Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080414
ISSN 2371-8838 Vol. 8 Num. 4

Enhancing the application of signal light recognition for the YOLOv8 model in complex traffic scenarios

Xinghe Chen[1,*], Guanchen Du[2]

[1]*School of Computer Science and Engineering, North Minzu University, Yinchuan, 750021, China*
[2]*College of Mathematics and Computer Science, Shantou University, Shantou, 515000, China*
[*]*Corresponding author: cxh_8080@163.com*

Keywords: Object detection, YOLOv8, attention mechanisms, traffic signals

Abstract: In intricate traffic environments, traffic lights, as pivotal signaling tools, are influenced by factors such as observational distance and lighting conditions. This article proposes an enhanced YOLOv8 model that integrates a hybrid attention mechanism to adapt signal light recognition to complex traffic scenarios. Particularly, the introduction of the Global Attention Mechanism (GAM) within the YOLOv8 model is highlighted. GAM leverages a three-dimensional arrangement and dual-layer MLPs (Multilayer Perceptrons) to emphasize and strengthen channel features that are advantageous for the task of traffic light detection, while also maintaining cross-dimensional channel-spatial dependencies. It concentrates and merges spatial information with channel information through convolutional layers, enabling interaction and avoiding information loss by excluding max-pooling operations. Experimental results demonstrate the exceptional signal light recognition capabilities of the YOLOv8 model enhanced by the GAM attention mechanism in complex traffic scenes, fulfilling practical application requirements across all metrics. Post enhancement, the average recognition rate (Map@50) reaches as high as 93%, demonstrating the model's stability and efficiency in complex environments. The proposed method, based on the improved YOLOv8 model combined with the GAM attention mechanism for signal light recognition, effectively enhances the accuracy and robustness of traffic light detection in complex traffic environments, offering valuable research findings for the advancement and implementation of intelligent transportation systems.

1. Introduction

In contemporary life, transportation is intricately interwoven with human existence, and the transportation system plays a pivotal role in the functionality of an entire city. Within the transportation system, traffic lights are universally employed as a traffic signaling tool across the globe. Individuals can determine whether to continue or stop based on the changing colors, a concept that even young children can easily understand. If computers could accurately identify traffic lights, they could significantly assist individuals in conducting more rational analyses of traffic and pedestrian flow for traffic signal control, facilitating intelligent vehicle driving or mobile map navigation, and predicting commuting times. However, correctly recognizing the patterns of traffic lights is no simple task for computers. Initially, as the viewing distance changes, traffic lights

transition in size from small to large. When observed from a considerable distance in images captured by cameras, they might occupy just a few pixels.

Accurately identifying the colors of traffic lights under such circumstances is challenging. Traditional machine learning methods, such as Support Vector Machines (SVM) and k-means, struggle with this task. As sample sizes increase, shallow neural networks gradually become inadequate for adapting to complex sample variations. This led to the development of deep neural networks building upon the foundation of shallow ones. Deep neural networks, by better emulating biological neural networks, can learn from shallow to deep, constructing crucial features and providing enhanced accuracy. Employing deep neural networks for object detection yields commendable detection outcomes. R. Gokul et al. [1] compared the performance of Faster R-CNN and YOLO, two deep learning models, in traffic light object detection. They concluded that YOLO outperforms Faster R-CNN in this domain. Liu et al. [2] utilized the YOLOv5 model for traffic signal object detection in railway settings, achieving exceptionally high accuracy to meet practical requirements. Qian et al. [3] further enhanced YOLOv5 by designing a Memory Feature Fusion Network, which improved the robustness of the traffic light detection algorithm. Following this, Li et al. [4] integrated various techniques, including a Coordinate Attention Layer, into the YOLOv5 backbone network to enhance its feature extraction capabilities. This integration led to higher average precision and improved detection capabilities in tasks related to traffic lights and obstacle objects. While numerous studies have focused on traffic light object detection, the models used have been relatively outdated. To further enhance detection accuracy, this article utilizes the latest YOLOv8 model [5] for superior performance. YOLOv8 exhibits notable improvements over previous models in the YOLO series. Lastly, to address the challenge of small-scale traffic light objects in complex traffic environments, a Global Attention Mechanism (GAM) [6] is incorporated to enable the model to focus on and learn the intricate details of these small targets. The aim is to achieve accurate identification of small-scale traffic light objects in complex traffic scenarios.

2. Construction of Improved YOLOv8 Model

2.1 The YOLOv8 Model

Presently, YOLOv8 exhibits superior performance compared to YOLOv5 and YOLOv7, making it the most cutting-edge model within the YOLO family. Drawing inspiration from YOLOv7, YOLOv5, and YOLOvx, YOLOv8 incorporates modifications and enhancements based on their foundations. In the Backbone section, YOLOv8 integrates elements from CSPNet (Cross Stage Partial Network) and the residual C3 module, replacing C3 with the C2F module. This change ensures improved gradient flow and information propagation while maintaining a lightweight design. In the Neck section, similar to the Backbone, a transition from C3 to C2F is implemented, which reduces convolutional and fully connected operations. Within the head section, the model transitions from holistic feature map processing to feature segmentation. This modification effectively reduces the number of parameters and computational complexity, thereby improving the model's ability to generalize and its robustness. Transitioning from the anchor-based method for object detection to an anchor-free approach, YOLOv8 directly classifies and regresses each position in an image, enabling the detection of objects of arbitrary sizes and proportions with increased recall rates.In essence, YOLOv8 demonstrates a significant improvement over YOLOv7. Within YOLOv8, similar to its predecessors, the model is subdivided into five distinct variants: YOLOv8-x, YOLOv8-l, YOLOv8-xm, YOLOv8-s, and YOLOv8-n. The enhanced version will leverage the YOLOv8-n iteration, which boasts a higher parameter count compared to YOLOv8-S, thus enhancing the model's capacity for efficient traffic light object detection.

2.2 Improvement Strategy

In order to improve the YOLOv8 model's capability to accurately detect distant traffic lights and focus on subtle details, a Global Attention Mechanism (GAM) is integrated into YOLOv8 to optimize traffic light object detection. This addition allows YOLOv8 to capture intricate details of traffic light information in complex traffic scenarios, further building upon the existing model foundation. The GAM attention model consists primarily of channel attention sub-modules and spatial attention sub-modules, each playing a crucial role in refining the model's attention mechanisms. Below, we delve into detailed explanations of these two modules. The integration of Global Attention Mechanism (GAM) within YOLOv8 is organized as illustrated in Figure 1.

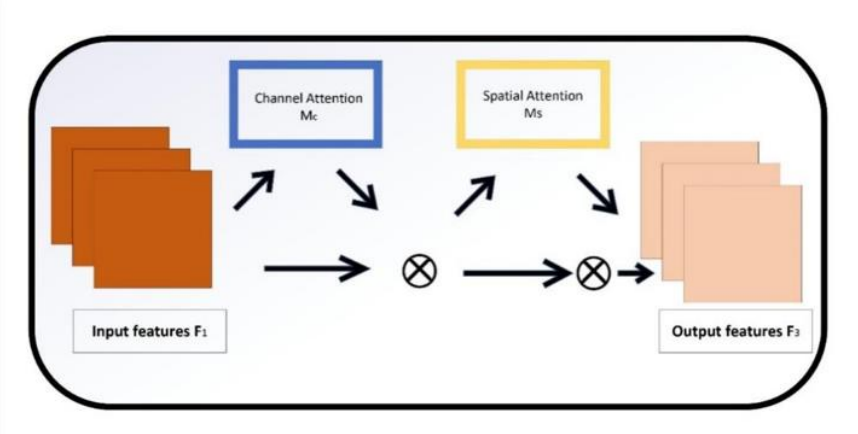

Figure 1: GAM Attention Model Diagram

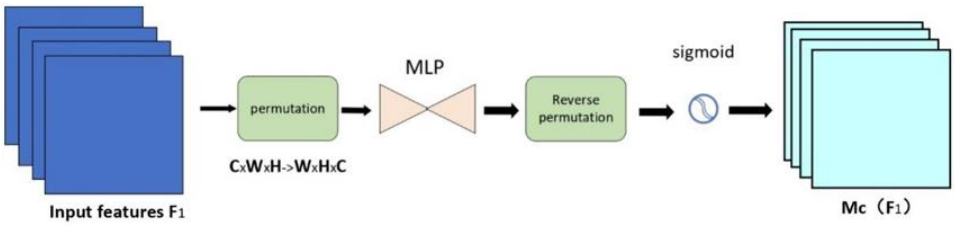

Figure 2: Channel Attention Submodule Diagram

2.2.1 Channel Attention Submodule

The primary objective of the Channel Attention Submodule is to focus on and enhance channel features that are most beneficial for the task of red-green light target detection. This is achieved by preserving information across three dimensions, specifically through a three-dimensional arrangement. This three-dimensional arrangement involves height, width, and channel dimensions, enabling the model to capture richer spatial-channel information. The structure of the Channel Attention Submodule is depicted in Figure 2.Within the Channel Attention Submodule, the three-dimensional arrangement better captures and comprehends the complexity of the input data. Given the multidimensionality of the input data in image processing or similar tasks, data typically possess three dimensions: height, width, and channels (such as the red, green, and blue channels in an RGB image). Each dimension contains crucial information essential for understanding the content of the image. The three-dimensional arrangement in the Channel Attention Submodule aims to operate simultaneously across these dimensions, rather than focusing solely on one or two dimensions. This approach allows the model to comprehensively grasp the input data, capturing more details and

contextual information.In this way, the three-dimensional arrangement plays a crucial role in the Channel Attention Submodule. It not only helps the model retain more original information but also enables the model to better understand and utilize this information. This is particularly effective in enhancing model performance and accuracy when dealing with complex tasks.The Channel Attention Submodule utilizes a two-layer Multilayer Perceptron (MLP) to enhance cross-dimensional channel-spatial dependencies. This implies that MLP considers not only inter-channel information but also spatial information, establishing dependencies between the two. The establishment of these dependencies is crucial for the model's performance in complex scenarios.The MLP is used here as an encoder-decoder structure. During the encoding phase, MLP compresses the input information, reducing its dimensionality and extracting key features. During the decoding phase, MLP expands these essential features back to their original dimensions to merge with other information. This encoding-decoding process enables the model to reduce computational burden while retaining information.In summary, the Channel Attention Submodule of GAM achieves effective integration and enhancement of channel and spatial information through a tridimensional arrangement and a two-layer MLP, enabling the model to meet the demands of complex tasks more effectively.

2.2.2 Spatial Attention Submodule

The Spatial Attention Submodule is equally essential in the GAM attention mechanism, with its primary task being to focus on and enhance spatial information within images. Spatial information is crucial for many computer vision tasks, particularly those demanding precise localization, such as red-green light target detection.Within the Spatial Attention Submodule, to concentrate on and integrate spatial information, two convolutional layers are typically utilized. The roles of these two convolutional layers are to extract and integrate spatial information from the input feature map. The first convolutional layer may capture local spatial features, while the second one is responsible for integrating these local features into global features. Through this approach, the Spatial Attention Submodule can enhance spatial features that are useful for specific tasks.Moreover, the Spatial Attention Submodule incorporates information from the Channel Attention Submodule. It adopts the same reduction ratio r as BAM (Bottleneck Attention Module) to retrieve information from the Channel Attention Submodule. This implies that the Spatial Attention Submodule can not only focus on spatial information but also engage in effective interaction and fusion with channel information.In GAM, to preserve more information and avoid information loss, the Spatial Attention Submodule eliminates the maximum pooling operation. Although max pooling, a common downsampling technique, is employed to help the model reduce computational burden and improve the robustness of feature maps, in complex traffic scenes, it may lead to information loss, especially with large pooling windows or strides. Therefore, the Spatial Attention Submodule chooses to remove the max pooling operation to further retain feature mappings.

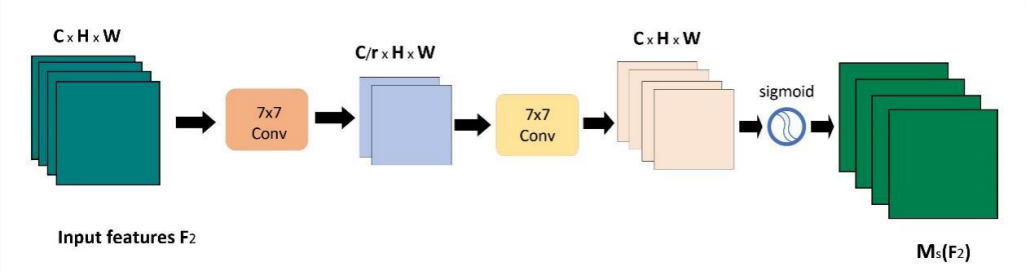

Figure 3: Spatial Attention Submodule Diagram

The Spatial Attention Submodule of GAM enhances the model's performance in recognizing red-green lights in complex traffic scenes by integrating spatial and channel information, and employing

appropriate convolution and pooling strategies. By effectively understanding and leveraging the spatial features of the input data, the model is better equipped to excel in challenging traffic scenarios. The structure of the Spatial Attention Submodule is illustrated in the above Figure 3.

2.2.3 Integrating GAM

GAM, a global attention mechanism, comprises a Spatial Attention Submodule and a Channel Attention Submodule. The formulation process is depicted in equations (1) and (2) as shown below.

$$F_2 = M_C(F_1) \otimes F_1 \quad (1)$$

$$F_3 = M_S(F_2) \otimes F_2 \quad (2)$$

Introducing an attention mechanism to YOLOv8-N enhances the neural network's ability to focus on traffic light information and refine local associations. This addition is integrated into the backbone section of the YOLOv8 network, following the convolutional layer of P5/32. The refinement enriches the attention capabilities of YOLOv8.

3. Analysis of Experimental Findings

3.1 Dataset

The dataset utilized in this experiment is S2TLD, comprising 3,785 images. The distribution involves 3,065 images in the training set, 314 images in the validation set, and 379 images in the test set. The following Figure 4 illustrates a selection of examples from the dataset.

Figure 4: Dataset Examples

3.2 The experimental setup employed was as follows

The experimental setup utilized the Windows 11 Professional operating system. The CPU processor employed was the 13th Gen Intel(R) Core(TM) i5-13400F, with a GeForce RTX 2080ti GPU boasting 22 GB of memory. The model training framework utilized PyTorch 2.1.2.

3.3 Evaluation metrics

To accurately assess the model's performance in detecting traffic lights, three metrics—R, Map@50, and Map@50-95—were employed to evaluate the constructed model. The Recall (R)

metric signifies the model's ability to correctly identify positive instances among all actual positive cases, as defined by the formula shown in Equation (3).

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{3}$$

Whereas True Positives represent the number of samples correctly predicted as positive by the model, and False Negatives indicate the number of samples erroneously predicted as negative.

On the other hand, Map@50 (Mean Average Precision at 50) and Map@50-95 (Mean Average Precision at 50-95) indicate the proportion of correct predictions in the top k predicted results. By calculating the average across all positions, Mean Average Precision is derived. Ultimately, computing MAP@50 entails the average precision of the first 50 predicted results. Conversely, Map@50-95 is a variant of Map@50, considering predictions from the 50th to the 95th positions.

3.4 Results

To validate the feasibility of the approach to improving the YOLOv8 model in this study, experiments were conducted on the dataset to assess the model's accuracy. After incorporating an attention mechanism into the YOLOv8 model, the detection precision and recall rate met the practical application requirements for recognizing traffic light signals. The specific numerical results are presented in the table below. Experimental analysis involving the addition of the GAM attention mechanism concludes that the enhanced YOLOv8 model fulfills the detection requirements for traffic light targets in complex traffic scenarios. The inference of the detection performance is illustrated in Fig. 5 and outlined in Table 1.

Figure 5: Detection Performance

Table 1: Inference Results

Class	R	map@50	map@50-95
all	0.879	0.933	0.685
red	0.955	0.99	0.737
yellow	0.89	0.912	0.703
green	0.962	0.965	0.676
off	0.852	0.867	0.623
off	0.852	0.867	0.623

By utilizing the open-source dataset S2TLD and preprocessing the training data, the images were randomly split into training, validation, and testing sets at a ratio of 80%, 20%, and 20% respectively. Achieving an average recognition rate of 93%, the improved model's performance exhibited a considerable enhancement, reaching a level suitable for practical applications. This advancement

holds significant practical implications for deploying vehicular monocular traffic signal recognition systems, aiding drivers in promptly identifying traffic lights during distractions, thereby promoting driver awareness and enhancing the driving environment.

4. Conclusion

In order to enhance proficiency in recognizing traffic lights in intricate traffic scenarios, this paper employs the cutting-edge YOLOV8 as the baseline model. Building upon this foundation, the GAM attention mechanism is introduced to endow YOLOV8 with enhanced learning of details and generalization capabilities. Experimentally, on the dataset, an average accuracy of 93% under the standard of MAP50 can be achieved, showcasing the further stable, efficient, and accurate recognition of traffic lights in complex traffic settings after the integration of GAM. This adaptation aligns the model well with the demands of contemporary traffic environments for the detection of traffic lights. Nevertheless, existing models still exhibit certain shortcomings that can be subject to further enhancement. For instance, post-improvement models may face challenges in deployment on lightweight in-vehicle recognition devices. Addressing these areas will be the focal point of our future endeavors. Our forthcoming efforts will persist in elevating and refining the model for enhanced deployment feasibility without compromising recognition efficiency and accuracy.

References

[1] Gokul R, Nirmal A, Bharath K M, et al. A comparative study between state-of-the-art object detectors for traffic light detection[C]//2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). IEEE, 2020: 1-6.

[2] Liu W, Wang Z, Zhou B, et al. Real-time signal light detection based on yolov5 for railway[C]//IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2021, 769(4): 042069.

[3] Wu Q, Guozhong W, Guoping L I. Improved YOLOV5 traffic light real-time detection robust algorithm[J]. Journal of Frontiers of Computer Science & Technology, 2022, 16(1): 231.

[4] Li Z, Zhang W, Yang X. An Enhanced Deep Learning Model for Obstacle and Traffic Light Detection Based on YOLOv5 [J]. Electronics, 2023, 12(10): 2228.

[5] Terven J, Córdova-Esparza D M, Romero-González J A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas[J]. Machine Learning and Knowledge Extraction, 2023, 5(4): 1680-1716.

[6] Liu Y, Shao Z, Hoffmann N. Global attention mechanism: Retain information to enhance channel-spatial interactions [J]. arXiv preprint arXiv:2112.05561, 2021.(GAM)