $1080 \times 1920$  pixels, selected from traffic footage freely available under the Creative Commons license—annotations, such as bounding boxes, aid vehicle recognition in traffic scenarios. The dataset primarily focuses on metropolitan road scenes to guarantee applicability to real-world autonomous driving applications. The 10,000 annotated photos comprising the training dataset are taken from  $1080 \times 1920$  pixel traffic footage available to the public. The locations of vehicles, primarily automobiles, are shown by labelled bounding boxes in each image. Scenes include residential streets, highways, urban junctions, and heavy traffic areas. Datasets include various lighting circumstances, including daylight, night, dusk, and inclement weather (rain and fog). Twenty thousand additional synthetic variants were generated using flipping, rotating, scaling, cropping, and brightness correction techniques. These strategies enhance the model's adaptability and robustness, guaranteeing that the dataset includes all possible traffic and environmental situations. Several data augmentation strategies were used during training to make the model more generalizable and resilient. Some are geometric transformations for inverting, rotating, and scaling objects to make them size-and-orientation-invariant. Contrast and brightness controls make it seem like daytime, midnight, or bad weather. Using cropping and padding, varying item sizes and partial occlusions are managed. The saturation and hues in the HSV colour space are changed to make colour identification more reliable. Synthetic variants imitate real-world road conditions by adding noise and random distortions.

In this study, vehicles on the road are the identified targets for detection. The algorithm's primary goal is vehicle detection and localization in complicated road sceneries. Such scenarios may have obstacles, including occlusions, changing illumination, and heavy traffic. Enhancing vehicle identification accuracy, improved multi-scale adaptive feature fusion technology (IMSAFFT) uses deep neural networks (DNNs) to manage the complexities of real-world road surroundings.

**Table 1.** Experimental setup.

Experimental setting	Configuration
Operating environment	Windows 10
CPU	Intel(R) Core(TM) i7-11700 CPU @2.50 GHz
Memory	16G
Programming Languages and Learning Framework	Python 3.6.8 and Tensor Flow 1.13.1
GPU	NVIDIA GeForce RTX 3090
Compiling Tool	PyTorch 1.11.0
LearningRate	0.001

Table 1 shows the experimental setup.

### Annotation details:

- **Classes:** The dataset is focused on car object detection, and car objects are labelled as the target class (aka one class only)
- **Bounding Boxes:** Each image frame contains annotated bounding boxes around car objects, marking their locations in the scene.
- **Annotation files** are provided in YOLO text format.
- **Object Detection:** This dataset can be utilized to train and assess object detection models, emphasizing detecting cars in road traffic scenarios.
- **Data Format:** Images are provided in JPEG format.

To ensure its adaptability, the model is trained and tested on various pictures from varied landscapes, traffic types, and weather circumstances (e.g., daylight, night-time, rain, and fog). Data augmentation helps the dataset be more diverse and reflect real-world volatility. Transfer learning with pre-trained weights allows the model to adjust to new scenarios quickly. We also test the model on unknown datasets and real-world events to confirm it is durable and functions well beyond the training set.

### 4.1 Accuracy ratio

With vehicle object recognition being a distinct object detection task, researchers are concentrating on two fronts: computer vision and hardware. When it comes to the equipment, it's all about the sensor system, which is responsible for gathering efficient traffic data and processing and reporting it. Despite that, sensor manufacturing is not scaleable due to its high costs. Within computer vision, algorithms are used to process visual data, provided an effective algorithm is developed to enhance detection accuracy and efficiency. That it may find widespread use, this work proposes a multi-dimensional detection model for traffic surveillance videos that improves the accuracy of foreground-object recognition. It used information entropy-based Histograms of Optical Flow (HOF) for complete visual feature analysis, making it possible to identify traffic congestion more efficiently. According to quantitative and visual assessment findings, the suggested model attains

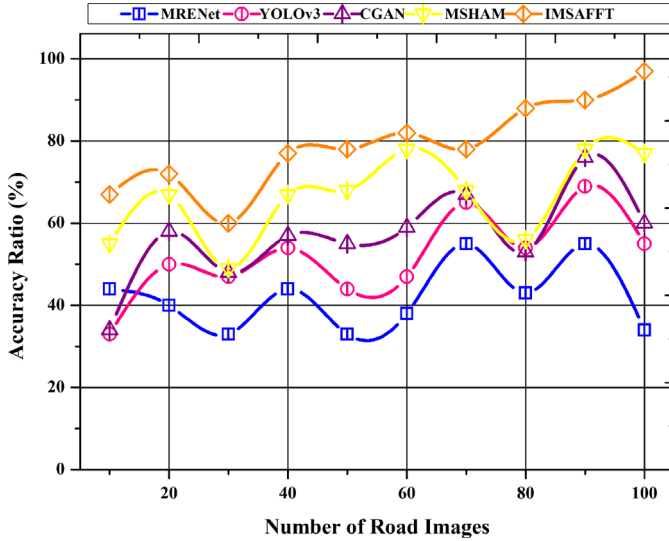


Fig. 5. Accuracy ratio.

better outcomes when evaluating traffic surveillance footage than the present approaches for detecting traffic congestion. When working with high-resolution images captured from complicated road scenes using the DNN model, object detection based on Region of Interest (ROI) may significantly improve execution speed and accuracy by excluding irrelevant regions during processing. Based on equation (7), the accuracy of traffic category prediction has been obtained. Figure 5 shows the accuracy ratio.

#### 4.2 F1-score ratio

The F1 score, determined by recall and precision, is used to evaluate lane border point classification. Recall is the percentage of positive ground truth images discovered in the sample, whereas precision is the fraction of all positive instances detected.

$$Precision = \frac{1}{m} \sum_{l=1}^m \frac{TP_l}{TP_l + FP_l} \quad (10)$$

$$Recall = \frac{1}{m} \sum_{l=1}^m \frac{TP_l}{TP_l + FN_l} \quad (11)$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (12)$$

As shown in equations (10), (11), and (12), where  $TP_l$  denotes the number of adequately categorized frames in the  $l$ th condition (i.e., the  $l$ th condition can be slow, free or congested);  $FP_l$  signifies the number of frames that are miscategorized to  $l$ th condition;  $TN_l$  indicates the number of adequately categorized frames not in  $l$ th condition and  $TN_l$  specifies the number of frames in  $l$ th status yet

miscategorized as not in  $l$ th condition. Here,  $n=3$  represents the 3 types. Figure 6 demonstrates the F1-score ratio.

#### 4.3 Target vehicle detection ratio

For the complex interaction features of the target vehicle, since it is difficult for dynamics and kinematics methods to represent the interaction behaviour fully, it is necessary to extract the temporal and spatial correlations in the data using deep learning-based methods. First, a unique framework integrating a lane tracker approach with cameras and radar forward vehicle tracker systems has been developed for this research to achieve strong road scene understanding. This framework is conducive to heavy traffic. This research proposed an occupancy-matching image template with an embedded vehicle tracker to prevent extracting unneeded lane features generated by forward-target vehicles and road markings. Second, the study shows that a tracking system considering neighbouring and ego lanes can identify several highly reliable lanes. It is essential to determine the collision risk of each target before choosing the target vehicle. Based on equation (1), the target vehicle has been identified. Figure 7 shows the target vehicle detection ratio.

#### 4.4 Processing speed ratio

The processing speed impact seems statistically significant when an object is compatible with the scene context, yet the accuracy effect is usually less. However, processing speed and accuracy are affected when an object's colour or shape differs from the background. There is a significant decrease in accuracy (around 13.9%), and the delay in background processing caused by salient and non-congruent items may approach 81 milliseconds. There is a processing speed of around 25 fps. Even if it's a little slow, it's still fast enough for a real-time processor and an autonomous driving system to optimize the suggested method. All four target lanes were regarded within 35 m of each other out of 3350 evaluated frames. Based on the equation (3), processing speed has been identified. Figure 8 shows the processing speed ratio.

With an average processing speed of 25 frames per second (fps), the model is highly effective in real-time applications. Because of this, driverless cars and traffic monitoring systems may benefit from the fast and accurate object detection they provide. Using efficient algorithms and hardware acceleration enhances the model's real-time applicability, facilitating seamless decision-making in ever-changing contexts.

The model's computational efficiency is measured by average inference time per frame and real-time fps. These metrics show the model's processing speed and real-time suitability for autonomous driving and traffic monitoring. Profiling tools may assess GPU, RAM, and latency in the detection pipeline phases. The model's design is improved by adding lightweight components and removing unnecessary layers to maximize efficiency and accuracy.

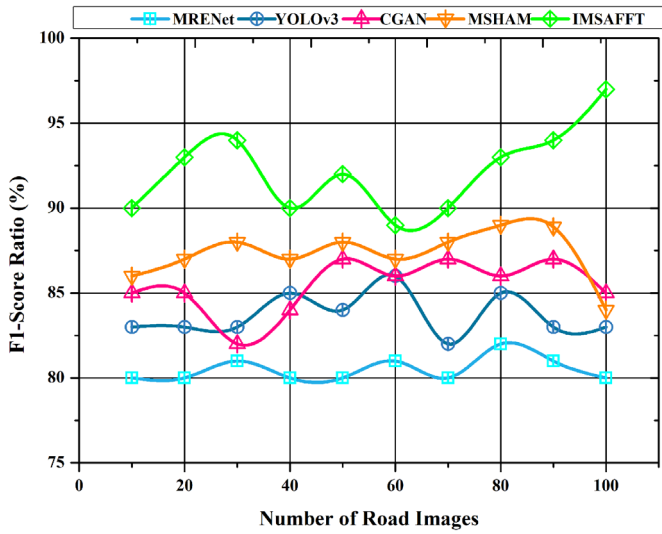


Fig. 6. F1-score ratio.

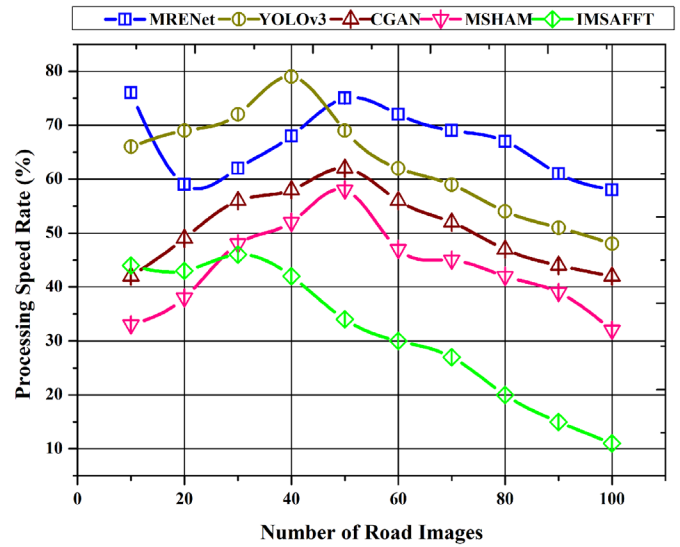


Fig. 8. Processing speed ratio.

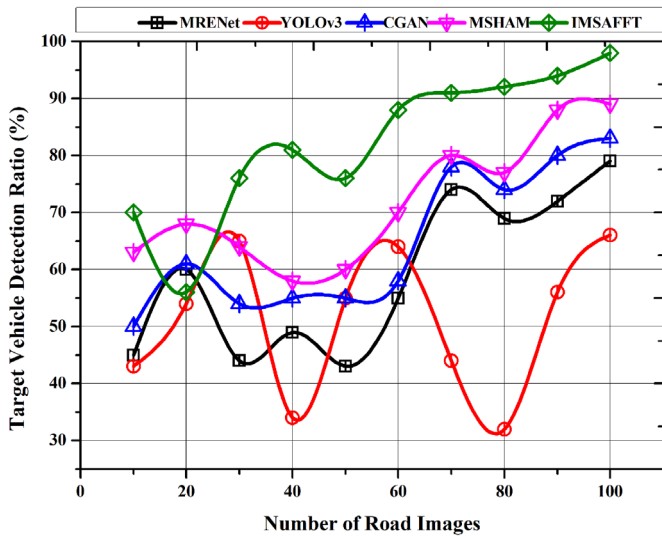


Fig. 7. Target vehicle detection ratio.

### 4.5 Traffic prediction ratio

Urban traffic efficiency improvements, including traffic direction and route planning, are an efficient method of reducing traffic congestion. Traffic prediction aims to predict the future of traffic conditions by analyzing historical data and the present state of traffic. Both data from fixed detectors and data from floating vehicles are utilized for traffic prediction, with the former being the more prevalent. Because they are simple to gather, traffic volume and vehicle speed are the most used indicators for traffic prediction. Dynamic traffic prediction uses the temporal characteristic model to forecast the amount of traffic on individual road segments. After that, the expected traffic density is used to determine the duration of the trip using the spatial characteristic model. Lastly, the model’s interpretability is examined, and real-world

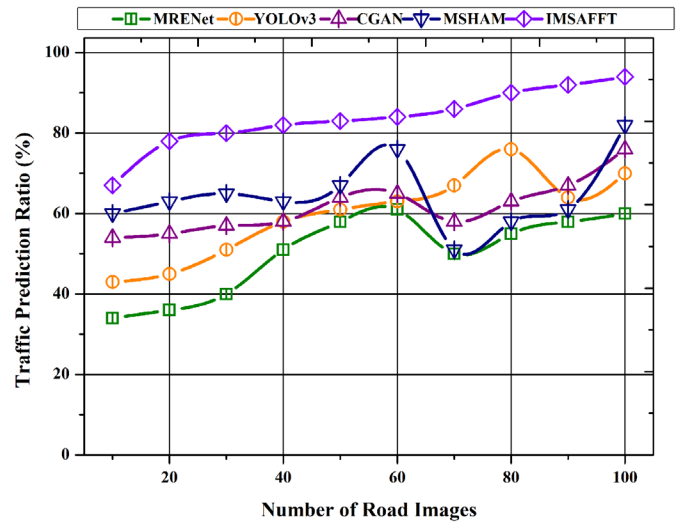


Fig. 9. Traffic prediction ratio.

data is used to confirm the correctness of the predictions. Based on equation (2), traffic density has been predicted. Figure 9 shows the traffic prediction ratio.

Figure 10 shows the precision-recall curve. When testing object identification algorithms on datasets skewed in one direction, a Precision-Recall (PR) graph is an indispensable tool. This is particularly true in cases when the negative samples outnumber the positive ones. Precision is the ratio of true positive detections to the total number of detections (true positives + false positives). Recall is the ratio of true positive detections to the total number of actual positives (true positives + false negatives). Figure 11 shows the overall performance analysis.

The model focuses on identifying vehicles, pedestrians, traffic signs, and lane markers in complicated road conditions. It prioritizes identifying objects of different sizes, distances, and occlusions in real-time traffic. The

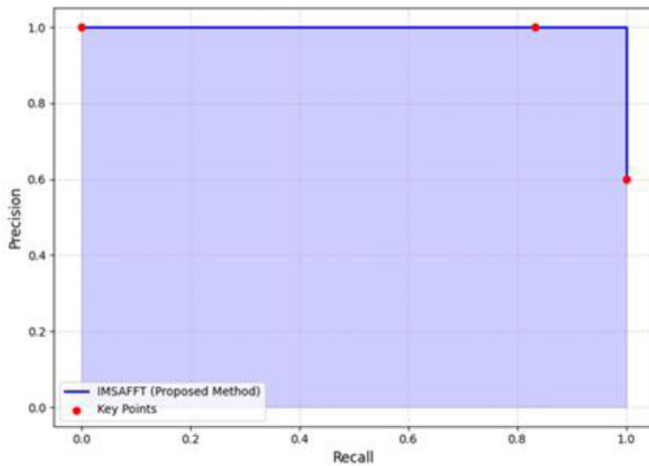


Fig. 10. Precision-recall curve.

present model does not directly identify wrecks but sets the groundwork for recognizing items needed for autonomous driving and traffic monitoring systems. Collisions and traffic irregularities might be detected in future extensions. Figure 12 shows an image of detection outputs with bounding boxes around cars, people, and traffic signs in a complicated road scenario. The graphic depicts the recommended model's object detection in a complex urban road scenario, including automobiles, pedestrians, and traffic lights. The model's object-segmentation capabilities are shown by bounding boxes and names for each item ("Car," "Pedestrian," and "Traffic Sign"). Multi-scale adaptive feature fusion is resilient to occlusions and accurately recognizes overlapping and faraway objects. Better training and adaptive learning make accurate localization and classification feasible independent of light. Visual evidence supports the proposed model's real-time traffic detection effectiveness.

The model's average inference time is achieved at 40 ms/frame. This method guarantees Timely results even in complicated road circumstances involving pre-processing, feature extraction, and object recognition. Applications requiring fast processing include live traffic analysis and collision avoidance for autonomous driving systems.

#### Advantages

1. In complicated situations, including overlapping, tiny, and faraway objects, the IMSAFFT attained a much greater detection accuracy (98.4%) than conventional object identification models.
2. Integrating an adaptive learning rate mechanism and a multi-scale adaptive feature fusion mechanism increases processing speeds and makes real-time traffic data analysis more efficient.
3. The technique surpasses state-of-the-art models in managing occlusions, fluctuating lighting conditions, and complicated metropolitan environments, demonstrating resilience in many scenarios.
4. The suggested system is scalable and runs on conventional GPUs with a minimal power need, making it ideal for use in autonomous vehicles and traffic monitoring systems in the real world.

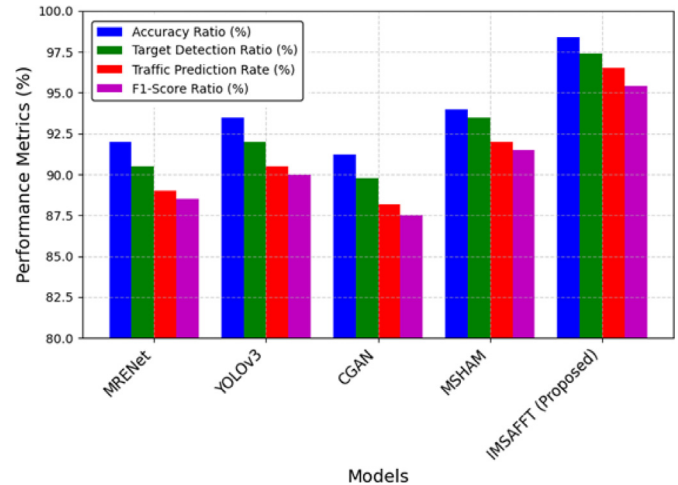


Fig. 11. Performance analysis.

#### Limitations

1. Despite IMSAFFT's promising results on the tested dataset, its potential use in other datasets that reflect vastly different geographic locations and traffic circumstances has not been investigated.
2. For smaller institutions or projects with limited budgets, the initial training of the deep neural network and multi-scale feature fusion method may be prohibitively resource-intensive.
3. Notwithstanding the model's robustness, it is necessary to fine-tune or modify it to achieve optimal performance in highly dynamic settings with many moving objects (such as congested pedestrian crossings).

## 5 Conclusion

This study presents the Improved Multi-Scale Adaptive Feature Fusion Technology (IMSAFFT) for a real-time traffic data identification method to fix the issues of low detection accuracy of road scenes and high false detection rates in panoramic video images. This study has intended a novel deep-learning technique for detecting traffic objects from complex road scene video images, explicitly concentrating on the rapid and accurate identification of automobiles within road scenes. The suggested model can avoid tedious foreground objects by building a DNN classifier that considers the factors impacting the congestion condition in multi-dimensional feature spaces. It extracts necessary visual features based on fine foreground pixels in the same space. This means that the suggested technique has the potential to provide more accurate and reliable detection findings. The feature fusion approach uses a hierarchical network architecture to adjust for varying-sized input photos. Features from many convolutional layers, each tuned to a different resolution, are combined in this structure. The model merges basic geographical information with advanced semantic characteristics using skip connections and attention processes. The system can achieve

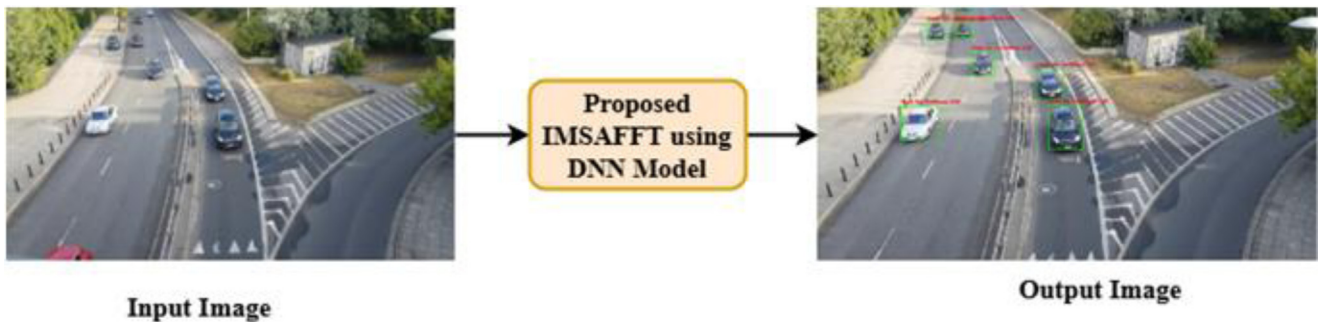


Fig. 12. Object detection visualization.

resilience in complex and dynamic traffic conditions by employing adaptive fusion while maintaining detection precision across different sizes and viewpoints. This research investigated the top-performing base networks and removed superfluous network layers by considering traffic sign characteristics to achieve realistic detection speed. Furthermore, our detector may operate on low-power mobile devices with a frame rate of 7 FPS by improving the resolution of the network input for the optimal speed-accuracy trade-off. The simulation outcomes illustrate that the recommended IMSAFFT model increases the accuracy ratio of 98.4%, target image detection ratio of 97.4%, traffic prediction rate of 96.5%, processing speed rate of 10.4% and F1-score ratio of 95.4% compared to other existing models.

#### Funding

This paper is the research result of the National Natural Science Foundation of China. (Project name: Research on randomized algorithms for solving several types of large-scale linear systems and their applications, project number: 12061048).

#### Conflicts of interest

The authors have nothing to disclose.

#### Data availability statement

This article has no associated data generated and/or analyzed.

#### Author contribution statement

Work concept or design - Z.X.; Conceptualization - Z.L.; Supervision - X.X.; Validation - S.A.; Methodology - N.M.D.; Formal Analysis - A.I.

#### References

1. K. Guo, X. Li, M. Zhang, Q. Bao, M. Yang, Real-time vehicle object detection method based on multi-scale feature fusion, *IEEE Access* **9**, 115126–115134 (2021)
2. L. Huang, C. Chen, J. Yun, Y. Sun, J. Tian, Z. Hao, H. Ma, Multi-scale feature fusion convolutional neural network for indoor small target detection, *Front. Neurobot.* **16**, 881021 (2022)
3. S. Tan, Z. Duan, L. Pu, Multi-scale object detection in UAV images based on adaptive feature fusion, *Plos one* **19**, e0300120 (2024)
4. M. Qiu, L. Huang, B.H. Tang, ASFF-YOLOv5: Multielement detection method for road traffic in UAV images based on multi-scale feature fusion, *Remote Sens.* **14**, 3498 (2022)
5. R. Shang, J. Zhang, L. Jiao, Y. Li, N. Marturi, R. Stolkin, Multi-scale adaptive feature fusion network for semantic segmentation in remote sensing images, *Remote Sens.* **12**, 872 (2020)
6. J. Zhang, MASFF: multi-scale adaptive spatial feature fusion method for vehicle recognition, *J. Comput.* **33**, 001–011 (2022)
7. X. Shen, H. Li, Y. Huang, Y. Wang, Vehicle detection method based on adaptive multi-scale feature fusion network, *J. Electr. Imag.* **31**, 043008 (2022)
8. A. Li, S. Sun, Z. Zhang, M. Feng, C. Wu, W. Li, A multi-scale traffic object detection algorithm for road scenes based on improved YOLOv5, *Electronics* **12**, 878 (2023)
9. J. Wu, G. Dai, W. Zhou, X. Zhu, Z. Wang, Multi-scale feature fusion with attention mechanism for crowded road object detection, *J. Real-Time Image Process.* **21**, 29 (2024)
10. Y. Zhang, L. Zhang, Y. Wang, W. Xu, AGF-Net: adaptive global feature fusion network for road extraction from remote-sensing images, *Complex Intell. Syst.* **10**, 1–18 (2024)
11. J. Dong, Y. Wang, Y. Yang, M. Yang, J. Chen, MCDNet: multi-level cloud detection network for remote sensing images based on dual-perspective change-guided and multi-scale feature fusion, *Int. J. Appl. Earth Observ. Geoinform.* **129**, 103820 (2024)
12. Y. Zhang, L. Li, C. Chun, Y. Wen, G. Xu, Multi-scale feature adaptive fusion model for real-time detection in complex citrus orchard environments, *Comput. Electr. Agric.* **219**, 108836 (2024)
13. R. Hasegawa, Y. Iwamoto, Y.W. Chen, Robust Japanese road sign detection and recognition in complex scenes using convolutional neural networks, *J. Image Graph.* **8**, 59–66 (2020)
14. Z. Shao, Z. Zhou, X. Huang, Y. Zhang, MRENet: simultaneous road surface and road centerline extraction in complex urban scenes from very high-resolution images, *Remote Sens.* **13**, 239 (2021)

15. L. Shen, H. Tao, Y. Ni, Y. Wang, V. Stojanovic, Improved YOLOv3 model with feature map cropping for multi-scale road object detection, *Measur. Sci. Technol.* **34**, 045406 (2023)
16. T.Y. Chow, K.H. Lee, K.L. Chan, Detection of targets in road scene images enhanced using conditional GAN-based dehazing model, *Appl. Sci.* **13**, 5326 (2023)
17. F. Yang, Y. Rao, Vision-based intelligent vehicle road recognition and obstacle detection method, *Int. J. Pattern Recogn. Artif. Intell.* **34**, 2050020 (2020)
18. R. Yousri, M.A. Elattar, M.S. Darweesh, A deep learning-based benchmarking framework for lane segmentation in the complex and dynamic road scenes, *IEEE Access* **9**, 117565–117580 (2021)
19. T. Deng, X. Liu, L. Wang, Occluded vehicle detection via multi-scale hybrid attention mechanism in the road scene, *Electronics* **11**, 2709 (2022)
20. L. Tang, L. Yun, Z. Chen, F. Cheng, HRYNet: a highly robust YOLO network for complex road traffic object detection, *Sensors* **24**, 642 (2024)
21. Z. Zhu, X. Li, J. Zhai, H. Hu, POdB: A learning-based polarimetric object detection benchmark for road scenes in adverse weather conditions, *Inform. Fusion* **108**, 102385 (2024)
22. A.V. Malawade, S.Y. Yu, B. Hsu, H. Kaeley, A. Karra, M.A. Al Faruque, Roadscene2vec: a tool for extracting and embedding road scene-graphs, *Knowl. Based Syst.* **242**, 108245 (2022)
23. <https://www.kaggle.com/datasets/boukrailyesali/traffic-road-object-detection-dataset-using-yolo>

**Cite this article as:** Zhaosheng Xu, Zhongming Liao, Xiaoyong Xiao, Suzana Ahmad, Norizan Mat Diah, Azlan Ismail, Target image detection algorithm of complex road scene based on improved multi-scale adaptive feature fusion technology, *Int. J. Simul. Multidisci. Des. Optim.* **16**, 6 (2025), <https://doi.org/10.1051/smdo/2025004>