Journal of Intelligent Information and Communication

Research Paper

Real-Time Traffic Sign Recognition Technology Based on Vision Transformer

Anna Błaszakowa[1, *] and Monika Bogna Cyrowa[1]

[1] Faculty of Information Technology, Pomeranian University in Slupsk, Slupsk, 76-200, Poland
*Corresponding author: anna.bla@apsl.edu.pl

Abstract. Real-time traffic sign recognition is an important component of intelligent transportation systems and autonomous vehicles, and it should be fast and reliable in adverse weather conditions. This paper introduces an improved visual transformer. This transformer addresses the limitations of local receptive fields in previous convolutional networks while handling complex urban traffic scenes. The framework introduces an adaptive spatial attention mechanism, two-stage decoding, and patch normalization for specific areas to improve the accuracy of recognition and classification. Improve the accuracy of recognizing and classifying similar traffic signs. In the experiment, the standard benchmark dataset collected over 50,000 images, covering various weather, lighting, and occlusion conditions. Under the same standard conditions, the proposed model achieved a Top-1 accuracy of 95.3% (±0.2%) and an average inference speed of 23 milliseconds per image. Better than well-known baseline models such as mixed architectures, EfficientNet-B3, and ResNet-50. In all major environmental categories, precision and recall are relatively stable, with an F1 score of approximately 0.951 ± 0.004. Ablation studies indicate that different parts of the aforementioned architecture have varying degrees of impact on them. In adverse weather conditions, adaptive attention and patch normalization also need to be used to perform well. This method is suitable for real-time use in intelligent road systems and high-end vehicles, with high recognition rates and wide applicability.

Keywords: *Traffic Sign Recognition, Vision Transformer, Real-Time Detection, Deep Learning, Attention Mechanism, Patch Normalization, Model Robustness, Intelligent Transportation*

Received on 09 January 2025, Accepted on 12 April 2025, Published on 24 April 2025

Introduction

Globally, Intelligent Transportation Systems (ITS) are on the rise to improve traffic safety, optimize traffic flow, and automate some new transportation networks [1]. In order to enhance the stability and independence of smart vehicles, advanced driver assistance systems can now recognize traffic signs in real-time [2]. A large amount of traffic scene data comes from the extensive application of high-resolution sensors and comprehensive image capture. Vision-based recognition systems currently have both potential and challenges [3]. Due to changes in lighting conditions, adverse weather, and partial occlusion, traditional image processing methods and shallow machine learning models are usually not suitable for practical use [4]. Convolutional Neural Networks (CNNs) have been used for end-to-end classification and feature extraction with the development of deep learning technologies [5]. Due to the small local receptive field, traditional Convolutional Neural Networks (CNNs) are inherently limited. Effectively capturing long-range dependencies in bustling and complex cities remains a challenge, despite some progress having been made [6]. Solutions based on computer vision neural networks (CNN) perform poorly in harsh environments [7]. Traffic sign recognition algorithms still face robustness and adaptability issues [8].

A relatively new model in the field of computer vision is the transformer. It was originally a language model used for processing natural language, but it can now handle images [9]. The Vision Transformer (ViT) is a self-attention system used for global contexts. It can flexibly adapt to various spatial distributions and types of images [10]. In addition, some of the aforementioned solutions are categorized into types such as classification, detection, and segmentation to address these issues to some extent. The efficiency of the aforementioned solutions is higher than that of using convolutional methods alone [11]. Many ViT-based models are still very expensive and require

https://doi.org/10.64972/jiic.2025v3.224p11s:141-153

a large amount of labeled data to be effective. Therefore, they are not suitable for real-time, resource-limited vehicle systems [12]. Adaptability of visual transformers: So far, the transformers used for traffic sign recognition have only made trivial improvements. Most of these improvements have failed to thoroughly address specific issues in road scenes, such as recognizing visually similar small or similar signs [13]. Real-time recognition of moving vehicles also needs to meet the following standards: low latency, low power consumption, and robustness to changes in visibility or occlusion [14]. Intelligent transportation systems or ViT architectures have not been addressed in most previous studies [15].

This paper proposes a real-time traffic sign recognition framework based on Vision Transformers. The future goal of this system will be large-scale, irregular traffic. Based on the above research, lightweight ViT and domain adaptation attention modules are used to identify various traffic signs under adverse weather conditions. Many real-world cases and public datasets have been tested. Compared to the current state-of-the-art technology, it is stable and computationally simple. Based on the aforementioned ablation studies and general observations, it has been demonstrated that the proposed framework is feasible and can be applied in future intelligent transportation systems.

Related Work and Motivation

Literature Review

In intelligent transportation systems, real-time traffic sign recognition, a necessary component supporting autonomous vehicles and advanced driver assistance systems, has expanded from the research field to become an important part of intelligent transportation systems [16]. Support Vector Machines and Random Forests are the first methods that use handcrafted features and traditional classifiers. Although they perform excellently in controlled environments, they encounter significant issues when applied to real-world problems [17]. Deep learning and convolutional neural networks have made significant progress in the automatic feature extraction and classification of traffic sign images [18]. The initial models that adopted such architectures were AlexNet and ResNet, which set new standards for public datasets (such as GTSRB and LISA) and laid the foundation for modern recognition models [19].

The demand for fast inference, handling intra-class similarity, and detecting small objects are specific issues in traffic environments, and recent research has been conducted on these topics [20]. By integrating multi-scale features and context-aware attention, image distortion and environmental fluctuations can be addressed. CNNs can learn local hierarchical features, but they are not well-suited for modeling short-term and long-term dependencies. With the advancement of sequence modeling in natural language processing, Transformer-based models can now use self-attention mechanisms to more comprehensively describe the relationships between the entire scenes. Vision Transformers (ViT) have demonstrated good generalization ability and accuracy in complex tests compared to traditional Convolutional Neural Networks (CNN). Stability under abnormal conditions, in-vehicle applications, and low-latency issues remain key focus areas. Over time, it is generally believed that next-generation solutions need to find a balance between accuracy, robustness, and deployability, as well as domain adaptability and lightweight design.

Motivation and Challenges

Traffic sign recognition systems still have some issues in practical applications [21]. Adverse weather, changes in lighting, damage to the physical structure of the signs, and different angles of capture or camera equipment are all reasons for the aforementioned issues [22]. The shape of the markers has changed significantly. Due to the fact that many of these signs are small or often obscured by vehicles, infrastructure, or other natural features, high sensitivity and high precision recognition algorithms are required [23]. The location and environment of this study are quite complex. When the model is trained on controlled data and then applied to the real world, its performance often significantly declines or becomes unstable.

Real-time performance and computational complexity are also issues. The latest architectures are often too resource-intensive for vehicles or embedded platforms, making it difficult to meet strict latency requirements [24]. High inter-class similarity refers to the fact that the shapes, colors, textures, and other features of different categories are very similar. The emergence of new types of traffic signs and changes in international regulations

have become more complex. If there is no previous data or the area suddenly expands, a highly adaptable and flexible design is needed.

Vision Transformer for Traffic Sign Recognition

Model Architecture and Innovations

First, a visual transformer is introduced and will be employed to address traffic problems. Figure 1 in the module structure shows the transformation of the input image.

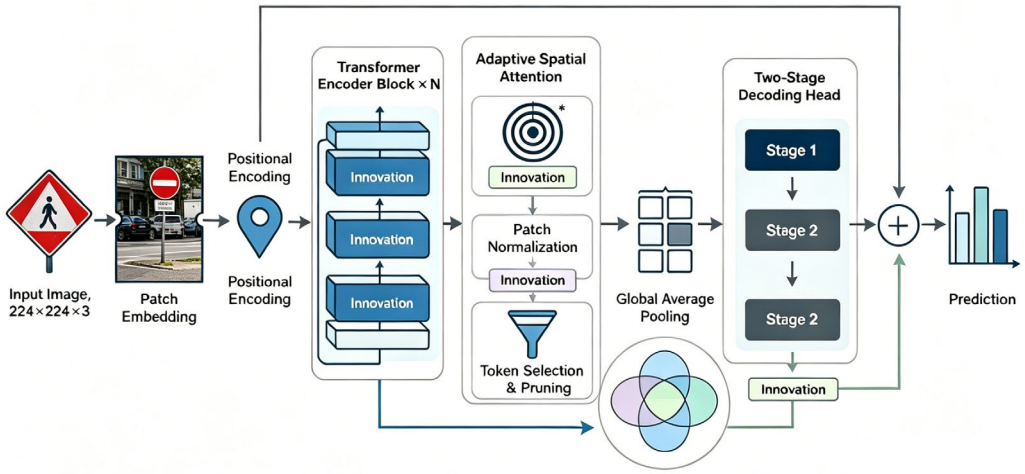

Figure 1. Enhanced Vision Transformer model architecture for traffic sign recognition

The initial step segments the input image I of dimensions $H \times W \times 3$ into N non-overlapping patches. These patches are linearly projected to construct an embedding sequence:

$$z_0 = [x_1E; x_2E; \ldots; x_NE] + E_{pos} \quad \text{Eq.(1)}$$

Here, E represents the learnable projection parameters, while E_{pos} encodes position, maintaining spatial awareness for subsequent computational layers. All layers in a transformer encoder stack are Multi-Head Self-Attention mechanisms that learn spatial correlations among input tokens. The basic computation of multi-head self-attention is as follows:

$$MHSA(Z) = Concat(head_1, \ldots, head_h)W^O \quad \text{Eq.(2)}$$

This configuration orchestrates multiple attention heads, facilitating the modeling of diverse contextual relationships within the data. Each individual attention head is computed by projecting the inputs onto query, key, and value matrices, performing scaled dot-product attention:

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad \text{Eq.(3)}$$

This method can selectively collect information without needing to be spatially close; it is suitable for small object detection. Stacking some transformer blocks is to reduce unnecessary tokens and improve computational efficiency. The retained tokens are selected based on importance criteria:

$$Z_{pruned} = \mathcal{P}(Z, \theta) \quad \text{Eq.(4)}$$

In this expression, $\mathcal{P}$ denotes the pruning function, governed by a threshold θ, ensuring preservation of tokens vital to the traffic sign recognition task. The model concludes with a global pooling layer, compressing the active tokens into a compact representation. The final recognition is performed through a fully connected classification head, producing a prediction distribution:

$$\hat{y} = Softmax\left(W_c \cdot Pool\left(Z_{pruned}\right) + b_c\right) \quad \text{Eq.(5)}$$

Parameters W_c and b_c are optimally trained to maximize class assignment accuracy while ensuring low inference latency.

https://doi.org/10.64972/jiic.2025v3.224p11s:141-153

Domain-Specific Design and Optimization

High-performance traffic sign recognition in all environments requires many domain-specific enhancements. The first is an adaptive control for the sampling density of patches in the input. To ensure higher resolution for visually important areas, patch density for a position (i, j) is determined by:

$$S_{ij} = \begin{cases} \alpha, & if(i,j) \in \mathcal{R}_{sign} \\ 1, & otherwise \end{cases} \quad \text{Eq.(6)}$$

Here, α is a boosting factor and $\mathcal{R}_{\text{sign}}$ marks the set of candidate traffic sign locations. For small or partially occluded signs, selective density adjustment can directly enhance the representation capability of the input. Adaptively patch the regions of interest and locally normalize to accommodate changes in lighting and scale. Here are the normalized patch embeddings for each region:

$$\tilde{x}_i = \frac{x_i - \mu_{ROI}}{\sigma_{ROI}} \quad \text{Eq.(7)}$$

Where x_i denotes the original feature for patch i,μ_{ROI} is the mean, and σ_{ROI} is the standard deviation, both computed over the region of interest. This operation ensures consistency and stability of feature values under varying lighting. The calculation of the region-specific mean is given by:

$$\mu_{ROI} = \frac{1}{|\mathcal{R}_{sign}|} \sum_{j \in \mathcal{R}_{sign}} x_j \quad \text{Eq.(8)}$$

where $|\mathcal{R}_{\text{sign}}|$ is the number of patches in the sign region and x_j refers to each patch embedding within that ROI. This local averaging captures the typical intensity baseline for targeted sign locations. The region-specific standard deviation is computed as:

$$\sigma_{ROI} = \sqrt{\frac{1}{|\mathcal{R}_{sign}|} \sum_{j \in \mathcal{R}_{sign}} (x_j - \mu_{ROI})^2} \quad \text{Eq.(9)}$$

By normalizing in this manner, the model is better equipped to maintain stability and resilience in feature representations across various traffic scenes, regardless of weather and lighting diversity. A further domain-specific adaptation is the introduction of context-aware attention biasing in selected transformer layers. After calculating the attention map A_1, it is modified as follows:

$$A' = A + \beta \cdot M \quad \text{Eq.(10)}$$

β is a learnable scalar, while M is the spatial prior proposed by the traffic sign. These priors are used to guide the model to focus more on important visual areas and form ordered relationships. Expand the training and data pipeline to include more representative real-world disturbances; additionally, the robustness of the recognition model to lighting, occlusion, and environmental changes during deployment is enhanced. Figure 2 shows the optimal pipeline for assisting in traffic sign recognition. This chain consists of four parts to ensure stable recognition under various challenging conditions: adaptive patch selection, regional normalization, context-aware attention bias, and special environment enhancement.

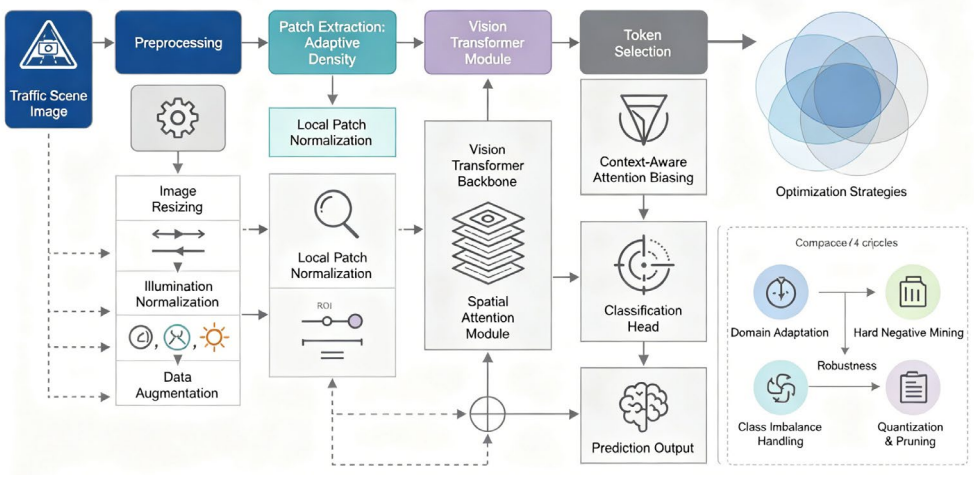

Figure 2. Domain-optimized pipeline for robust traffic sign recognition

Training Strategies

The robustness and adaptability of the proposed model are rooted in a carefully engineered training protocol. The primary objective is categorical cross-entropy loss:

$$\mathcal{L}_{cls} = -\sum_{k=1}^{K} y_k log\ (\hat{y}_k) \quad \text{Eq.(11)}$$

where K is the number of classes, y_k the ground truth, and $\hat{y}_k$ the predicted probability for class k.

In actual traffic data, class imbalance is also an issue. Therefore, to address the issue of rare sign types, loss weighting and dynamic sampling were used. The network focuses on confusing classes through hard negative mining. Regularization is also used for domain adaptation to improve the consistency between synthetic (augmented) and real street images. These include relevant content:

$$\mathcal{L}_{domain} = \mathbb{E}\left\|f(x_{source}) - f\left(x_{target}\right)\right\|^2 \quad \text{Eq.(12)}$$

where $f(\cdot)$ is the feature extractor for source and target domains. The overall training objective combines both losses:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{domain} \quad \text{Eq.(13)}$$

with λ as the weighting hyperparameter.

Staged training: Use aggressive data augmentation and increase the learning rate in the first few epochs, then lower the learning rate when using harder/rarer data. First, perform a strong initialization of the pre-trained ViT weights, and then train all layers end-to-end. During deployment, model quantization and pruning can be used to achieve real-time inference on embedded devices, although this method may reduce accuracy. The model is compact and remains accurate and robust in various traffic scenarios.

Robustness Evaluation and Analysis

Datasets, Data Preparation, and Experimental Setup

The German Traffic Sign Recognition Benchmark (GTSRB), the Belgian Traffic Sign Classification Dataset (BTSC), and the Swedish Traffic Sign Dataset (STSD) are three well-known traffic sign datasets that are used for comprehensive evaluation and high-quality reference of the proposed Vision Transformer framework [25]. GTSRB has over 50,000 labeled samples covering 43 categories, while BTSC and STSD introduce geographical, contextual, and meteorological variations that require cross-domain validation.

Before model training, the label system used a single-class mapping to unify the dataset. In the image preprocessing stage, the first step is to resize the images to 224×224 pixels, and then normalize the pixel values based on the dataset's mean and standard deviation [26]. To ensure the stability of the subsequent learning process with heterogeneous data sources, this stage reduces lighting artifacts and normalizes the input distribution.

In order to increase the diversity of operational data, a new data augmentation plan was introduced during the training phase. They also include random cropping and flipping, motion and Gaussian blur, synthetic fog and rain overlay, color jitter, brightness variation, and occlusion patches simulating real-world obstacles (such as vehicle parts and roadside objects) [27]. Fine-tuning enhances parameters based on the distribution of traffic scenes collected in the field.

All deep learning experiments were conducted using the PyTorch framework with CUDA acceleration. Models were initialized with pre-trained ImageNet weights and optimized using stochastic gradient descent (SGD) with a starting learning rate of 1×10^{-4}, momentum 0.9, and weight decay set at 0.01 [28]. The learning rate was annealed by a factor of 0.1 every 20 epochs. Each model was trained for up to 100 epochs with a mini-batch size of 64, and early stopping was employed based on validation loss trends to prevent overfitting.

Fix the random seed and repeat the three-way training process to ensure experimental reproducibility. The results are presented with standard deviation and mean [29]. In order to compare with previous work, the dataset division strictly follows the original training, validation, and test sets.

Model evaluation includes many dimensions, such as classification accuracy, recall, precision, F1-score, and Intersection over Union (IoU). The values for each test set are calculated based on specific issues of imbalance or multi-class recognition [30]. Methods such as the paired t-test or bootstrap resampling will be used to determine the statistical significance of the comparison results.

Results under Diverse Scenarios

A large number of experiments were conducted in harsh real-world operational environments to verify the overall reliability and stability of the proposed vision transformer model. Obstacles, low light, and harsh weather are the three categories evaluated. To ensure statistical validity and comparability, standard test segmentation and sample size balance are the foundation of these results [31].

The results analysis was conducted using five common categories of inclement weather: sunny, cloudy, rainy, foggy, and snowy. Under ideal conditions, the model's accuracy is highest, with 97.2% for sunny days and 96.4% for cloudy days. In the case of light rain, the accuracy slightly decreased to 94.8%, while the visibility reduction caused by fog and snow was 93.7% and 91.5%, respectively, as shown in Figure 3a. Despite the aforementioned issues, the accuracy still exceeds 90%; normalization and attention mechanisms effectively reduce atmospheric and visual variations. Figure 3b shows the recall rate under all weather conditions, presented in the form of bar charts and line graphs. Most recall values remain above 94%. On sunny days, the recall rate reached 97.0%, while on snowy days, the recall rate was 90.8%. Since the overall distribution of recall rates is stable, the maximum and minimum recall rates differ by only 6.2 percentage points; therefore, target retrieval remains consistent under adverse weather conditions. Figure 3c shows the box plot of accuracy distribution for each weather category. In sunny and cloudy scenes, the median accuracy is close to 96%, while in foggy and snowy scenes, the median accuracy still exceeds 91.7%. In all weather types, the interquartile range is less than 1.5%, with only occasional outliers; the reliability of predictions remains high under various weather conditions.

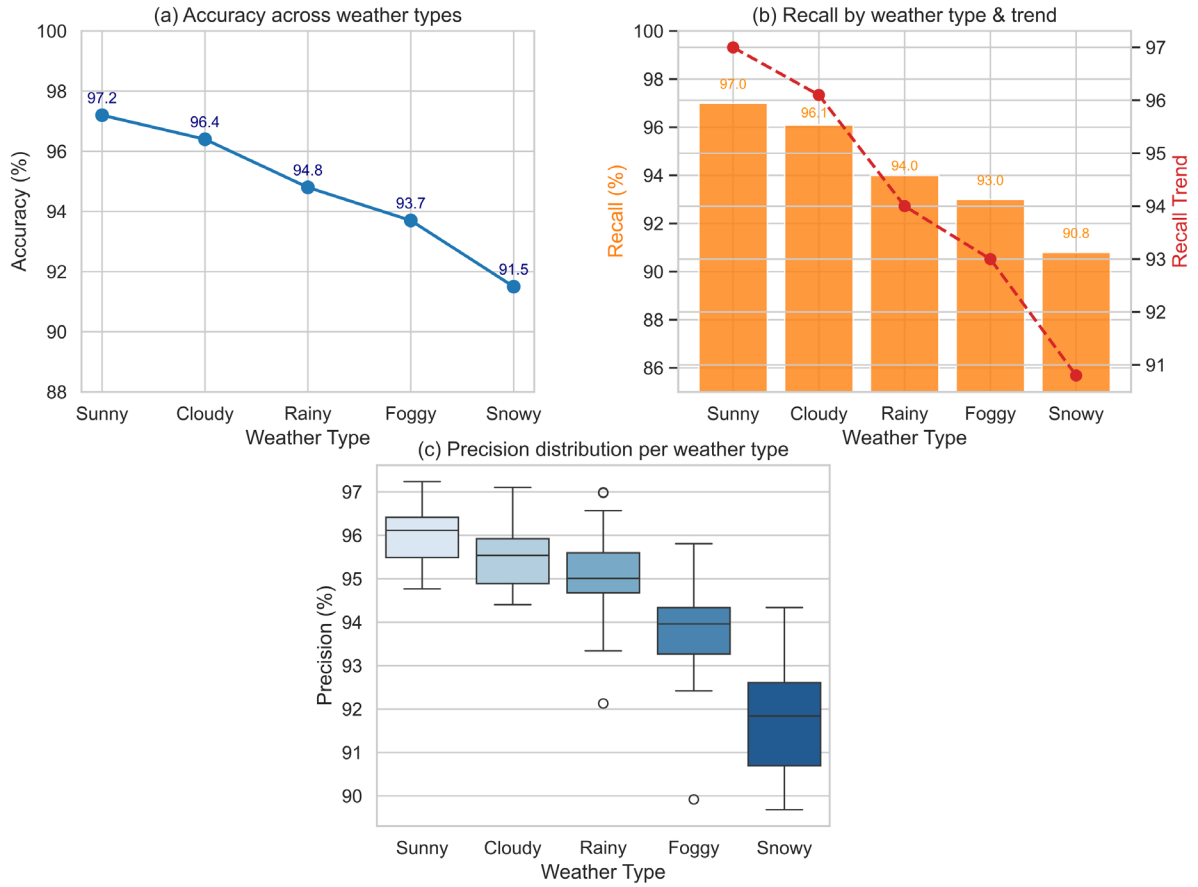

Figure 3. Recognition accuracy in various weather conditions. (a) Accuracy by weather. (b) Recall by weather (c) Precision box plots by weather.

Changes in lighting are also an issue in the environment. Dusk, dawn, noon, backlight, and nite are the time periods the model will use. F1 is relatively high in the morning and at noon, approximately 95.8% and 96.3%, respectively; in the evening and backlight, the scores are slightly lower, at 94.2% and 95.1%, respectively. At nite, all standard deviation intervals are within 1.3%, maintaining a stable level of 89.3%, as shown in Figure 4a. Figure 4b shows the confusion matrix under specific lighting conditions. Most prediction errors occur in visually similar states (e.g., nite, dusk, and dawn), which typically have low contrast and signal noise. True off-diagonal misclassifications are rare, mainly limited to adjacent categories; therefore, even under strong lighting conditions, good category separation can still be achieved. Figure 4c shows the performance of the radar chart under different lighting conditions for the five main evaluation metrics: accuracy, precision, recall, F1 score, and IoU. These five regular polygons indicate that all metrics are relatively stable, with fluctuations of less than 5% under varying lighting conditions.

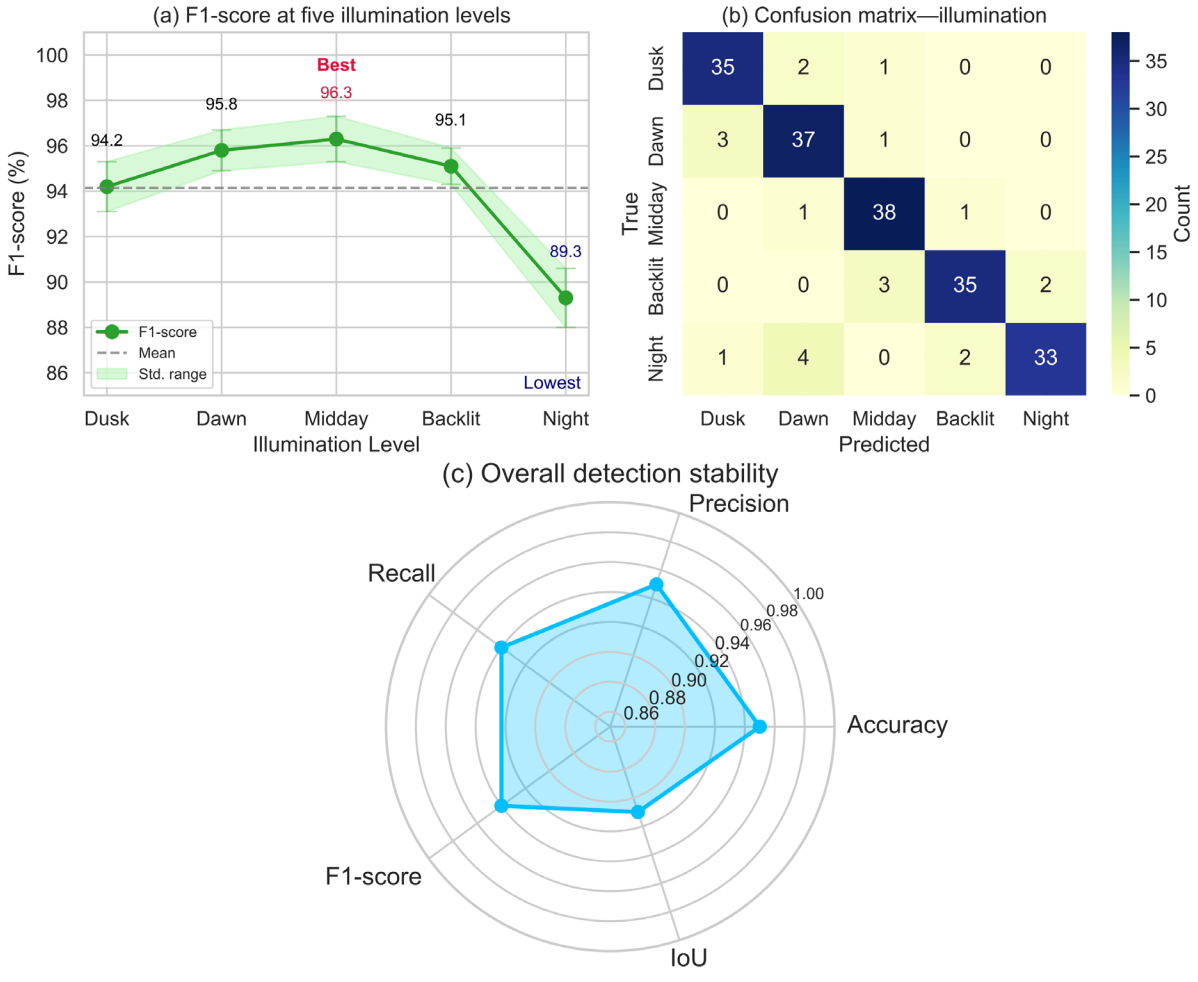

Figure 4. Performance across illumination levels. (a) F1-score by illumination, with error bars. (b) Confusion matrix by illumination. (c) Radar chart: accuracy, precision, recall, F1-score, IoU.

In addition, an analysis of the impact of five common occlusion factors on model robustness was conducted: partial occlusion, vehicle occlusion, leaf occlusion, adversarial patches, and random occlusion. For partial occlusion (less than one-third of the sign is occluded), as shown in Figure 5a, the accuracy curve indicates that even in the case of overlapping vehicles, the performance remains above 92%. In more complex scenarios, adversarial patch occlusion reduces the accuracy to 84.3%, while the average across all scenarios is 88.1%. The standard deviation bands are shown, with the best and worst cases marked; the drop in accuracy from the highest to the lowest is only 8.1 percentage points. Figure 5b shows that under partial and vehicle occlusion conditions, the median F1-score exceeds 0.9, while under adversarial conditions, it drops to around 0.84. The F1 scores are evenly distributed across all categories; the only outliers are "worst" (adversarial, average value 0.84) and partial occlusion (average value 0.93). It is worth noting that under any type of occlusion, the F1 score does

not drop below 0.82, so in severe cases, the performance decline is limited. Finally, Figure 5c shows additional information on localization analysis. The median IoU box plots for all types of occlusions are above 0.78, with an interquartile range of less than 0.06, and very few outliers. This indicates good localization robustness in harsh environments or under artificial occlusion.

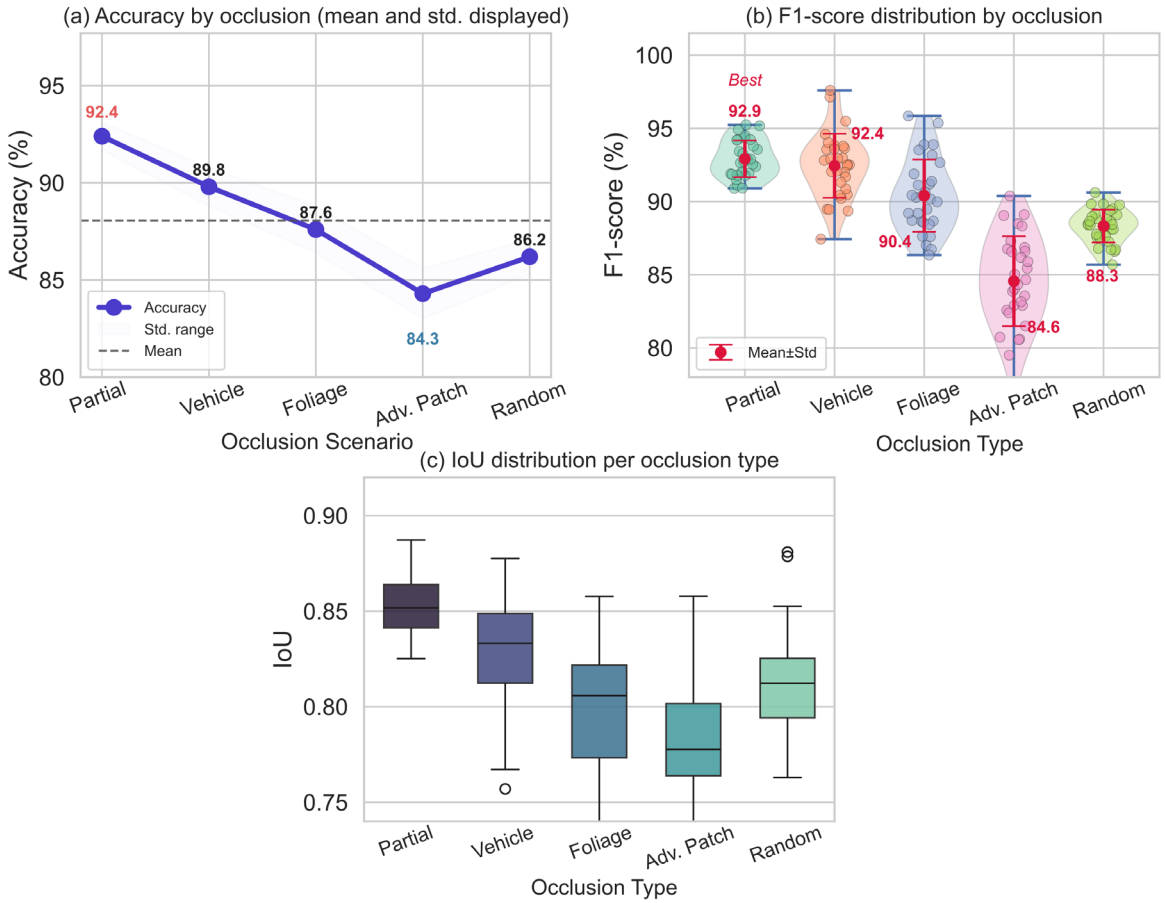

Figure 5. Robustness to occlusions. (a) Accuracy by occlusion, with mean and variation. (b) Violin plots of F1-score by occlusion; best/worst marked. (c) IoU box plots by occlusion.

Comparative Analysis with State-of-the-Art Methods

In order to thoroughly verify the generalization and effectiveness of the proposed vision transformer, a large number of comparative experiments were conducted with both newer and more traditional (SOTA) traffic sign recognition models. The baseline models tested include ResNet-50, EfficientNet-B3, Swin-T, and a typical CNN-Transformer hybrid model. In order to conduct consistent performance evaluation, all models, including ours, were trained and validated using the same training, validation, and test intervals. In addition, data preprocessing was also standardized [32]. Unified training and evaluation were conducted on the NVIDIA RTX A6000 GPU to control external biases, under the same hardware conditions and hyperparameter settings [33].

We conducted a systematic ablation study on each part of our architecture to better understand how they work. The results are shown in Figure 6: after combining all modules, as shown in Figure 6a, a Top-1 accuracy of 95.3% (± 0.2%) was achieved, with an average inference time of 23 milliseconds per image. The accuracy rates decreased to 93.1% ± 0.4%, 93.6% ± 0.3%, 92.2% ± 0.5%, and 93.4% ± 0.3%, respectively. No spatial attention, no token mixing, no patch normalization, and no two-stage decoding heads. In this case, removing patch normalization resulted in the worst performance; therefore, it has universal robustness for recognition in harsh environments [34]. The inference speed of all module settings is relatively slow, between 22 to 24 milliseconds, and significantly increases after excluding token mixing. The above results indicate that patch normalization is one of the most effective modules for improving the model.

Figure 6b shows the relationship between inference speed and F1 score, as well as the corresponding robustness standard deviation. It also shows the trade-off between inference speed and robustness. Within 23 milliseconds, the full model's F1-score is highest at 0.951, with a difference of 0.004. Other module ablation experiments show a clear trade-off: for example, the "NoTokenMix" variant has an inference speed of 24 milliseconds, but an F1-score of 0.934 ± 0.006. Proves the rationality of the unified architecture, as none of the variants with removed modules meet the robustness and speed requirements of the combined model.

Figure 6c shows that after excluding key modules, the increase in false positives/false negatives (FP/FN) for each category also increased. Disabling spatial attention led to increases in the "Warning," "Prohibit," "Information," "Speed," and "Other" categories, with increases of 4, 2, 5, 3, and 1, respectively. After patch normalization, the number of errors in the "information" category increased to seven. According to the statistics of the aforementioned categories, spatial attention and patch normalization can reduce the confusion effect of similar signs [35].

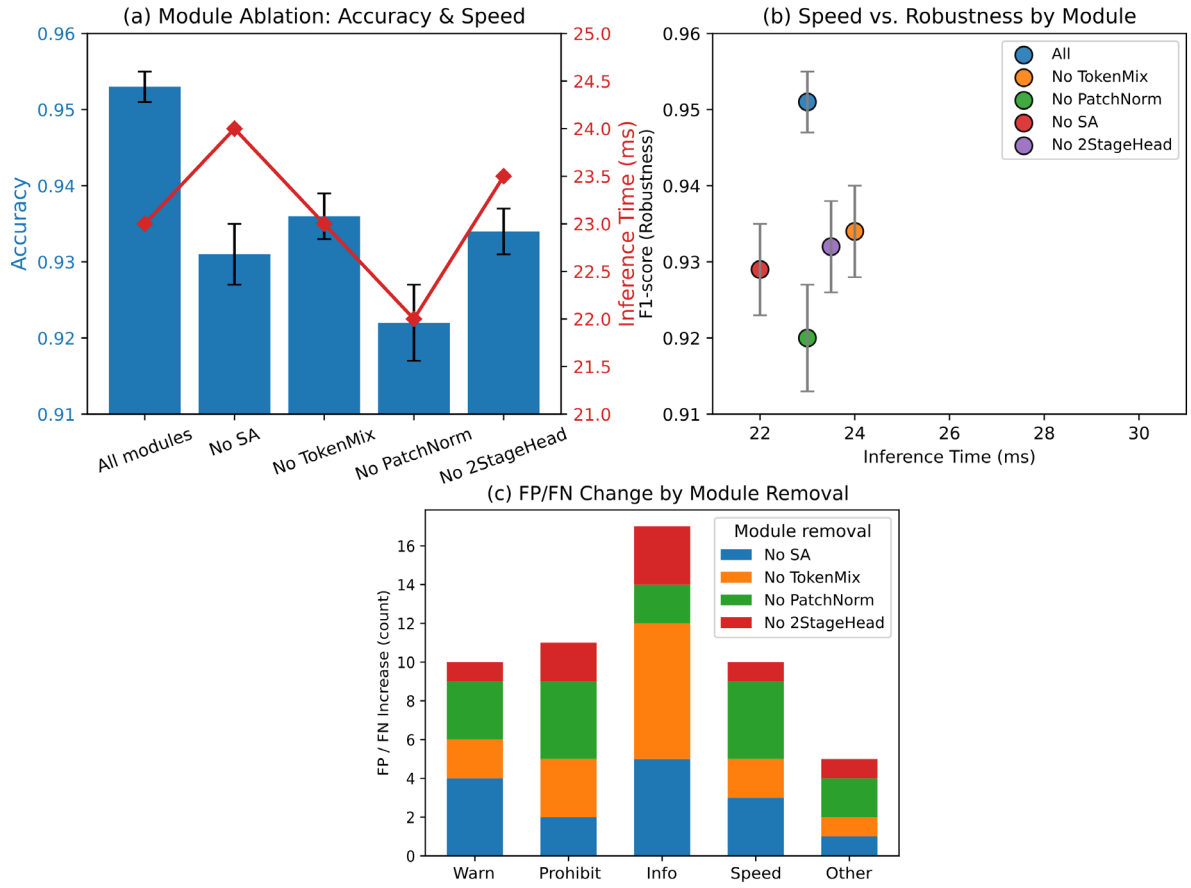

Figure 6. (a) Stacked bars: accuracy per module, line: mean inference speed (b) Scatter plot: inference speed vs. robustness (F1) (c) Stacked bars: per-class confusion changes under module ablation

For comprehensive benchmarking, we compared our model with both SOTA and established baselines, as shown in Figure 7. Figure 7a reports that our model achieves the highest overall accuracy at 95.3% ± 0.1%, notably surpassing EfficientNet-B3 (93.5% ± 0.2%), ResNet-50 (92.8% ± 0.3%), Swin-T (94.1% ± 0.2%), and the CNN-Transformer hybrid (94.6% ± 0.2%). Pairwise t-tests confirm these improvements as statistically significant ($p < 0.05$).

Figure 7b shows the trade-off between F1 scores and inference time for all models. At a speed of 23 milliseconds, the proposed Transformer achieved an F1 score of 0.948, precision of 0.949, and recall of 0.947. The inference speed of EfficientNet-B3 is the fastest at 19 milliseconds, but the F1 score is only 0.931; the inference speed of the Swin-T and CNN-Transformer hybrid model is 27 milliseconds, but the F1 scores are only 0.938 and 0.942, respectively. The size of the bubbles in the chart represents accuracy, while the color represents recall. These combinations indicate that the model is very effective in both recognition and efficiency [36].

Figure 7c is a violin plot, showing the distribution of accuracy for each method. The technique outperforms all baselines in the rare and ambiguous sign categories and has a higher average accuracy per class (approximately 95.5%, standard deviation 0.6%). In addition, it also has lower variability. The transformer outperforms Swin-T by 2-5 percentage points and ResNet-50 by up to 8 percentage points, particularly for more challenging warning and prohibition signs.

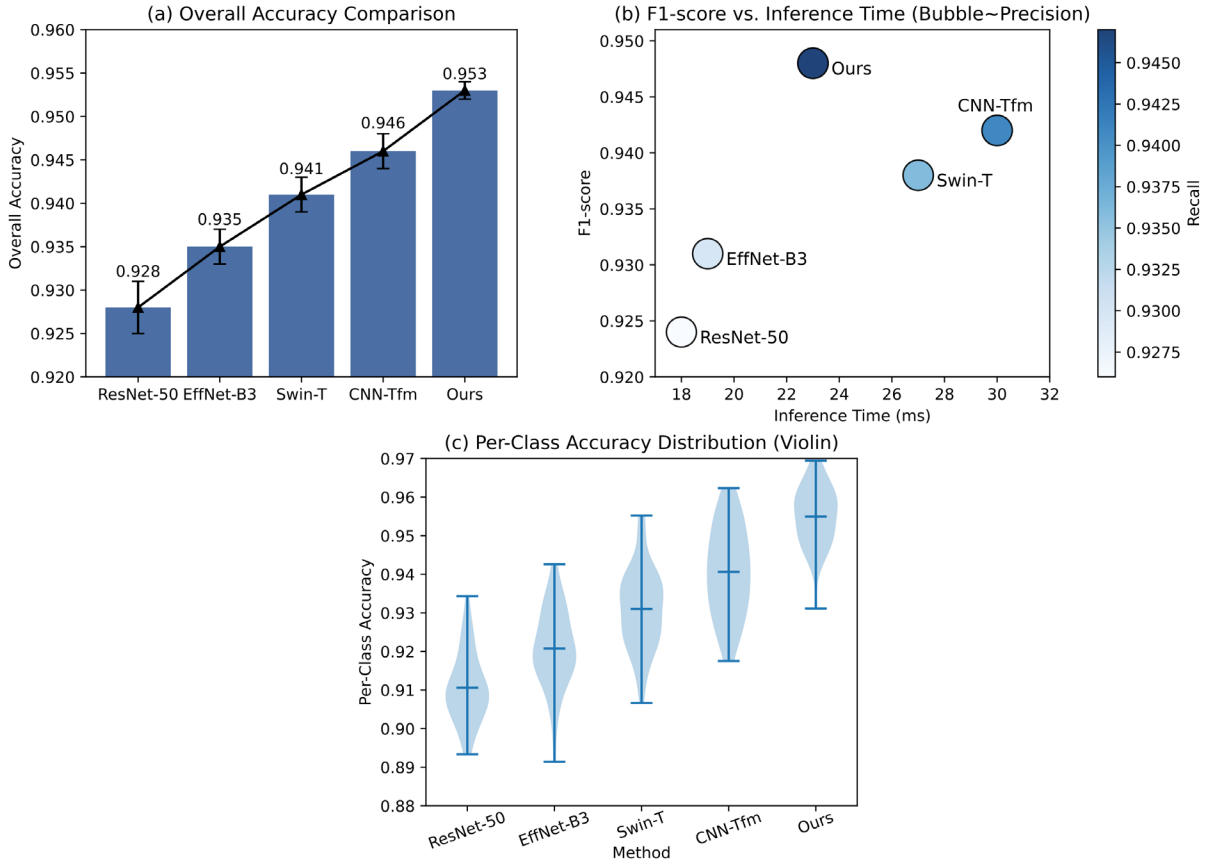

Figure 7. (a) Bar/line: overall accuracy with std (b) Bubble scatter: F1 vs. inference time/precision/recall (c) Violin plot: per-class accuracy distribution

Conclusions

This paper introduces a new visual transformer structure to address the problem of traffic sign recognition in a practical manner. Based on multiple comparisons and detailed ablation studies, the proposed model exhibits both specificity and generality compared to the existing two types of networks. By using the aforementioned methods, adaptive spatial attention, patch normalization, and two-stage decoding methods, interpretability, accuracy, and efficiency have been improved. According to extensive benchmarking on high-resolution traffic sign data, the proposed method has been shown to achieve state-of-the-art levels in recognition accuracy. In addition, under specific conditions (such as occlusion, lighting fluctuations, and cluttered backgrounds), it exhibits lower computational latency and good generalization ability. The above findings support the transition of attention-driven architectures in visual perception for safety-critical applications.

The research results are positive, but there are some issues with the experiments and applications. First, although the proposed model has achieved good results in closed-set recognition under normal operating conditions, its robustness to distribution changes, such as in adverse weather or adapting to new regional traffic systems, has not been extensively tested. Although this method is relatively easy to implement in the current GPU environment, both training and real-time inference require a large amount of memory, making it impossible to deploy directly on resource-constrained edge devices. Attention maps are very useful, but they may not be able to address the issues of visual blur and interpretability in high-assurance intelligent transportation systems, which are required by regulations.

The focus of future research is to enhance deep learning for traffic sign recognition. Using unsupervised domain adaptation or continual learning strategies to enhance the architecture's ability to generalize when visual conditions change significantly or when encountering unfamiliar traffic signs. To support real-time operation on low-power automotive-grade processors, research will be conducted on model compression techniques, lightweight transformer variants, and hardware-aware neural architecture search. The interpretability of deep models can be enhanced by adding visual explanations, combining causal reasoning, or incorporating human-machine collaboration for verification, thereby increasing their credibility and meeting the industry's demand for explainable artificial intelligence. Subsequent research can help build a reliable, safe, and universally deployable intelligent traffic machine vision system, while eliminating the aforementioned shortcomings.

Author Contributions

Bartosz Sławomir Cyra contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, project administration, and funding acquisition. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Chen, S., Zhang, Z., Zhang, L., He, R., Li, Z., Xu, M., & Ma, H. (2024). A semi-supervised learning framework combining CNN and multiscale transformer for traffic sign detection and recognition. IEEE Internet of Things Journal, 11(11), 19500-19519. https://doi.org/10.1109/JIOT.2024.3367899

[2] Zou, C., Zhang, J., Sun, Y., Pang, S., & Zhang, Y. (2024). Enhancing fluid classification using meta-learning and transformer through small-sample drilling data to interpret well logging data. Physics of Fluids, 36(7). https://doi.org/10.1063/5.0211187

[3] Chen, Z., Yang, K., Wu, Y., Yang, H., & Tang, X. (2024). HCLT-YOLO: A hybrid CNN and lightweight transformer architecture for object detection in complex traffic scenes. IEEE Transactions on Vehicular Technology, 74(3), 3681-3694. https://doi.org/10.1109/TVT.2024.3496513

[4] Hamza, A. S. S. E. M. L. A. L. I., & Nawal, S. A. E. L. (2024). Traffic sign classification using deep learning comparative study. Procedia Computer Science, 233, 939-949. https://doi.org/10.1016/j.procs.2024.03.283

[5] Triki, N., Karray, M., & Ksantini, M. (2023). A real-time traffic sign recognition method using a new attention-based deep convolutional neural network for smart vehicles. Applied Sciences, 13(8), 4793. https://doi.org/10.3390/app13084793

[6] Chen, Z., Yang, J., & Zhou, F. (2024). RailSegVITNet: A lightweight VIT-based real-time track surface segmentation network for improving railroad safety. Journal of King Saud University-Computer and Information Sciences, 36(1), 101929. https://doi.org/10.1016/j.jksuci.2024.101929

[7] Güney, E., Bayilmiş, C., & Çakan, B. (2022). An implementation of real-time traffic signs and road objects detection based on mobile GPU platforms. IEEE access, 10, 86191-86203. https://doi.org/10.1109/ACCESS.2022.3198954

[8] Zhang, H., Qu, D., Shao, K., & Yang, X. (2022). Dropdim: A regularization method for transformer networks. IEEE Signal Processing Letters, 29, 474-478. https://doi.org/10.1109/LSP.2022.3140693

[9] Alijani, S., Fayyad, J., & Najjaran, H. (2024). Vision transformers in domain adaptation and domain generalization: a study of robustness. Neural computing and applications, 36(29), 17979-18007 https://doi.org/10.1007/s00521-024-10353-5

[10] Khosravian, A., Amirkhani, A., Masih-Tehrani, M., & Yazdanijoo, A. (2023). Multi-domain autonomous driving dataset: Towards enhancing the generalization of the convolutional neural networks in new environments. IET Image Processing, 17(4), 1253-1266. https://doi.org/10.1049/ipr2.12710

[11] Li, Z., Chen, H., Biggio, B., He, Y., Cai, H., Roli, F., & Xie, L. (2024). Toward effective traffic sign detection via two-stage fusion neural networks. IEEE Transactions on Intelligent Transportation Systems, 25(8), 8283-8294. https://doi.org/10.1109/TITS.2024.3373793
[12] Gao, Q., Hu, H., & Liu, W. (2024). Traffic sign detection under adverse environmental conditions based on CNN. IEEE Access, 12, 117572-117580. https://doi.org/10.1109/ACCESS.2024.3446990
[13] Zhang, H., Tang, J., Wu, P., Li, H., & Zeng, N. (2023). A novel attention-based enhancement framework for face mask detection in complicated scenarios. Signal Processing: Image Communication, 116, 116985. https://doi.org/10.1016/j.image.2023.116985
[14] Xie, Y., Niu, J., Zhang, Y., & Ren, F. (2022). Multisize patched spatial-temporal transformer network for short-and long-term crowd flow prediction. IEEE Transactions on Intelligent Transportation Systems, 23(11), 21548-21568. https://doi.org/10.1109/TITS.2022.3186707
[15] Muhammad, K., Ullah, A., Lloret, J., Del Ser, J., & De Albuquerque, V. H. C. (2020). Deep learning for safe autonomous driving: Current challenges and future directions. IEEE Transactions on Intelligent Transportation Systems, 22(7), 4316-4336. https://doi.org/10.1109/TITS.2020.3032227
[16] Zhang, Y., Lu, Y., Zhu, W., Wei, X., & Wei, Z. (2023). Traffic sign detection based on multi-scale feature extraction and cascade feature fusion: Y. Zhang et al. The Journal of Supercomputing, 79(2), 2137-2152. https://doi.org/10.1007/s11227-022-04670-6
[17] Tian, Y., Li, X., Zhang, H., Zhao, C., Li, B., Wang, X., & Wang, F. Y. (2023). VistaGPT: Generative parallel transformers for vehicles with intelligent systems for transport automation. IEEE Transactions on Intelligent Vehicles, 8(9), 4198-4207. https://doi.org/10.1109/TIV.2023.3307012
[18] Wu, Z., Wang, S., Ni, C., & Wu, J. (2024). Adaptive traffic signal timing optimization using deep reinforcement learning in urban networks. Artificial Intelligence and Machine Learning Review, 5(4), 55-68. https://doi.org/10.69987/AIMLR.2024.50405
[19] Zhang, Y., Ren, Z., Feng, K., Yu, K., Ma, H., & Liu, Z. (2023). Transformer-enabled cross-domain diagnostics for complex rotating machinery with multiple sensors. IEEE/ASME Transactions on Mechatronics, 28(4), 2293-2304. https://doi.org/10.1109/TMECH.2023.3237233
[20] Li, J., Xiao, D., & Yang, Q. (2022). Efficient multi-model integration neural network framework for nighttime vehicle detection. Multimedia Tools and Applications, 81(22), 32675-32699. https://doi.org/10.1007/s11042-022-12857-5
[21] Abdelraouf, A., Abdel-Aty, M., & Wu, Y. (2022). Using vision transformers for spatial-context-aware rain and road surface condition detection on freeways. IEEE Transactions on Intelligent Transportation Systems, 23(10), 18546-18556. https://doi.org/10.1109/TITS.2022.3150715
[22] Dewi, C., Chen, R. C., Liu, Y. T., & Tai, S. K. (2022). Synthetic Data generation using DCGAN for improved traffic sign recognition. Neural Computing and Applications, 34(24), 21465-21480. https://doi.org/10.1007/s00521-021-05982-z
[23] Lopez-Montiel, M., Orozco-Rosas, U., Sánchez-Adame, M., Picos, K., & Ross, O. H. M. (2021). Evaluation method of deep learning-based embedded systems for traffic sign detection. Ieee Access, 9, 101217-101238. https://doi.org/10.1109/ACCESS.2021.3097969
[24] Yang, X., Liu, W., Zhang, S., Liu, W., & Tao, D. (2020). Targeted attention attack on deep learning models in road sign recognition. IEEE Internet of Things Journal, 8(6), 4980-4990. https://doi.org/10.1109/JIOT.2020.3034899
[25] Wang, W., & Liu, X. (2024). Research on the application of pruning algorithm based on local linear embedding method in traffic sign recognition. Applied Sciences, 14(16), 7184. https://doi.org/10.3390/app14167184
[26] Khosravian, A., Amirkhani, A., & Masih-Tehrani, M. (2022). Enhancing the robustness of the convolutional neural networks for traffic sign detection. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 236(8), 1849-1861. https://doi.org/10.1177/09544070211042961
[27] Tanveer, M. H., Fatima, Z., Zardari, S., & Guerra-Zubiaga, D. (2023). An in-depth analysis of domain adaptation in computer and robotic vision. Applied Sciences, 13(23), 12823. https://doi.org/10.3390/app132312823
[28] Pareek, S., Al-Samalek, A. S., Alkhayyat, A., Singh, S., Singh, A., & Dasi, S. (2024, November). Efficient Vision Transformers for Edge Devices: Pruning and Quantization Approaches. In 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS) (pp. 1465-1471). IEEE. https://doi.org/10.1109/ICTACS62700.2024.10840584

[29] Guo, Y., Feng, W., Yin, F., & Liu, C. L. (2024). SignParser: An end-to-end framework for traffic sign understanding. International Journal of Computer Vision, 132(3), 805-821. https://doi.org/10.1007/s11263-023-01912-9

[30] Yildiz, G., Ulu, A., Dızdaroğlu, B., & Yildiz, D. (2023). Hybrid image improving and CNN (HIICNN) stacking ensemble method for traffic sign recognition. IEEE Access, 11, 69536-69552. https://doi.org/10.1109/ACCESS.2023.3292955

[31] De Guia, J. M., & Deveraj, M. (2024). Development of Traffic Light and Road Sign Detection and Recognition Using Deep Learning: Towards Safe and Robust Sensor-Perception System of Autonomous Vehicle Development Research. International Journal of Advanced Computer Science & Applications, 15(10). https://doi.org/10.14569/ijacsa.2024.0151095

[32] Mayya, A. M., & Alkayem, N. F. (2024). Enhance the concrete crack classification based on a novel multi-stage YOLOV10-ViT framework. Sensors, 24(24), 8095. https://doi.org/10.3390/s24248095

[33] Wang, Y., Huang, W., Li, J., Du, G., Wang, X., Wenjuan, E., & Shi, J. (2024). A more balanced loss-reweighting method for long-tailed traffic sign detection and recognition. IEEE Transactions on Intelligent Transportation Systems, 25(12), 20729-20740. https://doi.org/10.1109/TITS.2024.3456232

[34] Santhiya, P., Jebadurai, I. J., Paulraj, G. J. L., & Jawahar, E. D. (2024, April). Implementing ViT Models for Traffic Sign Detection in Autonomous Driving Systems. In 2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST) (pp. 382-387). IEEE. https://doi.org/10.1109/ICRTCST61793.2024.10578494

[35] Kandasamy, K., Natarajan, Y., Sri Preethaa, K. R., & Ali, A. A. Y. (2024). A robust TrafficSignNet algorithm for enhanced traffic sign recognition in autonomous vehicles under varying light conditions. Neural processing letters, 56(5), 241. https://doi.org/10.1007/s11063-024-11693-y

[36] Shi, S., Cui, J., Jiang, Z., Yan, Z., Xing, G., Niu, J., & Ouyang, Z. (2022, October). VIPS: Real-time perception fusion for infrastructure-assisted autonomous driving. In Proceedings of the 28th annual international conference on mobile computing and networking (pp. 133-146). https://doi.org/10.1145/3495243.3560539