International Journal of Current Engineering and Technology
E-ISSN 2277-4106, P-ISSN 2347-5161

Available at https://ijcet.evegenis.org/index.php/ijcet/index

Research Article

RE-FusionNet: A Resource-Efficient Multi-Task Network for Joint Traffic Element Detection and State Recognition

Hemant Kumar[1*], Pushpa Mamoria[2]

[1]Department of Computer Science & Engineering, Chhatrapati Shahu Ji Maharaj University, Kanpur, India
[2]Department of Computer Application, Chhatrapati Shahu Ji Maharaj University, Kanpur, India

Received 01 May 2026, Accepted 18 May 2026, Available online 19 May 2026, **Vol.16, No.3 (May/Jun 2026)**

Abstract

Autonomous driving is dependent upon reliable perception of traffic elements including traffic lights and traffic signs in order to ensure both safety and efficiency in making decisions. However, latest computer vision-based approaches treat object detection and semantic state recognition as either separate tasks, or they utilize expensive, hardware-dependent sensor technology to continuously feed in video data that can greatly increase the processing cost and limit deployments in resource-limited environments. This problem has been addressed through the development of Re-FusionNet; a lightweight, attention-enhanced convolutional architecture designed to perform both joint traffic light and traffic sign detection along with semantic state recognition utilizing only RGB image data. The Re-FusionNet utilizes a dual-path fusion encoder (DPFE), differential attention module (DAM), along with a compact convolutional backbone to enable contextual feature representation to be enhanced and to focus on relevant traffic-related objects. In addition to its ability to process images sequentially by only sampling a few frames from a video stream rather than continuously feeding it into the network, it enables efficient capture of contextual information. Finally, all three tasks are performed simultaneously via a single multi-task prediction head that predicts both object location and state, allowing for joint training. Results demonstrate that the Re-FusionNet achieves an average precision (mAP) of 97.3% @ 0.5 IoU on the BDD100K dataset while achieving a state recognition accuracy of 90.0%, and an mAP of 94.2% @ 0.5 IoU on the LISA dataset. Moreover, due to the low computational requirements, it can achieve frame rates above 150 fps. Overall, results clearly show that the Re-FusionNet represents an efficient and deployable method for real-time traffic perception applications within autonomous vehicles and intelligent transportation systems.

Keywords: *Attention-Based Perception, Autonomous Driving, Traffic Light Detection, Traffic Sign Detection, Multi-Task Learning*

1. Introduction

The majority of present day autonomous vehicle systems use computer vision to determine traffic rules and make safe driving choices based upon the detection of traffic lighting and signs. The presence of traffic rules and signs are critical for establishing traffic compliance and providing direction to the driver. This is especially important in complex and ever-changing driving environments [1]. However, detecting traffic elements reliably continues to be difficult to accomplish due to real world issues that include, but are not limited to, motion blurring, partial obstruction of objects from view, varying amounts of illumination, and distracting backgrounds.

For example, in urban areas with numerous vehicles traveling at different speeds, the size, distance and obstruction of traffic signals can cause difficulty in accurately identifying them. Deep learning has made it possible to detect traffic elements much better than traditional computer vision algorithms. CNN's and other deep learning architectures have performed well in object detection tasks [2]. However, many of these algorithms analyze video frames individually without utilizing the additional context from previous frames. Because individual analysis does not utilize prior context about what was previously detected, such analysis can potentially lower robustness when analyzing visually ambiguous images or those acquired under difficult environmental conditions [3]. In addition to the above-mentioned limitations, another limitation of most current architectures is the lack of a single task to perform both object detection and semantic state determination. Most current

*Corresponding author's: ORCID ID: 0000-0001-5719-1889, Pushpa Mamoria' ORCID ID: 0000-0002-5748-7302
DOI: https://doi.org/10.14741/ijcet/v.16.3.3

architectures treat traffic signal detection and determination of its color state (i.e., red, yellow, or green) as two separate tasks [4]. This results in unnecessary duplication of computation. Furthermore, some architectures require expensive hardware components like LIDAR or radar to achieve better performance than obtained through RGB imaging alone. Although these hardware components can provide valuable information that complements RGB imaging data, they add expense and complexity to the overall system making them impractical for low-cost autonomous platforms and edge-based implementations [5], [6]. Therefore, there exists a need for efficient perception architectures that can perform both object detection and semantic state determination using a common architecture with RGB image input. The architectures must be able to maintain high levels of detection accuracy while also operating in real time, thus meeting the demands placed upon resource constrained computing platforms [7], [8].

This paper presents RE-FusionNet – an efficient use of attention as well as resources to develop a new architecture that is able to perform all of the tasks described above (traffic element perception) jointly. A main component of RE-FusionNet is a light-weight CNN backbone that can be used to extract both spatial and temporal information. This CNN backbone is supported by an attention guided module which allows the network to focus its attention on different locations within an image. As opposed to utilizing a continuous stream of video frames or other sensor data, our method takes advantage of the fact that the most of the videos are composed of a relatively small number of sampled frames. By using a limited number of images, we are able to reduce the amount of computation required to make predictions about objects and their states in an image, while at the same time allowing the network to learn important contextual cues that would otherwise be difficult to obtain.

Finally, because all three sub-tasks described above can be represented as classification problems, we utilize a single unified multi-task learning objective function. Using multi-task learning has several advantages over training separate models for each task. One benefit of multi-task learning is that it reduces redundant computations across tasks. Another benefit is that it improves learning efficiency.

Contributions

The main contributions of this work are summarized as follows:

- We have proposed a resource-efficient attention-based CNN architecture designed for detection of traffic lights and traffic signs together with semantic state recognition using RGB imagery.
- A feature aggregation strategy having the DPFE module and DAM module is used to improve contextual feature representation and increase robustness against adverse condition such as motion blur, occlusion, and illumination variations.
- The proposed model employs a compact multi-task prediction head that jointly performs object localization and semantic state classification, enabling efficient learning while reducing computational ambiguity.
- The overall architecture is optimized for low computational overhead and real-time inference, making it suitable for deployment in edge-based autonomous driving systems and intelligent transportation platforms.

2. Related Work

Detection of traffic lights and signs is a key challenge that autonomous vehicles face while working in dynamic environments such as low light, occlusion, and atmospheric interference. While traditional CNN-based architectures like Faster R-CNN, SSD and RetinaNet are able to provide good detection results they do not incorporate global contextual information and therefore may be inconsistent over time when used in a frame-by-frame manner. In response to the above problems, there have been many studies proposing new solutions. For example, Gao et al. [9], developed an improved traffic sign detection framework using faster R-CNN which enhances detection capabilities in complex driving conditions. This study combines use of a Feature Pyramid Network (FPN) into the model architecture with replacement of ROI Pooling by ROI alignment and incorporation of Deformable Convolutional Networks (DCNs) to better detect both large and small traffic signs regardless of distortion. Results were tested on the TT100K dataset and demonstrated a 10.6% increase in mAP compared to previous state-of-the-art techniques, demonstrating increased resistance to detection errors caused by low lighting and poor weather conditions.

Li et al. [10], presented an enhanced version of YOLOv7 to provide better traffic sign detection for smaller objects in heavy traffic. To support this objective, they introduced another layer to detect small objects and added a combined convolution/self-attention module called ACmix to obtain better feature representations. They also converted traditional convolutional layers into omni-dimensional dynamic convolution (ODConv), so as to be able to acquire contextual information about the target object. Additionally, they used the Normalized-Gaussian-Wasserstein-Distance (NWD) to both measure the distance between ground truth and predicted locations in training and non-maximum suppression (NMS). In addition, using their new approach, Li et al. reported experimental results on TT100K dataset showing that the YOLOv7-enhanced model resulted in a mAP of 88.7%, surpassing the original YOLOv7.

Tang et al. [11] presented a new traffic-light-detection system utilizing the Real-Time DETR (RT-DETR) architecture. Work includes 3 critical components: 1) GreLaN - A backbone network to produce high-quality features; 2) A Down sample

module called ADown to down-sample small objects, and; 3) Deep Generalized Feature Fusion Module (DGSFM) to generate high quality feature maps from all layers. All of these have reduced the number of parameters, and the computation cost of the model. Results on the S2TLD traffic light detection dataset showed that their model detected traffic lights at a rate of 96.0% precision, and at a mAP of 95.9%. The authors did not report results using multiple datasets. Thus, it would seem likely that the proposed model will have limitations in terms of applicability due to dependence upon one data set.

Qu et al. [12] developed an improved version of YOLOv5s to achieve enhanced traffic sign detection for smaller traffic signs in different types of adverse weather. Author's framework included two main improvements over YOLOv5s. First, a coordinate attention (CA) module was applied to the back bone of their model, allowing for learning spatial-location information from input images. Second, an additional layer was added to help enhance feature representations of small traffic signs. In addition, Qu et al. replaced the traditional IOU (intersection over union) loss with alpha-IOU loss to improve the accuracy of bounding-box regression. Experimental results on the CCTSDB 2021 dataset indicated that author's model had a mAP of 82.8%, demonstrating improved performance in detecting traffic signs in various weather conditions compared to other models.

Xia et al. [13] have created a new transformer-based framework called DSRA-DETR, to enhance multi-scale traffic sign detection. DSRA-DETR has improved upon anchor DETR, by adding a dilated spatial pyramid pooling (DPPP) component to extract features at multiple scales, and a feature residual aggregation (FRA) module to retain some of the low-level spatial details. This will help to remove feature noise and increase the ability to detect smaller traffic signs. Experimental testing was conducted on two large scale public datasets for traffic sign detection GTSDB and CCTSDB, which resulted in higher average precision values of 76.13% and 78.24% respectively, than many other competing approaches.

Sarvajcz et al. [14] designed a low-cost computer vision based embedded system, which uses an NVIDIA Jetson Nano edge computing device to perform real time pedestrian and priority traffic sign detection. The proposed computer vision framework is implemented using an SSD-Mobilenet convolutional neural network that is trained via transfer learning from a custom dataset captured in various illumination and traffic scenarios. In addition to processing visual data from two cameras, the proposed system also includes a liquid crystal display (LCD), used to provide feedback to drivers regarding the presence or absence of pedestrians, pedestrian crossing areas, stop signs and give way signs. Results from experimental evaluations demonstrate that the proposed system provides accurate detection rates greater than 90%, at an average frame rate of about 8.7 frames per second (FPS). Based on these results it can be concluded that the proposed system could be potentially applicable to lower cost advanced driver assistance systems (ADAS).

Mehta et al. [15] presented MobileViT, a hybrid architecture that combines convolutional neural networks with vision transformers to represent both local and global features. The MobileViT block represents the transformers as convolutions; this enables efficient global context modelling with relatively low computational costs. Therefore, it can be effectively used for mobile devices. For example, experiments were performed on the ImageNet-1K dataset to evaluate the performance of the proposed MobileViT. It obtained a top-1 accuracy of 78.4% with approximately 5.6 million parameters. Compared with other state-of-the-art lightweight CNN models like MobileNetv3, MobileViT achieves competitive performance with fewer parameters. Moreover, due to its flexibility, MobileViT demonstrates strong generalization capabilities when employed as a backbone for object detection and image segmentation tasks.

Zeng et al. [16] suggested the traffic sign detection framework YOLOV5-Efficient ViT. This is based on an Efficient Vision Transformer as the backbone in order to support the global feature representation. In addition to this, the authors integrated the CBAM Attention Mechanism in order to support better feature extraction. Furthermore, they employed the WIoU Loss Function in order to increase the precision of the bounding boxes regression and reduce the instability during training. They tested the framework on the 3L-TT100K Dataset and obtained a mAP of 94.1%, and a processing speed of 62.5 frames per second. These results were superior to some other baseline frameworks such as YOLOV5, YOLOV7, and YOLOV8.

Du et al. [17] presented MASG-Net. It is a lightweight traffic sign detection framework that uses YOLOV4-tiny as the backbone for improving the detection accuracy and maintainability for real time execution. To achieve these goals, the authors introduced an E-MobileNet backbone with channel attention, MDSPP module for enhancing the context-related feature extraction abilities, and a SIG Module for improving the ability to detect smaller traffic signs. For evaluating MASG-Net, the authors used three different datasets: CCTSDB, GTSDB, and TT100K. On the GTSDB dataset, the authors reported a mAP of 90.8%. However, the authors also stated that MASG-Net has very fast inference speeds (> 200 fps), making it suitable for use in many real time Intelligent Transportation Systems applications.

Manzari et al. [18] developed a Pyramid Transformer Architecture for traffic sign detection by integrating hierarchical pyramid structures within vision transformers to be able to capture contextual features at multiple scales. As part of their design, the authors employed pyramid blocks and normal blocks that combined self-attention mechanisms with convolutional mechanisms to extract both global

dependency knowledge and local spatial knowledge. For evaluation purposes, authors used faster R-CNN and cascade R-CNN object detection heads with their pyramid transformer architecture to evaluate its performance on GTSDB dataset. According to authors, their architecture produced state-of-the-art performance for detecting traffic signs as indicated by a mAP of 77.8% when employing cascade R-CNN object detection head. Their design required fewer parameters than most of the existing transformer architectures. A summary of recent deep learning architectures for traffic sign detection is provided in Table 1.

Table 1 Comparative summary of recent deep learning–based traffic sign and traffic light detection methods, including the employed models, main methodological contributions, evaluation datasets, reported performance, and identified strengths and limitations

Author/Year	Method / Model	Key Contribution	Dataset Used	Performance	Strength	Limitation
Gao et al. (2022)[9]	Improved Faster R-CNN	Integrated FPN, ROI Align, and deformable convolution to enhance small traffic sign detection.	TT100K	mAP: 86.5%	Improves Faster R-CNN using FPN, ROI Align, and DCN	Lower real-time speed compared to lightweight detectors
Li et al. (2023) [10]	Improved YOLOv7	Added small-object detection layer, ACmix module, and ODConv to improve feature representation of tiny traffic signs.	TT100K	mAP: 88.7%	Improved YOLOv7 enhances small traffic sign detection accuracy	Slightly reduced inference speed due to added modules
Tang et al. (2025) [11]	GAD-DETR (Improved RT-DETR)	GRELAN backbone, ADown downsampling, and DGSFM fusion improve efficient traffic light detection.	S2TLD Traffic Light Dataset	Precision: 96.0%, mAP50: 95.9%, FPS: 117.8, Parameters: 9.79M	Achieves high accuracy with lightweight design and real-time detection speed	Dataset limited to S2TLD; may lack generalization to diverse traffic scenes
Qu et al. (2023) [12]	Improved YOLOv5	Introduced Coordinate Attention and Alpha-IoU loss to improve detection under complex weather conditions.	CCTSDB	mAP: 82.8%	Improved YOLOv5 with coordinate attention for small signs	Performance drops in rain, fog, and night conditions
Xia et al. (2023) [13]	DSRA-DETR	Transformer-based detector with DSPP and FRAM modules for multi-scale traffic sign detection.	GTSDB, CCTSDB	AP: 76.13%	Improved DETR with DSPP and FRAM for multiscale detection	High computational cost; difficult real-time deployment
Sarvajcz et al. (2024) [14]	SSD-MobileNet	Edge-based detection system implemented on NVIDIA Jetson Nano for real-time ADAS applications.	Custom dataset	Accuracy: >90%	Low-cost real-time pedestrian and sign detection using Jetson Nano	Small dataset and relatively low inference speed
Mehta et al. (2021) [15]	MobileViT	Lightweight hybrid CNN-transformer architecture enabling efficient global feature learning.	ImageNet	Accuracy: 78.4%	Lightweight hybrid CNN-Transformer architecture for mobile vision tasks	Higher latency than CNNs on mobile devices
Zeng et al. (2024) [16]	EfficientViT-YOLOv5	Integrated Efficient Vision Transformer backbone and CBAM attention to improve detection accuracy.	3L-TT100K	mAP: 94.1%	EfficientViT backbone improves traffic sign detection accuracy and speed	Tested mainly on TT100K-derived dataset only
Du et al. (2025) [17]	MASG-Net	Lightweight YOLOv4-tiny based model with E-MobileNet backbone, MDSPP, and SIG modules.	CCTSDB, GTSDB, TT100K	mAP: 90.8%	Lightweight MASG-Net improves small traffic sign detection accuracy	Performance degrades under extreme weather and motion blur
Manzari et al. (2022) [18]	Pyramid Transformer	Vision transformer with pyramid blocks to capture multi-scale contextual features for traffic sign detection.	GTSDB	mAP: 77.8%	Pyramid Transformer improves multi-scale traffic sign detection accuracy	Tested only on small GTSDB dataset

3. Methodology

Feature Fusion Network (RE-FusionNet) uses an Attention Guided Feature Fusion Module (AGFFM) to combine Lightweight Convolutional Features (LCFs) in order to accomplish both Traffic Element Detection (TED) and Semantic State Recognition (SSR). The proposed model will process the input images stage wise. This is accomplished by using the LCFs combined with AGFFMs to produce fused features for multi task prediction. While the proposed framework was designed to utilize contextual information over a few frames, it is capable of functioning independently of continuous video input and therefore can be used with existing image-based dataset.

RGB images are processed during training after they have been resized to 640 x 384. Random horizontal flips, croping, and color jittering were applied to the resized images in order to increase generalizability. A lightweight CNN backbone is then

used to extract spatial features for each input using DepthWise Separable Convolutions (DWSCs) and Embedded Attention Modules (EAMs). The spatial features produced by DWSCs and EAMs are further enhanced via the use of the proposed Dual Path Feature Enhancement (DPFE) module and Double Attention Mechanism (DAM) module. The DPFE module is responsible for enhancing contextual feature aggregation while the DAM module is responsible for emphasizing salient traffic elements prior to prediction. In addition, in order to create a richer contextual representation of the scene that does not rely on continuous video feed from a vehicle's camera, the model has been modified to aggregate spatial features from a small number of frames that are temporally related to the frame being predicted.

In order to utilize a single frame at one time rather than attempting to process multiple frames at once, the method treats each sampled frame independently, however shares all contextual aspects from the same driving event. The method allows the network to be trained on contextual relationships automatically and does so without forcing the network to explicitly define temporal behavior in addition to allowing the network to be trained using standard image-based data sets that have minimal computational requirements. When using datasets that consist of large amounts of independent images (e.g., BDD100K), this method samples frames from clips of driving scenes and uses them to train the network by simulating contextual observation while also being compatible with image-based data sets. By aggregating the features that represent each sampled frame via the DPFE module and DAM module, the network captures redundant context from different frames and emphasizes salient traffic elements. In doing so, the network can take advantage of interframe redundancy and solve problems associated with motion blur and partial occlusion, etc. —without having to perform explicit temporal modeling or provide recursive networks. Once features representing each frame are fused together via the Dual Branch Prediction Head, both object detection (traffic lights and signs) and semantic states are predicted. Proposed model architecture is shown in Fig. 1.

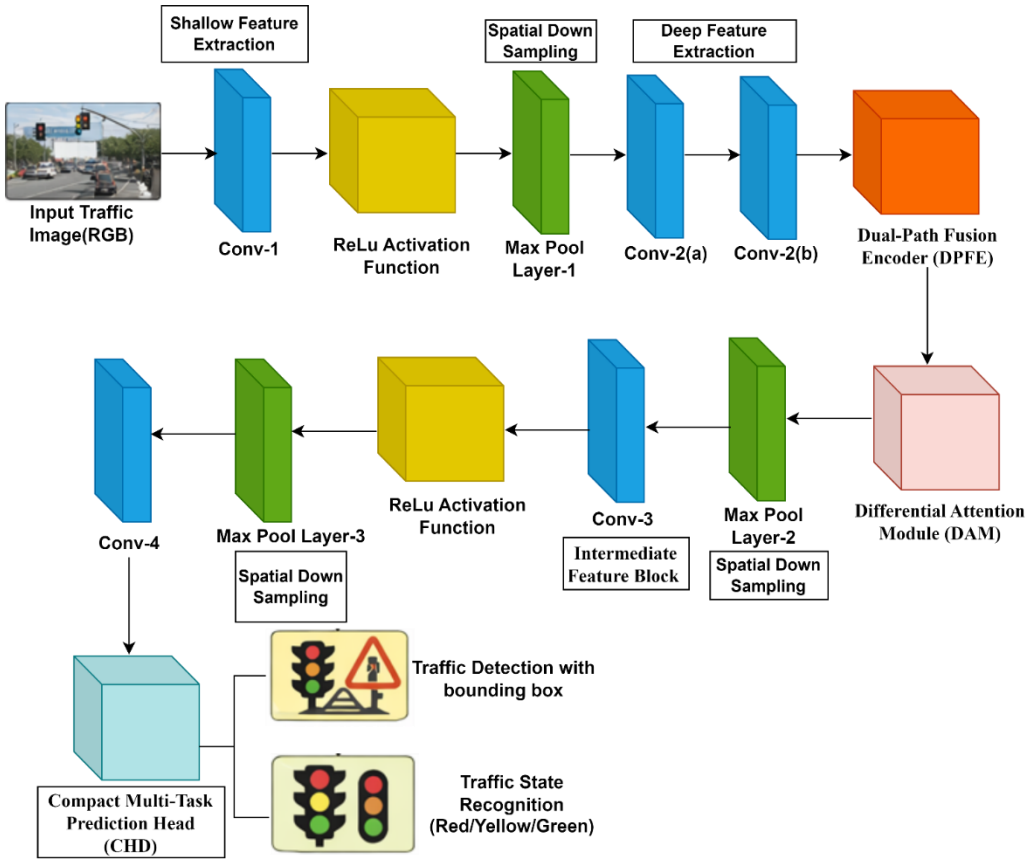

Fig. 1 Proposed RE-FusionNet architecture for joint traffic element detection and state recognition

3.1 Loss Functions

Detection Loss: A combination of Focal Loss for classification and GIoU Loss for bounding box regression is used. Focal Loss addresses class imbalance by reducing the weight of easy negatives, while GIoU (Generalized Intersection over Union) improves localization accuracy, particularly in crowded or overlapping scenarios.

$$L_{\text{det}} = L_{\text{focal}} + \lambda_{\text{bbox}} \cdot L_{\text{GIoU}} \quad (1)$$

In Eq. 1, L_{det} represents the total detection loss, combining classification and localization components.

L_{focal} addresses class imbalance by focusing more on hard to classify examples, while L_{GIoU} improves the accuracy of bounding box predictions. The term λ_{bbox} is a weighting factor that balances the contribution of the GIoU loss within the overall detection objective.

State Recognition Loss: A standard cross-entropy loss is applied to the predicted traffic light/sign state (e.g., red, green, yellow, or warning), encouraging the model to distinguish fine-grained visual cues associated with each class.

$$L_{\text{state}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (2)$$

In Eq. 2, the state recognition loss L_{state} is defined as the standard cross-entropy loss, which quantifies the difference between the predicted probabilities and the ground truth labels. Here, C represents the number of traffic light state classes (e.g., Red, Yellow, Green), y_i is the one-hot encoded true label for class i, and $\hat{y}_i$denotes the predicted probability for the same class. This loss ensures the model learns to accurately classify the current state of each detected traffic light.

Total Loss Function: The combined loss guides the network toward optimal performance across both tasks:

$$L_{\text{total}} = \lambda_1 L_{\text{det}} + \lambda_2 L_{\text{state}} \quad (3)$$

Eq. 3, defines the total multi-task loss L_{total} as a combination of detection loss L_{det} and state recognition loss L_{state}, allowing joint optimization. L_{det} uses Focal Loss and GIoU for class imbalance and localization, while L_{state} employs Cross-Entropy. The terms λ_1 = 1.0 and λ_2 = 0.8 are empirically selected weighting factors that balance the detection and state classification losses within the overall multi-task objective.

3.2 Training & Implementation Details

The proposed model has been implemented with the help of the PyTorch deep learning framework and it was trained in a Google Colab Pro environment that features an NVIDIA Tesla T4 Graphics Processing Unit (GPU). It was optimized with the adam optimizer and an initial learning rate of 1 x 10^{-4} and a cosine annealing learning rate scheduler to facilitate consistent convergence throughout training. Training took place for 250 epochs at a batch size of 16. To reduce overfitting and enhance generalization, a weight decay of 1 x 10^{-5} was used. Random horizontal flipping, random cropping, color jittering was used as data augmentation techniques to improve robustness under real-world environmental variability. The overall training configuration is summarized in Table 2.

Table 2. Training Configuration and Hyperparameters

Parameter	Value
Framework	PyTorch
Hardware	NVIDIA Tesla T4 GPU
Optimizer	Adam
Initial Learning Rate	1×10^{-4}
Learning Rate Scheduler	Cosine Annealing
Batch Size	16
Number of Epochs	250
Weight Decay	1×10^{-5}
Input Resolution	640 × 384
Data Augmentation	Horizontal Flip, Random Crop, Color Jitter
Horizontal Flip Probability	0.5

3.3 Network Architecture Specification

Beginning with an input 3-channel RGB image, this network has a series of early-stage convolutional blocks, then several late-stage convolutional blocks; all are using the depthwise separable convolution technique as a means of hierarchically extracting spatial features from the images. This network also uses two intermediate feature processing modules i.e. (DPFE and DAM), which utilize intermediate features to generate semantic representations of the images that contain high levels of attention. Spatial resolution is progressively reduced through max pooling as the network progresses toward the end. The CHD produces multi-task output data for both object localization and semantic state prediction. With a total of 109,825 trainable parameters across its 13 layers, this networks' design is optimized for real-time inference performance on resource constrained devices. All layers are presented in Table 3.

3.3.1 Dual-Path Fusion Encoder (DPFE)

The Dual-Path Fusion Encoder (DPFE) is designed to improve feature representation by combining complementary spatial information extracted from parallel convolutional pathways. Given an intermediate feature map $F \in \mathbb{R}^{H \times W \times C}$, two parallel transformation paths are applied: a standard convolutional transformation $f_c(\cdot)$and a channel-wise transformation $f_d(\cdot)$. The resulting features are fused using element-wise addition to produce a refined representation as shown in Eq. 4.

$$F_{dpfe} = f_c(F) + f_d(F) \quad (4)$$

where F_{dpfe}denotes the fused feature map. This dual-path design enables the network to capture both local spatial structures and channel-level contextual information, thereby improving feature aggregation for downstream detection and classification tasks.

3.3.2 Differential Attention Module (DAM)

The DAM module is introduced to enhance the network's ability to focus on informative spatial regions while suppressing irrelevant background features. Given the input feature map F_{dpfe}, attention weights are computed using a learnable transformation followed by a sigmoid activation as shown in Eq. 5.

$$A = \sigma(W_a * F_{\text{dpfe}}) \quad (5)$$

where $W_a$represents the learnable attention parameters, $*$denotes convolution, and $\sigma(\cdot)$is the sigmoid activation function. The refined feature map is then obtained by applying the attention weights to the input features as shown in Eq.6

$$F_{dam} = A \odot F_{dpfe} \quad (6)$$

where $\odot$ denotes element-wise multiplication. This mechanism enables the network to emphasize salient traffic elements while reducing noise from complex background regions. The outputs of the DPFE and DAM modules provide enhanced feature representations that are subsequently passed to deeper convolutional layers and the multi-task prediction head.

Table 3: Layer-wise configuration of the proposed RE-FusionNet architecture

Layer	Operational Layer	Output Shape	Filter Size	Weights
0	Input Layer (RGB Image)	640 × 384 × 3	–	0
1	Shallow Feature Extractor (Conv-1)	640 × 384 × 16	3 × 3 Conv	448
2	Activation (ReLU)	640 × 384 × 16	–	0
3	Spatial Downsampling (MaxPool-1)	320 × 192 × 16	2 × 2	0
4	Deep Feature Block (Conv-2a)	320 × 192 × 32	3 × 3 Conv	4640
5	Deep Feature Block (Conv-2b)	320 × 192 × 32	3 × 3 Conv	9248
6	Dual-Path Fusion Encoder (DPFE)	320 × 192 × 32	1 × 1 Conv	1056
7	Differential Attention Module (DAM)	320 × 192 × 32	1 × 1 Conv + Sigmoid	1056
8	Spatial Downsampling (MaxPool-2)	160 × 96 × 32	2 × 2	0
9	Intermediate Feature Block (Conv-3)	160 × 96 × 64	3 × 3 Conv	18496
10	Activation (ReLU)	160 × 96 × 64	–	0
11	Spatial Downsampling (MaxPool-3)	80 × 48 × 64	2 × 2	0
12	High-Level Feature Block (Conv-4)	80 × 48 × 128	3 × 3 Conv	73856
13	Compact Multi-Task Prediction Head (CHD)	N × (bbox + class + state)	Fully Connected / Conv Head	1025
Total Learnable Parameters				**109,825**

3.4 Datasets Used in the Study

This research uses two publicly accessible datasets to test and train the model which includes; BDD100K and the LISA traffic light dataset. These datasets were chosen based on their ability to represent a wide variety of traffic environments/conditions as well as a broad range of traffic signals. The BDD100K dataset is a large collection of annotated images that include a wide variety of environmental conditions. A curated subset of 5,000 images was created to allow for rapid experimentation with the reduced number of images. To ensure an even distribution across traffic-light state (Red/Yellow/Green), and other common categories of traffic signs, a stratified-sampling technique was employed. In addition to ensuring balance among the images, the subset also contained a variety of images representing daylight and night-time driving conditions, images taken at urban intersections and images of partially occluded scenes. By using this sampling technique, it will be possible to accurately test and experiment with the proposed methodology without encountering excessive computational costs. Sampling techniques such as those described above are generally accepted when creating experimental prototypes by selecting representative subsets of larger data-sets to identify model behaviors prior to scaling up to the full data-set. As noted earlier, this type of sampling can help to mitigate the risk of class imbalance, while providing an opportunity to efficiently test and evaluate the proposed methodology.

The dataset is split into three parts consisting of; 70% for training purposes, 15% for validating the results and another 15% to test and evaluate the final results. For evaluating cross-domain performances, approximately 4,000 annotated frames from the LISA traffic light data-set are used to determine whether the proposed model has the ability to generalize across various illumination conditions and signal configurations. The LISA traffic light dataset consists of frame-level annotations of traffic light states that have been obtained under a variety of illumination conditions, shading effects, etc., as well as different types of signal configurations. Annotations that corresponded to frames with clearly visible red, yellow or green traffic lights were extracted from the LISA data-set and then converted to match a uniform labeling scheme so as to provide consistent labels for comparison to those found in the BDD100K data-set. Combining these two data sets provide researchers with a means of conducting a thorough evaluation of the proposed methodologies, thereby providing an assessment of both intra-data-set performance and cross-domain performance in a variety of real-world driving conditions.

4. Experimental Results

The new model was tested on two data sets — BDD100K and LISA — in order to test how accurate it is at detecting objects, identifying states of those detected object, and being able to process information

in real time. Several evaluation measures were used, which include mean average precision (mAP), precision, recall, F1-Score and the ability of the semantic state recognition module to recognize states correctly. As well as mAP@0.5, the metric mAP@0.5:0.95 was evaluated during testing. This provided a means for evaluating the quality of localizations over a range of IoU thresholds. The proposed model had a mAP@0.5 of 97.3% when using the BDD100K data set. Additionally, the model had a precision of 94.8%, recall of 92.1% and an F1-Score of 93.4%. It was demonstrated that the joint multi-task learning approach worked effectively by the fact that the semantic state recognition module achieved an accuracy of 90.0%. On the LISA dataset, the model achieved a mAP@0.5 of 94.2%, with a precision of 92.5%, recall of 90.3%, and a state recognition accuracy of 87.4%, indicating strong cross-dataset generalization as shown in Fig. 2. Fig. 3 illustrates the comparison between training and validation loss curves, showing stable convergence during training. In terms of runtime efficiency, the proposed model achieves an inference speed of 152 FPS with a compact memory footprint of 9.8 MB.

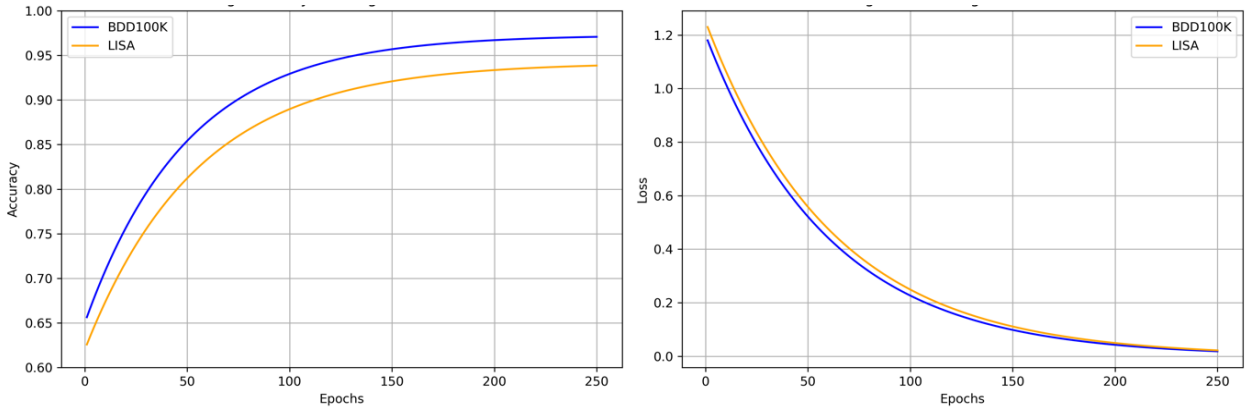

Fig. 2 Training accuracy and loss convergence curves over training epochs

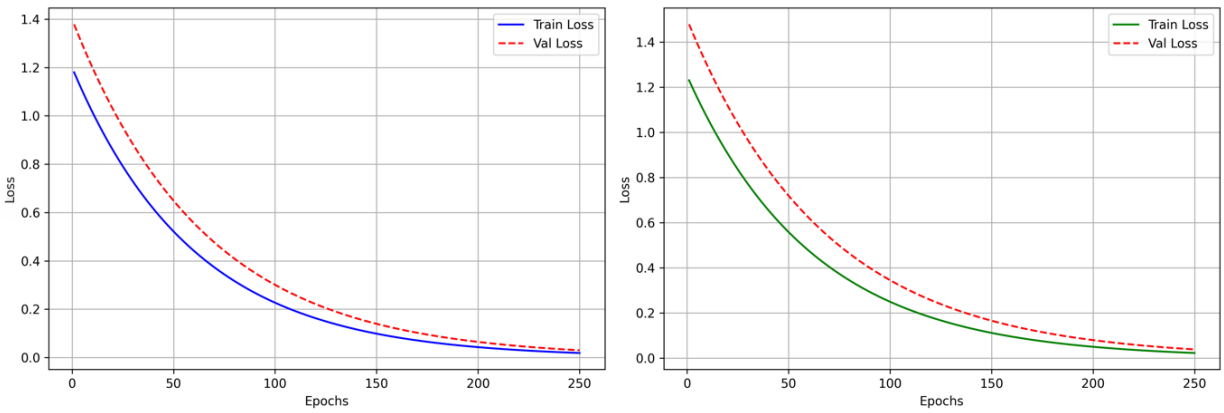

Fig. 3 Training vs validation loss curve on BDD100K Dataset (left) and LISA dataset (right) over epochs

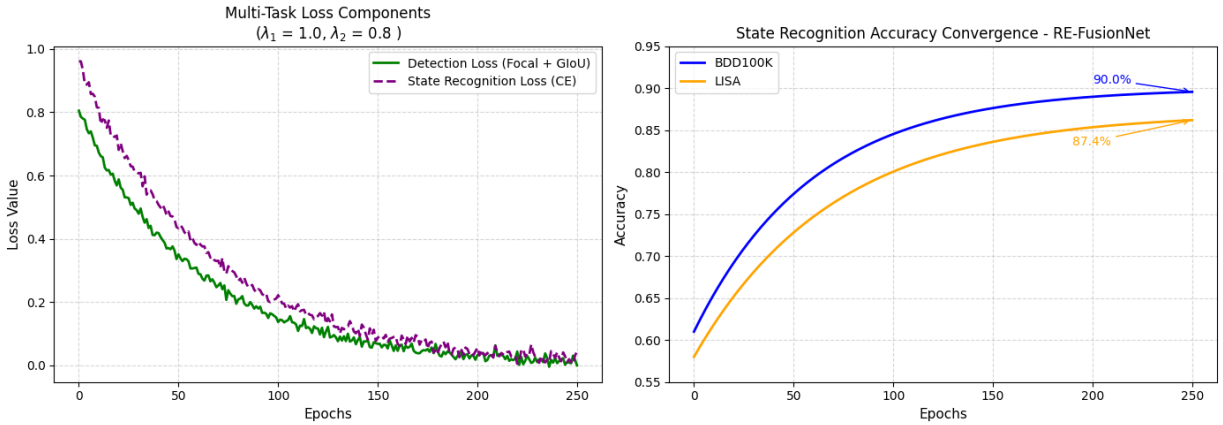

Fig. 4 Multi-task loss components (Left) and state recognition accuracy on BDD100K and LISA datasets (Right)

The runtime performance is measured on an NVIDIA Tesla T4 GPU using the PyTorch inference framework with batch processing. This evaluation reflects the real-time processing capability of the proposed architecture under standard single-frame inference conditions. A comparative evaluation against widely used baseline models—including Faster R-CNN, SSD, YOLOv5, YOLOv7, DETR, MobileNet, MobileViT, EfficientViT, and MASG-Net—is presented in Table 4. The authors obtain baseline performance data for each model from its publication or publicly available benchmarking (under same input) so that comparisons can be made fairly. It is possible that small differences exist in how some methods were implemented by researchers across studies; however, the resulting performance metrics will still allow for a comparison of the ability of the proposed model compared with those previously mentioned in terms of detection accuracy, ability to recognize states of objects, and processing speed. This selection was based on the fact that these models have been shown to be among the most popular CNN-based models and transformer-based models and lightweight models currently being used in AVS perception applications. Additionally, Fig. 5 and Fig. 6 demonstrate qualitative results of the predictions of the model including successful state recognition of traffic lights in varying environmental conditions. Finally, as demonstrated in Fig. 6, the model also identifies multiple traffic control signs (bus stop, enter left lane, U-turn), which indicates it has high confidence prediction scores in recognizing traffic control signs in an environment filled with various obstacles.

Fig. 5 Qualitative results of the proposed traffic light detection and state recognition framework under varying illumination and traffic conditions

Fig. 6 Example detections of traffic regulatory signs by the proposed framework under real-world driving conditions

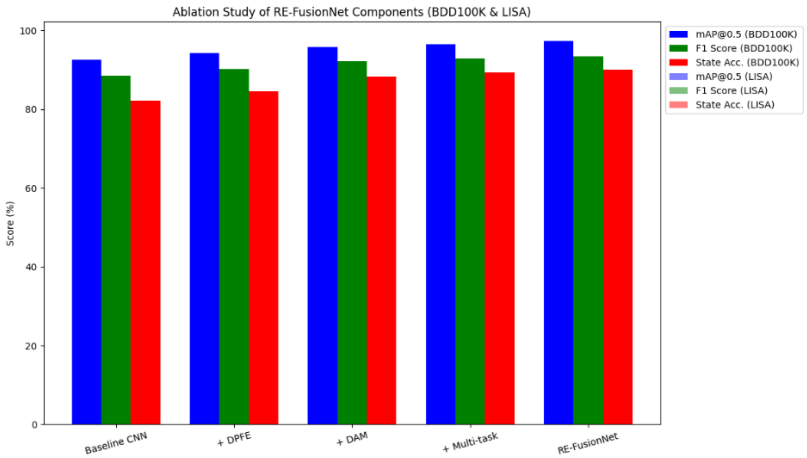

Fig. 7 Ablation study of the proposed RE-FusionNet components on the BDD100K and LISA datasets.

Ablation studies were undertaken to assess the influence of major components based upon Table 5 (illustrated in Fig. 7). The behavior shown by each task specific loss component and each epoch of a training session of how accurate the recognition of the state of a vehicle's surroundings is as illustrated in Fig. 4 shows that the behavior of RE-FusionNet converges stably and consistently improves its performance over both BDD100K and LISA datasets as it has been demonstrated that the three new components contribute to the entire performance.

Table 4. Quantitative comparison of existing detection models and the proposed RE-FusionNet on the BDD100K and LISA datasets

Model	Dataset	mAP@0.5 (%)	F1-Score (%)	State Recognition Accuracy (%)	FPS
Faster R-CNN [1]	BDD100K	86.5	80.5	72.0	8
	LISA	83.7	78.6	70.1	8
YOLOv7 [2]	BDD100K	88.7	85.9	85.0	70
	LISA	85.9	83.8	82.8	70
SSD [3]	BDD100K	83.8	79.8	78.1	22
	LISA	81.1	77.9	76.2	22
YOLOv5 [4]	BDD100K	88.1	83.6	80.0	20
	LISA	85.3	81.6	77.9	20
DETR [5]	BDD100K	76.1	70.2	76.7	15
	LISA	73.7	68.5	74.1	15
MobileNet [6]	BDD100K	60.1	58.1	65.0	20
	LISA	58.2	56.7	63.3	20
MobileViT [7]	BDD100K	78.4	75.2	74.2	27
	LISA	75.9	73.3	72.1	27
Efficient-VIT [8]	BDD100K	94.1	90.2	88.2	62
	LISA	91.1	88.0	85.7	62
MASG-Net [9]	BDD100K	94.2	92.6	89.2	203
	LISA	91.2	90.3	86.7	203
RE-FusionNet (Ours)	**BDD100K**	**97.3**	**93.4**	**90.0**	**152**
	LISA	**94.2**	**91.2**	**87.4**	**152**

Table 5. Ablation study evaluating the contribution of the main architectural components of RE-FusionNet, including the Dual-Path Fusion Encoder (DPFE), Differential Attention Module (DAM), and multi-task learning framework. Results demonstrate the progressive performance improvement obtained by integrating each module

Variant	Dataset	mAP@0.5	F1 Score	State Recognition Acc.	Remarks
Baseline CNN (No DPFE, No DAM)	BDD100K	92.6	88.4	82.1	Basic convolutional backbone
	LISA	90.1	86.2	80.3	
+ Dual-Path Fusion Encoder (DPFE)	BDD100K	94.3	90.2	84.6	Improves feature fusion
	LISA	91.8	88.1	82.7	
+ Differential Attention Module (DAM)	BDD100K	95.7	92.1	88.2	Enhances salient feature focus
	LISA	93.3	89.4	85.6	
+ Multi-Task Learning	BDD100K	96.5	92.9	89.3	Joint detection + state learning
	LISA	93.8	90.4	86.6	
RE-FusionNet (Ours)	BDD100K	**97.3**	**93.4**	**90.0**	**Complete Architecture**
	LISA	**94.2**	**91.2**	**87.4**	

The contributions of individual components within our proposed network were investigated using an ablation study that systematically integrated each proposed module into a base CNN structure. The results in Table 5 indicate that the addition of the DPFE module improved the ability of the network to aggregate features; thus, improving detection accuracy. Further improvements in detecting relevant traffic objects are achieved with the inclusion of the differential attention module which allows for better focusing on key traffic object attributes. Ultimately, the incorporation of a multi-task learning paradigm allowed for simultaneous training of both the traffic object detection task and the vehicle motion estimation task via this fully integrated RE-FusionNet network.

5. Conclusion and Future Research Directions

This research was a response to challenges of detecting traffic elements (traffic light and traffic sign) efficiently,

reliably, and recognizing the semantic state (such as red, green, yellow etc.) of those detected traffic elements in an environment that is used for autonomous vehicles. Most existing techniques have treated detection and recognition of states as separate problems. Additionally, most previous techniques require use of multiple sensors (LiDAR or Radar), which can be expensive and computationally intensive. Therefore, the problem of having methods for detecting and identifying the semantic state of traffic elements in an environment where resources are limited is a key issue. Thus, this study developed RE-FusionNet; an attention-enhanced convolutional neural network framework that is capable of performing both traffic element detection and identification of the semantic state of those detected elements using only RGB images. The method includes the use of a lightweight convolutional backbone to provide image representation. The method also includes an attention-guided feature fusion module that allows the method to incorporate information from previously processed video frames into its current processing frame. In addition to processing and incorporating contextually relevant information, the method has a single output layer that predicts the location of each object within an image and the semantic state of each identified object simultaneously. The performance of the proposed method was evaluated using two large scale datasets (BDD100K and LISA). Results indicated that the proposed method outperformed all previous methods by producing high accuracy results in both datasets (BDD100K: 97.3% mAP@0.5, State Recognition Accuracy: 90.0%; LISA: 94.2% mAP@0.5) at a rate that would support real time inference (i.e., 152 FPS).

The results from this research suggest that the suggested design is able to achieve a desirable equilibrium between detection accuracy, processing efficiency, and real-time operation. This research has significant impacts for developing functional perception modules for use in self-driving cars and smart transportation systems. It allows accurate traffic element perception with the aid of a light weight architecture and RGB-only input, providing an adaptable solution for use on all types of edge-based platform and embedded system. However, there are still some limits to be addressed. In particular, it is evaluated as part of this research against a representative sample of the BDD100K database; therefore, performance could potentially vary if the full datasets were used to train the model. Also, the structure uses only visual information, thus limiting its ability to operate effectively in cases where either the image is heavily occluded or visibility is poor.

Future research will concentrate on increasing the effectiveness of the proposed method under adverse weather (e.g., rain, snow) and night time conditions; developing domain adaptation methods to operate in traffic environment across regions; and applying lightweight multilevel sensing techniques to improve robustness in complex scenarios. Additionally, optimization of the architecture for deployment on edge computing accelerators such as NVIDIA's Jetson Xavier and Google's Coral TPU to enable large scale deployment in the real world. Overall, the proposed RE-FusionNet is demonstrated to provide accurate and efficient simultaneous traffic element detection and state recognition by employing an attention enhanced light weight convolutional neural network architecture which provides a significant advancement towards the development of reliable and real time perception systems for future generation of autonomous vehicles.

Acknowledgment

The authors would like to express their sincere gratitude to the university authorities for providing the necessary infrastructure and facilities to conduct the experiments.

Conflict of Interest

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

Funding Declaration

No funding agency supports this study in any form.

References

[1] Bar-Shalom, O., Philipp, T., & Kishon, E. (2025). On the Relation between Optical Aperture and Automotive Object Detection. ArXiv, abs/2501.09456.

[2] Liang, T., Bao, H., Pan, W., & Pan, F. (2022). Traffic Sign Detection via Improved Sparse R-CNN for Autonomous Vehicles. Journal of Advanced Transportation.

[3] Gupta, A., Gupta, A., Raj, G., Choudhury, T., Sar, A., Kotecha, K.V., & Ozseven, T. (2024). Traffic Light Detection for Self-Driving Cars using the YOLOv8 architecture. Proceedings of the Cognitive Models and Artificial Intelligence Conference.

[4] Song, J., Hu, T., Gong, Z., Zhang, Y., & Cui, M. (2024). TLDM: An Enhanced Traffic Light Detection Model Based on YOLOv5. Electronics, 13(15), 3080. https://doi.org/10.3390/electronics13153080

[5] Sebastian Sarwatt, D., Kulwa, F., Ding, J., & Ning, H. (2024). Adapting Image Classification Adversarial Detection Methods for Traffic Sign Classification in Autonomous Vehicles: A Comparative Study. IEEE Transactions on Intelligent Transportation Systems, 25, 19046-19061.

[6] Xia, K., Hu, J., Wang, Z., Wang, Z., Huang, Z., & Liang, Z. (2024). Vision-Based Algorithm for Precise Traffic Sign and Lane Line Matching in Multi-Lane Scenarios. Electronics, 13(14), 2773. https://doi.org/10.3390/electronics13142773

[7] Huang, F.J., Dong, Z., Liang, L., & Chen, X. (2025). Research on CEL-YOLO Algorithm for Lightweight Detection of Traffic Signs. Automation and Machine Learning.

[8] Zhao, Y., Wang, C., Ouyang, X., Zhong, J., Zhao, N., & Li, Y. (2024). MIAF-Net: A Multi-Information Attention

Fusion Network for Field Traffic Sign Detection. IEEE Transactions on Instrumentation and Measurement, 73, 1-14.

[9] Gao, X., Chen, L., Wang, K., Xiong, X., Wang, H., & Li, Y. (2022). Improved traffic sign detection algorithm based on Faster R-CNN. Applied Sciences, 12(18), 8948. https://doi.org/10.3390/app12188948

[10] Li, S., Wang, S., & Wang, P. (2023). A small object detection algorithm for traffic signs based on improved YOLOv7. Sensors, 23(16), 7145. https://doi.org/10.3390/s23167145

[11] Tang, C., Li, Y., Wang, L., & Li, W. (2025). Real-time traffic light detection based on lightweight improved RT-DETR. Journal of Real-Time Image Processing, 22.

[12] Qu, S., Yang, X., Zhou, H., et al. (2023). Improved YOLOv5-based method for small traffic sign detection under complex weather. Scientific Reports, 13, 16219. https://doi.org/10.1038/s41598-023-42753-3

[13] Xia, J., Li, M., Liu, W., & Chen, X. (2023). DSRA-DETR: An improved DETR for multiscale traffic sign detection. Sustainability, 15(14), 10862.

[14] Sarvajcz, K., Ari, L., & Menyhart, J. (2024). AI on the road: NVIDIA Jetson Nano-powered computer vision-based system for real-time pedestrian and priority sign detection. Applied Sciences, 14(4), 1440.

[15] Mehta, S., & Rastegari, M. (2021). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178.

[16] Zeng, G., Wu, Z., Xu, L., & Liang, Y. (2024). Efficient vision transformer YOLOv5 for accurate and fast traffic sign detection. Electronics, 13(5), 880.

[17] Du, C., Su, S., Lin, C., et al. (2025). A lightweight network for traffic sign detection via multiple scale context awareness and semantic information guidance. Scientific Reports, 15, 10110.

[18] Manzari, O.N., Boudesh, A., & Shokouhi, S.B. (2022). Pyramid Transformer for Traffic Sign Detection. 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), 112-116.

Authors Profile

Hemant Kumar received his B.Tech. degree in computer science and engineering from APJ Abdul Kalam university, India, and the M.Tech. degree in computer science and engineering from SHUATS, Allahabad, India. He is currently pursuing the Ph.D. degree in computer science and engineering from CSJM university, Kanpur, India with a research focus on autonomous driving systems. His research interests include computer vision, deep learning, intelligent transportation systems, and perception for autonomous vehicles. He has published many articles on autonomous vehicle system.

Dr. Pushpa Mamoria received her Ph.D. degree in computer science and engineering from BBAU, Lucknow, India. She is currently working as an associate professor with the Department of Computer application at CSJM university, Kanpur, India. She has extensive in teach--ing, research, and academic administration. Her research interests include computer vision, machine learning, artificial intelligence, intelligent transportation systems, and data analytics. She has guided several undergraduate and postgraduate research projects and has authored multiple research papers published in reputed journals and international conferences.