MLD-DETR: A Method for Detecting Small Traffic Signs in Complex Scenarios

1st Shengchun Yang
School of Computer Science
Northeast Electric Power University
Jilin, China
foxysc@foxmail.com

2nd Lixun Zhou*
School of Computer Science
Northeast Electric Power University
Jilin, China
3179064028@qq.com

3rd Huixiang Wang
School of Computer Science
Northeast Electric Power University
Jilin, China
1003885366@qq.com

4th Ruiqi Li
School of Computer Science
Northeast Electric Power University
Jilin, China
lrq328634694@163.com

5th Chiyuan Wang
School of Computer Science
Northeast Electric Power University
Jilin, China
wangchiyuan1204@163.com

6th Yuhuan Wang
School of Computer Science
Northeast Electric Power University
Jilin, China
13694039975@163.com

***Abstract*—Traffic sign detection technology is a crucial component of intelligent driving systems. To address challenges such as small pixel areas, susceptibility to background interference, and low detection accuracy during traffic sign detection, the MLD-DETR traffic sign detection model is proposed. First, the MSGC module was introduced to replace the feature extraction component in the backbone network. By employing a dual-path parallel convolution strategy, it simultaneously aggregates local detail information and global contextual information, thereby enhancing the network's ability to represent features of objects at different scales. Subsequently, to reduce background interference, the LASS model was proposed. Finally, the DACFPN module was introduced to fuse features at different scales, further improving detection accuracy. Experimental results demonstrate that the MLD-DETR model achieves outstanding performance on the TT100K traffic sign dataset. Compared to the baseline model, it achieves an mAP50 of 83.5%, representing a 7.9% improvement, and an mAP50-95 of 61.4%, showing a 7.3% increase. The number of parameters decreases from 19.9M to 18.2M, a reduction of 1.7M, outperforming other state-of-the-art methods.**

***Index Terms*—RT-DETR, small object detection, traffic sign detection, multi-scale feature fusion.**

I. Introduction

With the rapid advancement of intelligent transportation systems, traffic sign recognition has been widely integrated into autonomous driving and driver assistance systems. Accurate detection and recognition of traffic signs are critical to ensuring road safety. Failure to detect traffic signs in a timely manner significantly increases the risk of traffic accidents, violations, and navigation errors, posing a serious threat to the lives of drivers, passengers, and pedestrians. Therefore, achieving efficient and reliable traffic sign detection is of paramount importance. Currently, traffic sign detection primarily relies on deep learning algorithms [1], [2]. The application of deep learning in traffic sign detection has significantly improved system accuracy in complex real-world road environments, enabling efficient localization and recognition of various traffic signs. This provides core perception capabilities for advanced driver assistance systems, autonomous vehicles, and intelligent traffic management, effectively addressing the shortcomings of traditional methods and substantially reducing accident risks caused by driver inattention to signs. However, challenges remain. Small traffic signs are prone to false positives and false negatives, while complex backgrounds—including road,vehicles, and buildings—further complicate detection [3].

We propose a novel traffic sign detection method—MLD-DETR,specifically designed to address the challenges in traffic sign detection. Our proposed MLD-DETR incorporates three key modules to tackle the aforementioned issues.

- **Multi-scale gated convolutional block.** We introduced a Multi-scale gated convolutional block(MSGC) into the backbone network of RT-DETR. It simultaneously aggregates local detail information and global contextual information through a dual-path parallel convolution strategy, thereby enhancing the network's ability to represent features of targets of varying sizes. The goal is to overcome the limitations of a single convolution kernel's receptive field and improve detection accuracy in complex scenes.
- **Locality-enhanced adaptive sparse self-attention.** We apply an locality-enhanced adaptive sparse self-attention(LASS) mechanism that enhances locality, enabling the model to learn the complementary relationship between local spatial context and global semantics. This allows it to more effectively capture subtle textures and boundary information within traffic signs, reducing feature smoothing and background noise interference caused by global attention. Consequently, the model's detection performance in complex scenes is improved.
- **Dual-attention context feature pyramid network.** The dual attention context feature pyramid network (DACFPN) employs a dual attention mechanism to concurrently extract global semantics and local details. It

utilizes channel attention to filter background noise while incorporating a concatenation strategy to preserve original feature information. This enables it to more accurately capture the geometric contours of traffic signs, effectively leverage contextual information to suppress noise interference from complex backgrounds, and reduce information loss of small target signs during multi-scale feature fusion.

II. Related Work

Early traffic sign detection research primarily relied on color segmentation and shape features to construct candidate region separation frameworks. Traditional methods based on prior knowledge and machine learning typically detect signs using their color and shape [4], [5] , demonstrating applicability in scenarios with clear signs and simple backgrounds. However, such approaches struggle to handle variations in lighting, occlusions, blurring, and small target scale variations present in real-world road images, limiting their detection performance in complex, dynamic, and variable environments.

To overcome the limitations of traditional methods in modeling complex scenes and achieving generalization capabilities, researchers began introducing deep learning models for end-to-end detection of traffic signs. Zhu et al. [6] proposed a two-stage framework based on Faster R-CNN, utilizing a region proposal network to enhance multi-scale sign localization accuracy; Zhang et al. [7] constructed a single-stage YOLOv3 detector, significantly improving real-time detection speed; Snegireva et al. [8] introduced a YOLOv5 variant, optimizing small object anchor designs to mitigate scale imbalance issues. While these approaches substantially enhance models' representation capabilities in complex road environments, they still struggle to effectively address small object loss, severe occlusion, and nonlinear interference from adverse weather conditions due to feature extraction directly on raw images.

Against this backdrop, combined approaches integrating multi-scale fusion and lightweight models have emerged as a key direction for enhancing detection stability. Related research has introduced small object detection layers, multi-scale feature pyramids, and attention mechanisms to reduce the difficulty of modeling small objects and improve adaptability to signs of varying scales. Wang et al. [9] proposed an improved YOLOv8 architecture by adding a small object detection head and optimizing the loss function, thereby boosting recall in heavily occluded scenarios. Wu et al. [10] introduced a lightweight traffic sign detection model based on YOLOv11 to enhance lightweight performance. Linardi et al. [11] addressed dataset quality and generalization limitations; Zhang et al. [12] proposed the YOLO-BS model, integrating channel attention to enhance feature representation in complex backgrounds. These approaches significantly improve detection reliability for small objects and low-light scenarios, providing a more robust feature extraction foundation for real-world road sign detection.

However, most of the aforementioned studies focus on pixel-level feature fusion and multi-scale processing, typically assuming independent distribution of traffic signs. This approach struggles to capture the global correlations arising from dynamic occlusions, extreme weather conditions, and the coexistence of multiple signs on real roads. To address the limitations in modeling long-range dependencies, research has further introduced Transformers and hybrid attention mechanisms to construct efficient frameworks. Chen et al. [13] proposed a triple attention mechanism to enhance small object detection under non-uniform distributions; Deng et al. [14] developed Heterogeneous Attention YOLO, dynamically adjusting channel and spatial weights to reflect real-world road interference; Qiu et al. [15] created the lightweight DP-YOLO model, integrating Transformers to reduce parameters while improving small object detection. Carion et al. [16] introduced the end-to-end object detection framework DETR, addressing the reliance of traditional detectors on numerous hand-crafted components; Zhao et al. [17] proposed the real-time end-to-end detector RT-DETR, resolving the high computational complexity of traditional DETR variants that hindered real-time detection. These approaches significantly enhance modeling capabilities for sign degradation relationships in complex real-world scenarios and boost engineering applicability.

III. Methodology

We improved the model based on RT-DETR. The overall framework of MLD-DETR is shown in Fig 1. We integrated the proposed MSGC module into the backbone network of RT-DETR to enhance both local and global contextual information. Additionally, LASS learns the complementary relationship between local spatial context and global semantic information, improving the model's detection performance in complex scenes. Finally, DACFPN effectively utilizes contextual information to suppress noise interference from complex backgrounds, minimizing information loss of small traffic signs during multi-scale feature fusion.

A. MSGC

The backbone network of RT-DETR lacks the ability to capture long-range dependencies and struggles to distinguish between background and traffic signs. To address these issues, we propose the MSGC module and integrate it into the backbone network, as illustrated in Fig 2. First, the input features undergo a 1×1 convolution projection and are split into a gating branch and a feature branch. The gating branch utilizes a Sigmoid function to generate weights that suppress background noise and enhance the target region. The feature branch is further split and processed through a 3×3 deep convolution to extract local texture details and a 5×5 deep convolution to extract global contextual information. The specific calculation formulas are as follows:

$$F = BatchNorm(X) \tag{1}$$

$$F_1 = Sigmoid(PWConv_1(F)) \tag{2}$$

$$\begin{aligned} F_2 = {} & DWConv_{3\times3}(PWConv_2(F)) + \\ & DWConv_{5\times5}(PWConv_2(F)) \end{aligned} \tag{3}$$

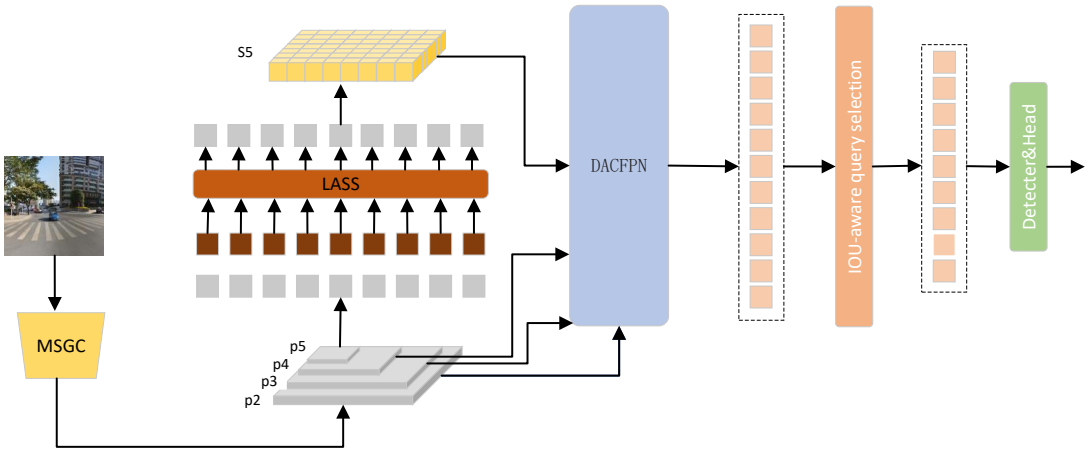

Fig. 1: MLD-DETR overall structure diagram.

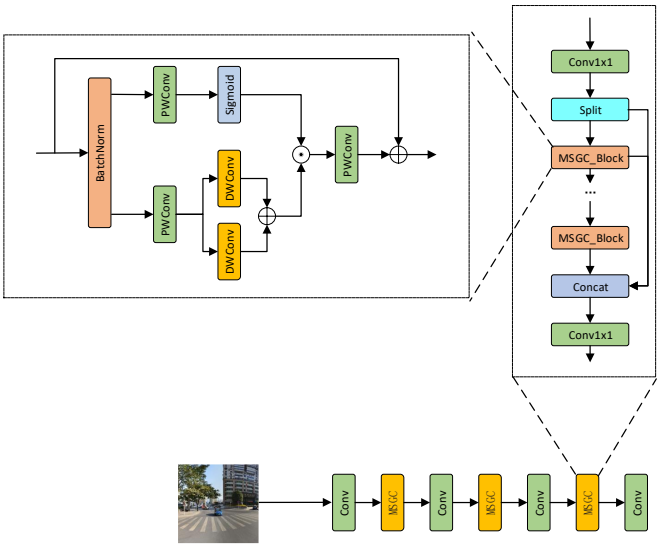

Fig. 2: MSGC module structure diagram.

where, X represents the feature map, PWConv denotes pointwise convolution, and DWConv denotes depthwise convolution. Subsequently, F_1 serves as the gating weight for F_2, which then undergoes projection through another layer of PWConv. The resulting output is added to X via residual connection, as detailed in the following formula:

$$\hat{F} = X + PWConv_3(F_1 \cdot F_2) \tag{4}$$

This approach not only leverages multi-scale convolutions to overcome the limitations of fixed receptive fields, enabling the network to simultaneously detect both minute targets and massive objects, but also introduces lightweight pixel attention through a gating mechanism, allowing it to actively suppress complex background noise.

B. LASS

The AIFI module employs a multi-head self-attention mechanism to perform global context modeling of high-level features, aiming to capture long-range dependencies. However, since it requires computations for all feature points, it introduces noise interference from irrelevant background regions, leading to background areas being mistaken for targets. To address this issue, we propose the LASS module, which incorporates strategies from ASSA [18] , as shown in Fig 3. LASS introduces sparse branches to ignore background noise, focusing solely on regions relevant to the target. It also employs convolutional residual branches to force the model to capture edge and texture information within pixel neighborhoods, thereby enhancing its perception of local fine-grained structures.

The input feature X undergoes LayerNorm and linear projection to obtain the query matrix Q, key matrix K, and value matrix V. The attention calculation can be defined as:

$$S = \frac{QK^T}{\sqrt{d}} \tag{5}$$

where S represents the similarity matrix. After obtaining S, a dual-branch structure is constructed, comprising the self-attention branch and the sparse self-attention branch. This dual-branch architecture ensures global information exchange while highlighting key features. The specific computational process is as follows:

$$A_1 = Softmax(S) \tag{6}$$

$$A_2 = ReLU^2(S) \tag{7}$$

A_1 and A_2 are the attention matrices obtained from the two branches, respectively. To balance the contributions of both branches, we introduce adaptive weights ω_1 and ω_2 to adjust their respective contributions. Finally, the final blended attention matrix A is generated.

$$A = \omega_1 * A_1 + \omega_2 * A_2 \tag{8}$$

Since the standard value matrix V is merely a linear mapping of the input, it neglects the influence of surrounding information. To address this, we introduce a local enhancement module within the value matrix path. We restore the serialized

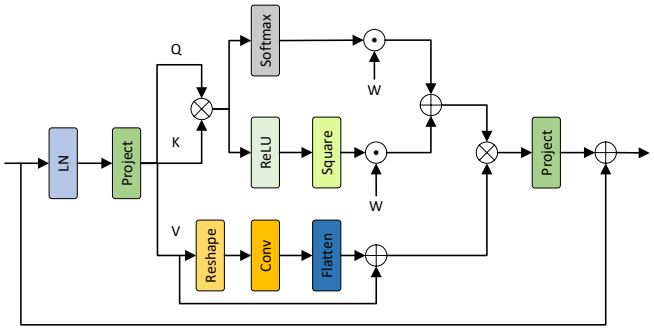

Fig. 3: LASS module structure diagram.

V to a two-dimensional spatial structure, extract local information through convolutional operations, and then fuse it back into the original feature stream:

$$V_{local} = Flatten(Conv(Reshape(V))) \tag{9}$$

$$V_{out} = V + V_{local} \tag{10}$$

Ultimately, the output feature vector is obtained by aggregating the enhanced value matrix V_{out} through the attention matrix A.

$$X_{out} = Project(A * V_{out}) + X \tag{11}$$

This design not only effectively suppresses background noise through a dual-branch mechanism but also compensates for AIFI's shortcomings in capturing local details via a convolutional enhancement path.

C. DACFPN

Traditional RT-DETR employs a crossscale context fusion module (CCFM) as its feature fusion network. However, it fails to effectively suppress background interference and is prone to inducing detail distortion, resulting in the dilution of weak target features. To address these issues, we propose the DACFPN module. This module incorporates the CGR [19] to enhance the feature fusion process, reducing background interference while preserving small target features.

The DACFPN architecture is illustrated in Fig 4. After feature extraction through the backbone network, feature maps of varying scales are obtained. Layers $P2$, $P3$, $P4$, and $S5$ are downsampled and averaged to reduce them to the same size, then concatenated and fed into multiple RCM modules to extract multi-scale information. The RCM dynamically weights and nonlinearly transforms features across spatial and channel dimensions, generating context-enhanced feature representations that effectively focus on target regions while suppressing background noise. Finally, the features undergo segmentation and multiscale feature fusion with the original features. This process can be represented as:

$$P = RCM(AP(P_2, 8), AP(P_3, 4), AP(P_4, 2), AP(S_5, 1)) \tag{12}$$

$AP(P, x)$ denotes average pooling with downsampling of feature P at a reduction factor of x. P represents the feature with a pyramidal background.

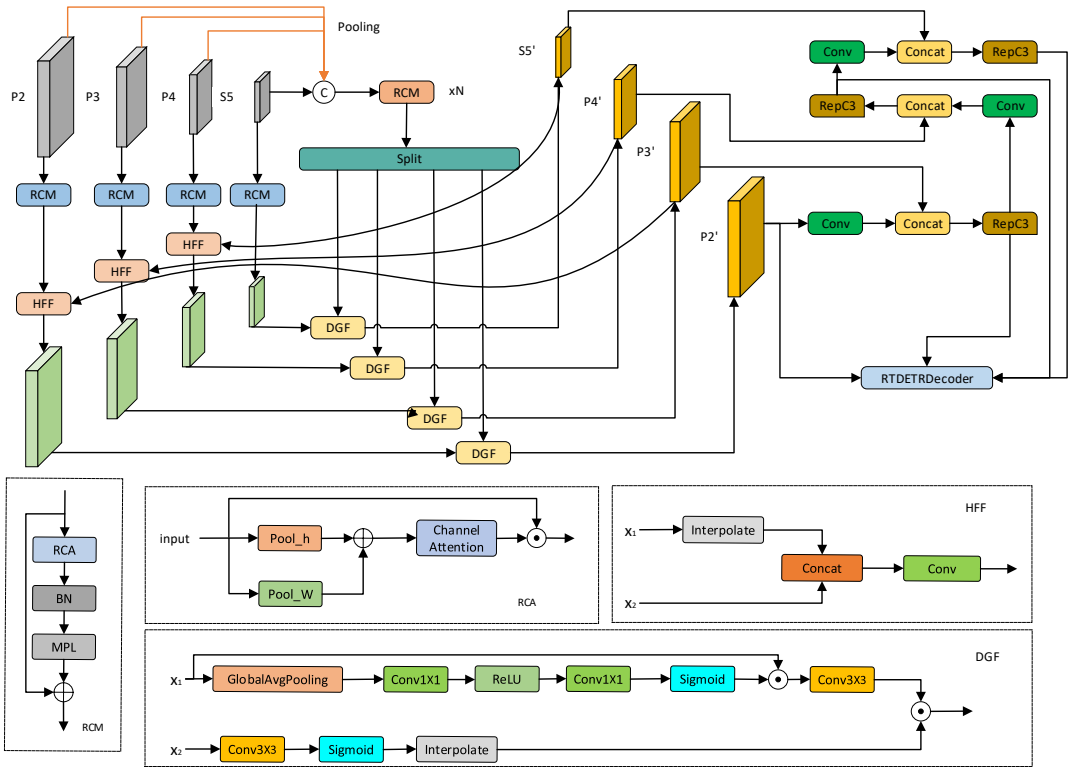

Fig. 4: DACFPN module structure diagram.

RCM is a crucial component of DACFPN, responsible for enhancing contextual feature representations while preserving original information. RCA serves as the core of RCM, capturing long-range spatial positional information of features through dimension-decomposing pooling operations. The specific implementation process of RCA is as follows:

$$RCA(X) = X \odot \sigma(F_{1\times1}(\delta(F_{1\times1}(Pool_H(X) + Pool_W(X))))) \quad (13)$$

Among these, $Pool_H(X)$ and $Pool_W(X)$denote pooling operations along the height H and width W directions, respectively. δ represents the nonlinear activation function, σ denotes the Sigmoid activation function, $\odot$ signifies element-wise multiplication, and $F_{1\times1}$ indicates a 1×1 convolution.

IV. Experiments

A. Datasets and Evaluation Metric

The experimental dataset used in this paper is the TT100K traffic sign dataset jointly released by tsinghua university and tencent. It comprises over 100,000 high-resolution traffic sign images and includes more than 100 categories of traffic signs. In this study, we selected 12 common traffic sign categories totaling 8,000 images, with 6,000 images allocated to the training set, 1,200 to the validation set, and 800 to the test set. The primary evaluation metrics adopted in this study are mAP50, mAP50-95,GFLOPs, and the params (M).

B. Experimental Environmen

The experimental setup utilized Ubuntu 22.04 with an NVIDIA GeForce RTX 3090 GPU. The programming framework employed was PyTorch 2.6.0, running on Python 3.10 with CUDA 12.4. Training parameters included a learning rate of 0.001, input image dimensions of 640×640, 150 training iterations, and a batch size of 8.

TABLE I
Ablation Study Results

MSGC	LASS	DACFPN	mAP50	mAP50-95	Params(M)
	RT-DETR		75.6	54.1	19.9
✓	×	×	74.3	53.6	14.2
×	✓	×	77.2	57.2	20.8
×	×	✓	76.5	55.4	19.3
✓	✓	×	79.4	58.2	17.2
✓	×	✓	80.1	58.7	16.6
×	✓	✓	81.2	59.4	18.9
✓	✓	✓	**83.5**	**61.4**	**18.2**

C. Ablation Study

To verify the contribution of each module in the MLD-DETR we proposed, we set up eight groups of comparative experiments. The experimental results are shown in Table I . According to Table 1, in the single-module analysis, only adding the MSGC module caused mAP50 and mAP50-95 to decrease by 1.3% and 0.5% respectively, but its parameter quantity dropped to 14.23M, a reduction of 5.74M compared to the original 19.97M. It achieved the effect of lightweighting with almost no impact on accuracy. Adding only the LASS module increased mAP50 by 1.6% and mAP50-95 by 3.1%. Adding only the DACFPN module increased mAP50 by

Fig. 5: RT-DETR and MLD-DETR visualization results diagram.

0.9 and mAP50-95 by 1.3%. The combination of the three modules led to varying degrees of improvement in mAP50 and mAP50-95. After integrating all the improvements, the updated model achieved increases of 7.9% and 7.3% in mAP50 and mAP50-95 respectively, compared to the original model, while reducing the parameter count by 1.68M. This demonstrates the effectiveness of the improved algorithm proposed in this paper.

TABLE II
COMPARISON RESULTS WITH OTHER METHODS ON THE TT100K DATASET

Method	mAP50	mAP50-95	Params(M)	GFLOPs
Faster R-CNN	78.4	55.2	41.5	91.0
YOLOX-M	79.7	56.4	25.3	73.8
YOLOv10-M	81.3	58.5	**16.5**	59.7
YOLOv11-M	80.6	58.9	20.1	68.8
RT-DETR	75.6	54.1	19.9	**57.3**
MLD-DETR (Ours)	**83.5**	**61.4**	18.2	80.3

D. Compared With the Algorithm

To evaluate the differences between the improved algorithm in this paper and current mainstream traffic sign detection models, comparative experiments were conducted under identical experimental conditions between the improved model and a series of state-of-the-art models. The results are shown in Table II. First, compared to the original RT-DETR model, mAP50 increased from 75.6% to 83.5%, representing a 7.9% improvement. mAP50, which reflects high localization accuracy, also saw a significant boost of 7.3% to 61.4%. Furthermore, the total number of segments was further optimized from 19.9 million to 18.2 million. These results fully demonstrate the effectiveness of the improved strategy. mAP50-95 also significantly improved by 7.3% to 61.4%. Furthermore, the total number of parameters was optimized from 19.9 M to 18.2 M, fully demonstrating that the improvement strategy enhances feature representation capabilities without introducing redundant computational overhead. Second, compared to the classic Faster R-CNN, our method achieves a 5.1% higher detection accuracy while significantly reducing the number of parameters, demonstrating the high efficiency advantage of MLD-DETR. Compared to YOLOv11M, our method achieves a 2.9% higher mAP50 while using fewer parameters. Compared to the extremely low-parameter YOLOv10M, our method, though slightly higher in parameters, achieves a significant 2.2% accuracy improvement.The visualized results are shown in Fig 5.

V. CONCLUSION

In this study, the MSGC module to replace the feature extraction component in the original backbone network. This addresses the limited receptive field of single convolutional kernels and enhances detection accuracy in complex scenes. Next, we design the LASS module to mitigate feature smoothing and background noise interference caused by global attention. Finally, we incorporate the DACFPN module to minimize information loss of small object features during multi-scale feature fusion. Experimental results demonstrate that the proposed MLD-DETR enhances detection accuracy, providing a reliable technical solution for traffic sign detection tasks. Future work will further optimize the model architecture to handle more complex traffic scenarios and improve performance in practical applications.

REFERENCES

[1] X. R. Lim, C. P. Lee, K. M. Lim, T. S. Ong, A. Alqahtani, and M. Ali, "Recent advances in traffic sign recognition: approaches and datasets," *Sensors*, vol. 23, no. 10, p. 4674, 2023.

[2] H. Chen, M. A. Ali, Y. Nukman, B. Abd Razak, S. Turaev, Y. Chen, S. Zhang, Z. Huang, Z. Wang, and R. Abdulghafor, "Computational methods for automatic traffic signs recognition in autonomous driving on road: A systematic review," *Results in Engineering*, p. 103553, 2024.

[3] S. Li, S. Wang, and P. Wang, "A small object detection algorithm for traffic signs based on improved yolov7," *Sensors*, vol. 23, no. 16, p. 7145, 2023.

[4] E. Kim, H. Li, and X. Huang, "A hierarchical image clustering cosegmentation framework," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2012, pp. 686–693.

[5] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural networks*, vol. 32, pp. 323–332, 2012.

[6] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2110–2118.

[7] H. Zhang and V. M. Patel, "Density-aware single image de-raining using a multi-stream dense network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 695–704.

[8] D. Snegireva and A. Perkova, "Traffic sign recognition application using yolov5 architecture," in *2021 International Russian Automation Conference (RusAutoCon)*. IEEE, 2021, pp. 1002–1007.

[9] G. Wang, P. Jin, Z. Qi, and X. Li, "Traffic sign detection method based on improved yolov8," *Scientific Reports*, vol. 15, no. 1, p. 19385, 2025.

[10] Y. Wu, T. Zhang, J. Niu, Y. Chang, and G. Liu, "Yolo-based lightweight traffic sign detection algorithm and mobile deployment," *Optoelectronics Letters*, vol. 21, no. 4, pp. 249–256, 2025.

[11] R. Linardi, N. Soo, A. M. Nabua, C. G. S. Adhi, and V. J. L. Engel, "Traffic sign detection using yolo: Evaluating failures on localized data," *Procedia Computer Science*, vol. 269, pp. 1494–1501, 2025.

[12] H. Zhang, M. Liang, and Y. Wang, "Yolo-bs: a traffic sign detection algorithm based on yolov8," *Scientific Reports*, vol. 15, no. 1, p. 7558, 2025.

[13] W.-Y. Dong, Y.-X. Chen, and G.-Y. Zou, "Autotrinet yolo triple attention framework for robust traffic sign detection," *Scientific Reports*, 2025.

[14] Y. Deng, L. Huang, X. Gan, Y. Lu, and S. Shi, "A heterogeneous attention yolo model for traffic sign detection," *The Journal of Supercomputing*, vol. 81, no. 6, p. 765, 2025.

[15] J. Qiu, W. Zhang, S. Xu, and H. Zhou, "Dp-yolo: A lightweight traffic sign detection model for small object detection," *Digital Signal Processing*, p. 105311, 2025.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[17] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 965–16 974.

[18] S. Zhou, D. Chen, J. Pan, J. Shi, and J. Yang, "Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 2952–2963.

[19] Z. Ni, X. Chen, Y. Zhai, Y. Tang, and Y. Wang, "Context-guided spatial feature reconstruction for efficient semantic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 239–255.