vehicles

MDPI

Article

A Deep Learning Framework for Traffic Accident Detection Based on Improved YOLO11

Weijun Li [1], Liyan Huang [1,*] and Xiaofeng Lai [2]

[1] Department of Forensic Science, Fujian Police College, Fuzhou 350007, China
[2] Public Security Bureau, Sanming 365000, China
* Correspondence: huangliyan@fjpsc.edu.cn

Abstract

The automatic detection of traffic accidents plays an increasingly vital role in advancing intelligent traffic monitoring systems and improving road safety. Leveraging computer vision techniques offers a promising solution, enabling rapid, reliable, and automated identification of accidents, thereby significantly reducing emergency response times. This study proposes an enhanced version of the YOLO11 architecture, termed YOLO11-AMF. The proposed model integrates a Mamba-Like Linear Attention (MLLA) mechanism, an Asymptotic Feature Pyramid Network (AFPN), and a novel Focaler-IoU loss function to optimize traffic accident detection performance under complex and diverse conditions. The MLLA module introduces efficient linear attention to improve contextual representation, while the AFPN adopts an asymptotic feature fusion strategy to enhance the expressiveness of the detection head. The Focaler-IoU further refines bounding box regression for improved localization accuracy. To evaluate the proposed model, a custom dataset of traffic accident images was constructed. Experimental results demonstrate that the enhanced model achieves precision, recall, mAP50, and mAP50–95 scores of 96.5%, 82.9%, 90.0%, and 66.0%, respectively, surpassing the baseline YOLO11n by 6.5%, 6.0%, 6.3%, and 6.3% on these metrics. These findings demonstrate the effectiveness of the proposed enhancements and suggest the model's potential for robust and accurate traffic accident detection within real-world conditions.

Keywords: accident detection; computer vision; YOLO11; deep learning; emergency response optimization

Academic Editors: Lie Yang and Xiangkun He

Received: 24 June 2025
Revised: 25 July 2025
Accepted: 30 July 2025
Published: 4 August 2025

Citation: Li, W.; Huang, L.; Lai, X. A Deep Learning Framework for Traffic Accident Detection Based on Improved YOLO11. *Vehicles* **2025**, *7*, 81. https://doi.org/10.3390/vehicles7030081

1. Introduction

Traffic accident detection plays a pivotal role in contemporary traffic management and public safety systems [1]. The continued expansion of road networks and increasing vehicle density have heightened the urgency for effective and reliable accident detection technologies [2,3]. Traditional traffic monitoring methods, reliant on manual observation, suffer from limitations including human error and a lack of scalability. Consequently, these shortcomings are driving the adoption of automated solutions [4]. The automation of traffic accident detection has emerged as an inevitable trend, with the accurate and timely identification of accidents during monitoring remaining a central challenge [5].

Driven by the growing demand for automated traffic accident detection, this technology has emerged as a critical component of modern intelligent transportation systems (ITSs) [6]. Early detection methods were predominantly based on traditional machine learning techniques which typically required image preprocessing and the manual extraction of

features such as shape, texture, and color, followed by the training of classification models to detect traffic accidents [7]. However, manual feature engineering is highly susceptible to environmental variations and subjective biases, thereby limiting the robustness and scalability of these methods in real-world traffic conditions [8,9].

In recent years, the rapid advancement of deep learning has provided promising solutions to the limitations of traditional methods [10]. Supported by improvements in computational hardware and algorithmic optimization, deep learning techniques have attracted significant attention in the domain of intelligent visual recognition [11]. Unlike conventional approaches, deep learning models are capable of processing raw image data in an end-to-end manner, automatically learning hierarchical feature representations from the data themselves [12,13]. This capability eliminates the need for manual feature extraction and mitigates the risk of error propagation inherent in multi-stage processing pipelines. Moreover, deep learning models demonstrate strong scalability and can achieve superior performance when trained on large-scale datasets, thereby enabling their application across diverse traffic environments and real-world scenarios [14,15].

Despite these advantages, a comprehensive review of the existing literature reveals several notable limitations. Many deep learning models have been evaluated predominantly under idealized or controlled conditions, with limited consideration for the complexities of real-world traffic environments [16]. Challenging factors such as variable lighting, occlusions, object overlaps, and the detection of small objects are often insufficiently addressed, thereby constraining the models' generalizability and robustness in practical applications [17].

This study builds upon the YOLO11n, introducing several key modifications aimed at enhancing its effectiveness in traffic accident detection. The primary contributions of this work are summarized as follows:

- A dataset was constructed, covering a diverse range of accident types and environmental conditions. This dataset facilitates robust feature learning and significantly improves the model's generalization capability in real-world traffic scenarios.
- MLLA (Mamba-Like Linear Attention) Module: A lightweight linear attention mechanism was integrated into the network architecture, enhancing feature representation and improving detection accuracy while incurring minimal computational overheads.
- AFPN (Asymptotic Feature Pyramid Network): An optimized detection head structure was implemented and an asymptotic feature fusion strategy was employed to progressively integrate low-level, high-level, and top-level features. This design enhances detection performance, particularly for small and overlapping objects.
- Focaler-IoU Loss Function: The Focaler-IoU loss function was employed to improve localization accuracy and enhance the model's robustness in complex and challenging detection scenarios.

2. Related Works

Traffic accident detection has been an active area of research within intelligent transportation systems (ITSs), evolving significantly from traditional computer vision pipelines to modern deep learning-based approaches. While methods such as that proposed by Kussia et al. [18], which employed convolutional neural networks (CNNs) to classify traffic conditions, including accidents, achieved classification accuracies of up to 94.4%, their effectiveness was constrained by a reliance on hand-engineered features and high sensitivity to environmental variations. To address computational inefficiency and improve generalizability, lightweight deep learning models have gained traction. Tamagusko et al. [19] employed transfer learning with MobileNetV2 and EfficientNetB1 architectures, supplemented by synthetic image augmentation, to detect abnormal traffic events on Nordic roads, achieving

mean average precision (mAP) scores ranging from 88% to 89%. Although these models exhibited competitive performance, they were primarily implemented using two-stage detection frameworks, which inherently introduce latency and reduce responsiveness in real-time applications.

More complex frameworks have sought to address these limitations by integrating multiple modules for detection, tracking, and behavior analysis. Ghahremannezhad et al. [20] proposed a real-time traffic accident detection pipeline which combines YOLOv4 for object detection, a Kalman filter for vehicle tracking, and a trajectory conflict detection module. Although the system demonstrated promising results under controlled conditions, its multi-stage structure increased system complexity and reduced robustness in scenarios involving occlusion, visual clutter, or poor visibility. To address these challenges, attention mechanisms have recently emerged as effective solutions for enhancing feature learning in lightweight networks. Lin et al. [21] incorporated a spatial attention module into a MobileNetV2 backbone for traffic congestion detection, achieving an accuracy of 98.58% on a highway dataset while maintaining real-time efficiency. Similarly, temporal modeling has been explored through CNN–RNN hybrid architectures, which analyze sequential data to capture motion dynamics. One such approach reported an accuracy exceeding 98% for accident classification using video sequences, underscoring the importance of modeling spatiotemporal patterns in dynamic traffic scenarios.

Although significant advancements have been made in traffic accident detection, the existing literature continues to reveal critical limitations in generalization capability and real-world robustness. Most models are trained and evaluated on datasets which lack sufficient diversity in accident types, environmental conditions, and road scenarios. Key challenges, such as small object detection, object occlusion, low-light environments, and multi-class accident classification remain inadequately addressed. Fang et al. [22] conducted a comprehensive survey of visual accident detection techniques, emphasizing the limited scalability and adaptability of many deep learning models when applied to dynamic traffic environments. These challenges underscore the urgent need for unified, end-to-end architectures which can achieve efficient feature representation, accurate detection across varied scenarios, and rapid inference suitable for real-time deployment.

To address these gaps, this study proposes an enhanced YOLO11-based framework, incorporating attention mechanisms, optimized feature fusion strategies, and a novel dataset encompassing diverse accident conditions.

3. Materials and Methods

3.1. Dataset Construction

To support the training and evaluation of the YOLO11-AMF for traffic accident detection, a comprehensive and diverse image dataset was constructed by aggregating data from multiple publicly available online sources. The dataset was carefully curated to reflect real-world scenarios, thereby facilitating the development of a model capable of achieving robust and reliable performance across heterogeneous traffic environments. A total of 594 images were collected, exhibiting significant variation in resolution, image quality, object scale, and lighting conditions. To simulate realistic operational challenges, the dataset includes various weather conditions such as clear, rainy, foggy, and overcast scenes, all of which are known to influence the accuracy of visual recognition systems. Additionally, each image was tagged with its traffic scene type, urban road, rural road, or highway, based on contextual features. This information was not used during model training but was recorded to ensure dataset diversity and support future scene-aware research.

For effective model training and evaluation, the dataset was systematically divided into three subsets: 411 images for training, 35 for validation, and 148 for testing. This parti-

tioning ensured a balanced distribution across scene types and supported both parameter optimizations. To further enhance model robustness and mitigate overfitting, a three-stage data augmentation strategy was employed, involving random 90-degree rotations to simulate camera orientation variations, Gaussian blurring to emulate motion-induced image degradation, and the injection of stochastic noise to mimic artifacts from low-quality sensors or unstable transmission conditions. This augmentation pipeline effectively expanded the diversity of the training data, enabling the model to develop stronger feature generalization and improved resilience to noise, occlusion, and distortion, factors which are critical for ensuring accurate and dependable accident detection.

3.2. YOLO11-AMF

Given the complexity of real-world road environments, conventional object detection frameworks often face difficulties in maintaining an optimal balance between detection accuracy and computational efficiency. To address these challenges, a comprehensive evaluation of state-of-the-art object detection models was undertaken. Following a rigorous comparative analysis, YOLO11n, the most recent advancement in the YOLO series, was selected as the baseline architecture for this study, owing to its favorable trade-off between performance and real-time inference capability.

The YOLO series of object detectors, including the recent YOLO11, is renowned for its exceptional balance of inference speed and detection accuracy. However, a persistent challenge remains in its application to environments with high object density, such as traffic accident scenes, where the detection of small, partially occluded, and overlapping objects is critical. In response to this limitation, we propose YOLO11-AMF, an advanced variant of the YOLO11n architecture. To enhance the model's ability to discern complex contextual relationships, we integrate a Mamba-Like Linear Attention (MLLA) module. Concurrently, an Asymptotic Feature Pyramid Network (AFPN) is employed to optimize the fusion of features across different scales, which is vital for detecting objects of varying sizes. Finally, to specifically address localization inaccuracies for challenging objects, a novel Focaler-IoU loss function is introduced to refine bounding box regression.

As illustrated in Figure 1, the YOLO11-AMF architecture maintains the conventional three-stage structure, consisting of the backbone, neck, and head modules. To improve adaptability to diverse traffic scenarios, each module has been systematically enhanced. The feature extraction backbone integrates a modified C2PSA-MLLA module, which extends the original C2PSA block by incorporating a Mamba-Like Linear Attention (MLLA) mechanism. This integration strengthens the network's capacity for robust multiscale feature representation. Furthermore, a Spatial Pyramid Pooling Fast (SPPF) module is employed to aggregate global contextual information, thereby improving the model's ability to detect distant and small-scale objects effectively.

The backbone and neck components of the proposed YOLO11-AMF architecture are specifically designed to enhance multiscale feature extraction and semantic representation, both of which are critical for accurate traffic accident detection. The neck incorporates a redesigned C3k2 module, which replaces the traditional bottleneck structure with a serialized pair of C3k blocks (with hyperparameter N = 2), thereby increasing the depth of feature learning while maintaining a lightweight computational footprint suitable for real-time deployment. As illustrated in Figure 2, the improved C3k2 configuration (c3k = True) adopts deeper C3k pathways compared to the original structure (c3k = False), enabling richer and more expressive feature representations. Additionally, the architecture integrates a cross-scale feature fusion strategy through upsampling operations and lateral skip connections, facilitating the hierarchical aggregation of low-level spatial information with high-level

semantic features. This design leads to the generation of more discriminative feature maps across multiple resolution levels, improving detection robustness and accuracy.

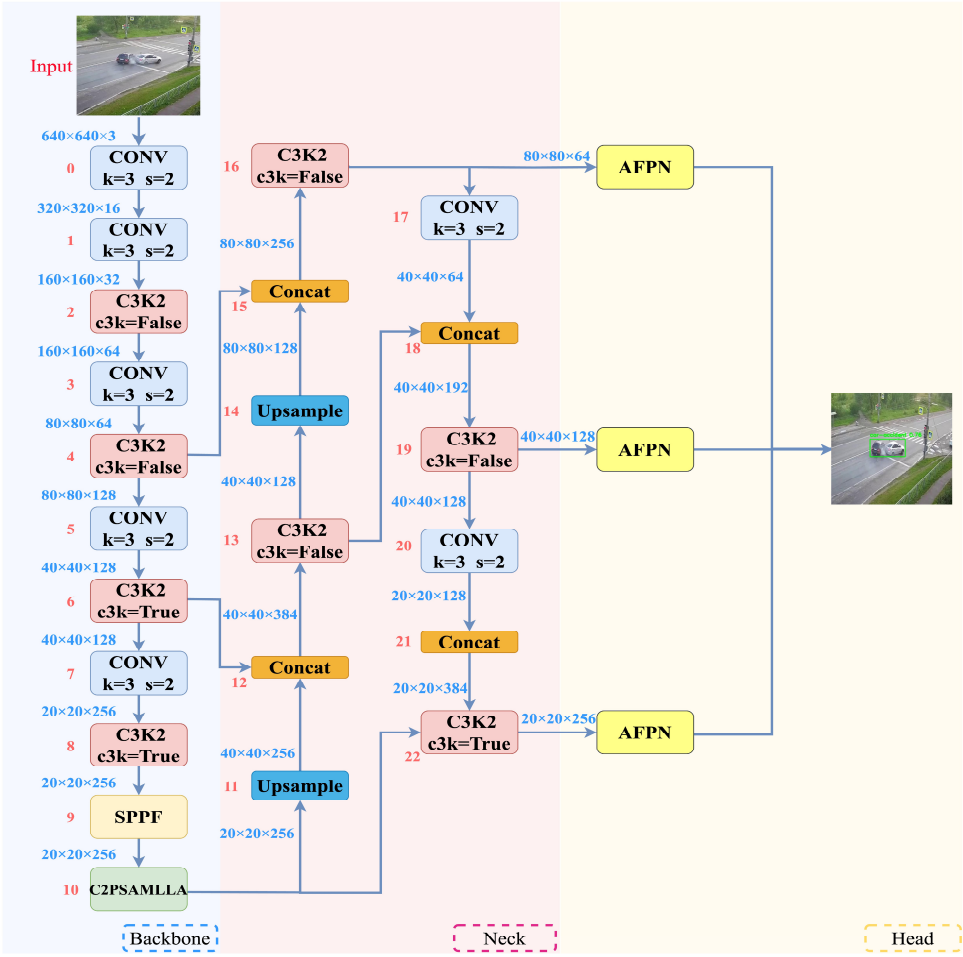

Figure 1. Model architecture of YOLO11-AMF. In the backbone, the Mamba-Like Linear Attention (MLLA) module is integrated to enhance contextual representation. In the head, an Asymptotic Feature Pyramid Network (AFPN) is employed to improve feature fusion. The red numbers indicate the number of layers in each component.

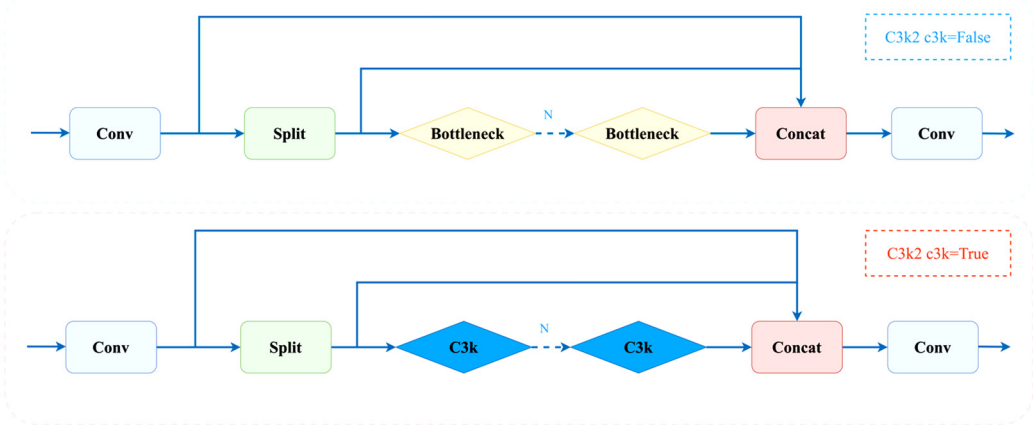

Figure 2. Comparison of the original and improved C3k2 module structures. The upper architecture (c3k = False) uses bottleneck blocks, while the lower version (c3k = True) serializes two C3k blocks with N = 2 for deeper feature extraction.

To further enhance the model's capability in detecting objects of varying scales, a novel Asymptotic Feature Pyramid Network (AFPN) is introduced to replace the conventional detection head. Unlike traditional feature pyramid structures which rely on fixed or heuristic fusion strategies, the AFPN incorporates learnable gating mechanisms which enable the network to adaptively adjust the contribution of feature maps from different pyramid levels based on the semantic content of the input. This dynamic weighting mechanism allows the model to emphasize the most relevant spatial and contextual features, thereby improving its scale-invariant detection performance.

3.2.1. MLLA

The Mamba-Like Linear Attention (MLLA) mechanism introduces a novel, Transformer-inspired attention module specifically tailored for visual recognition tasks in real-time traffic accident detection frameworks [23,24]. As depicted in Figure 3, MLLA diverges from conventional self-attention mechanisms, which are often computationally expensive and exhibit poor scalability with increasing input dimensions, by employing a streamlined architecture which integrates a forget gate and a linearized attention structure. This architectural simplification effectively alleviates the memory and latency bottlenecks typically associated with traditional Transformer models.

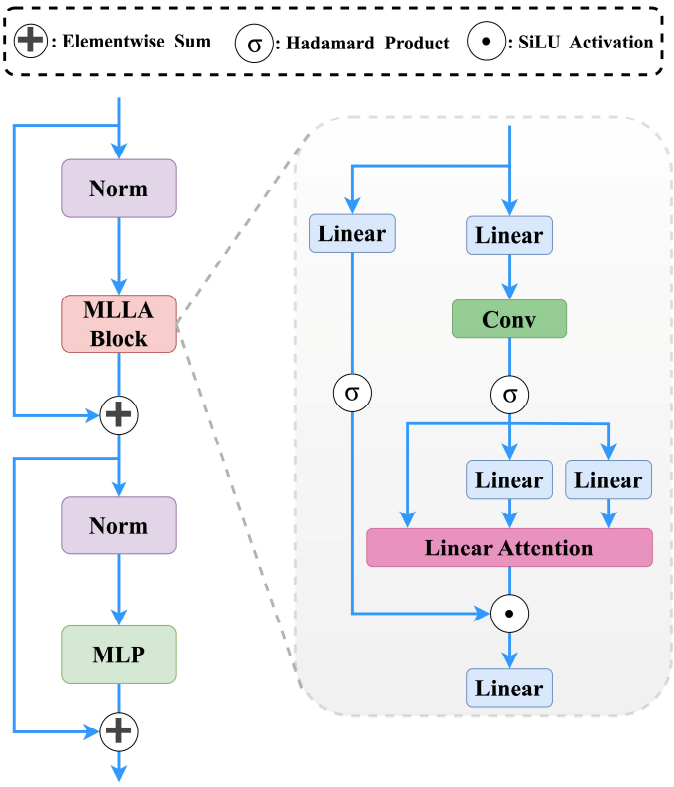

Figure 3. Structure of the proposed MLLA (Multi-Level Linear Attention) block. The block consists of a normalization layer, the MLLA attention module, and an MLP. The MLLA module integrates linear and convolutional projections with Hadamard product and SiLU activation to compute attention.

The forget gate plays a crucial role in dynamically modulating the influence of historical and current input features, allowing the model to selectively preserve salient spatial and temporal information across sequential feature maps. This capability is particularly beneficial in traffic accident scenarios, where subtle, transient cues, such as sudden deceleration, skidding, or impact shadows, may be key indicators of an event. By enhancing the model's sensitivity to such contextual patterns, MLLA significantly improves detection accuracy and robustness.

Furthermore, the linear formulation of the attention mechanism ensures that the computational complexity scales linearly with respect to the input size, in contrast to the quadratic growth seen in standard self-attention models. This makes the MLLA mechanism highly suitable for deployment in edge computing environments or resource-constrained systems, where real-time inference and computational efficiency are critical requirements.

3.2.2. AFPN

The Adaptive Feature Pyramid Network (AFPN) significantly enhances traditional feature fusion strategies by introducing a dynamic, content-aware mechanism for multiscale feature representation [25,26]. As illustrated in Figure 4, in contrast to conventional Feature Pyramid Networks (FPNs) which apply fixed, uniform fusion rules across all scales, the AFPN leverages a learnable gating function to dynamically modulate the information flow between different feature levels. This adaptive control enables the network to selectively amplify salient features while suppressing irrelevant or redundant information based on the input image's context and complexity. Architecturally, the AFPN employs a two-stage pyramidal fusion pipeline where lower-resolution, semantically rich features are progressively integrated with high-resolution spatial features.

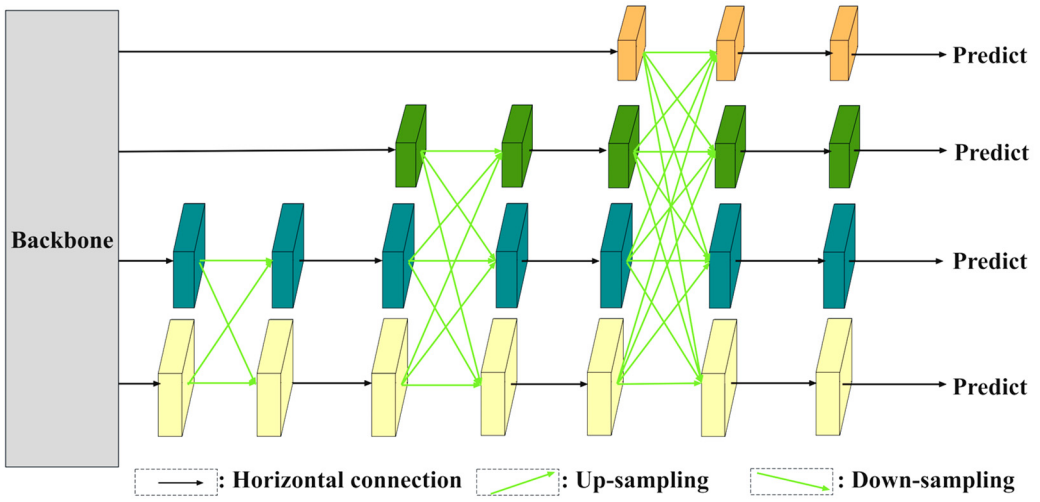

Figure 4. Architecture of Asymptotic Feature Pyramid Network (AFPN). Low-level features are fused initially, followed by higher-level and top-level features. Black arrows denote convolutions; aquamarine arrows indicate adaptive spatial fusion.

The hierarchical integration is fine-tuned by the gating mechanism, which allows the network to dynamically adjust its receptive field and feature representation in response to objects of varying sizes and occlusion levels. Such flexibility is particularly advantageous in traffic accident detection, a scenario characterized by complex scenes, overlapping vehicles, and significant scale variation. By emphasizing contextually relevant features while preserving high-fidelity spatial information, the AFPN improves both localization accuracy and object classification robustness. Furthermore, empirical results from benchmark evaluations confirm that the AFPN achieves superior detection precision, outperforming static pyramid-based methods in generalizability and performance under challenging visual conditions. Ultimately, the AFPN strengthens the YOLO11-AMF framework by providing a scalable, adaptive mechanism for more intelligent and resilient feature learning across diverse traffic environments.

3.2.3. Focaler-IoU

To enhance the regression accuracy of object boundaries, particularly in complex traffic scenarios, this study introduces a refined loss function termed Focaler-IoU. This function refines the conventional Intersection over Union (IoU) loss by implementing a linear interval mapping strategy [27,28]. The primary objective of this method is to enable a more nuanced optimization process by modulating the loss contributions from regression samples across a spectrum of difficulty. The underlying mechanism involves mapping the original IoU value to a recalibrated interval $[d, u] \subseteq [0, 1]$, where d and u are tunable hyperparameters. By adjusting these bounds, the Focaler-IoU function dynamically shifts its focus toward samples near critical decision boundaries. This targeted weighting mitigates the risk of the model overfitting to dominant, easily detected regions and enhances its sensitivity to subtle spatial variations, such as those involving occlusions, ambiguous edges, or small objects, which are critical for precise localization.

$$\mathrm{IoU}^{\mathrm{focaler}} = \begin{cases} 0 & \mathrm{Iou} < \mathrm{d} \\ \frac{\mathrm{IoU}-\mathrm{d}}{\mathrm{u}-\mathrm{d}} & \mathrm{d} \le \mathrm{Iou} \le \mathrm{u} \\ 1 & \mathrm{Iou} > \mathrm{u} \end{cases} \tag{1}$$

The reconstructed loss function is formally expressed in Equation (1), where the original IoU is remapped using the hyperparameters d and u to produce the final $\mathrm{IoU}^{\mathrm{focaler}}$ value. By strategically adjusting d and u, the Focaler-IoU method reconstructs the loss landscape to prioritize both challenging and easily identifiable samples, thereby fostering a more balanced training dynamic. Consequently, the incorporation of Focaler-IoU into the training pipeline contributes to more stable model convergence, improved boundary alignment, and enhanced detection robustness, especially under diverse and noisy visual conditions.

3.2.4. Evaluation Indicators

Evaluating the performance of the YOLO-AMF model for traffic accident detection necessitates a multifaceted assessment which extends beyond conventional object detection metrics [29,30]. In this study, key evaluation indicators include precision (P), recall (R), average precision (AP), and mean average precision (mAP), specifically mAP50 and mAP50–95, which together provide a comprehensive measure of model accuracy and generalization. Among these, precision holds particular significance in the context of accident detection, as it quantifies the proportion of correctly predicted accident instances among all cases labeled as accidents by the model. A high precision value indicates a low false positive rate, which is essential to ensure that the system does not erroneously flag normal scenes as accidents, improving trust, minimizing unnecessary alerts, and enhancing operational reliability in real-world deployment scenarios.

Formally, precision is calculated as:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

where TP (true positive) denotes correctly identified accident instances, and FP (false positive) represents cases incorrectly flagged as accidents.

Complementary to precision, recall measures the proportion of actual accidents, which are successfully detected by the model, reflecting the system's ability to minimize false negatives (FNs), real accidents, which go undetected. Recall is computed as:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

In the context of traffic accident detection, a false negative is particularly critical, as it corresponds to a failure to detect a real incident, potentially delaying emergency response or mitigation efforts. Therefore, maintaining a balance between precision and recall is crucial.

To capture this balance, average precision (AP) is introduced, which integrates both precision and recall over varying thresholds, offering a unified measure of detection performance. It is defined as the area under the precision–recall curve (P–R curve):

$$AP = \int_0^1 P(r)dr \tag{4}$$

where P (r) denotes precision as a function of recall.

For multi-class evaluation or multi-scenario assessments, mean Average Precision (mAP) is used, defined as:

$$mAP = \frac{1}{k}\sum_{k=1}^{k} AP_k \tag{5}$$

Here, k is the number of categories and AP_k is the AP value of the category k.

Another critical metric is the F1-score, which serves as the harmonic mean of precision and recall:

$$F1 - score = 2\frac{PR}{P + R} \tag{6}$$

Here, 'P' represents precision, and 'R' represents recall.

The F1-score provides a balanced evaluation metric, particularly valuable in imbalanced datasets where the cost of both false positives and false negatives is high. Unlike standalone precision or recall, it captures the trade-off between the two and is especially relevant in safety-critical applications like traffic accident detection. These metrics not only quantify the model's capability in localizing and classifying accident scenes but also shed light on its robustness, reliability, and readiness for deployment in complex real-world environments.

3.3. Experimental Environment and Parameter Settings

In order to ensure the fairness and reproducibility of the training and evaluation process, a diverse traffic accident image dataset was constructed and randomly divided into training, validation, and testing sets with a ratio of approximately 8:1, comprising 411 training images and 35 validation images. The training set covered a wide range of realistic traffic scenarios, including normal flow, minor collisions, and severe accident scenes, thereby enhancing the model's generalization ability.

All training procedures were conducted on Google Colab, utilizing a GPU-enabled cloud environment. The details of the experimental configuration are provided in Table 1, while the selected hyperparameter settings are summarized in Table 2.

Table 1. Experimental environment configuration.

Environment Configuration	Parameter
Operating system	Ubuntu 20.04 (Google Colab Environment)
CPU	Intel Xeon @ 2.20 GHz
GPU	NVIDIA Tesla T4 (12 GB GDDR5)
RAM	13 GB
Development environment	Google Colab
Programming language	Python 3.9

Table 2. Hyperparameter settings.

Hyperparameter	Value
Epochs	500
Batch size	16
Num workers	2
Initial learning rate	0.01
Optimizer	Adam
Input image size	640 × 640

To evaluate detection performance under complex real-world conditions, the mean Average Precision (mAP) at an IoU threshold of 0.5 was used as the primary evaluation metric. The final hyperparameter configuration was selected to optimize both convergence speed and detection accuracy while ensuring real-time inference efficiency.

4. Results

4.1. Experimental Results of the Improved YOLO11 Model

As shown in Table 3, we compare the performance of several representative object detection models, including DETR, Faster R-CNN, YOLOv5n, YOLOv8n, YOLO11, and our improved model YOLO11-AMF, all evaluated under the same batch size of 16. DETR achieves a precision of 91.4% and an mAP50–95 of 63.4%, while Faster R-CNN obtains slightly higher accuracy with an mAP50–95 of 66.1%. DETR, as a transformer-based one-stage detector, benefits from global attention mechanisms but suffers from slow convergence and high computational cost. Faster R-CNN, a classical two-stage detection framework, often achieves high accuracy but at the expense of increased inference time and model complexity.

Table 3. Comparative experimental results of the different models.

Algorithm	Batch Size	Precision/%	mAP50–95/%	Parameters/m	GFlops
DETR	16	91.4	63.4	36.7	36.81
Faster R-CNN	16	82.9	66.1	28.2	37.52
YOLOV5n	16	94.9	58.6	2.5	7.2
YOLOV8n	16	87.2	59.4	3.0	8.2
YOLO11	16	90.0	59.7	2.6	6.4
YOLO-AMF	16	96.5	66	2.7	6.8

However, both models come with substantial computational overheads, with DETR having 36.7 million parameters and 36.81 GFlops, and Faster R-CNN having 28.2 million parameters and 37.52 GFlops. This significantly limits their practical use in real-time scenarios. In contrast, YOLO-based models strike a better balance between accuracy and efficiency. For example, YOLOv5n and YOLOv8n greatly reduce model size and complexity, with parameter counts around 2.5 to 3 million and GFlops below 8, while still maintaining competitive detection performance. Notably, the proposed YOLO11-AMF achieves the highest precision at 96.5% and an mAP50–95 of 66.0%, comparable to Faster R-CNN, but with a much smaller footprint of 2.7 million parameters and 6.8 GFlops.

Given the strict requirements of real-time traffic accident detection, such as rapid response, low latency, and efficient deployment on edge devices (e.g., roadside cameras or embedded systems), YOLO-based models, particularly YOLO11-AMF, demonstrate clear advantages. The model's high precision helps reduce false alarms in critical safety systems, while its lightweight architecture ensures fast inference and ease of deployment.

Following 500 epochs of training, the proposed YOLO11-AMF model underwent a comprehensive quantitative evaluation on the test set. The model demonstrated exceptional performance, achieving a precision of 96.5% and a recall of 82.9%. On more comprehensive metrics, it obtained a mean average precision at an IoU of 0.5 (mAP50) of 90.0% and a mAP50–95 of 66.0%. Furthermore, the model achieved a high F1-score of 89.2%, indicating a well-balanced trade-off between precision and recall. These quantitative results collectively validate the model's high accuracy and robustness, confirming its suitability for deployment in complex, real-time traffic monitoring systems.

Figure 5 illustrates the training and validation performance curves of the proposed traffic accident detection model. As observed, all loss metrics, including box loss, cls_loss, and dfl_loss, demonstrate a clear and consistent decline as the number of training iterations increases. This downward trend indicates that the model is effectively minimizing prediction errors and achieving progressive convergence during training. Notably, the loss curves for both the training and validation sets are closely aligned, suggesting that the model generalizes well without signs of overfitting.

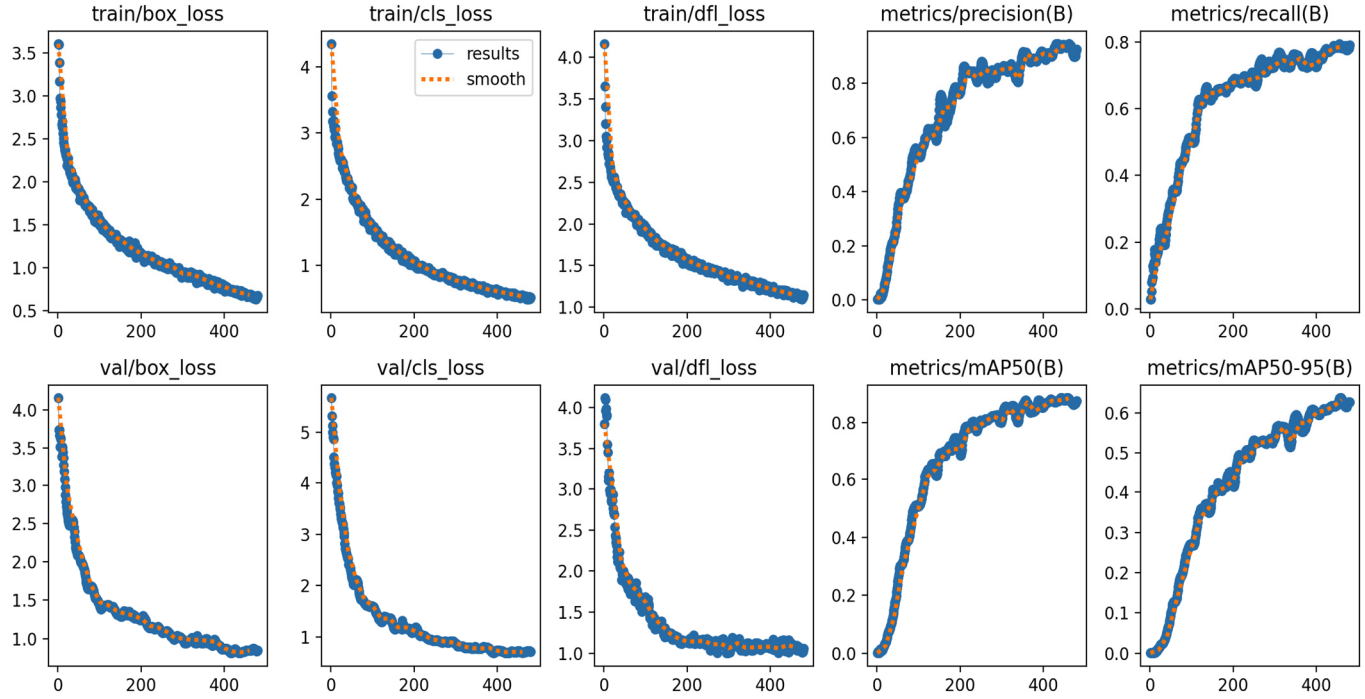

Figure 5. Experimental results of the YOLO11-AMF.

In parallel, the evaluation metrics namely precision, recall, mAP50, and mAP50–95 exhibit continuous improvement throughout the training process. These results indicate that the model is able to incrementally extract more representative and discriminative features from the augmented dataset, thereby enhancing its detection capability in complex traffic accident scenarios. Overall, the convergence of the loss curves alongside the steady increase in accuracy metrics confirms the model's robustness and effectiveness in handling the given task.

The quantitative findings are further substantiated by qualitative analysis. As visualized in Figure 6, the YOLO11-AMF model effectively detects traffic accidents across a range of challenging real-world scenarios, including nighttime environments, adverse weather conditions (e.g., rain), and complex road settings. In these instances, the model consistently yields high confidence scores (typically >0.70), reflecting a high degree of certainty in its predictions. Although minor variations in bounding box localization are occasionally observed, particularly in cases of partial occlusion or complex visual backgrounds, the

overall detection performance remains highly reliable. This visual evidence underscores the model's significant potential for practical application.

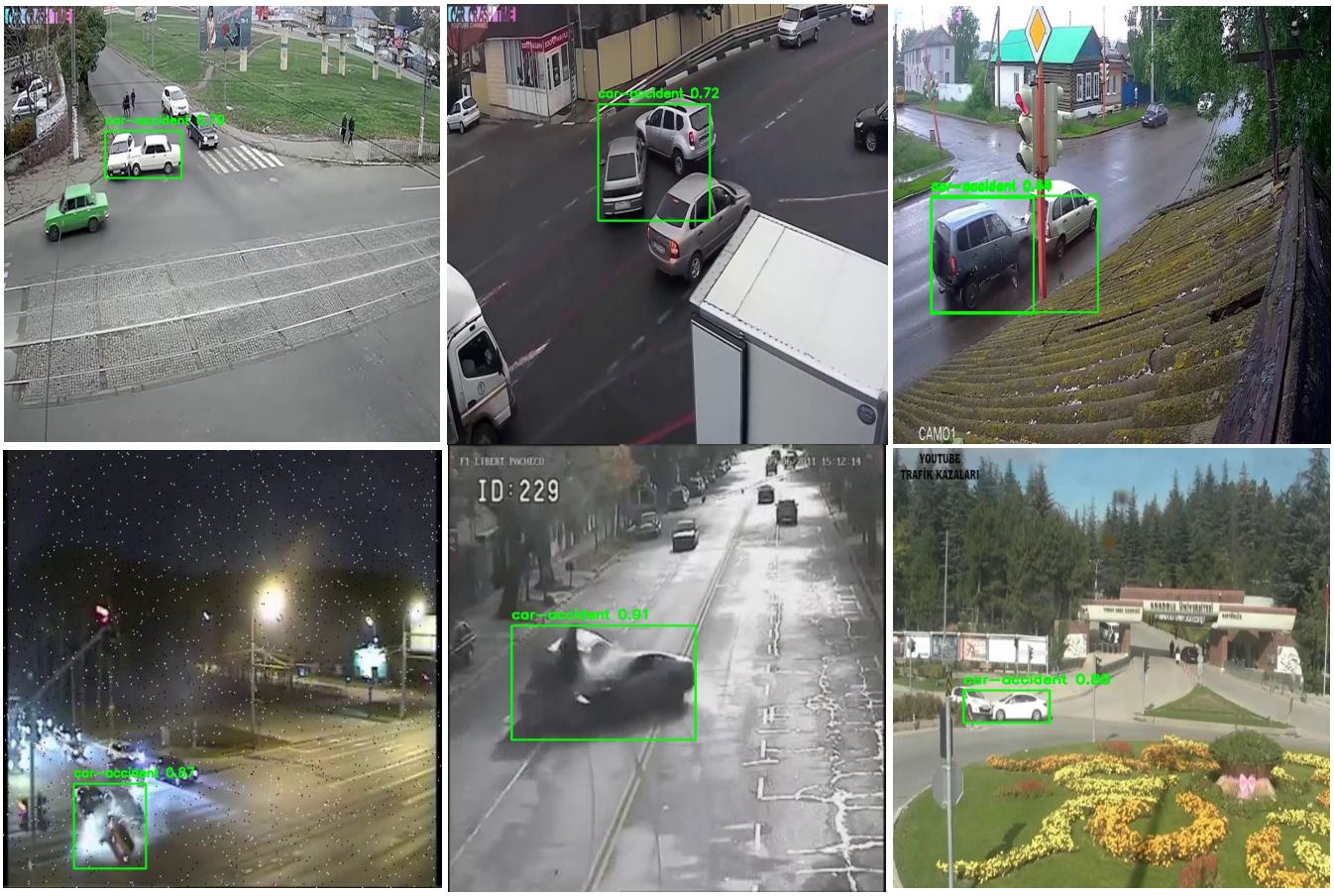

Figure 6. Qualitative detection results of the proposed YOLO-AMF model across diverse and challenging scenarios, including adverse weather, low-light conditions, and partial occlusions. The green bounding boxes with confidence scores demonstrate the model's robust and accurate performance in detecting traffic accidents in real-world environments.

While these qualitative results are promising, they also highlight the necessity for a more rigorous, quantitative benchmark to validate the performance of the YOLO11-AMF model. The occasional variations in localization, especially under challenging conditions, motivate a systematic comparison against other state-of-the-art models. Such an analysis is crucial for establishing a performance baseline which will guide future optimizations, including the exploration of advanced data augmentation and dataset expansion to enhance robustness and generalization.

4.2. Comparison of Different Models Experiment

To systematically evaluate performance, a comparative analysis was conducted on four object detection models: YOLOv5n [31], YOLOv8n [32,33], YOLO11 [34], and our proposed YOLO11-AMF. Each model was tested across three batch sizes (16, 32, and 64). Performance was evaluated using standard metrics, including precision, recall, F1-score, mean average precision at an IoU threshold of 0.5 (mAP50), and mean average precision across IoU thresholds from 0.5 to 0.95 (mAP50–95). To ensure a fair comparison, all models were trained under identical conditions, isolating the batch size as the sole variable to assess its impact on detection accuracy and model robustness.

The baseline models exhibited significant sensitivity to batch size variations. As presented in Table 4, YOLOv5n achieved its highest precision (94.9%) at a batch size of 16; however, its mAP50–95 degraded as the batch size increased, reaching a low of 56.9% at size 64. Similarly, YOLOv8n performed optimally at a batch size of 32, achieving a peak mAP50 of 86.7% and a notably high precision of 99.1%. Nevertheless, this was accompanied by a reduced recall of 68.6%, suggesting a trade-off which may lead to missed detections. Furthermore, its performance dropped substantially at a batch size of 64, indicating limitations in its scalability.

Table 4. Comparative performance of different models under varying batch sizes.

Algorithm	Batch Size	Precision/%	Recall/%	mAP50/%	mAP50–95/%	F1-Score/%
YOLOV5n	16	94.9	71.4	82.4	58.6	81.5
	32	84.1	80.0	83.8	63.6	82.0
	64	91.9	71.4	78.8	56.9	80.4
YOLOV8n	16	87.2	78.2	86.4	59.4	82.5
	32	99.1	68.6	86.7	64.1	81.0
	64	96.0	68.6	79.3	56.2	80.0
YOLO11	16	90.0	76.9	83.7	59.7	82.9
	32	92.5	74.6	89.1	61.8	82.6
	64	92.4	77.1	82.0	60.1	84.1
YOLO11-AMF	16	96.5	82.9	90.0	66.0	89.2
	32	90.6	77.1	89.0	60.8	83.3
	64	83.6	87.7	86.9	63.8	85.6

In contrast, the YOLO11 model displayed more stable and consistent performance across all batch sizes. It recorded its optimal results at a batch size of 32, achieving an mAP50 of 89.1% and an F1-score of 82.6%; Figure 5 reflects a well-balanced trade-off between precision and recall. These results establish YOLO11 as a reliable baseline, demonstrating consistent detection capabilities under varying training configurations.

The proposed YOLO11-AMF model consistently outperformed all other models across the majority of evaluation metrics. It achieved its peak performance at a batch size of 16, registering the highest scores among all configurations: 96.5% precision, 82.9% recall, 90.0% mAP50, 66.0% mAP50–95, and an F1-score of 89.2%. Critically, even at a larger batch size of 64, the model maintained robust performance, including a recall of 87.7% and an mAP50–95 of 63.8%. This demonstrates not only superior accuracy but also stronger generalization and resilience to changes in training parameters.

While the performance of all evaluated models exhibits sensitivity to batch size, the proposed YOLO11-AMF architecture consistently surpasses its counterparts in accuracy, scalability, and robustness. This sustained high performance across different training configurations validates the efficacy of its integrated architectural enhancements. Consequently, these findings confirm the model's suitability for deployment in demanding, high-stakes applications such as real-time traffic monitoring, where both precision and reliability are paramount.

4.3. Ablation Experiment

To isolate and quantify the contributions of the core components within the proposed YOLO11-AMF architecture, a systematic ablation study was conducted. The study involved progressively integrating three key modules, the Mamba-Like Linear Attention (MLLA) mechanism, the Adaptive Feature Pyramid Network (AFPN), and the Focaler-IoU loss function into the YOLO11 baseline. As detailed in Table 5, the performance of each

configuration was evaluated using standard metrics to measure the isolated and combined impact of these components on model accuracy and robustness.

Table 5. Ablation results of modules in YOLO11-AMF performance.

MLLA	AFPN	Focaler-Iou	Precision/%	Recall/%	mAP50/%	mAP50–95/%	F1-Score/%
-	-	-	89.9	76.9	83.6	59.7	82.9
√	-	-	87.9	77.1	83.1	58.4	82.2
-	√	-	76.7	80.0	82.2	56.6	78.3
-	-	√	91.0	71.4	83.4	57.8	80
√	√	-	68.9	76.1	79.5	52.5	72.4
√	-	√	92.7	74.3	84.1	55.0	82.5
-	√	√	97.0	68.6	83.1	60.9	80.3
√	√	√	96.5	82.9	90.0	66.0	89.2

The baseline YOLO11n model established a strong starting point with an 82.9% F1-score and 83.6% mAP50. The introduction of single modules revealed distinct trade-offs. The MLLA module alone caused a marginal performance decrease, suggesting that its feature refinement capabilities require synergistic integration with other components to be effective. While the AFPN module improved recall to 80.0%, this came at a significant cost to precision (76.7%), indicating that enhanced multi-scale fusion without refined feature selection can amplify noise. Conversely, the Focaler-IoU loss function substantially boosted precision to 91.0% at the expense of recall (71.4%), highlighting its effectiveness in penalizing low-quality predictions to reduce false positives.

The integration of dual modules revealed important interaction effects. The combination of MLLA and Focaler-IoU yielded a balanced performance uplift, achieving a 92.7% precision and an 84.1% mAP50, demonstrating a complementary relationship between attention-based feature refinement and a robust loss function. In contrast, the AFPN + Focaler-IoU pairing achieved the highest precision (97.0%) of any variant but suffered from poor recall (68.6%), resulting in a highly selective but less comprehensive detector. Notably, the MLLA + AFPN combination performed poorly across all metrics, confirming that, without a guiding loss function like Focaler-IoU, the combined complexity of these modules fails to converge effectively.

The full YOLO11-AMF model, integrating all three components, achieved the best overall performance, demonstrating a clear synergistic effect. This final configuration recorded the highest scores across all key metrics: 96.5% precision, 82.9% recall, 90.0% mAP50, 66.0% mAP50–95, and a top F1-score of 89.2%. This synergy allows the model to overcome the trade-offs observed in partial configurations: MLLA enhances feature representation, AFPN provides robust multi-scale fusion, and Focaler-IoU refines the optimization landscape. Collectively, these components create a state-of-the-art detector, which is both highly accurate and resilient in complex, real-world traffic environments.

5. Discussion

This study introduced and validated YOLO11-AMF, an enhanced object detection architecture, which demonstrates significant advancements in traffic accident detection. By integrating a Mamba-Like Linear Attention (MLLA) mechanism, an Adaptive Feature Pyramid Network (AFPN), and a Focaler-IoU loss function, our model achieves state-of-the-art performance. The experimental results confirm that this integrated approach delivers a final precision of 96.5%, recall of 82.9%, and an F1-score of 89.2%. These findings validate

the synergistic contributions of the proposed modules in enhancing feature representation and optimizing the loss landscape, resulting in a robust and highly accurate framework for identifying collisions in challenging real-world scenarios.

Despite its high accuracy, a primary limitation of the YOLO11-AMF model is the potential trade-off with real-time computational efficiency. The model's complexity, while integral to its performance, may impede deployment on resource-constrained edge devices where low latency is critical for immediate incident response. Therefore, future research will focus on bridging this gap. We plan to investigate model optimization techniques such as pruning, quantization, and knowledge distillation to develop a lightweight yet powerful version of the model, seeking an optimal balance between detection accuracy and inference speed for practical deployment.

A second limitation pertains to the dataset used for training and evaluation. While focusing on a specific dataset streamlined initial development, it inherently constrains the model's generalizability across a wider spectrum of accident types, vehicle models, and environmental conditions. This introduces a potential for sampling bias, where the model may be less effective on underrepresented scenarios. To address this, future work will prioritize the expansion and diversification of our training data. Furthermore, we will explore advanced generalization strategies, including leveraging transfer learning to adapt the model to new, smaller datasets of rare accident types, and employing domain adaptation techniques to enhance its portability across different geographical locations and camera systems. These efforts will be crucial for developing a truly universal and robust traffic accident detection system.

6. Conclusions

This study introduces YOLO11-AMF, an enhanced object detection framework designed for robust traffic accident recognition. The framework improves upon the YOLO11 baseline by integrating three key modules: a Mamba-Like Linear Attention (MLLA) mechanism to refine attention modeling, an Adaptive Feature Pyramid Network (AFPN) to strengthen multi-scale feature representation, and a Focaler-IoU loss function to improve bounding box regression. The efficacy of this integrated approach was validated through extensive experimentation. In comparative evaluations, YOLO11-AMF consistently outperformed baseline models, achieving a peak performance with 96.5% precision, 82.9% recall, and an 89.2% F1-score. Furthermore, a comprehensive ablation study confirmed the synergistic contributions of the MLLA, AFPN, and Focaler-IoU modules, demonstrating that the combined integration of these components is critical for achieving optimal performance.

In conclusion, the YOLO11-AMF framework establishes a new benchmark for accuracy and robustness in traffic accident detection, validating the proposed architectural modifications as an effective strategy for overcoming the challenges posed by complex, real-world visual environments. Building on these results, future research will pursue several key directions. Efforts will be directed toward optimizing the model for real-time deployment by exploring lightweight inference strategies. Concurrently, its generalization capabilities will be enhanced through the curation of more extensive and diverse datasets. These advancements will be crucial for broadening the model's applicability and ensuring its practical utility in intelligent transportation systems and other critical monitoring applications.

Author Contributions: Conceptualization, W.L.; methodology, W.L.; software, W.L.; validation, W.L. and L.H.; formal analysis, W.L.; investigation, W.L.; resources, W.L.; data curation, L.H.; writing original draft preparation, W.L.; writing review and editing, L.H.; visualization, L.H.; supervision, X.L.; project administration, X.L.; funding acquisition, L.H. All authors have read and agreed to the published version of the manuscript.

Funding: Funds for the Innovation of Policing Science and Technology, Fujian province (Grant number: 2024Y0072&2024Y0069).

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors utilized OpenAI's ChatGPT (GPT-4) to assist with improving the linguistic clarity, grammar, and consistency of the technical descriptions. The AI tool was strictly used for language enhancement purposes; all scientific concepts, experimental designs, data analyses, and interpretations were developed and validated by the authors. The authors acknowledge the use of OpenAI's ChatGPT (GPT-4, OpenAI, San Francisco, CA, USA) to assist with language editing and improving the clarity of the manuscript. The authors take full responsibility for the content, scientific accuracy, and conclusions presented in this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
2. Luo, J.; Li, Y.; Wei, L.; Nie, G. High-Precision Traffic Sign Detection and Recognition Using an Enhanced YOLOv5. *J. Circuits Syst. Comput.* **2025**, *34*, 2550118. [CrossRef]
3. Zhang, J.; Yi, Y.; Wang, Z.; Alqahtani, F.; Wang, J. Learning multi-layer interactive residual feature fusion network for real-time traffic sign detection with stage routing attention. *J. Real-Time Image Process* **2024**, *21*, 176. [CrossRef]
4. Liu, L.; Wang, L.; Ma, Z. Improved lightweight YOLOv5 based on ShuffleNet and its application on traffic signs detection. *PLoS ONE* **2024**, *19*, e0310269. [CrossRef]
5. Wang, Q.; Li, X.; Lu, M. An Improved Traffic Sign Detection and Recognition Deep Model Based on YOLOv5. *IEEE Access* **2023**, *11*, 54679–54691. [CrossRef]
6. Yan, H.; Pan, S.; Zhang, S.; Wu, F.; Hao, M. Sustainable utilization of road assets concerning obscured traffic signs recognition. *Proc. Inst. Civ. Eng.-Eng. Sustain.* **2024**, *178*, 124–134. [CrossRef]
7. Zhao, S.; Gong, Z.; Zhao, D. Traffic signs and markings recognition based on lightweight convolutional neural network. *Vis. Comput.* **2024**, *40*, 559–570. [CrossRef]
8. Wang, J.; Chen, Y.; Dong, Z.; Gao, M. Improved YOLOv5 network for real-time multi-scale traffic sign detection. *Neural Comput. Appl.* **2023**, *35*, 7853–7865. [CrossRef]
9. Qu, S.; Yang, X.; Zhou, H.; Xie, Y. Improved YOLOv5-based for small traffic sign detection under complex weather. *Sci. Rep.* **2023**, *13*, 16219. [CrossRef] [PubMed]
10. Elassy, M.; Al-Hattab, M.; Takruri, M.; Badawi, S. Intelligent transportation systems for sustainable smart cities. *Transp. Eng.* **2024**, *16*, 100252. [CrossRef]
11. Suganuma, N.; Yoneda, K. Current status and issues of traffic light recognition technology in Autonomous Driving System. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2022**, *105*, 763–769. [CrossRef]
12. Yang, L.; He, Z.; Zhao, X.; Fang, S.; Yuan, J.; He, Y.; Li, S.; Liu, S. A Deep Learning Method for Traffic Light Status Recognition. *J. Intell. Connect. Veh.* **2023**, *6*, 173–182. [CrossRef]
13. Li, Z.; Zhang, W.; Yang, X. An Enhanced Deep Learning Model for Obstacle and Traffic Light Detection Based on YOLOv5. *Electronics* **2023**, *12*, 2228. [CrossRef]
14. Zhou, Q.; Zhang, D.; Liu, H.; He, Y. KCS-YOLO: An Improved Algorithm for Traffic Light Detection under Low Visibility Conditions. *Machines* **2024**, *12*, 557. [CrossRef]
15. Li, K.; Wang, Y.; Hu, Z. Improved YOLOv7 for Small Object Detection Algorithm Based on Attention and Dynamic Convolution. *Appl. Sci.* **2023**, *13*, 9316. [CrossRef]
16. Kamble, S.J.; Kounte, M.R. Machine Learning Approach on Traffic Congestion Monitoring System in Internet of Vehicles. *Procedia Comput. Sci.* **2020**, *171*, 2235–2241. [CrossRef]
17. Ashraf, I.; Hur, S.; Kim, G.; Park, Y. Analyzing Performance of YOLOx for Detecting Vehicles in Bad Weather Conditions. *Sensors* **2024**, *24*, 522. [CrossRef] [PubMed]
18. Kumeda, B.; Fengli, Z.; Oluwasanmi, A.; Owusu, F.; Assefa, M.; Amenu, T. Vehicle Accident and Traffic Classification Using Deep Convolutional Neural Networks. In Proceedings of the 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, Chengdu, China, 14–15 December 2019; pp. 276–280.
19. Tamagusko, T.; Correia, M.G.; Huynh, M.A.; Ferreira, A. Deep Learning applied to Road Accident Detection with Transfer Learning and Synthetic Images. *Transp. Res. Procedia* **2022**, *64*, 90–97. [CrossRef]

20. Ghahremannezhad, H.; Shi, H.; Liu, C. Real-Time Accident Detection in Traffic Surveillance Using Deep Learning. In Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST), Kaohsiung, Taiwan, 15–17 October 2022; pp. 1–6.
21. Lin, C.; Hu, X.; Zhan, Y.; Hao, X. MobileNetV2 with Spatial Attention module for traffic congestion recognition in surveillance images. *Expert Syst. Appl.* **2024**, *255*, 14. [CrossRef]
22. Fang, J.; Qiao, J.; Xue, J.; Li, Z. Vision-Based Traffic Accident Detection and Anticipation: A Survey. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 1983–1999. [CrossRef]
23. Fu, Y. Combining Mamba and Attention-Based Neural Network for Electric Ground-Handling Vehicles Scheduling. *Systems* **2025**, *13*, 155.
24. Li, Z. Mamba with split-based pyramidal convolution and Kolmogorov-Arnold network-channel-spatial attention for electroencephalogram classification. *Front. Sens.* **2025**, *6*, 2673–5067. [CrossRef]
25. Yang, G.; Lei, J.; Tian, H.; Feng, Z. Asymptotic Feature Pyramid Network for Labeling Pixels and Regions. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 7820–7829. [CrossRef]
26. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. Augfpn: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12595–12604.
27. Zhang, Y.F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and Efficient IOU Loss for Accurate Bounding Box Regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
28. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Dollar, Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
29. Liu, T.; Meidani, H. End-to-end heterogeneous graph neural networks for traffic assignment. *Transp. Res. Part C Emerg. Technol.* **2024**, *165*, 104695. [CrossRef]
30. Drliciak, M.; Cingel, M.; Celko, J.; Panikova, Z. Research on Vehicle Congestion Group Identification for Evaluation of Traffic Flow Parameters. *Sustainability* **2024**, *16*, 1861. [CrossRef]
31. Zhao, Y.; Ju, Z.; Sun, T.; Dong, F.; Li, J.; Yang, R.; Fu, Q.; Lian, C.; Shan, P. TGC-YOLOv5: An Enhanced YOLOv5 Drone Detection Model Based on Transformer, GAM & CA Attention Mechanism. *Drones* **2023**, *7*, 446. [CrossRef]
32. Huang, Z.; Li, L.; Krizek, G.C.; Sun, L. Research on Traffic Sign Detection Based on Improved YOLOv8. *J. Comput. Commun.* **2023**, *11*, 226–232. [CrossRef]
33. Talaat, F.M.; ZainEldin, H. An improved fire detection approach based on YOLO-v8 for smart cities. *Neural Comput. Appl.* **2023**, *35*, 20939–20954. [CrossRef]
34. Huang, Y.; Wang, D.; Wu, B.; An, D. NST-YOLO11: ViT Merged Model with Neuron Attention for Arbitrary-Oriented Ship Detection in SAR Images. *Remote Sens.* **2024**, *16*, 4760. [CrossRef]