Enhancing Intelligent Road Target Monitoring: A Novel BGS-YOLO Approach Based on the YOLOv8 Algorithm

Xingyu Liu, Yuanfeng Chu, Yiheng Hu, and Nan Zhao, *Member, IEEE*

Abstract Road target detection is essential for enhancing vehicle safety, increasing operational efficiency, and optimizing user experience. It also forms a crucial part of autonomous driving and intelligent monitoring systems. However, current technologies face significant limitations in multi-level feature fusion and the accurate identification of key targets in complex data environments. To address these challenges, this paper proposes an innovative algorithmic model called BiFPN GAM SimC2f-YOLO (BGS-YOLO), aimed at improving detection performance. Initially, this paper employs the Bidirectional Feature Pyramid Network (BiFPN) to effectively integrate multi-level features. This integration overcomes the limitations in feature extraction and recognition found in existing target detection algorithms. Following this, this paper introduces the Global Attention Module (GAM), which markedly improves the efficiency and accuracy of extracting key target information in complex data environments. Additionally, this paper innovatively designs the SimAM-C2f (SimC2f) network, further advancing feature expressiveness and fusion efficiency. Experiments on the public COCO dataset demonstrate that the BGS-YOLO model significantly outperforms the existing YOLOv8n model. Notably, it shows a 7.3% increase in mean average precision (mAP) and a 2.4% improvement in accuracy. These results highlight the model's high precision and swift response in detecting road targets in complex traffic scenarios. Consequently, the BGS-YOLO model has the potential to significantly enhance road safety and contribute to a considerable reduction in traffic accident rates.

***Index Terms*—**Road target detection; autonomous driving; BiFPN; GAM; YOLO.

I. INTRODUCTION

ROAD target detection technology plays a crucial role in autonomous driving [1], intelligent surveillance [2] systems, and robotics [3]. However, it faces significant challenges in various complex environments. The differences in scale between vehicles and pedestrians, traffic congestion, and rapid changes in dynamic backgrounds all directly impact the precision and efficiency of detection [4].

Approximately 1.35 million fatalities are attributed to traffic accidents each year. These incidents incur economic losses, encompassing both immediate expenses and sustained potential costs, which may exceed hundreds of billions of dollars [5]. The variance of environments, along with the likelihood of vehicles and pedestrians infringing traffic regulations due to temporal constraints, markedly elevates the risk of traffic collisions [6]. Consequently, the precise and swift identification of pedestrians, vehicles, and bicycles is essential in the context of autonomous driving and intelligent surveillance systems [7].

Early research in target detection primarily relied on traditional methods such as manual observation, where traffic police and drivers predict potential dangers through visual inspection. Although intuitive, this approach has limitations in terms of manpower consumption and accuracy. Moreover, surveillance systems and digital image processing technologies were widely utilized in early investigations [8]. For instance, Timo Ojala et al. introduced the Local Binary Pattern (LBP) [9], a technique for extracting texture features by transforming local pixel values into binary codes. Navneet Dalal and Bill Triggs developed the Histogram of Oriented Gradients (HOG) [10], a method for depicting shapes and textures through the analysis of gradient histograms in specific image areas. David Lowe proposed the Scale-Invariant Feature Transform (SIFT) [11], aimed at identifying image key points and extracting feature descriptors that remain constant regardless of scale and rotation. Furthermore, Herbert Bay et al. improved upon this with the Speeded Up Robust Features (SURF) [12], which enhanced computational efficiency and robustness. Nevertheless, these traditional approaches exhibited limitations in precision and processing speed in complex environments. Particularly, in scenarios with varied target postures and high scene complexity, their adaptability was constrained, resulting in slow processing times and reduced detection accuracy. Therefore, methods based on computer vision have increasingly gained prominence in target detection [13].

Initial computer vision applications were heavily dependent on image processing technologies and basic pattern recognition algorithms, incorporating conventional machine learning techniques [14]. Machine learning, rooted in statistical analysis for prediction and decision-making [15], boasts a rich history with a variety of classic approaches. The Bayesian method, introduced by 18th-century British mathematician Thomas Bayes [16], utilizes posterior probabilities for classification or regression

tasks. The K-Nearest Neighbors (KNN) algorithm, proposed by Evelyn Fix and Joseph Hodges [17], performs classification or regression by leveraging distance metrics and feature analysis. Vapnik and Cortes developed the Support Vector Machine (SVM) [18], which operates by identifying an optimal hyperplane in the feature space for classification purposes. Furthermore, Breiman introduced the Random Forest algorithm [19], which employs multiple decision trees to execute classification or regression. Despite the success of machine learning algorithms across various domains, they encounter challenges related to parameter tuning and computational intensity. Consequently, deep learning approaches, with their automated and efficient learning capabilities, have progressively become prevalent in computer vision.

Deep learning, utilizing multi-layer neural networks, focuses on modelling and solving complex problems [20]. This approach often employs Convolutional Neural Networks (CNN) for feature extraction and classification localization [21], achieving significant detection results in processing complex datasets. Deep learning techniques are primarily divided into two categories: two-stage and single-stage algorithms [22].

The core of two-stage algorithms lies in first generating candidate boxes and then classifying and regressing these boxes' positions. For instance, R-CNN (Region-based Convolutional Neural Networks) [23] is a classic representative of this type of algorithm, generating candidate boxes and then extracting features and classifying each box. Shaoqing Ren et al. introduced the Faster R-CNN algorithm [24], which integrates a Region Proposal Network (RPN) [25]. This innovation significantly accelerates the process of generating and classifying candidate boxes. Mask R-CNN, proposed by Kaiming He et al. [26], added flexibility and robustness, particularly excelling in detecting targets of various sizes and shapes. However, the limitation of these methods is that they require identifying the target before proceeding with regional detection, leading to large model parameters and extended inference time. To address this issue, in 2015, Joseph Redmon et al. introduced the single-stage algorithm YOLO [27].

YOLO achieves fast, direct object detection by dividing images into grids and predicting object categories and locations in each grid, effectively capturing global information. Despite these improvements, its accuracy still did not fully meet the real-time detection needs of intelligent systems. YOLOv3 [28] adopted Darknet-53 as the feature extractor and combined it with a feature pyramid network (FPN) and cross-layer connections to predict at different scales, thereby improving the capabilities of feature extraction and multi-scale object detection. However, YOLOv3 experienced a slight decrease in speed compared to YOLOv2. NAS-YOLOX [29], an upgrade from baseline YOLOX, replaces PAFPN with NAS-FPN for better multi-scale feature fusion. It integrates a dilated convolution module (DFEM) and multi-scale channel-spatial attention (MCSA), enhancing target information extraction and focus. These improvements boost detection accuracy, yet distinguishing similar targets in dense scenes remains challenging. YOLOv5 [30] integrated data augmentation, deep feature extraction, and feature fusion techniques to achieve fast and accurate detection at three different scales. It introduced Mosaic data augmentation and multi-scale preset anchors, along with a PANet-inspired feature fusion strategy, to improve the model's generalization ability and detection accuracy. LEF-YOLO [31], based on YOLOv5, integrates MobileNetv3's bottleneck structure and uses depthwise separable convolution for a lightweight design. It employs multiscale feature fusion and Coordinate Attention with Spatial Pyramid Pooling-Fast to enhance feature extraction, improving detection accuracy. However, distinguishing subtle differences among dynamic targets still needs improvement. Subsequently, the YOLOv7 algorithm [32] was proposed, focusing on optimizing the model's structural design and training process. This method improved target detection accuracy and training efficiency by introducing optimized modules and methods, while keeping inference costs manageable. Following this, Ultralytics released the YOLOv8 model [33], innovating on the basis of YOLOv5 by introducing a new backbone network, anchorless detection head, and loss function. These innovations enhanced the model's performance and adaptability, enabling it to support various visual tasks, including image classification, object detection, and instance segmentation, while being compatible with multiple hardware platforms.

Although previous research has made significant progress in terms of accuracy and speed in object detection, there are still deficiencies in multi-scale object recognition and object detection against complex backgrounds. This paper introduces BGS-YOLO, an innovative road object detection algorithm model leveraging YOLOv8. The model's principal innovations comprise:

1) This paper introduces the Weighted Bidirectional Feature Pyramid Network (BiFPN) as an innovative solution. Through deep, multi-level feature fusion, BiFPN markedly improves the semantic representation of features. It does so by effectively minimizing information loss and controlling parameter growth. Implementing this technology significantly enhances the precision and speed of detecting vehicles and pedestrians. Consequently, it substantially boosts the efficacy of road safety monitoring and autonomous driving systems.
2) The paper incorporates a Global Attention Module (GAM), which substantially enhances the capability of traditional models to capture information in dynamic settings. By incorporating spatial, channel, and temporal analyses, GAM significantly boosts both the accuracy and the interpretability of detection outcomes. Consequently, this advancement markedly improves the performance of traffic monitoring systems in forecasting and reacting to varying road conditions.
3) Drawing inspiration from SimAM, this paper introduces

a novel architecture for the SimC2f network to overcome the constraints identified in YOLOv8's C2f module. By conducting a comprehensive analysis of the C2f module's output features, this strategy substantially improves feature fusion efficiency and addresses the scale variation challenges inherent in feature extraction and fusion processes. Such a significant enhancement considerably augments the network's detail detection capability, leading to remarkable advancements in the accuracy and real-time responsiveness of vehicle and pedestrian detection.

II. Materials and Methods

A. Network Structure of YOLOv8n Algorithm

YOLOv8 represents the most advanced model in the realm of object detection algorithms, with YOLOv8n especially distinguished within this series due to its superior detection efficacy and precision [34]. This paper focuses on the application of the YOLOv8n model for object detection assignments, offering an exhaustive analysis of its network architecture. The architecture comprises four core components: Input layer, Backbone network, Neck network, and Head network.

1) The input layer facilitates the reception and preprocessing of images, ensuring compatibility with the YOLOv8n network's specifications. This preprocessing encompasses scaling, normalization, and augmentation of the data. Specifically, YOLOv8n incorporates several data augmentation techniques, including rotation, cropping, and flipping. Moreover, it adopts the YOLOX strategy of deactivating Mosaic augmentation during the last 10 epochs of training, a measure aimed at bolstering the model's capacity for generalization.
2) The backbone network utilizes a meticulously optimized CSP-Darknet53 for feature extraction. This architecture includes five convolutional layers, two C2f modules, and one SPPF module. The introductory 3x3 convolutional layer achieves a significant reduction in computational load by employing a fourfold down-sampling of resolution, while still preserving essential features. Inspired by the ELAN structure's design principles, the C2f module boosts gradient flow through cross-layer connections and reduces computational requirements by eliminating convolutional layers and introducing split operations. Meanwhile, the SPPF module skillfully ensures a balance between computational efficiency and multi-scale feature fusion, leveraging both sequential and parallel pooling methods.
3) The neck network utilizes the path Aggregation Network Feature Pyramid Network (PAN-FPN) architecture for feature down-sampling and up-sampling, which significantly improves feature fusion across layers, thus enhancing information flow and interchange. Moreover, incorporating the C2f module as a residual block substantially boosts learning efficiency, leading to further refinement and optimization of features extracted by the backbone network.
4) The head network integrates a decoupled head structure and anchor-free detection technology, enabling precise feature extraction via distinct branches for classification and DFL regression. It utilizes multi-scale detectors for accurate bounding box predictions. Furthermore, the implementation of the Task-Aligned Assigner strategy optimizes sample selection, substantially improving detection accuracy and efficiency.

This paper is organized as follows: the innovations of the BGS-YOLO model are presented in Section (2). While Section (3) presents the simulation and experimental results. Discussion is covered in Section (4), followed by the conclusion in Section (5).

B. Improved Network Architecture of YOLOv8n Algorithm

1) Bidirectional Feature Pyramid Network (BiFPN)

Optimizing the exchange of information between road

FIGURE 1. Comparison of PAN-FPN and BiFPN network structure.

target image features at different scales is crucial for improving feature extraction and recognition capabilities. The Feature Pyramid Network (FPN) employed by researchers demonstrates significant effectiveness in processing targets with substantial scale differences, such as pedestrians and vehicles, owing to its efficient multi-scale feature fusion capability. However, FPN encounters issues of information loss across layers as the network deepens. To address this, YOLOv8n introduced the PAN-FPN structure, which facilitates more effective fusion and information flow with higher-level (lower-resolution) features, though challenges remain in information acquisition and parameter control. Consequently, this paper incorporates the BiFPN [35], [36] within YOLOv8n's neck network to achieve more efficient multi-level feature fusion. BiFPN enhances fusion by integrating both top-down and bottom-up feature fusion pathways and introducing weighted context information edges, leading to more effective multi-level feature fusion. This design not only enhances the semantic richness of features but also effectively addresses the issues of information loss and parameter increase, thereby significantly improving the overall performance of the network. The comparison between the PAN-FPN and BiFPN network structures is shown in Figure 1.

BiFPN utilizes a path-enhanced bidirectional fusion strategy to effectively establish feature fusion channels both from top to bottom and bottom to top, facilitating bi-directional cross-scale connections. In the fusion of multi-scale feature maps, BiFPN dynamically adjusts weights according to the significance of input features across different resolutions, thereby ensuring a balance of multi-scale feature information and substantially improving detection accuracy. By removing single input nodes and reinforcing connections among nodes on the same layer, the network integrates a broader spectrum of feature information. Moreover, BiFPN accomplishes deeper feature integration through the recurrent application of bidirectional paths at the feature layer.

Considering the significant differences in the importance of different input features for network learning, this paper assigns learnable weights to each input feature and adopts a rapid normalization feature fusion strategy. This network enables the network to adaptively learn the importance of features, effectively overcoming the issue of neglecting resolution differences in traditional feature fusion methods. The process of weight allocation involves softmax processing for each feature weight, with the specific formula as shown in Equation (1).

$$O = \sum_{\mathrm{i}} \frac{w_i}{\varepsilon + \sum_j^{w_j}} I_i \tag{1}$$

Where O signifies the output feature, w_i refers to the node weights, and I_i represents the input features. The learning rate is set to $\varepsilon = 0.0001$ to maintain numerical stability. After processing, the final feature map is generated through bidirectional scale connections and efficient normalization fusion. For example, the expression for node 6 is provided by Equations (2) and (3).

$$P_6^{td} = Conv\left(\frac{w_1 \cdot P_6^{in} + w_2 \cdot Resize(P_7^{in})}{w_1 + w_2 + \varepsilon}\right) \tag{2}$$

$$P_6^{out} = Conv\left(\frac{w_1' \cdot P_6^{in} + w_2' \cdot P_6^{td} + w_3' \cdot Resize(P_5^{out})}{w_1' + w_2' + w_3' + \varepsilon}\right) \tag{3}$$

Where *Resize* represents either a down-sampling or up-sampling operation, and w_i denotes learnable parameters. P_6^{td} indicates the intermediate features at the 6th layer in the top-down pathway, while P_6^{out} refers to the output features at the 6th layer in the bottom-up pathway. Moreover, to enhance efficiency, BiFPN employs depthwise separable convolutions [37] for feature fusion and applies batch normalization and activation functions after each convolution operation.

In conclusion, BiFPN enhances object detection capabilities by effectively combining features of varying scales through the use of bidirectional connections and a mechanism for weighted feature fusion, resulting in a substantial improvement in performance.

2) Global Attention Module (GAM)

Extracting crucial target information from complex data remains a significant challenge. Traditional attention mechanisms like Squeeze-and-Excitation (SE) focus primarily on the variance in channel significance but inadequately capture spatial information, overlooking spatial dimension correlations. Furthermore, CBAM (Convolutional Block Attention Module), while merging attention weights for channel and spatial dimensions, does not adequately account for the interplay of global information, restricting cross-dimensional information extraction. To address these issues, this paper proposes the GAM [40]. GAM integrates spatial and channel dimensions with time series analysis to rectify traditional models' deficiencies in dynamic settings. Implementing GAM within the backbone network minimizes information loss and boosts interaction among various global dimensions. This mechanism successfully navigates the constraints posed by limited receptive fields and dimensionality separation, leading to improved efficiency in information extraction and noise filtering, thereby substantially improving deep neural networks' overall efficacy.

GAM comprises two main components: channel and spatial modules. The channel module facilitates efficient processing of features across various channel dimensions, thereby markedly diminishing information loss. Concurrently, the spatial module amplifies the interaction among features spanning different spatial dimensions, culminating in focused attention across dimensions. The collaborative functionality of these modules enables GAM to thoroughly process features within both spatial and channel dimensions. This mechanism enhances information processing and feature representation effectiveness across

diverse datasets and classification tasks.

Considering the input feature labeled as F_1, it initially undergoes processing via the channel attention mechanism M_c, followed by an element-wise multiplication with the original feature F_1 to produce the feature F_2 after channel attention processing. Subsequently, this product undergoes another element-wise multiplication with the output of the spatial attention mechanism M_s to yield the final output feature F_3. In the diagram, the symbol Y denotes the element-wise multiplication operation. The specific formulas for F_2 and F_3 are provided as follows (Equations (4) and (5)).

$$F_2 = M_c(F_1) \otimes F_1 \tag{4}$$

$$F_3 = M_s(F_2) \otimes F_2 \tag{5}$$

In the GAM's channel attention submodule, features undergo initial channel-wise transposition followed by processing via a dual-layer Multilayer Perceptron (MLP). This step aims to amplify their relevance across both spatial and channel dimensions. Concurrently, the spatial attention submodule employs two 7×7 convolutional layers for semantic reinforcement, effectively capturing detailed information within the spatial dimensions.

GAM optimizes the capture of comprehensive image information through superior global perception capability, thereby enhancing target processing. By aggregating features of key regions through weighted methods, the network's performance in information recognition and feature expression is significantly enhanced. Furthermore, the flexible architecture of GAM and its intuitive weight distribution mechanism further improve the model's adjustability and interpretability.

3) SimC2f Network

Improving scale fusion task performance is crucial for thoroughly integrating features from varying scales. The scale-to-frequency (C2f) conversion has demonstrated notable benefits, particularly in boosting fusion efficiency. Yet, C2f's adaptability is constrained in complex and dynamically changing scenarios. Motivated by the Simultaneous Attention Module (SimAM) [38], this paper introduces an innovative restructuring of C2F into the SimC2f network. This network focuses on effectively handling highly variable scenes and dynamic features, aiming to overcome the limitations of traditional methods in dealing with complex environments, thereby enhancing overall performance and adaptability. With the advanced attention mechanism of SimAM, the network achieves a deep integration of spatial and channel information without adding extra parameters, generating three-dimensional attention weights. Within the YOLOv8n framework, the SimC2f network greatly enhances semantic recognition of road targets and processing of low-salience features, significantly improving the model's recognition efficiency and accuracy. The structural framework of the SimC2f network and the improved structure of YOLOv8n are shown in Figure 2.

SimAM precisely evaluates the linear separability between individual features and other features within the same channel by defining an energy function, thereby

FIGURE 2. Improved Structure of YOLOv8n.

accurately determining the importance of each feature. The energy function is defined as Equation (6).

(6)

Where and represent the target feature information and other feature information in the channel, respectively, with and being the linear transformation weight and bias for . The index denotes the spatial dimension order, is a hyperparameter, and is the number of all feature information on a single channel.

The SimAM structure is designed based on neuroscientific theories to define an energy function that identifies key neurons and calculates attention weights accordingly. In neuroscience, neurons rich in information often exhibit different firing patterns compared to surrounding neurons and can inhibit neighboring neurons, known as spatial suppression effects [39]. Therefore, neurons with significant spatial suppression effects should be given higher priority. These important neurons are identified through the energy function defined by Equation (6).

(7)

(8)

Equations (7)-(8) demonstrate that as energy decreases, the distinctiveness between neuron t and its surrounding neurons gradually increases, leading to a higher specificity and importance of neuron t at lower energy levels. The importance of a neuron can be calculated through . After assessing the neuron's importance using the energy function, key features are further refined through a scaling operation.

In summary, the application of the SimAM weighting method significantly improves the efficient management of channel information during feature fusion and effectively prevents the random distribution of features across different channels, thereby substantially enhancing feature extraction performance.

III. Experiments and Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

A. Dataset

This research employs the COCO [40] dataset, a benchmark for vehicle detection, comprising over ten object types. Specifically, three categories were chosen for analysis. We compiled 3,600 training images and 400 testing images, focusing on bicycles, vehicles, and pedestrians. The dataset was segmented into training, validation, and testing sets at an 8:1:1 ratio. Renowned for its detailed annotations and diverse sensor data, the COCO dataset is instrumental across various domains, especially in road object detection.

Refer to the subsequent figure for dataset specifics. Figures 3(a) and 3(b) illustrate the dataset's categories, with x and y indicating the bounding boxes' center points. Figure 3(c) depicts the bounding boxes' dimensions.

B. Experimental Environment and Evaluation Metrics

The experiments were executed in a Linux setting, employing an NVIDIA GeForce RTX 4090 Ti graphics card with 32 GB VRAM, alongside PyTorch 1.7.0 and Python 3.8 as the computational framework. Key experimental parameters included a starting learning rate of 0.01 over 300 epochs, with a momentum of 0.937 and weight decay set to 0.0005. The chosen batch size was 16.

For a thorough evaluation of the BGS-YOLO model's

(a) Data set category. (b) Center point of the bounding box. (c) Width and height of the bounding box.

FIGURE 3. Performance results for the road target dataset.

performance, we relied on four principal metrics to gauge the algorithm's efficacy comprehensively:

1) Precision (P) quantifies the accuracy of positive predictions made by the model. Its formula is defined as Equation (9):

(9)

True Positives (TP) refers to correctly identified targets, while False Positives (FP) indicates instances where non-targets are mistakenly classified as targets.

2) Recall (R) is the ratio of correctly predicted positive instances by the model to the total number of actual positive instances. The equation is given in Equation (10).

(10)

False Negatives (FN) represent instances where actual targets were not detected.

3) The F1- Score is the harmonic mean of Precision and Recall. The formula is given in Equation (11).

(11)

4) The Mean Average Precision (mAP) represents the average of the Average Precision (AP) for each category in multi-class detection scenarios. Assuming there are N categories, with the area under the curve for each category denoted as AP and the AP for each category represented as , the formula for is as follows:

(12)

C. Experimental Results and Analysis

1) Comparative Evaluation Before and After Algorithmic Enhancement

This paper rigorously assesses the improved algorithm's efficacy through a series of comparative tests involving the YOLOv8n and BGS-YOLO models. The outcomes of these experiments are detailed in Table 1. For a visual representation of model performance, the Precision-Recall (PR) curves are depicted in Figures 4 and 5.

TABLE 1. Comparative experimental results before and after algorithm improvement.

Model	Precision (%)	Recall (%)	F1 (%)	mAP (%)
YOLOv8n	85.4	83.2	84.3	85.8
BGS-YOLO	87.8	89.6	88.7	93.1

FIGURE 4. Precision-Recall (PR) curve for the YOLOv8n experiment.

FIGURE 5. Precision-Recall (PR) curve for the BGS-YOLO model.

The results presented in Table 1 reveal that, compared to YOLOv8n, the BGS-YOLO model proposed in this paper achieved an improvement of 2.4% in Precision, 6.4% in Recall, and 4.4% in F1 Score. Furthermore, the Mean Average Precision (mAP) of the BGS-YOLO model, as indicated by the area under the curve, increased by 7.3% compared to YOLOv8n, with detailed information available in the previously mentioned figures. These significant performance enhancements can be attributed to the introduction of the BiFPN network, which significantly boosted the model's capabilities in detecting multi-scale road target features. In addition, the incorporation of a global attention module has effectively improved the model's ability to recognize and focus on key features. The inclusion

of the newly designed SimC2f network has enabled the model to efficiently process and identify image features across various scales and complexities, further enhancing model performance. Overall, the model proposed in this study outperforms the original model across all evaluated metrics.

2) Ablation Experiment

This paper conducted ablation experiments to transition from the YOLOv8n model to the BGS-YOLO model and test the effectiveness of the improved model. The results are presented in Figure 6 and Table 2. The designations YOLOv8n+G, YOLOv8n+G+B, and YOLOv8n+G+B+S sequentially represent the integration of GAM, BiFPN, and the restructured SimC2f network into the base YOLOv8n, respectively.

TABLE 2. Results of Ablation Experiments

Model	Precision (%)	Recall (%)	F1 (%)	mAP (%)
YOLOv8n	85.4	83.2	84.3	85.8
YOLOv8n+G	86.2	85.4	85.8	87.6
YOLOv8n+ G+B	87.1	87.9	87.5	90.2
YOLOv8n+ G+B+S	87.8	89.6	88.7	93.1

FIGURE 6. Graphical Results of the Ablation Study.

The results depicted in Figure 9 and Table 2 show a 7.3% enhancement in the Mean Average Precision (mAP) of the model, primarily due to the innovative jump connection mechanism incorporated in the BiFPN, which surpasses the capabilities of the PAN. This method efficiently enables the bidirectional transfer and integration of features across various layers, leading to a notable increase in the precision of identifying distracted driving behaviors.

With the sequential addition of new structures to the model, precision saw respective increases of 0.8%, 0.9%, and 0.7%. In a similar vein, the recall rate experienced gradual enhancements of 2.2%, 2.5%, and 1.7%, correspondingly. These advancements can primarily be attributed to the integration of GAM, which markedly enhances the model's efficacy in intricate scenarios by improving its capability to identify and focus on key features, thereby increasing overall detection precision and efficiency. Additionally, the implementation of the SimC2f network grants the BGS-YOLO model the vital ability to effectively process and discern image features across a broad spectrum of scales and complexities. This integration also led to a significant 4.4% improvement in the F1 score.

3) Comparison Experiments with Other Algorithms

To thoroughly evaluate the performance of the improved algorithm proposed in this paper, comparative experiments with other mainstream algorithms were conducted. The specific results are detailed in Table 3.

This comparative experiment evaluates BGS-YOLO and nine additional algorithms, including Faster-RCNN, SSD, MobileNet-SSD, YOLOv3, RetinaNet, CenterNet, YOLOv5, DETR, and YOLOv7. Visual comparisons of the experimental outcomes are presented in Figure 7.

TABLE 3. Experimental results in comparison with other algorithms.

Model	Precision (%)	Recall (%)	F1 (%)	mAP (%)
Faster-RCNN	79.4	83.2	81.3	83.8
SSD	80.7	84.7	82.7	84.6
MobileNet-SSD	81.4	83.3	82.3	84.2
YOLOv3	82.5	85.4	83.9	86.4
RetinaNet	82.7	85.6	84.1	86.7
CenterNet	83.6	85.9	84.7	87.3
YOLOv5	84.2	86.9	85.5	88.9
DETR	85.3	87.0	86.1	89.4
YOLOv7	86.1	87.3	86.7	91.2
BGS-YOLO	87.8	89.6	88.7	93.1

(a) Experimental Group 1.

(b) Experimental Group 2.

FIGURE 7. Experimental results comparing with other algorithms.

The data shown in Table 3, Figure 7(a), and Figure 7(b) indicate that Faster R-CNN significantly lags behind other algorithms in terms of precision. In comparison with Faster R-CNN, the SSD algorithm excels in the domain of rapid and precise target detection, thanks to its superior detection speed and accuracy, evidencing an enhancement of 1.5% in recall and 1.4% in F1 score. MobileNet-SSD retains the high-speed processing benefits of the SSD algorithm and additionally demonstrates a 0.7% enhancement in precision compared to the original SSD algorithm. YOLOv3, notable for its quick detection speed and effective handling of targets across a range of sizes, exceeds the performance of MobileNet-SSD in identifying and classifying targets across various dimensions. The experimental data indicate that YOLOv3 achieves improvements of 1.1% in precision and 2.2% in mAP compared to MobileNet-SSD. Compared to YOLOv3, RetinaNet demonstrates enhancements in various performance metrics. Simultaneously, CenterNet achieves an additional 0.9% increase in accuracy relative to RetinaNet. Distinguished by its remarkable detection speed and accuracy, YOLOv5, enhanced by a lightweight and deployable framework, significantly surpasses CenterNet, especially in detection efficiency and model compactness, with a substantial rise in mAP from 87.3% to 88.9%. DETR (Detection Transformer) accurately processes complex objects and demonstrates a 1.1% improvement in precision over YOLOv5. Demonstrating excellence in detection accuracy and speed, YOLOv7 proves adept at navigating diverse and complex scenarios, with a 0.8% increase in precision and a 1.8% enhancement in mAP over DETR. The BGS-YOLO model, as proposed in this paper, excels beyond existing frameworks, characterized by significant advancements in feature representation, data capture, and processing efficiency, thus markedly enhancing target detection within dynamic environments. Comparative analysis shows that, relative to YOLOv7, BGS-YOLO records a 2.3% improvement in recall and a 2.0% gain in the F1 score, affirming its superior performance across essential metrics.

IV. Discussion

Addressing the challenges of scale diversity, data background complexity, and feature fusion efficiency in road target detection, this paper proposes a new model, BGS-YOLO. Through innovative design, this model significantly improves the overall performance of target detection.

The paper introduces a Weighted BiFPN that effectively integrates multi-level features, considerably enhancing semantic analysis capabilities and image data processing efficiency. This improvement optimizes the operational efficiency of the model, which is crucial for establishing a safe and efficient road traffic environment. Furthermore, the paper incorporates the GAM, enhancing the traditional model’s ability to capture key information in dynamic environments. This enhancement not only increases the accuracy and interpretability of target detection but also improves efficiency in high-speed dynamic scenarios, contributing to road safety and traffic fluidity. Based on the Simultaneous Attention Module (SimAM), the paper innovatively modifies the C2f module of YOLOv8n, creating the newly designed SimC2f network. By conducting an in-depth analysis of feature fusion, the model effectively tackles the challenges posed by changes in target scale, significantly enhancing the detection precision and response speed for various targets.

This research integrates the Weighted BiFPN, effectively amalgamating multi-level features to substantially boost semantic analysis capabilities and image data processing efficiency. This advancement streamlines the model's operational efficiency, pivotal for fostering a safe and effective road traffic environment. Additionally, this paper incorporates the GAM to augment the model's ability to capture essential information in dynamic settings. This improvement not only elevates the precision and interpretability of target detection but also augments efficiency in high-speed dynamic scenarios, thereby enhancing road safety and traffic flow. Inspired by the Simultaneous Attention Module (SimAM), the study innovatively refines YOLOv8n's C2f module, leading to the creation of the SimC2f network. Through a thorough analysis of feature fusion, the model adeptly addresses the challenges associated with target scale variability, significantly boosting detection accuracy across a variety of targets.

Traditional methods for road target detection, such as SURF, demonstrate effectiveness in specific scenarios but frequently lack the flexibility needed for dynamic environments. Data presented in Table 3 reveal that two-stage algorithms, including Faster-RCNN and SSD, enhance detection accuracy within intricate contexts yet underperform in terms of processing speed and real-time capabilities. Conversely, single-stage algorithms, notably the YOLO series, exhibit rapid processing advantages but

necessitate improvements in accurately identifying small targets and navigating complex environments.

The BGS-YOLO algorithm presented in this paper merges the strengths of two-stage and single-stage detection methods, markedly boosting accuracy and excelling in speed and real-time capabilities, thus providing a sophisticated and effective approach for road target detection. However, opportunities for enhancement exist in crucial domains such as dataset diversity, network architecture optimization, and deployment methodologies. The performance of the model is significantly influenced by the comprehensiveness of the training data, especially when facing extreme climatic conditions and complex lighting environments, necessitating further reinforcement of its generalization capability. To bolster the model's adaptability and generalization abilities, future research endeavors should prioritize the diversification of datasets to encompass a wider array of scenarios, leveraging synthetic data augmentation techniques or real data from diverse locales.

Despite the advancements in feature fusion capabilities facilitated by BiFPN technology, the detection and processing efficiency for diminutive targets remain challenging, necessitating the exploration of more sophisticated architectures. Optimizing the deployment strategy for models, especially for applications like autonomous driving that require rapid response, is exceptionally crucial. This involves ensuring stable and efficient operation across various platforms and enhancing performance in complex and diverse environments.

Additionally, future research should focus on deep synergistic optimization between algorithms and hardware to achieve a balance between energy consumption and performance, and conduct thorough evaluations of model robustness in extreme conditions. These sustained efforts will not only significantly increase the application value of the BGS-YOLO model but also contribute to the advancement of autonomous driving and intelligent transportation system technologies.

V. Conclusions

To enhance planning and control processes in autonomous driving and intelligent surveillance domains, this paper introduces the novel BGS-YOLO model. The integration of the BiFPN facilitates multi-level feature fusion, markedly boosting the algorithm's efficiency in feature extraction and recognition during target detection tasks. The incorporation of the GAM refines global feature representation, while the innovative SimC2f network architecture further hones feature representation, thereby elevating fusion efficiency. Although the model demonstrates superior performance across various metrics, it still imposes a significant computational load. Consequently, forthcoming research endeavors will aim to discover strategies that reconcile high efficiency with swift execution speeds.

References

[1] J. Jeong, Y. H. Yoon, and J. H. Park, "Reliable Road Scene Interpretation Based on ITOM with the Integrated Fusion of Vehicle and Lane Tracker in Dense Traffic Situation," *Sensors*, vol. 20, no. 9, p. 2457, May 2020.

[2] B. Yang and H. Zhang, "A CFAR Algorithm Based on Monte Carlo Method for Millimeter-Wave Radar Road Traffic Target Detection," *Remote Sens.*, vol. 14, no. 8, p. 1779, Apr. 2022.

[3] Y. Li, W. Liu, L. Li, W. Zhang, J. Xu, and H. Jiao, "Vision-Based Target Detection and Positioning Approach for Underwater Robots," *IEEE Photonics J.*, vol. 15, no. 1, p. 8000112, Feb. 2023.

[4] F. Hong, C. H. Lu, W. Tao, and W. Jiang, "Improved SSD Model for Pedestrian Detection in Natural Scene," *Wirel. Commun. Mob. Comput.*, vol. 2022, p. 1500428, Nov. 2022.

[5] F.-H. Wang, L.-Y. Li, Y.-T. Liu, S. Tian, and L. Wei, "Road traffic accident scene detection and mapping system based on aerial photography," *Int. J. Crashworthiness*, vol. 26, no. 5, pp. 537–548, Sep. 2021.

[6] H. Wang, J. Zhao, H. Wang, C. Hu, J. Peng, and S. Yue, "Attention and Prediction-Guided Motion Detection for Low-Contrast Small Moving Targets," *IEEE T. Cybern.*, vol. 53, no. 10, pp. 6340–6352, Oct. 2023.

[7] L. Liu *et al.*, "Yolo-3DMM for Simultaneous Multiple Object Detection and Tracking in Traffic Scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 8, pp. 9467–9481, Aug. 2024.

[8] X. Lu, Y. Zhong, and L. Zhang, "Open-Source Data-Driven Cross-Domain Road Detection From Very High Resolution Remote Sensing Imagery," *IEEE Trans. Image Process.*, vol. 31, pp. 6847–6862, 2022.

[9] L. Lei, D.-H. Kim, W.-J. Park, and S.-J. Ko, "Face Recognition Using LBP Eigenfaces," *IEICE Trans. Inf. Syst.*, vol. E97D, no. 7, pp. 1930–1932, Jul. 2014.

[10] R. Wang, L. Tang, and T. Tang, "Fast Sample Adaptive Offset Jointly Based on HOG Features and Depth Information for VVC in Visual Sensor Networks," *Sensors*, vol. 20, no. 23, p. 6754, Dec. 2020.

[11] M. Mahamdioua and M. Benmohammed, "Automatic adaptation of SIFT for robust facial recognition in uncontrolled lighting conditions," *IET Comput. Vis.*, vol. 12, no. 5, pp. 623–633, Aug. 2018.

[12] R. Srivastava, R. Tomar, A. Sharma, G. Dhiman, N. Chilamkurti, and B.-G. Kim, "Real-Time Multimodal Biometric Authentication of Human Using Face Feature Analysis," *CMC-Comput. Mat. Contin.*, vol. 69, no. 1, pp. 1–19, 2021.

[13] Z. He, Q. Li, H. Feng, and Z. Xu, "Class agnostic moving target detection by color and location prediction of moving area," *Optik*, vol. 251, p. 168002, Feb. 2022.

[14] G. Lian, Y. Wang, H. Qin, and G. Chen, "Towards unified on-road object detection and depth estimation from a single image," *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 5, pp. 1231–1241, May 2022.

[15] S. Babbar and J. Bedi, "Real-time traffic, accident, and potholes detection by deep learning techniques: a modern approach for traffic management," *Neural Comput. Appl.*, vol. 35, no. 26, pp. 19465–19479, Sep. 2023, doi: 10.1007/s00521-023-08767-8.

[16] U. Ahmed, G. Srivastava, Y. Djenouri, and J. C.-W. Lin, "Knowledge graph based trajectory outlier detection in sustainable smart cities," *Sust. Cities Soc.*, vol. 78, p. 103580, Mar. 2022, doi: 10.1016/j.scs.2021.103580.

[17] P. Wan, X. Deng, L. Yan, X. Jing, L. Peng, and X. Wang, "A Double-Layered Belief Rule Base Model for Driving Anger Detection Using Human, Vehicle, and Environment Characteristics: A Naturalistic Experimental Study," *Comput. Intell. Neurosci.*, vol. 2022, p. 5698393, Jan. 2022.

[18] T. Xue, Z. Zhang, W. Ma, Y. Li, A. Yang, and T. Ji, "Nighttime Pedestrian and Vehicle Detection Based on a Fast Saliency and Multifeature Fusion Algorithm for Infrared Images," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16741–16751, Sep. 2022.

[19] S. Singh and S. K. Verma, "Congestion and Accident Alerts Using Cloud Load Balancing & Random Forest in VANET," *Wirel. Pers. Commun.*, vol. 128, no. 1, pp. 43–65, Jan. 2023.

[20] B. Mahaur, N. Singh, and K. K. Mishra, "Road object detection: a comparative study of deep learning-based algorithms," *Multimed. Tools Appl.*, vol. 81, no. 10, pp. 14247–14282, Apr. 2022.

[21] G. Zhang, Y. Peng, and H. Wang, "Road Traffic Sign Detection Method Based on RTS R-CNN Instance Segmentation Network," *Sensors*, vol. 23, no. 14, p. 6543, Jul. 2023.
[22] Y. Zhang, Z. Shi, and Y. Zhang, "Adioc loss: An Auxiliary descent IoC loss function," *Eng. Appl. Artif. Intell.*, vol. 116, p. 105453, Nov. 2022.
[23] B. Zeng *et al.*, "Top-Down aircraft detection in large-scale scenes based on multi-source data and FEF-R-CNN," *Int. J. Remote Sens.*, vol. 43, no. 3, pp. 1108–1130, Feb. 2022.
[24] J. Luo, H. Fang, F. Shao, Y. Zhong, and X. Hua, "Multi-scale traffic vehicle detection based on faster R-CNN with NAS optimization and feature enrichment," *Def. Technol.*, vol. 17, no. 4, pp. 1542–1554, Aug. 2021.
[25] J. Ma, H. Guo, S. Rong, J. Feng, and B. He, "Infrared Dim and Small Target Detection Based on Background Prediction," *Remote Sens.*, vol. 15, no. 15, p. 3749, Aug. 2023.
[26] Q. Wu *et al.*, "Improved Mask R-CNN for Aircraft Detection in Remote Sensing Images," *Sensors*, vol. 21, no. 8, p. 2618, Apr. 2021.
[27] F. Zhou, H. Zu, Y. Li, Y. Song, J. Liao, and C. Zheng, "Traffic-Sign-Detection Algorithm Based on SK-EVC-YOLO," *Mathematics*, vol. 11, no. 18, p. 3873, Sep. 2023.
[28] Q. Xu, R. Lin, H. Yue, H. Huang, Y. Yang, and Z. Yao, "Research on Small Target Detection in Driving Scenarios Based on Improved Yolo Network," *IEEE Access*, vol. 8, pp. 27574–27583, 2020.
[29] H. Wang, D. Han, M. Cui, and C. Chen, "NAS-YOLOX: a SAR ship detection using neural architecture search and multi-scale attention," *Connection Science*, vol. 35, no. 1, pp. 1–32, Dec. 2023.
[30] L. Li, L. Jiang, J. Zhang, S. Wang, and F. Chen, "A Complete YOLO-Based Ship Detection Method for Thermal Infrared Remote Sensing Images under Complex Backgrounds," *Remote Sens.*, vol. 14, no. 7, p. 1534, Apr. 2022.
[31] J. Li, H. Tang, X. Li, H. Dou, and R. Li, "LEF-YOLO: a lightweight method for intelligent detection of four extreme wildfires based on the YOLO framework," *Int. J. Wildland Fire*, vol. 33, no. 1, Dec. 2023.
[32] S. Liu, Y. Wang, Q. Yu, H. Liu, and Z. Peng, "CEAM-YOLOv7: Improved YOLOv7 Based on Channel Expansion and Attention Mechanism for Driver Distraction Behavior Detection," *IEEE Access*, vol. 10, pp. 129116–129124, 2022.
[33] L. Shen, B. Lang, and Z. Song, "DS-YOLOv8-Based Object Detection Method for Remote Sensing Images," *IEEE Access*, vol. 11, pp. 125122–125137, 2023.
[34] Y. Zhang and C. Liu, "Real-Time Pavement Damage Detection With Damage Shape Adaptation," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2024.
[35] T. Li, Y. Zhang, Q. Li, and T. Zhang, "AB-DLM: An Improved Deep Learning Model Based on Attention Mechanism and BiFPN for Driver Distraction Behavior Detection," *IEEE Access*, vol. 10, pp. 83138–83151, 2022.
[36] Y. Du, X. Xu, and X. He, "Optimizing Geo-Hazard Response: LBE-YOLO's Innovative Lightweight Framework for Enhanced Real-Time Landslide Detection and Risk Mitigation," *Remote Sensing*, vol. 16, no. 3, p. 534, Jan. 2024.
[37] W. Guo, W. Li, Z. Li, W. Gong, J. Cui, and X. Wang, "A Slimmer Network with Polymorphic and Group Attention Modules for More Efficient Object Detection in Aerial Images," *Remote Sens.*, vol. 12, no. 22, p. 3750, Nov. 2020.
[38] V. C. Mahaadevan, R. Narayanamoorthi, R. Gono, and P. Moldrik, "Automatic Identifier of Socket for Electrical Vehicles Using SWIN-Transformer and SimAM Attention Mechanism-Based EVS YOLO," *IEEE Access*, vol. 11, pp. 111238–111254, 2023.
[39] Y. Du, X. Liu, Y. Yi, and K. Wei, "Optimizing Road Safety: Advancements in Lightweight YOLOv8 Models and GhostC2f Design for Real-Time Distracted Driving Detection," *Sensors*, vol. 23, no. 21, p. 8844, Nov. 2023.

Xingyu Liu received the MSc in Engineering Control and Instrumentation from the University of Huddersfield, UK in 2019. She is currently pursuing a Ph.D. in Electrical Engineering at the University of Huddersfield, UK.
Her current research interests include machine learning, intelligent detection and electrical system control and optimization.

Yuanfeng Chu received the M.Eng. degree in Control Engineering from Shenyang Aerospace University, China, in 2019. He is currently a Senior Electrical and Electronics Architecture Engineer at Li Auto Inc. in China. His current research interests include deep learning, drone formation, and automotive fault diagnosis.

Yiheng Hu (S'18-M'21) received the Ph.D. degree in Electrical Engineering from University of Huddersfield, Huddersfield, UK, in 2021. She was a Senior Power System Researcher at University College Dublin, Dublin, Ireland about two years. She is currently a Lecturer in electronics and electrical engineering, at University of Huddersfield, Huddersfield. UK. Her research interests include electrical machines and energy storage systems for electric vehicles and renewable energy systems, as well as the associated AI technologies.

Nan Zhao (S'15-M'17) received the Ph.D. degree in Electrical Engineering from McMaster University, Ontario, Canada, in 2017. He was a Sessional Lecturer with the Department of Electrical and Computer Engineering, McMaster University, from 2017 to 2018. He was an Assistant Professor with the School of Electrical and Electronic Engineering, University College Dublin, Ireland, from 2018 to 2022. In 2022, he joined the School of Engineering, Lancaster University, Lancaster, UK, as a Senior Lecturer. His research interests include transportation electrification, renewables integration, and the associated machine learning and AI technologies.