WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Development of Vehicle Digital Image Model for Traffic Light Control Applications Using Machine Vision

Jibril Danladi Jiya [a], Abdulateef Alhaji Salihu [a], Mohammed Aminu [a*], Samuel Dayo Adesola [b]

[a,b] Department of Mechatronics and Systems Engineering, Abubakar Tafawa Balewa University, Bauchi, 740272, Nigeria*

[a] Department of Electrical/Electronics Engineering, Abubakar Tafawa Balewa University, Bauchi, 740272, Nigeria

A B S T R A C T

Traffic congestion remains a significant challenge in urban areas, leading to increased environmental pollution and time wastage on congested roads. To address this issue, intelligent traffic control methods have emerged as efficient alternatives to conventional approaches, particularly at high-density intersections. In this paper, design, simulation and implementation of a Raspberry Pi-based traffic monitoring system equipped with three cameras to enhance real-time traffic management. The system integrates Vision Transformers and ResNet techniques, specifically, to build aa digital image model that can recognize and count vehicles. By leveraging these technologies, this approach demonstrates the potential for reducing traffic congestion and improving overall traffic efficiency through real-time capturing and counting of vehicles for traffic light control applications. In simulation experiment, the approach was found successful and the unit implemented using the Raspberry Pi microcontroller.

Keywords: Vision Transformers, ResNet, Traffic, Intelligent Traffic Control,

1. Introduction

The management of road traffic is a critical challenge faced by major cities worldwide. During peak hours, roads are often overwhelmed with long vehicle queues, leading to significant delays. This issue is not confined to developing nations but is also prevalent in developed countries. Contributing factors include the growing demand for automobiles, increasing affordability, and the need for daily commutes related to work and other activities (Meng et al.'2020). Traffic congestion at intersections causes a range of problems, including environmental pollution, mental stress, increased fuel consumption, and financial burdens.

Traffic control system can mitigate flow of traffic and safety in transportation systems. This can be achieved through the use of microcontroller and cameras to draw the number of vehicles to provide time based dynamic flow monitoring systems. Image processing based adaptive traffic signal control and machine vision can adapt to sweer-changing the real-time traffic situations. Timing can be calculated automatically according to the traffic volume.

In an undisciplined traffic environment, vehicle detection is a difficult task as different vehicle categories travel very close to each other and do not follow traffic rules. Several convolutional neural network (CNN)-based deep learning (DL) and Vision Transformer-based techniques for vehicle and object detection have been developed (Deshmukh et al., 2023). These techniques are complex and do not extract multi-scale features as a result of involvement of existing CNN feature extraction backbones. Furthermore, most techniques failed to account for an undisciplined traffic environment as a result of unavailability of labelled vehicle datasets.

Despite the growing importance of Vision Transformer (ViT) models for optimizing traffic flow, there remains a significant gap in the available literature regarding their comprehensive application in this domain. Most current studies do not adequately address the full impact and potential of ViT models for traffic flow prediction and urban mobility management.

The aim of this study is to develop a hybrid ViT and ResNet model to enhance traffic flow identification by harnessing the potential of ViT and ResNet to overcome the limitations of CNNs. The proposed system is made up of fixated converses along the traffic flow placed in specific lanes to capture live video feed at the intersection. Integration of ResNet model with ViT model preserves shallow features along with perception of global information. Thus, enhances the feature representation capability and consequently improves detection accuracy of the model.

Vehicle detection plays a significant role in Intelligent Traffic Management Systems (ITMS). It results in reduced average waiting time at traffic intersections, serves fuel consumption control, mitigates traffic congestion, reduces accident rates, and consequently builds up human safety. Recent developments in artificial intelligence (AI) have increased the use of Intelligent Vehicle Detection (IVD) in traffic environments. ViT and ResNet models can accurately predict traffic conditions, identify potential problems and recommend optimal routes through integration of real-time information from global positioning system (GPS), traffic cameras, social media and other sensors. This will enable the adaptive adjustment of traffic signals and rerouting

of vehicles to prevent congestion. This will bring about smoother traffic flow, less idling time and reduce environmental impact to achieve the goals of sustainability and enhanced overall functionality and resilience of urban transport systems. This will also lead to higher fuel savings, reduce vehicle emissions and consequently cleaner air as well as healthier urban environment. (Boppuru et al. 2023). These models leverage advanced Machine Learning to capture intricate spatiotemporal patterns to provide valuable insights for traffic light control applications, urban planners and traffic management centers.

In this paper, a Vison Transformers (ViT) and ResNet models are integrated to accurately recognize and count various types of vehicles. This approach enables the system to dynamically capture and evaluate vehicle densities based on input image. By leveraging this integration, the model achieves improved accuracy in representing and predicting traffic density, making it more suitable for real-time applications. This paper proposes a vehicle detection model based on ViT- ResNet. This method transforms traffic flow detection issues into image classification tasks.

2. Literature Review

Initially, traffic prediction models focused on statistical methods such as regression in time-series and forecasting (Gomes et al., 2023). However, the advent of deep learning produced models such as Artificial Neural Networks (ANNs) and Recurrent Neural Networks (RNNs) with improved predictive performance (Zhang et al., 2025). Furthermore, while RNNs struggle with long-range dependencies, this led to the rise of Vision Transformer (ViT) models (Boppuru et al. 2023).

2.1 Fundamental Concept

This section is concerned with the theoretical and conceptual frame work of the digital vehicle traffic model which is based on vision transformer and residual network.

2.1.1 Vision Transformers

These are types of neural networks that use self-attention mechanisms for image classification, object detection, and semantic image segmentation tasks (Mao et al., 2022; Li et al., 2022). Its architecture has been shown to outperform CNNs in classification tasks via capturing global dependencies within an image (Thisanke et al., 2023). ViT model utilizes transformer-like architecture over patches of image for image classification. It uses a self-attention mechanism to model the relationships between several image patches to capture global relationships between image patches (Isinkaye et al., 2024). The ability of ViT to capture global relationships between image patches and its lightweight architecture is used as a promising approach for traffic flow identification.

As a result of self-attention mechanisms and parallel processing capabilities, ViTs excel at capturing long term dependencies in spatiotemporal data. Thus, they significantly enhance traffic prediction for optimizing traffic light control. Adaptability to complex data patterns has made ViT applicable for control engineers to optimize traffic flow and create efficient, sustainable environments.

In recent years, researchers have studied ViT models in various computer vision tasks as a result of their excellent performance (Alshammari et al., 2022; Wang, et al., 2025). The ViT uses self-attention mechanisms to process image data by dividing the image into N patches. Each patch is flattened and embedded into a vector as:

$$x_1 = linear\ (patch) + positional\ encodings \quad (1)$$

Where; patch i is the i-th image patch, and positional encoding adds positional information. The multi-head self-attention mechanism calculates attention scores to capture global dependencies as:

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_g}}\right)V \quad (2)$$

Where; Q (query), K (key), and V (value) are matrices, and d_k is the dimension of the keys.

The output of the self-attention mechanism is passed through a feed-forward network and layer normalization to form the final classification logits. Vision transformer model is capable to capture both global context and local details in images.

2.1.2 Residual Network

ResNet model uses residual connections to extract deep features while retaining shallow information. The residual modules of ResNet can propagate gradients effectively to alleviate the gradient vanishing issue during the training of deep networks (He et al. 2015). ResNet can extract details and positional information of images while preserving low-level feature information, as it has the ability to capture local features.

2.2 Review of Related Works

Researchers have explored several methods to evaluate and predict traffic congestion levels (Bhardwaj et al., 2021). Spatial and temporal traffic data have been analyzed using fuzzy logic to assess continuous and discrete congestion patterns. Hidden Markov Models (HMMs) have also been employed to identify distinct traffic congestion states by analyzing traffic images and videos. In many cases, traffic parameters such as vehicle speed and density are input into fuzzy inference systems (FIS) to estimate and model congestion at intersections (Gandhi et al., 2020). While these approaches have shown potential in estimating traffic conditions, their accuracy in real-time environments still leaves room for improvement.[5] Intelligent traffic control methods provide efficient solutions as against the conventional methods especially in traffic intersections with high traffic density. The essential component of smart city initiative is the smart traffic infrastructure as traffic congestion is a severe issue which grows with city development. Smart traffic management includes intelligent transport systems integrated with components as such adaptive traffic signal controls, free – way management, emergency management services and roadside units (Mo et al., 2020). Meng et al. (2020) proposed an approach which streamlines the process of detecting and calculating multiple vehicles on the highway. This involved the development of a new correlated vehicle tracking algorithm to address tracking-point instability issues, cutting and non-linearity issues. Accuracy of over 93% and an average speed of 25 fps (frames per second) were achieved. Chung and Solin (2018) proposed a deep CNN model under supervised learning (deep convolutional neural networks) for counting the number of vehicles on a lane based on video imagery. The proposed deep CNN provided higher accuracy and smaller time delay but could not distinguish between moving vehicles and stationary vehicles. Bhardwaj et al. (2021) proposed an image-processing based smart traffic control system for smart cities using Sobel and Canny edge detection techniques. However, this technique requires more complexity and computational time. Gandhi et al. (2020) proposed smart control of traffic lights using machine learning (YOLO). However, this technique has low recall compared with CNN as well as struggles to detect close and small objects.

2.3 Research Gap

Existing research often fails to explore the unique capabilities of ViTs in handling complex spatiotemporal traffic data. This study seeks to fill this gap by systematically analyzing the effectiveness of Vision Transformer (ViT) models in conjunction with ResNet to address traffic flow issues. The integration of superficial intelligence such as ViT and ResNet models play a significant role in enhancing the predictive capabilities for traffic flow determination. In this study, application of ViT and ResNet models to predict and optimize traffic flow is proposed.

3. Materials and Methods

This section comprises both the materials used and methods proposed in achieving the objectives of the research. This covered the urban traffic network study area, data collection and the development of the vehicle digital image model using ViT and ResNet.

3.1 Study Area

The study area is the traffic intersection at the entrance of Abubakar Tafawa Balewa University, Bauchi, Nigeria, Yelwa campus, which has a T structure (T junction) as schematically shown in figure 1.

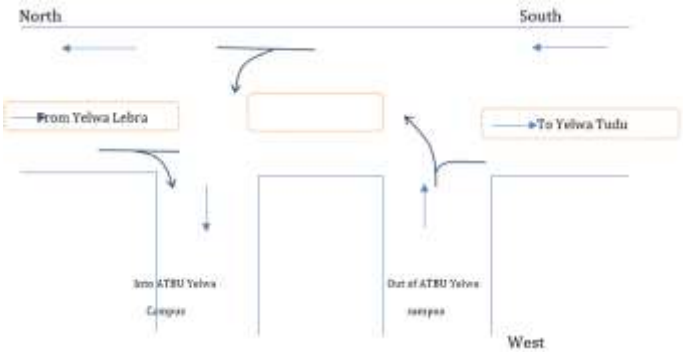

Figure 1 Schematic Diagram of the Study Area

The road network is associated with on and off ramps from yelwa Tudu, yelwa lebura and Sabon Kaura village. This field of study is chosen because of its location to Abubakar Ta fawa Balewa University, Bauchi Yelwa campus and adjoining settlements which has high vehicular movement.

3.2 Traffic Data

3.2.1 Data Collection

To develop an effective traffic regulation and control system, a comprehensive dataset of vehicle images were collected to train the object detection model. The primary objective of data collection was to ensure the system's ability to accurately detect and count common vehicles present on roads,

specifically cars, motorcycles, and tricycles. These images were taken under varying lighting conditions, weather scenarios, and different times of the day to enhance the model's robustness.

Video samples, each lasting 2 minutes, were collected from the Abubakar Tafawa Balewa University gate. Four lanes were considered, and the samples were recorded at different times of the day to capture variations in traffic density. A mobile phone was used to record video samples.

The collected images were manually annotated. Each image was annotated to mark the bounding boxes around vehicles, specifying their class (Car, Motorcycle, or Tricycle). This annotation process was crucial for training the object detection model to identify and distinguish between different vehicle types. The dataset was divided into three subsets: Training, Validation, and Test sets. The split ratio was maintained at 70% for training, 15% for validation, and 15% for testing to ensure the model's performance could be thoroughly evaluated on unseen data. Figure 2 shows sample images from the dataset.

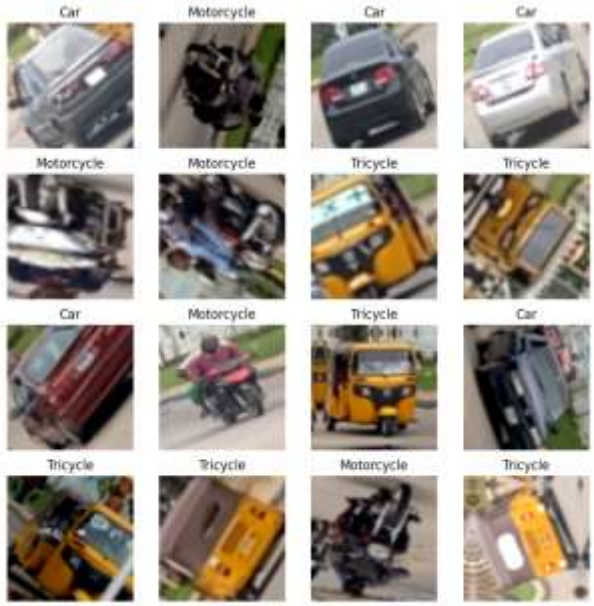

Figure 2: Sample Images of Dataset

3.2.2 Data Processing

The images were resized to a fixed resolution compatible with the ViT model, typically 224x224 pixels. Each image was then divided into smaller non-overlapping patches, such as 16x16 pixels, which were flattened into vectors to serve as input tokens. Image pixel values were normalized to a standard range, typically [0, 1] or [-1, 1], to ensure consistent input to the model and improve convergence during training.

3.2.3 Data Size

Hourly data size is used in the context of this work. Each hour of the day is assigned accordingly to each input.

3.3 Machine Vision Techniques

In this study, a hybrid model that combines ResNet and Vision Transformer (ViT) is utilized to address the challenge of traffic flow classification from traffic images. The primary aim is to harness the strengths of both models to streamlines the high-dimensional data without losing crucial traffic-related information. ViT enhances classification accuracy by handling the complex class variability in traffic flow effectively through its global self-attention mechanism. This approach provides a promising solution to the challenges posed by traffic flow detection. This is vital for the advancement of traffic light control and sustainable smart cities.

The one-dimensional datasets are transformed into two-dimensional datasets of images which are preserved time-series features through image blocks to achieve effective classification of different traffic flow categories. The integration of ResNet and ViT models combines the capability of ResNet to extract local detailed features with ViT's capability of self-attention mechanism to capture global contextual information. The fusion of features extracted by both models (ResNet and ViT) across stages and models enables the acquisition of richer feature information and consequent comprehensive feature learning. Thus, enhanced detection accuracy is achieved.

3.3.1 Vision Transformer Model

The Vision Transformer (ViT) model is employed in this research to detect and classify vehicles in traffic images. Unlike traditional Convolutional Neural Networks (CNNs), ViT leverages the power of Transformers, a deep learning architecture that has revolutionized natural language processing, for image classification tasks. The ViT model uses self-attention mechanisms to capture global dependencies and patterns in the input images, making it particularly effective in complex visual recognition tasks (Soudeep et al., 2024).

ViT divides the input image into a grid of patches, similar to how words are tokenized in text processing. Each patch is flattened into a vector and then embedded using a linear projection. These patch embeddings are passed through a series of Transformer layers that consist of multi-head self-attention mechanisms and feed-forward networks. The ViT model has several advantages over traditional CNNs, such as the ability to model long-range dependencies between different parts of an image, reduced inductive biases, and superior performance on large-scale datasets when trained properly.

3.3.2 ResNet Model

To train the ResNet Model from scratch, an identity block function was set to perform a series of operations including convolution, batch normalization, activation function, and skip connections. Three convolution layers were defined; each layer takes in an input tensor and outputs the resulting tensor. Lastly, the major peculiarity of ResNet was implemented, skip connection or shortcut branch, this allows the model to bypass the convolution layers and directly add the original input tensor X to the output of the third convolution layer to have an output defined by (1).

$$X_{Output} = ReLU(X3 + X_{shortcut}) \qquad (3)$$

The convolution block was implemented to alter the dimensionality of the feature map and enable the model to learn more complex representations. The block implements both the convolution later operations and the "shortcut" operation which enables the model to bypass certain layers as shown in (3) to eliminate the vanishing gradient problem. After passing through three layers, the resulting tensor is passed through the shortcut connection and then to the output layer. The ResNet models use these two blocks to extract features from the datasets of images, after the final convolution layer, pooling is applied to reduce the mentions further. The output of the pooling layer is further flattened and passed through a fully connected layer where a SoftMax activation is applied.

4. Results and Discussions

The Vision Transformer (ViT) model's performance was evaluated using multiple metrics, including training loss, accuracy, Receiver Operating Characteristic (ROC) curve, and the Confusion Matrix. These metrics provide a comprehensive assessment of the model's ability to detect and classify vehicles such as cars, motorcycles, and tricycles in real-time traffic images.

Figure 5 shows the training loss curve, this illustrates the model's learning progress over successive epochs. A steady decline in the loss value was observed as the model adjusted its weights to minimize classification errors. The convergence of the loss function indicates that the model effectively learned the underlying features necessary for vehicle detection. The model achieved a training accuracy of 75%, which demonstrates its competence in identifying vehicle types from images. The accuracy metric highlights the proportion of correctly classified images out of the total samples, serving as a key indicator of the model's overall effectiveness. Although 75% accuracy is a promising result, further tuning and optimization may enhance the model's precision in real-world deployment.

The ROC curve is a graphical representation of the model's diagnostic ability, plotting the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various threshold settings. The ROC curve illustrates the trade-off between sensitivity and specificity, providing insights into the model's ability to distinguish between different vehicle classes. The AUC value derived from the ROC curve quantifies the overall performance of the model across all classification thresholds. An AUC closer to 1 indicates excellent model performance, while an AUC closer to 0.5 suggests no better than random chance. The ViT model's AUC score reflects its reasonable capacity to correctly identify vehicle types in traffic images. The ROC Curve is shown in Figure 6.

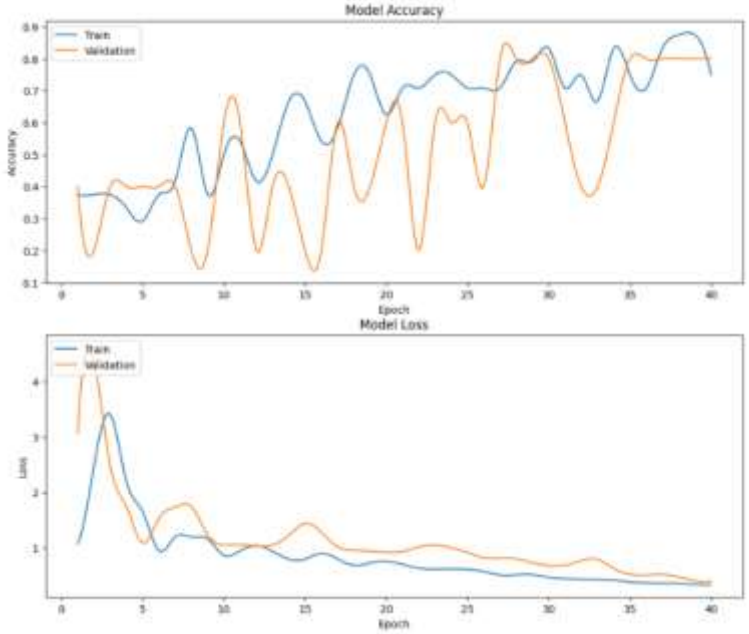

Figure 5: Training Loss Curve

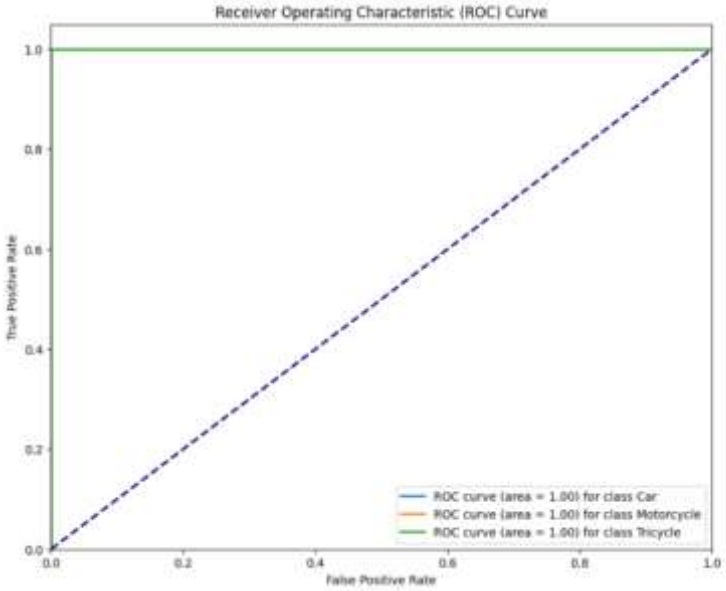

Figure 6: Receiver Operating Characteristics

The confusion matrix shown in figure 7 provides a detailed breakdown of the model's classification results across different vehicle classes (cars, motorcycles, and tricycles). Each cell in the matrix represents the count of true positives, false positives, true negatives, and false negatives for each class, allowing for an in-depth analysis of the model's strengths and weaknesses.

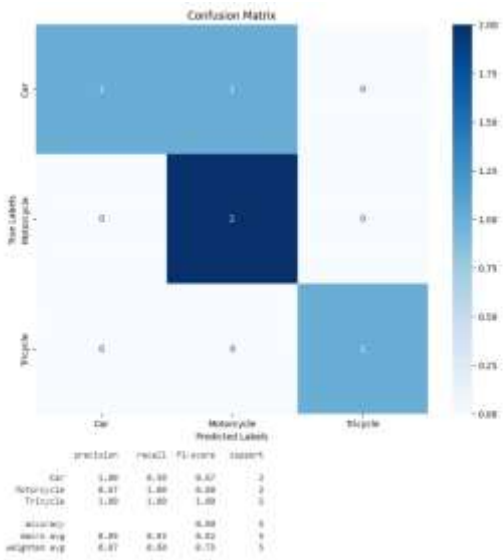

Figure 7: Confusion Matrix

To further demonstrate the effectiveness of the Vision Transformer (ViT) model in detecting and classifying vehicles, sample prediction images are presented as part of the evaluation as shown in figure 8. These images provide visual evidence of the model's performance, highlighting its ability to accurately identify and count various vehicle types—including cars, motorcycles, and tricycles—even in complex, real-time traffic scenarios. By showcasing the precision and robustness of the ViT model across diverse traffic densities and lighting conditions, these predictions validate its potential for practical deployment in intelligent traffic management systems.

Figure 8: Samples of Prediction Images

5. Conclusion

This research successfully developed an advanced intelligent vehicle density evaluation system. The system leverages the strengths of Vision Transformers (ViT) to evaluate the density of vehicles, this system can then allow the system to dynamically recognize real-time vehicle motion, assess

traffic density, and adjust traffic light phases accordingly to minimize delays and improve traffic flow efficiency. This study demonstrates how ViT and ResNet models can learn complex spatiotemporal patterns from real-time traffic data as well as historical data to enhance prediction accuracy. This improved predictive capability aids the development of efficient traffic light control systems. This study also highlights the transformative potential of predictive modeling using ViT and ResNet models for traffic light control applications. These models are known for their ability to process and interpret sequential data, can be used to understand complex patterns and long-term dependencies within traffic data.

Acknowledgements

The authors would like to acknowledge the financial support provided by Tertiary Education Trust Fund (TETFund) with reference No. TETF/DR&D/CE/UNIBAUCHI/IBR/2022/VOLI and the management of Abubakar Tafawa Balewa university, Bauchi, Nigeria for given access for the use of Laboratories in the Department of Electrical and Electronics Engineering as well as Department of Mechatronics and Systems Engineering.

References

Alshammari, H., Gasmi, K., Ltaifa, I.B., Krichen, M., Ammar, L.B., and Mahmood, M.A. (2022), "Olive disease classification based on vision transformer and CNN models," Computational Intelligence and Neuroscience.

Bhardwaj, V., Ramestti, Y and Valsan, V (2021), “Image-processing based smart traffic control system for smart city,” Proceedings of the 12th Int. Conference on Computing, Communication and Networking Technologies (ICCCNT), p. 1–6.

Chung, J and Sohn, K. (2018), “Image-based learning to measure traffic density using deep Convolutional neural network,” IEEE Trans. Intell. Transp. System, 19(5), 1670–1675.

Deshmukh, P., Satyanarayana, G. S. P., Majhi, S., Sahoo, U. K., & Das, S. K. (2023). Swin Transformer Based Vehicle detection in unregulated traffic environments. Expert Systems with Applications, Volume 213(Part B), Article 118992.

Gandhi, M. M., Solanki, D. S., Daptardar, R. S and Balborkar N. S. (2020), “Smart control of traffic light using Artificial Intelligence,” Proceedings of the 5th IEEE Int. Conf. on Recent Advances and Innovations in Engineering (ICRAIE), p. 1–6.

Gomes, B., Coelho, J and Aidos, H. (2023). A Survey on Traffic Flow Prediction and Classification. Intelligent Systems with Applications, vol. 20, article 200268.

Gupta,M., Miglani, H.,Deo, P and Barkhade, A. (2020), "Real-time Traffic Control and Monitoring," E-Prime – Advances in Electrical Engineering, Electronics and Energy.

He, K., Zhang, X., Ren, S and Sun, J. (2015). Deep Residual Learning for Image Recognition, Cornell University, arXiv: 1512.03385.

Isinkaye, F. O., Olusanya, M. O., & Akinyelu, A. A. (2025). A multi-class Hybrid Variational Autoencoder and Vision Transformer Model for Enhanced Plant Disease Identification. Intelligent Systems with Applications, 26, CE02 Article 200490.

Isinkaye, F.O., Olusanya, M. O., & Singh P. K (2024). A Deep Learning and Contend-Based FilteringTechniques for Improving Plant Disease Identification and Treatment A ComprehensiveReview. Heliyon 10(9), P. e29583.

Li, Y., Mao, H., Girshick, R., and He, K. (2022), "Exploring plain vision transformer backbone for object detection," Proceedings of European Conference on Computer Vision, pp. 280–296, Springer Nature Switzerland.

Mao, X., Qi, C., Chen, Y., Li, X., Duan, R., Ye, S., and Xue, H. (2022), "Towards Robust Vision Transformer," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12042–12051.

Meng, Q., Song, H., Zhang, Y., Zhang, X., Li, G and Yang, Y. (2020), “Video-based vehicle counting for expressway: A novel approach based on vehicle detection and correlation-matched tracking using image data from PTZ Cameras,” Math. Probl. Eng., 1–16.

Soudeep, S., Aunthy, L. N., Jim, J. R., Mridha, M. F., & Kabir, M.(2024) . Enhancing road traffic flow in sustainable cities through transformer models: Advancements and challenges. Sustainable Cities and Society, Volume 16, Article 105882.

Thisanke, H., Deshan, C., Chamith, K., Seneviratne, S., Vidanarachchi, R. and Herath, D. (2023), "Semantic segmentation using vision transformer – A survey," Engineering Applications of Artificial Intelligence, 126, Article 106669.

Wang, Y., Deng, Y., Zheng Y., Chattopadhyay, P & Wang, L. (2025). Vision Transfomer for Image Classification: A Comparative Survey, Technologies, volume 13, issue 1, article 10.3390.

Zhang, J., Peng, J., Kong, X., Wang, S., & Hu, J. (2025). Vehicle spatiotemporal distribution identification in low-light environments based on Image Enhancement and Object Detection. Advanced Engineering Informatics, Article 103165.