

Article

Lightweight Vehicle Damage Detection Using GSConv-Based Slim-Neck and Bi-Level Routing Attention

Liyan Huang ¹, Xiaofeng Lai ², Peiteng Lin ¹ and Weijun Li ^{1,*}¹ Department of Forensic Science, Fujian Police College, Fuzhou 350007, China; huangliyan@fjpsc.edu.cn (L.H.); linpeiteng@outlook.com (P.L.)² Public Security Bureau, Sanming 365000, China; aleidad454@gmail.com

* Correspondence: xjlwj@fjpsc.edu.cn

Abstract

Vehicle damage detection is an important task in intelligent transportation systems and insurance assessment, yet it remains challenging due to the subtle appearance, irregular shapes, and spatial dispersion of damage regions in complex environments. We propose a specialized structural synergy that organically integrates a GSConv-based Slim-Neck, a dynamic Bi-Level Routing Attention mechanism, and an orientation-aware SIOU loss. Rather than a superficial architectural combination, this cooperative design introduces a novel methodological framework engineered specifically to resolve the fundamental conflict between edge-deployment efficiency and fine-grained feature preservation in vehicle inspection. The method is evaluated on the publicly available Car Damage Detection dataset and compared with representative two-stage and one-stage detectors, including DETR, Faster R-CNN, YOLOv5n, YOLOv8n, and YOLO11n. Experimental results show that the proposed approach achieves a mAP50 of 67.9% and mAP50–95 of 53.8%, outperforming the baseline YOLO11n and other lightweight YOLO variants with only a moderate increase in computational cost. These results indicate that the proposed framework offers a favorable trade-off between detection accuracy and efficiency, showing potential for vehicle damage inspection under resource-constrained conditions.

Keywords: intelligent transportation systems; object detection; vehicle damage detection; vehicle inspection



Academic Editor: Michael Fowler

Received: 22 April 2026

Revised: 27 May 2026

Accepted: 28 May 2026

Published: 29 May 2026

Copyright: © 2026 by the authors.

Published by MDPI on behalf of the World Electric Vehicle Association.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Vehicle damage detection plays a critical role in intelligent transportation systems, insurance claim assessment, and vehicle maintenance services. With the continuous growth of the global automobile fleet, traffic accidents and wear-related damage have become increasingly common, leading to a rising demand for accurate, efficient, and automated vehicle inspection solutions [1]. Traditional manual inspection approaches are often time-consuming, subjective, and highly dependent on expert experience, which limits their scalability and consistency in practical applications. Consequently, computer vision-based methods have attracted considerable attention due to their potential to provide fast, objective, and reliable vehicle damage detection [2].

In recent years, deep learning techniques, particularly convolutional neural networks, have achieved remarkable success in object detection tasks. Among these approaches, the You Only Look Once series has been widely adopted because of its favorable balance between detection accuracy and real-time inference speed [3]. By employing end-to-end

detection pipelines and multi-scale feature learning strategies, YOLO-based frameworks have demonstrated strong performance in various application domains, including traffic sign recognition, road anomaly detection, and vehicle perception tasks. Nevertheless, vehicle damage detection remains challenging due to the subtle visual characteristics, irregular shapes, and diverse spatial distributions of damage types such as scratches, dents, and collision-induced deformations [4]. These damaged regions often exhibit strong visual similarity to background textures, particularly under complex illumination conditions or in cluttered environments.

Despite the strong baseline performance of existing YOLO-based detectors, several limitations restrict their effectiveness in vehicle damage detection [5]. First, conventional neck structures for multi-scale feature fusion, such as feature pyramid networks and path aggregation networks, often introduce feature redundancy and increased computational burden, which is unfavorable for lightweight deployment scenarios. Second, the interaction among features across different spatial regions and scales is often insufficient, limiting the ability to capture global contextual information. This limitation becomes more pronounced when detecting fine-grained and spatially scattered damage patterns, where long-range dependencies and selective attention to informative regions may play an important role [6]. Standard YOLO models frequently experience performance degradation when detecting small, irregular, or low-contrast vehicle damages due to their inherent structural design limitations. Specifically, the aggressive spatial downsampling mechanisms employed by generic detectors to compress computational overhead cause a rapid loss of high-frequency geometric information, which practically eliminates the subtle feature footprints of small cracks in deeper feature layers. Furthermore, irregular and elongated damage zones do not conform to uniform rectangular receptive fields, leading to significant feature dilution and localized misalignment. For low-contrast anomalies, conventional convolutional layers apply uniform feature aggregation across dense matrices, which lacks the capacity to dynamically isolate subtle foreground paint deformations from dominant background contexts such as specular reflections on glossy vehicle bodies. As a result, there is a clear need for lightweight detection frameworks that simultaneously reduce computational complexity and enhance feature representation capability.

To address these challenges, this paper proposes a lightweight vehicle damage detection framework that integrates a GSConv-based Slim-Neck architecture with bi-level routing attention. The Slim-Neck design leverages GSConv to improve channel mixing efficiency while significantly reducing computational cost, and incorporates VoVGSCSP modules to mitigate feature redundancy and facilitate gradient propagation during multi-scale feature fusion [7,8]. In addition, bi-level routing attention is introduced to strengthen contextual representation by selectively modeling interactions at both regional and token levels. This attention mechanism allows the network to place greater emphasis on informative damage-related regions while reducing the influence of irrelevant background information, thereby helping improve robustness to subtle and irregular damage patterns [9,10].

The proposed method is evaluated on the publicly available Car Damage Detection dataset to ensure reproducibility and enable fair comparison with existing approaches. Experimental results demonstrate that the proposed framework achieves improved detection performance while maintaining a lightweight model design, making it suitable for real-world vehicle damage inspection scenarios where both accuracy and computational efficiency are required.

The main contributions of this work can be summarized as follows:

A lightweight Slim-Neck architecture based on GSConv and VoVGSCSP is designed to enhance multi-scale feature fusion efficiency while effectively reducing computational redundancy.

Bi-level routing attention is integrated into the detection framework to further refine contextual information and improve sensitivity to fine-grained vehicle damage patterns.

Extensive experiments conducted on the Car Damage Detection dataset validate the effectiveness of the proposed method and demonstrate a favorable balance between detection accuracy and computational efficiency.

The remainder of this paper is structured as follows. Section 2 reviews the related literature on intelligent vehicle damage detection and efficient deep learning architectures. Section 3 delineates the proposed methodology, detailing the targeted structural synergy of the GSConv-based Slim-Neck, the Bi-Level Routing Attention mechanism, and the orientation-aware loss function. Section 4 presents the comprehensive experimental evaluations, quantitative comparisons, ablation studies, and category-wise performance analyses. Section 5 concludes the paper and outlines prospective directions for future research.

2. Related Works

The application of deep learning to vehicle damage detection has gained increasing attention in recent years, driven primarily by the growing demand for automated insurance claim assessment and intelligent vehicle inspection systems. Recent studies have demonstrated that convolutional neural network-based detection frameworks can effectively replace manual inspection processes by improving inspection efficiency and reducing subjectivity, thereby enhancing scalability and consistency in practical applications [11]. With continuous advancements in deep learning, vehicle damage detection systems have shown promising performance and robustness in real-world deployment scenarios.

This systematic taxonomy provided in Table 1 consolidates the baseline trajectories evaluated across contemporary literature. The tabulated comparison reveals that prior architectural adjustments either introduce immense parameter inflation or fail to represent fine-grained surface defect variations. The proposed framework structurally distinguishes itself by coupling an efficient GSConv-based Slim-Neck feature purifier with a dynamic Bi-Level Routing Attention mechanism.

Table 1. Summary of different object detection paradigms.

Methodology	Core Mechanism	Key Limitation	Efficiency
FRCNN	Two-stage proposal	Low inference speed	High memory overhead
DETR	Global Transformer	High baseline complexity	Intensive computing demand
Traditional YOLO	Uniform convolution	Info loss of small defects	Prone to degradation
OURS	Sparse routing	Affected by intense reflections	Lightweight

Object detection architectures for vehicle damage detection can generally be divided into two-stage and one-stage paradigms. Recent comparative studies indicate that two-stage detectors, such as Faster R-CNN variants, remain competitive in accuracy-oriented tasks but suffer from relatively low inference efficiency, which limits their applicability in real-time inspection systems [12]. In contrast, one-stage detectors, including YOLO-based and RetinaNet-based models, achieve faster inference while maintaining competitive detection accuracy, making them more suitable for online vehicle damage inspection and edge deployment scenarios [13]. Consequently, recent research has increasingly focused on improving the accuracy–efficiency trade-off of one-stage detectors.

Among one-stage detectors, the YOLO family has undergone rapid evolution in recent years. YOLOv8 and its successors introduce improved backbone and neck designs to

enhance feature representation and multi-scale fusion efficiency [14]. More recent versions further integrate advanced attention mechanisms and optimized convolutional structures to improve sensitivity to small and irregular targets while reducing parameter redundancy [15]. These improvements are particularly relevant for vehicle damage detection, where damage regions often exhibit subtle visual cues and strong background interference. Recent attention-enhanced YOLO-based approaches have demonstrated notable performance gains in detecting fine-grained vehicle damage patterns under complex imaging conditions [16].

In recent years, the literature on vehicle damage detection has increasingly focused on the deployment of lightweight YOLO-based frameworks and the integration of advanced attention mechanisms to capture irregular defect geometries [17]. Various efficient architectures have attempted to optimize parameter efficiency by replacing standard convolutions with depthwise separable operations or applying localized attention blocks to capture surface anomalies [18]. While these recent attention-enhanced variants improve detection performance for prominent damages, they frequently suffer from severe computational inflation when processing high-resolution deployment inputs [19], or they experience a dramatic drop in fine-grained feature representation due to over-simplified lightweighting strategies [20]. The proposed framework directly addresses this gap in the current literature landscape by coupling an efficient GSConv-based Slim-Neck feature purifier with a dynamic Bi-Level Routing Attention mechanism. This structural alignment achieves a sparse routing paradigm that captures long-range defect dependencies without the computational penalties typical of recent efficient damage detection architectures. This systematic taxonomy provided in Table 1 consolidates the baseline trajectories evaluated across contemporary literature. The tabulated comparison reveals that prior architectural adjustments either introduce immense parameter inflation or fail to represent fine-grained surface defect variations. The proposed framework structurally distinguishes itself by coupling an efficient GSConv-based Slim-Neck feature purifier with a dynamic Bi-Level Routing Attention mechanism.

3. Materials and Methods

3.1. Dataset Construction

The proposed method was evaluated using the CarDD dataset, a publicly available large-scale benchmark specifically designed for vision-based vehicle damage detection and segmentation. The dataset contains approximately 4000 high-resolution images of damaged vehicles, with more than 9000 accurately annotated instances covering six common categories of external damage, including dent, scratch, crack, glass shatter, tire flat, and lamp broken. Detailed information regarding dataset construction and statistical analysis is provided in the original dataset publication.

The images in CarDD were collected under real-world conditions and exhibit substantial diversity in viewpoints, illumination, and background complexity, making the dataset representative of practical vehicle inspection scenarios. The annotated damage instances vary significantly in scale, shape, and visual appearance, which introduces additional challenges for accurate detection, particularly for small and low-contrast damage regions. Furthermore, the dataset supports multiple computer vision tasks, such as object detection and instance segmentation, highlighting its suitability for comprehensive evaluation of vehicle damage analysis methods.

To ensure consistency with the detection framework adopted in this study, all images were resized to a fixed resolution of 640×640 pixels prior to training and evaluation. The dataset was randomly divided into training and validation subsets while preserving the original class distribution. This data partitioning strategy ensures balanced representation

of damage categories and enables fair and reproducible performance comparison. Overall, the CarDD dataset provides a standardized and publicly accessible benchmark for evaluating vehicle damage detection methods in realistic application scenarios.

3.2. Proposed YOLO11-Based Detection Framework

The proposed detection framework is built upon the YOLO11 architecture and adopts a classical three-stage design consisting of a backbone, a neck, and multi-scale detection heads. The backbone follows the standard YOLO11 configuration and is composed of stacked convolutional layers and C3k2 blocks, followed by a spatial pyramid pooling fast module and a C2PSA block at the deepest stage. This architecture enables hierarchical feature extraction through progressive downsampling while preserving essential spatial information via partial feature aggregation and spatial attention mechanisms. As a result, the backbone is able to capture both fine-grained texture details and high-level semantic representations, which are important for modeling subtle vehicle damage patterns such as fine scratches and minor dents.

Building upon the features generated by the backbone, a staged Slim-Neck structure is employed to perform structured multi-scale feature fusion. As illustrated in Figure 1, high-level features are progressively upsampled and fused with corresponding lower-level features, followed by VoVGSCSP blocks to enhance feature reuse and facilitate gradient propagation. Lightweight GSConv layers are selectively integrated into the downsampling paths to reduce computational redundancy while preserving effective channel interaction [21]. This stage-wise fusion strategy allows shallow, high-resolution features to retain detailed spatial cues that are helpful for detecting small or low-visibility damage regions, while deeper features provide stronger semantic consistency and contextual information.

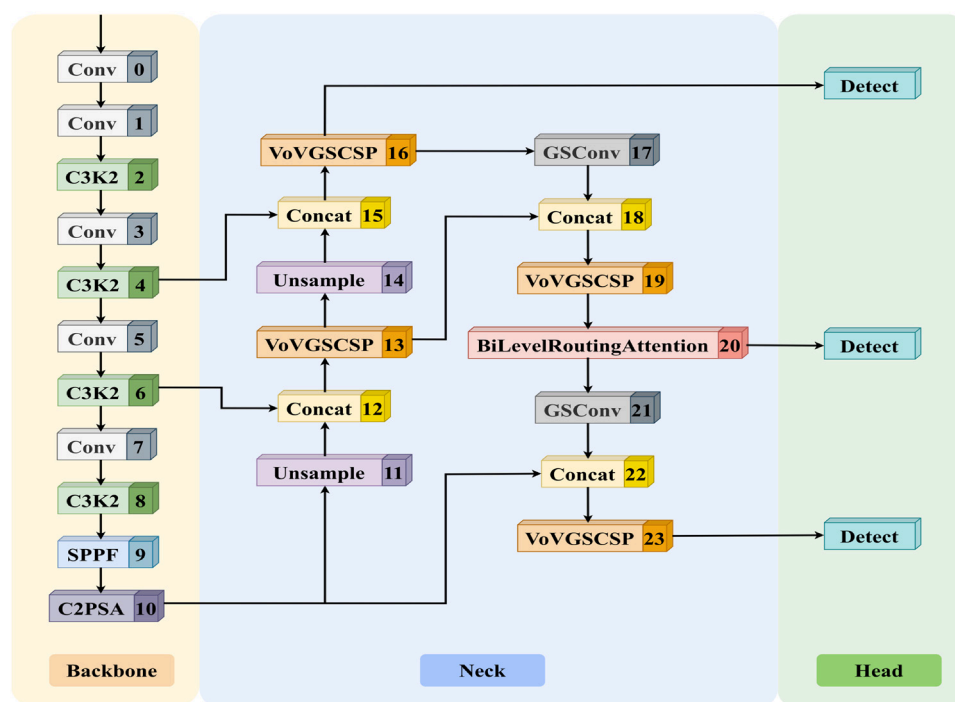


Figure 1. Overall architecture of the proposed YOLO11-based vehicle damage detection framework, illustrating the backbone, staged Slim-Neck with bi-level routing attention, and multi-scale detection heads for efficient and accurate damage localization.

To further enhance feature representation, a bi-level routing attention module is integrated into the medium-scale branch of the neck to refine fused features by selectively modeling semantically relevant regions and performing fine-grained interactions among

tokens within these regions [22]. This hierarchical attention mechanism strengthens the interaction between local and global features, helping reduce the influence of background interference such as reflections and shadows while highlighting irregular or spatially dispersed damage patterns. The refined multi-scale features are subsequently delivered to parallel detection heads at the P3, P4, and P5 levels, enabling simultaneous detection of damage instances across different spatial scales.

To improve localization robustness, the framework adopts Scylla Intersection over Union as the bounding box regression objective. This loss function introduces angle-aware and shape-consistent geometric constraints, which may help improve optimization under perspective variations and complex vehicle geometries [23]. Together, these components constitute a coherent and efficient enhancement of the YOLO11 architecture, supporting accurate and reliable vehicle damage detection in real-world automotive inspection scenarios.

3.2.1. GSConv-Based Slim-Neck

To improve the efficiency of multi-scale feature fusion while preserving sufficient representational capacity, a staged Slim-Neck is introduced into the YOLO11-based framework [24]. This design is motivated by the Slim-Neck concept, which seeks to reduce redundancy in feature aggregation by selectively replacing computationally expensive convolutional operations in the neck with lightweight alternatives, while keeping the backbone unchanged to retain its strong feature extraction capability.

Unlike conventional feature pyramid networks that apply uniform convolutional operations across all fusion stages, the proposed staged Slim-Neck adopts a stage-wise optimization strategy. As illustrated in Figure 2b, high-resolution stages that are sensitive to spatial detail retain structured aggregation blocks based on VoVGSCSP, which integrate GS bottlenecks within a cross-stage partial topology to maintain effective feature interaction and stable gradient propagation. These stages play a critical role in capturing fine-grained visual cues associated with small or low-contrast damage regions. In contrast, deeper fusion stages employ GSConv-based lightweight operations, as shown in Figure 2a, to reduce computational overhead while preserving sufficient semantic representation for medium- and large-scale damage patterns.

The efficiency of the proposed neck design is largely attributed to the internal structure of GSConv, which is illustrated in Figure 3. GSConv integrates standard convolution and depthwise separable convolution through parallel branches, followed by feature concatenation and channel shuffling. This hybrid structure enables effective mixing of spatial and channel-wise information, preserving feature diversity more effectively than purely depthwise separable convolution while substantially reducing computational complexity [25]. Building upon GSConv, the GS bottleneck further incorporates shortcut connections to facilitate stable feature propagation. When embedded within the VoVGSCSP structure, the GS bottleneck allows the network to achieve a balanced trade-off between representational capacity and computational efficiency in a modular and scalable manner.

By selectively embedding GSConv and GS bottleneck structures at appropriate stages of the neck, the proposed staged Slim-Neck mitigates the representational limitations that lightweight convolutions may introduce during early feature extraction, while fully leveraging their efficiency in multi-scale feature fusion.

This design aligns computational complexity with different levels of feature abstraction, enabling the network to preserve fine-grained damage cues at shallow scales while maintaining robust semantic consistency at deeper [26]. Such a structured fusion strategy is particularly well suited for vehicle damage detection tasks, in which small defects require high-resolution spatial detail, whereas larger damage regions benefit from stable and efficient semantic aggregation.

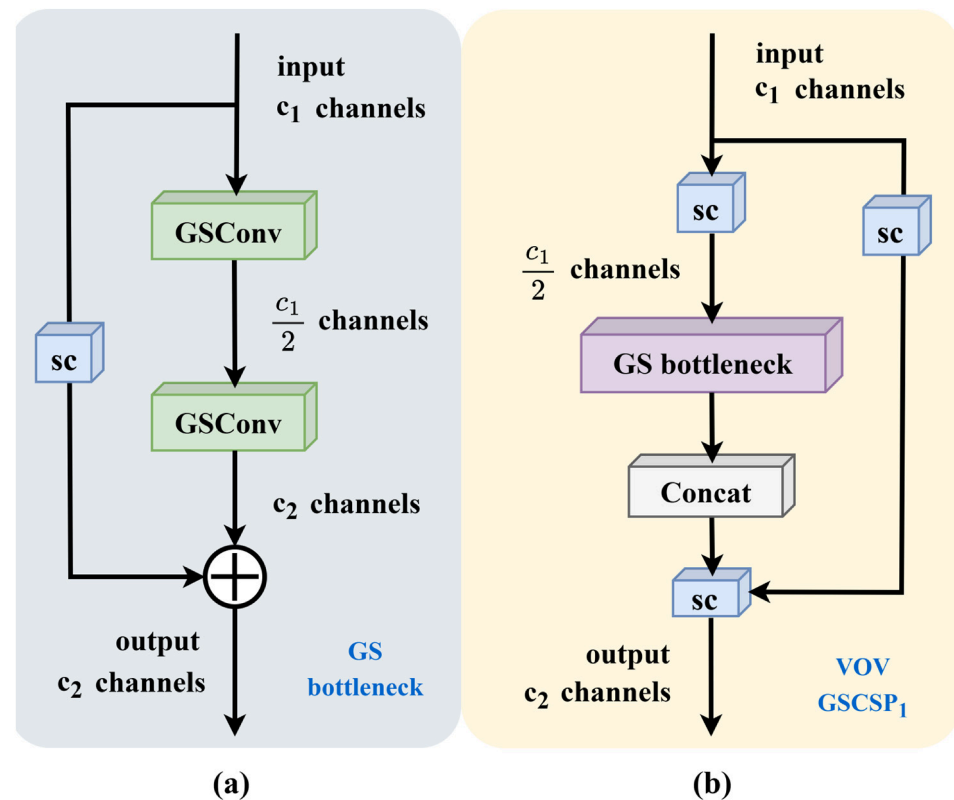


Figure 2. Structural diagrams of the GS bottleneck and the VoV-GSCSP module: (a) GS bottleneck and (b) VOV-GSCSP module.

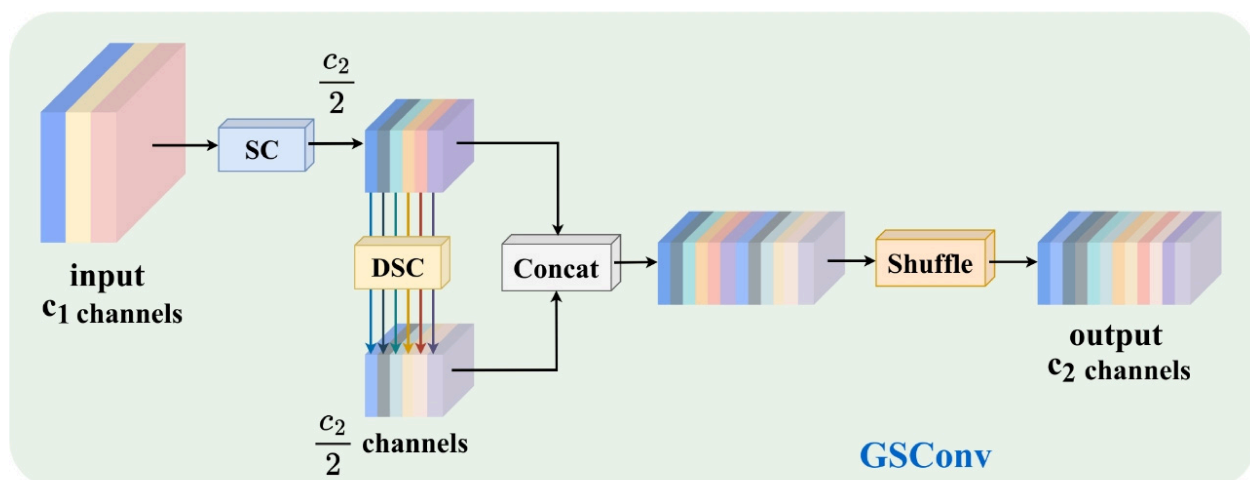


Figure 3. Structural illustration of the GSConv module for efficient channel mixing and feature aggregation.

3.2.2. Bi-Level Routing Attention

The core concept of bi-level routing attention is to decompose the attention computation into two complementary stages, namely region-level routing and token-level refinement, as illustrated in Figure 4. In the first stage, the input feature map is partitioned into non-overlapping spatial regions using a fixed stride. A gather operation is subsequently applied to aggregate features within each region, resulting in compact region-level representations [27]. These aggregated representations are employed to construct global key and value matrices, which encode coarse-grained semantic information across different spatial regions. In parallel, query features are generated directly from the original feature map at the token level, thereby preserving fine-grained spatial details.

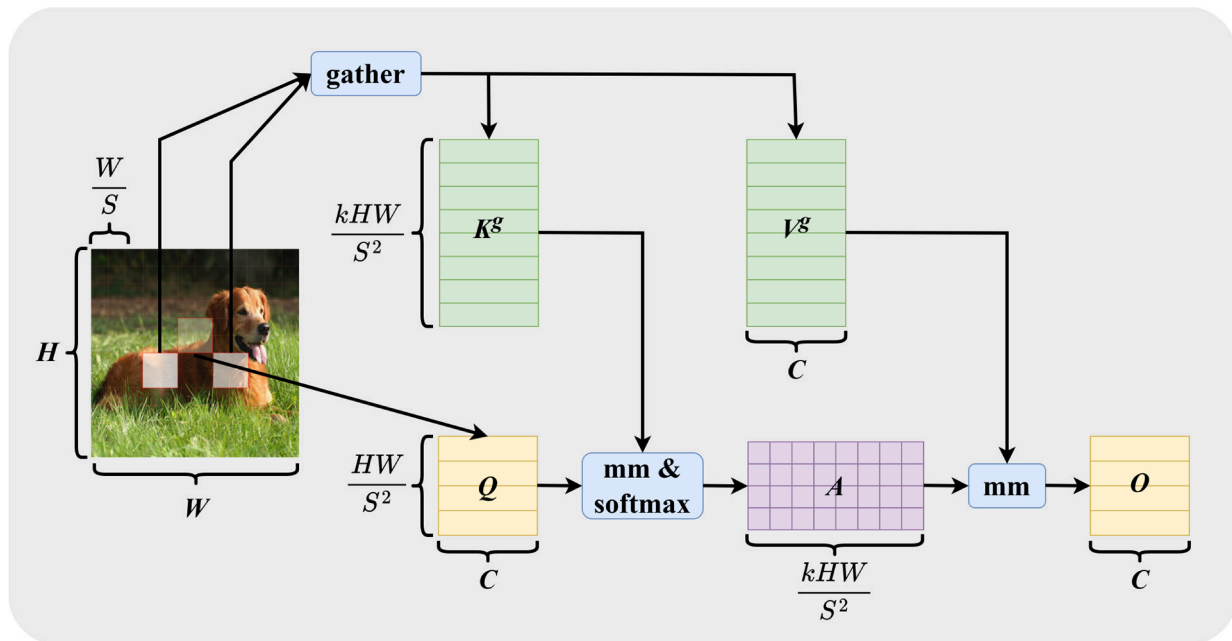


Figure 4. Illustration of the bi-level routing attention mechanism.

Based on the similarity between region-level query representations and the global keys, a routing function dynamically selects a small subset of semantically relevant regions for each query region [28], which can be expressed as:

$$R(q) = \text{Top} - K(\text{sim}(r_q, r_k)) \quad (1)$$

where r_q and r_k denote region-level feature embeddings and $\text{sim}(\cdot)$ represents a similarity measure. This routing mechanism allows the model to focus attention on informative global context while suppressing irrelevant background regions.

As shown in Figure 4, this routing step significantly reduces the attention search space by focusing computation on a limited number of informative regions, while suppressing irrelevant background areas.

In the second stage, fine-grained attention is performed exclusively within the routed regions. Specifically, the query features Q interact with the selected global keys through matrix multiplication followed by normalization, yielding an attention map A . The refined output features are then obtained by aggregating the corresponding value features:

$$y_i = \sum_{j \in R(q)} \alpha_{ij} v_j \quad (2)$$

where α_{ij} denotes the normalized attention weight and v_j represents the value features. The normalized attention weight α_{ij} is formally calculated via a token-to-token exponential expansion matrix as follows:

$$\alpha_{ij} = \frac{\exp\left(\frac{q_i k_j^T}{\sqrt{d}}\right)}{\sum_{m \in R(q)} \exp\left(\frac{q_i k_m^T}{\sqrt{d}}\right)} \quad (3)$$

This localized aggregation mechanism, as visualized in Figure 4, enables precise token-level refinement while avoiding the quadratic complexity of dense global self-attention.

Within the proposed YOLO11-based architecture, the bi-level routing attention module is integrated into the medium-scale feature branch of the neck. At this stage, multi-scale

features have already been aggregated through the staged Slim-Neck, and the attention module further refines these representations by enhancing the interaction between local details and global context [29]. By selectively emphasizing damage-relevant regions and suppressing background interference such as reflections and shadows, the attention mechanism improves the discriminative quality of the fused features.

The normalization protocol and computational complexity of the bi-level routing attention block are explicitly detailed within the text narrative. Layer Normalization is implemented immediately before the linear projections to stabilize internal gradient distributions. For an input feature map entering the attention block, the tensor dimension initiates as a standard four-dimensional spatial layout consisting of batch size, height, width, and channel depth. This structure is subsequently flattened into a two-dimensional sequence matrix defined by batch size and total token length before being projected into the channel spaces. During the regional partitioning phase, the tensor is further reshaped to isolate the non-overlapping spatial regions for coarse-grained routing. The theoretical computational complexity of this sparse routing mechanism scales linearly with respect to the channel dimension and retains a strictly bounded quadratic relationship with the token sequence length. This design significantly reduces the operational overhead compared with traditional dense global self-attention mechanisms, ensuring high computational efficiency.

3.2.3. SIOU-Based Localization Loss

Accurate localization is essential for vehicle damage detection, as damage regions often exhibit irregular shapes, elongated structures, and appear under oblique viewing angles. Conventional IoU-based localization losses, including IoU, GIoU, DIoU, and CIoU, mainly emphasize overlap area, center distance, and aspect ratio consistency [30]. Although effective for general object detection, these losses provide limited guidance when predicted and ground-truth bounding boxes are spatially close but geometrically misaligned, a situation frequently encountered in vehicle damage scenarios.

To address this issue, the proposed framework adopts the Scylla Intersection over Union loss for bounding box regression. SIOU extends traditional IoU formulations by incorporating geometric constraints that jointly consider overlap, center distance, directional alignment, and shape consistency [31]. In particular, SIOU introduces an angle-aware optimization strategy that models the relative orientation between the center displacement vector of the predicted and ground-truth boxes. By leveraging directional information, the regression process is guided along a more efficient geometric path, which improves convergence behavior when bounding boxes partially overlap or exhibit asymmetric alignment.

The computation of the angle is defined as follows:

$$\alpha = \arcsin \left(\frac{|y_2 - y_1|}{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + \varepsilon}} \right) \quad (4)$$

where (x_1, y_1) and (x_2, y_2) represent the centers of the predicted and ground-truth boxes. A smaller α value indicates closer alignment with the principal axis, reflecting enhanced orientation consistency. To complement this angle modeling, SIOU incorporates distance and shape constraints. Specifically, the distance term adaptively reweights center-point deviations based on the angular error, thereby imposing more stringent penalties on misaligned predictions. Concurrently, the shape constraint enforces structural consistency between the dimensions of the predicted and ground-truth bounding boxes, which is defined as follows:

$$shape_cost = \left(1 - e^{-\omega_\omega^\theta}\right) + \left(1 - e^{-\omega_h^\theta}\right) \quad (5)$$

where ω_w and ω_h represent normalized differences in width and height, and θ is a scaling factor. The final SIOU loss function extends the conventional IoU penalty by incorporating orientation-aware and shape-aware regularization terms, formulated as:

$$L_{SIOU} = 1 - IoU + 0.5 \cdot (\delta + shape_cost) \quad (6)$$

where δ denotes the angle-guided distance cost.

The structural integration of the GSConv-based Slim-Neck, Bi-Level Routing Attention, and SIOU loss within the YOLO11 backbone is not an arbitrary combination of previous components, but a targeted architectural synergy specifically engineered to address the unique physical challenges of vehicle damage detection. Generic lightweight YOLO variants often reduce parameters at the cost of fine-grained feature degradation, leading to high missed detections of subtle defects under complex ambient backgrounds. In our framework, the Slim-Neck serves as an efficient feature purifier that compresses redundant channels while preserving high-frequency structural details. This purified representation directly facilitates the Bi-Level Routing Attention module to execute dynamic sparse routing mechanisms, allowing the network to capture long-range contextual dependencies of highly irregular damage zones without being distracted by specular reflections on glossy car bodies. Additionally, the SIOU loss introduces a direction-aware vector alignment that directly accelerates the boundary convergence of elongated or tilted defects such as long scratches. This organic coupling effectively resolves the inherent conflict between deployment efficiency and fine-grained representation capacity, distinguishing OURS from general-purpose detection frameworks.

To ensure the reproducibility of the proposed framework, the key implementation parameters of the Bi-Level Routing Attention module are explicitly defined. Specifically, the input feature map is first divided into 7 by 7 non-overlapping routing regions. The top-k selection strategy employs a fixed value of $k = 4$, meaning that each routing region only attends to the top 4 most relevant key-value regions, which effectively filters out irrelevant background noise and limits the computational overhead. In terms of architectural integration, the attention module is strategically positioned at the output feature layers of the Slim-Neck immediately prior to the detection heads. This specific integration position was quantitatively determined through pilot comparative experiments, which revealed that embedding the module at the deep feature fusion stage provides the optimal balance between multi-scale contextual aggregation and parameter efficiency, compared with embedding it within the backbone layers.

3.2.4. Evaluation Indicators

A comprehensive evaluation of vehicle damage detection performance requires not only conventional object detection metrics but also careful consideration of robustness and practical applicability. Given that vehicle damage detection is closely associated with safety-critical scenarios such as insurance assessment, intelligent transportation monitoring, and vehicle maintenance, the evaluation framework should effectively reflect the model's ability to achieve high detection accuracy, reduce false alarms, avoid missed detections, and maintain reliable performance in real-world conditions.

Among the fundamental evaluation indicators, precision (P) and recall (R) are widely adopted to assess detection reliability and completeness. Precision measures the proportion of correctly detected damage instances among all predicted damage regions, as defined in Equation (3), where TP and FP denote the numbers of true positives and false positives, respectively. A higher precision value indicates that the detector is less likely to misclassify undamaged regions as damage, which is essential for preventing unnecessary interventions. In contrast, recall evaluates the proportion of actual damage instances that are successfully

detected, as expressed in Equation (4), where FN represents false negatives [32]. Recall is particularly important in vehicle damage detection, as missed detections may lead to incomplete inspections, underestimated repair costs, or potential safety risks.

To jointly evaluate detection accuracy across different confidence thresholds, the average precision (AP) is defined as the area under the precision–recall curve, as shown in Equation (5), where precision is treated as a function of recall. For multi-class vehicle damage detection tasks involving categories such as collision-related damage, dents, and scratches, the mean average precision (mAP) is employed, as defined in Equation (6),

$$P = \frac{TP}{TP + FP} \quad (7)$$

where denotes the total number of damage categories. Together, AP and mAP provide standardized and comprehensive metrics for assessing both localization and classification performance across diverse damage types.

$$R = \frac{TP}{TP + FN} \quad (8)$$

where FN represents false negatives, recall measures the proportion of actual vehicle damage instances successfully identified by the model; this metric is uniquely vital in automotive defect assessment because missed detections—such as overlooking deep dents or structural compromises—can result in incomplete appraisals, undervalued repair estimates, or severe post-repair safety hazards if critical damage goes unnoticed.

To unify these complementary measures, the average precision (AP) is calculated as the area integrated under the precision–recall curve:

$$AP = \int_0^1 P(r)dr \quad (9)$$

where $P(r)$ denotes precision as a function of recall. When dealing with multiple vehicle damage categories including accident-related structural damage, dents and scratches, the mean average precision (mAP) is utilized:

$$mAP = \frac{1}{n} \sum_{n=1}^n AP_n \quad (10)$$

With n representing the total number of categories. Collectively, AP and mAP provide a rigorous and standardized evaluation of both detection and classification accuracy, offering a comprehensive perspective on the model's performance across diverse damage types.

4. Results

4.1. Training and Evaluation Setup

To ensure fair and reproducible evaluation, experiments were conducted on the publicly available Car Damage Detection dataset, where the entire benchmark is partitioned into training, validation, and testing sets with a standard ratio of 80 percent, 10 percent, and 10 percent, respectively, yielding an 8 to 1 ratio between the 3427 training images and 403 validation images. These samples cover common vehicle damage categories such as dents, scratches, cracks, and accident-related deformations. During the training phase, online data augmentation techniques are comprehensively deployed to enhance reproducibility and model robustness. Specifically, the augmentation pipeline incorporates mosaic augmentation with a probability of 1.0 to construct complex multi-object scenes, mixup augmentation with a probability of 0.15 to blend structural boundaries, random horizontal flips with a probability of 0.5, and random hue-saturation-value color space

transformations to simulate diverse outdoor illumination environments, supporting consistent training across diverse visual conditions.

To evaluate the capability of the framework under class imbalance, the class-wise sample distribution of the CarDD dataset is explicitly examined. The dataset exhibits a typical long-tailed distribution, where prominent categories such as scratches and dents possess thousands of instances, whereas fine-grained categories such as cracks and tire damage contain significantly fewer samples. This inherent imbalance is addressed within our training pipeline through the application of advanced data augmentation to artificially increase the exposure of rare defects, combined with the orientation-aware loss function that penalizes localization errors more strictly on complex bounding boxes, preventing the majority classes from dominating the gradient updates.

All experiments were performed in a high-performance cloud computing environment equipped with an NVIDIA A100 GPU with 40 GB of memory. The computational setup provided sufficient resources to support model training and evaluation efficiently. Detailed hardware and software configurations are summarized in Table 2. To ensure absolute reproducibility, all models were trained under a strict deterministic configuration with a fixed random seed, eliminating stochastic optimization variance.

Table 2. Computational environment configuration.

Environment Configuration	Value
Operating system	Ubuntu 20.04 (Cloud Environment)
CPU	Intel Xeon @ 2.20 GHz
GPU	NVIDIA A100 (NVIDIA A100-SXM4)
RAM	40 GB
Development environment	Cloud-based Jupyter
Programming language	Python 3.12

During training, the model was optimized for 500 epochs using a batch size of 64, with all input images resized to a resolution of 640×640 pixels. Stochastic gradient descent was employed as the optimization algorithm with an initial learning rate of 0.01. This optimizer is widely adopted in YOLO-based detection frameworks and is recognized for its stable convergence behavior and strong generalization capability. To improve data loading efficiency, 16 worker threads were utilized during the training process.

For performance evaluation, mean average precision at an intersection over union threshold of 0.5 was selected as the primary metric, as it effectively reflects detection accuracy across diverse vehicle damage scenarios. The overall training configuration was carefully designed to achieve a balance between convergence stability, detection accuracy, and computational efficiency, thereby supporting the practical applicability of the proposed vehicle damage detection framework.

4.2. Comparison with Mainstream Methods

The proposed method is compared with several mainstream object detection algorithms, including DETR, Faster R-CNN, YOLOv5n, YOLOv8n, and YOLO11n. Quantitative results in terms of detection accuracy, computational complexity, and inference speed are summarized in Table 2, where the inference speed (FPS) is measured under single-image inference on an NVIDIA A100 GPU.

As shown in Table 3, DETR and Faster R-CNN exhibit higher model complexity, with parameter sizes of 28.30 M and 36.74 M, respectively, and relatively lower inference speeds. Although Faster R-CNN achieves a precision of 67.3%, its recall and mAP50-95 remain lower than those of the one-stage detectors. In contrast, lightweight YOLO-based models demonstrate substantially higher inference speeds, with YOLOv5n and YOLOv8n

exceeding 450 FPS while maintaining lower computational costs. Among them, YOLOv8n achieves 66.5% mAP50 and 52.3% mAP50-95, whereas YOLO11n attains the highest recall of 67.2% with the lowest GFLOPs and a compact parameter size of 2.58 M.

Table 3. Trade-off between accuracy and efficiency across detection models.

Algorithm	Precision/%	Recall/%	mAP50/%	mAP50-95/%	GFLOPs	Parameters/M	FPS
DETR	64.7	63.7	65.3	50.9	36.8	28.30	118.6
Faster R-CNN	67.3	58.3	64.9	45.5	37.5	36.74	85.4
YOLOv5n	73.0	62.4	65.9	49.7	7.1	2.50	454.5
YOLOv8n	74.0	62.2	66.5	52.3	8.1	3.07	476.2
YOLO11n	69.5	67.2	67.4	52.8	6.3	2.58	434.8
OURS	70.3	65.8	67.9	53.8	8.9	2.65	416.6

Notably, while the proposed framework increases the computational complexity to 8.9 GFLOPs compared to the YOLO11n baseline it strictly remains below the 10 GFLOPs threshold, which is widely recognized as an edge-level lightweight deployment scale. In fine-grained vehicle damage detection, target regions often exhibit highly irregular shapes and overlapping category features, inducing a natural trade-off or competitive relationship between Precision and Recall. Therefore, evaluating a single metric in isolation does not fully reflect the actual generalization capacity. The simultaneous improvement in both mAP50 and mAP50-95 underscores the superior comprehensive localization and classification capacity of our model, effectively lowering the joint risk of both missed detections and false alarms.

The proposed method achieves strong overall performance across the evaluated metrics. Specifically, OURS obtains 70.3% precision, 65.8% recall, and 67.9% mAP50, while achieving the highest mAP50-95 of 53.8% among the compared lightweight models. At the same time, OURS maintains a compact model size of 2.65 M parameters and achieves an inference speed of 416.6 FPS, indicating a favorable balance between detection accuracy and computational efficiency. The low parameter count and constrained GFLOPs of the proposed framework inherently guarantee its architectural efficiency and readiness for practical deployment on resource-limited edge devices, independent of specific hardware training environments. By maintaining the algorithmic complexity strictly below the ten GFLOPs threshold, the framework demonstrates a strong theoretical readiness for embedded migration, ensuring that the network can be efficiently mapped onto commodity edge processors without suffering from structural memory bottlenecks.

To further compare the performance of different methods across multiple accuracy metrics, Figure 5 presents a normalized heatmap of Precision, Recall, mAP50, and mAP50-95, where each metric is normalized with respect to the best-performing method. This visualization highlights the relative strengths and weaknesses of different detectors in a unified manner.

As shown in Figure 5, DETR and Faster R-CNN exhibit lower normalized scores across several metrics, particularly in recall and mAP50-95. Among the YOLO-based lightweight models, YOLOv5n and YOLOv8n demonstrate high normalized precision and mAP50 values, while YOLO11n achieves the highest normalized recall score and maintains strong performance on both mAP-related metrics.

The proposed method achieves consistently high normalized scores across all evaluated metrics. Notably, OURS attains the highest normalized values in both mAP50 and mAP50-95, while remaining competitive in precision and recall. These results indicate that the proposed method offers relatively stable and balanced detection performance across different evaluation criteria.

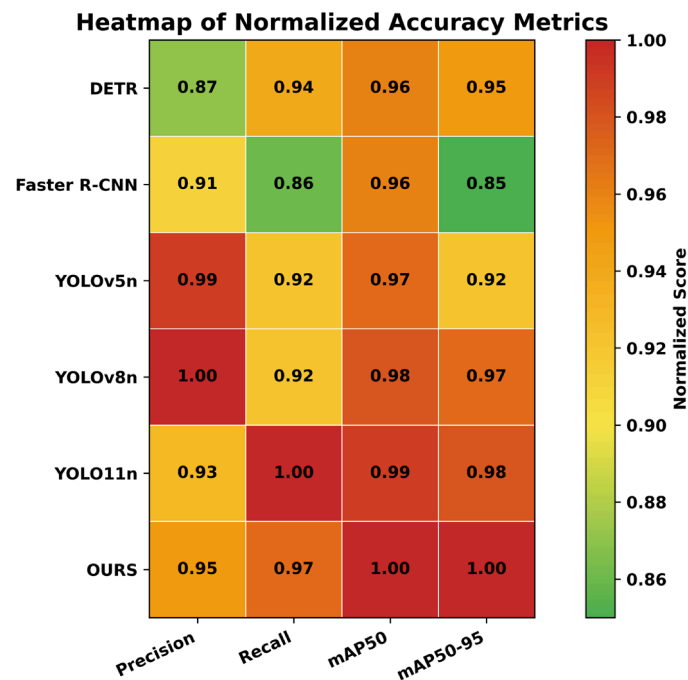


Figure 5. Heatmap of normalized detection accuracy metrics for different object detection methods.

Figure 6 illustrates the training and validation curves of the proposed method, including box loss, classification loss, and distribution focal loss, as well as the corresponding evaluation metrics across training epochs. Both training and validation losses exhibit a clear downward trend and gradually converge as training progresses, indicating stable optimization behavior.

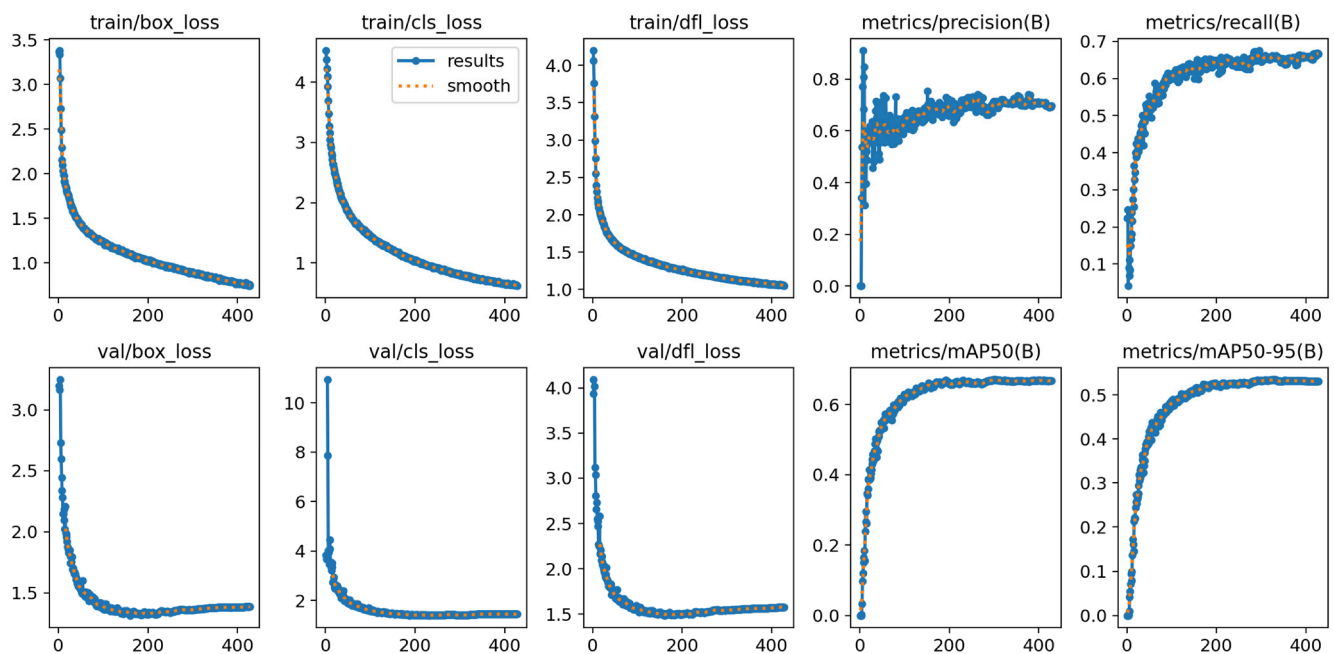


Figure 6. Training and validation curves of the proposed vehicle damage detection model.

Specifically, the training losses decrease rapidly during the early epochs and continue to decline smoothly thereafter. The validation losses follow a similar trend, remaining close to the training losses throughout the training process without significant divergence. This consistency suggests stable convergence and does not indicate obvious overfitting during training.

In terms of detection performance, precision, recall, mAP50, and mAP50-95 increase steadily with training epochs and reach stable values in later stages. Both mAP-related metrics show smooth convergence, with mAP50-95 consistently improving alongside mAP50. These results demonstrate that the proposed method achieves stable training dynamics and reliable performance convergence.

To further evaluate the fine-grained classification performance across distinct damage categories, Figure 7 presents a normalized confusion matrix of the proposed framework. The empirical results demonstrate that the proposed method achieves strong performance on macro-level structural damages. Specifically, OURS obtains an outstanding accuracy of 99% for glass shatter, 92% for tire flat, and 81% for lamp broken, indicating that the framework successfully captures the distinctive geometric features of these prominent categories. Conversely, the framework faces challenges with fine-grained surface anomalies, where OURS achieves a lower accuracy of 52% for dents, 58% for scratches, and 30% for cracks. The primary source of error stems from false negatives, where 69% of cracks, 42% of dents, and 40% of scratches are misclassified as background. This is heavily attributed to the subtle physical footprints of surface anomalies, which are frequently masked by complex ambient lighting, specular reflections on the vehicle body, or low contrast against intact panels. Additionally, a slight cross-misclassification of 6% exists between dents and scratches due to their overlapping feature boundaries. This visualization highlights the precise operational boundaries and category-wise capabilities of the proposed model in a unified manner, providing crucial insights for real-world deployment.

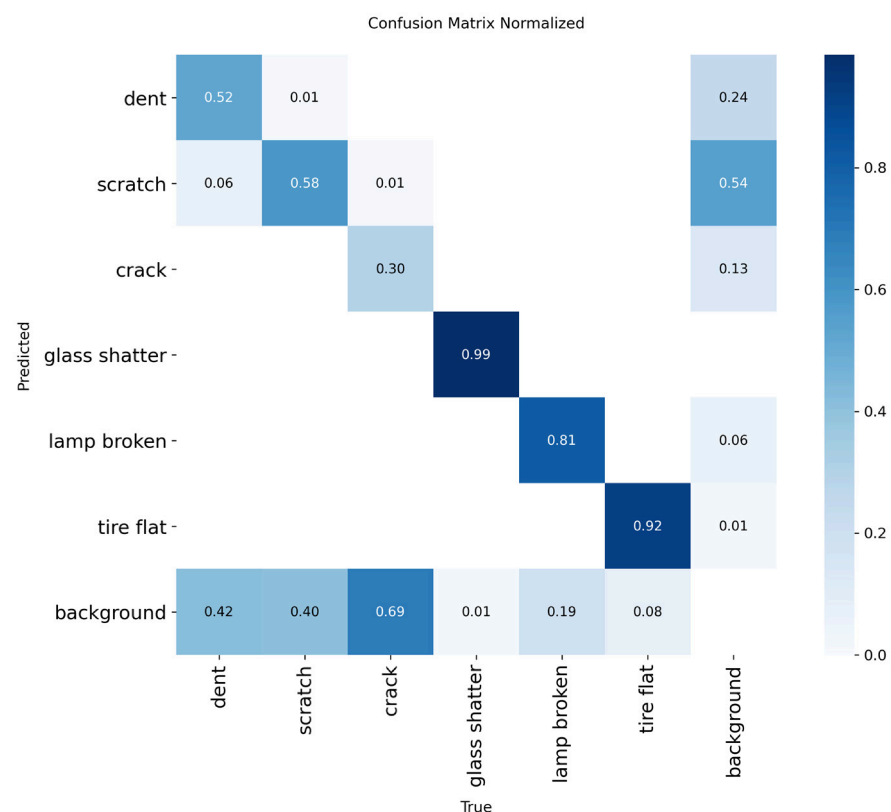


Figure 7. Category-Wise Performance Analysis.

Qualitative detection results on representative test images are provided to further evaluate the effectiveness of the proposed method across different damage categories. The visual examples, as shown in Figure 8, cover common vehicle damage types, including dent, scratch, crack, glass shatter, lamp broken, and tire flat.



Figure 8. Qualitative detection results of the proposed method on representative vehicle damage images.

The proposed method is able to localize damage regions with compact bounding boxes and assign category labels with relatively high confidence scores. For surface-level damages such as dents, scratches, and cracks, the model successfully captures fine-grained visual characteristics, even when the damaged areas are relatively small or located in complex background regions. For structural damages such as shattered glass, broken lamps, and flat tires, the detection results remain precise and visually consistent.

Overall, the qualitative results suggest that the proposed method maintains reliable detection performance across diverse damage categories and challenging visual conditions. The accurate localization and stable classification outcomes further confirm the effectiveness of the model in practical vehicle damage detection applications.

4.3. Ablation Experiment

To evaluate the contribution of each proposed component, ablation experiments are conducted by progressively enabling the Slim-neck structure, the BRA module, and the SIoU loss. The quantitative results are summarized in Table 4.

As shown in Table 4, the baseline model achieves 67.4% mAP50 and 52.8% mAP50-95. Introducing the Slim-neck alone leads to an improvement in precision, increasing it from 69.5% to 71.6%, while mAP50-95 is improved to 53.9%. When only the BRA module is applied, the overall performance shows a slight decline compared to the baseline, suggesting that this module may require complementary structural optimization to fully

realize its benefit. In contrast, using SIoU alone improves both precision and mAP50-95, indicating its positive contribution to localization accuracy.

Table 4. Ablation results of modules in the proposed framework's performance.

Slim-Neck	BRA	SIoU	Precision	Recall	mAP50	mAP50-95	GFLOPs
×	×	×	69.5	67.2	67.4	52.8	6.3
✓	×	×	71.6	64.8	66.8	53.9	6.3
×	✓	×	69.6	64.5	66.2	52.5	8.9
×	×	✓	71.1	65.5	67.4	53.0	6.3
✓	✓	×	70.0	65.0	67.2	53.7	8.9
✓	×	✓	69.0	65.8	67.4	53.1	6.3
×	✓	✓	70.8	65.8	66.8	53.5	8.9
✓	✓	✓	70.3	65.8	67.9	53.8	8.9

To provide a transparent evaluation of architectural efficiency, the module-wise computational complexity quantified via GFLOPs has been systematically integrated into the ablation matrix in Table 4. The baseline model establishes an initial computing benchmark of 6.3 GFLOPs. Empirical data indicates that incorporating the Slim-Neck module maintains the structural complexity at exactly 6.3 GFLOPs, quantitatively validating the exceptional lightweight channel-mixing capability of the GSConv framework. The primary computational variation occurs with the integration of the dynamic bi-level routing attention module, which scales the computational load to 8.9 GFLOPs. This incremental overhead remains strictly bounded below the ten GFLOPs threshold optimization curve required for edge hardware deployment, proving that our integrated configuration achieves an optimized trade-off between semantic representation accuracy and real-time processing constraint parameters.

When combining two components, further performance gains are observed. The combination of Slim-neck and SIoU maintains stable accuracy across all metrics, while the combination of BRA and SIoU improves precision and mAP50-95 to 70.8% and 53.5%, respectively. The joint use of Slim-neck and BRA results in balanced performance across precision, recall, and mAP metrics.

When all three components are integrated, the model achieves the best overall performance, reaching 67.9% mAP50 and 53.8% mAP50-95. Precision and recall are maintained at 70.3% and 65.8%, respectively. These results demonstrate that the proposed components are complementary and collectively contribute to improved detection performance.

To further analyze the impact of each component, the relative performance improvements over the baseline model are illustrated in Figure 9. The figure presents the variations in precision, mAP50, and mAP50-95 under different component combinations, providing a clearer understanding of the contribution of each module.

As shown in the figure, introducing the Slim-Neck structure leads to a notable improvement in precision, with an increase of 2.1 percent, while mAP50 exhibits a slight decrease and mAP50-95 improves by 1.1 percent. The bi-level routing attention module alone produces only a marginal improvement in precision but results in a reduction in mAP50, whereas the use of SIoU leads to simultaneous improvements in precision and mAP50-95. When two components are combined, different performance trade-offs can be observed. The combination of Slim-Neck and bi-level routing attention improves mAP50-95 by 0.9 percent, Slim-Neck combined with SIoU slightly reduces precision while enhancing mAP50-95, and the combination of bi-level routing attention and SIoU provides more balanced improvements in both precision and mAP50-95.

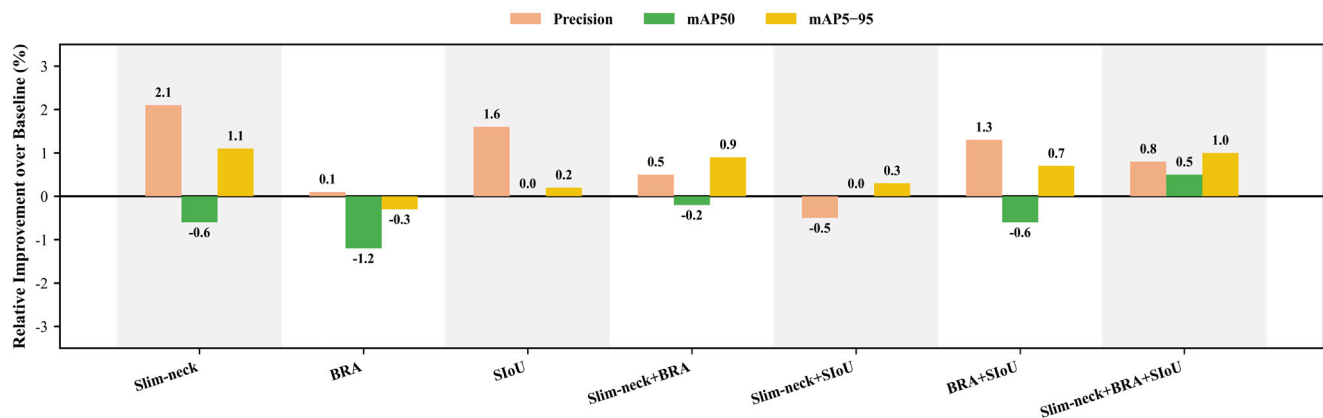


Figure 9. Relative performance improvements of different module combinations over the baseline model.

When all three components are integrated, the model achieves consistent performance gains across all evaluated metrics. Precision, mAP50, and mAP50–95 increase by 0.8 percent, 0.5 percent, and 1.0 percent, respectively. Compared with configurations using only single or paired components, the full configuration yields the most stable and comprehensive performance improvement.

As summarized in Table 4, isolating individual modules initially causes minor fluctuations in baseline performance. Specifically, incorporating the Slim-Neck alone reduces feature redundancy but slightly restricts the initial channel capacity, resulting in a marginal drop in mAP50 to 66.8% while refining localization accuracy. Conversely, applying the BRA module alone introduces dynamic sparse attention over unpurified features, leading to background noise interference and a drop in mAP50 to 66.2%. Crucially, when combined, the Slim-Neck acts as a structural purifier that filters background noise, allowing the BRA module to precisely capture long-range contextual dependencies. This synergistic effect eventually achieves the optimal comprehensive performance.

5. Discussion

The experimental results demonstrate that the proposed method delivers competitive or superior performance on mAP50 and mAP50-95 compared with mainstream lightweight YOLO-based models, while maintaining a compact model size and high inference speed. This indicates that the proposed design effectively improves overall detection quality without introducing excessive computational overhead, which is critical for real-time and resource-constrained deployment scenarios.

The normalized heatmap analysis further highlights the robustness of the proposed method across multiple evaluation metrics. Unlike some baseline models that exhibit strengths in individual metrics but weaknesses in others, the proposed method shows consistently high normalized scores in precision, recall, and mAP-related metrics. This balanced performance suggests improved generalization across different IoU thresholds and damage categories, rather than optimization toward a single evaluation criterion.

Training and validation curves provide additional evidence of stable optimization behavior. The smooth convergence of losses and the consistent improvement of precision, recall, mAP50, and mAP50-95 indicate reliable training dynamics without obvious overfitting. Such stability is particularly important for practical applications, as it reflects robustness to dataset variations and training conditions.

Qualitative visualization results further support the quantitative findings. OURS demonstrates accurate localization and reliable classification across diverse vehicle damage types. The ability to precisely detect small or subtle damage regions under complex

backgrounds suggests strong feature representation capability, which aligns with the observed improvements in mAP50-95.

The ablation experiments provide insight into the contribution of individual components and their combinations. While single modules offer limited or metric-specific improvements, combining multiple components leads to more consistent gains. In particular, the full configuration achieves the most balanced improvement across precision, mAP50, and mAP50-95, indicating that the proposed components complement each other effectively. The relative improvement analysis further confirms that the joint design mitigates performance trade-offs observed in partial configurations.

Despite the measured performance improvements, a critical analysis of the framework reveals specific operational limitations and failure modes. As illustrated by the category-wise performance analysis, the framework experiences degradation when detecting subtle surface anomalies such as cracks, scratches, and dents. The primary failure mode stems from false negatives, where these fine-grained deformations are misclassified as background panels. This vulnerability is highly pronounced under challenging environmental conditions, including intense specular reflections on glossy car bodies, non-uniform ambient lighting, and instances of extremely low contrast between the defect and intact paint. Furthermore, a minor computational trade-off exists within the design. The integration of advanced feature refinement modules introduces a slight increase in GFLOPs and a marginal reduction in frame rate compared with the native baseline. Although the remaining throughput of 416.6 FPS is vastly sufficient for standard real-time deployment, this trade-off highlights that further structural optimization is necessary for extreme hardware-constrained scenarios. Acknowledging these specific scenarios where the framework struggles provides a realistic boundary for its real-world application.

Although the experimental results demonstrate stable performance under diverse environmental constraints, validating the network on completely external cross-domain datasets represents a meaningful direction for our future work to further map the universal generalization boundaries of the framework.

6. Conclusions

In this paper, we propose an efficient object detection method for vehicle damage detection by integrating multiple lightweight design strategies. Extensive quantitative comparisons, normalized metric analyses, training convergence behaviors, qualitative visualizations, and ablation studies thoroughly confirm the effectiveness and robustness of the proposed framework. The empirical results demonstrate that the individual components jointly contribute to improved detection accuracy and stable performance across different evaluation categories while preserving high inference speed and a compact model size. Future work will focus on extending the method to more complex damage categories, deploying the framework onto physical low-power edge platforms for hardware-in-the-loop optimization, and exploring its applicability to broader real-world inspection tasks.

Author Contributions: Conceptualization, L.H.; methodology, W.L.; software, L.H.; validation, W.L. and L.H.; formal analysis, P.L.; investigation, X.L.; resources, X.L.; data curation, P.L.; writing—original draft preparation, W.L.; writing—review and editing, L.H.; visualization, P.L.; supervision, L.H.; project administration, X.L.; funding acquisition, W.L. 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).

Data Availability Statement: The dataset used in this study is publicly available. Specifically, the Car Damage Detection (CarDD) dataset can be accessed via github at: <https://github.com/CarDD-USTC/CarDD-USTC.github.io> (accessed on 8 January 2023).

Acknowledgments: The authors would like to acknowledge the authors of the Car Damage Detection dataset introduced in the study “Car Damage Detection and Segmentation Using Deep Learning” for making their dataset publicly available. The dataset was directly used in this work for model training and evaluation without additional reorganization or modification. The availability of this dataset provided essential support for the experimental validation of the proposed method and facilitated reproducible research in vehicle damage detection.

Conflicts of Interest: The authors declare that they have no known conflicts of interest or personal relationships that could have affected the work reported in this paper.

References

1. Katsamenis, I.; Bakalos, N.; Lappas, A.; Protopapadakis, E.; Martín-Portugués Montoliu, C.; Doulamis, A.; Doulamis, N.; Rallis, I.; Kalogeras, D. DORIE: Dataset of Road Infrastructure Elements—A Benchmark of YOLO Architectures for Real-Time Patrol Vehicle Monitoring. *Sensors* **2025**, *25*, 6653. [\[CrossRef\]](#)
2. Zhao, S.; Gong, Z.; Zhao, D. Traffic signs and markings recognition based on lightweight convolutional neural network. *Vis. Comput.* **2024**, *40*, 559–570. [\[CrossRef\]](#)
3. Elassy, M.; Al-Hattab, M.; Takruri, M.; Badawi, S. Intelligent transportation systems for sustainable smart cities. *Transp. Eng.* **2024**, *16*, 100252. [\[CrossRef\]](#)
4. Zhu, L.; Wang, B.; Yan, Y.; Guo, S.; Tian, G. A novel traffic accident detection method with comprehensive traffic flow features extraction. *Signal Image Video Process.* **2023**, *17*, 305–313. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Sundaresan Geetha, A.; Alif, M.A.R.; Hussain, M.; Allen, P. Comparative Analysis of YOLOv8 and YOLOv10 in Vehicle Detection: Performance Metrics and Model Efficacy. *Vehicles* **2024**, *6*, 1364–1382. [\[CrossRef\]](#)
6. Merolla, D.; Latorre, V.; Salis, A.; Boanelli, G. Improving Road Safety with AI: Automated Detection of Signs and Surface Damage. *Computers* **2025**, *14*, 91. [\[CrossRef\]](#)
7. Wang, H.; Zhao, J. Research on Defect Detection on Steel Rails Based on Improved YOLO11n Algorithm. *Appl. Sci.* **2026**, *16*, 842. [\[CrossRef\]](#)
8. Wu, W.; Chen, L. GS-BiFPN-YOLO: A Lightweight and Efficient Method for Segmenting Cotton Leaves in the Field. *Agriculture* **2026**, *16*, 102. [\[CrossRef\]](#)
9. Qin, Y.; Zhao, S. Lightweight YOLO-Based Online Inspection Architecture for Cup Rupture Detection in the Strip Steel Welding Process. *Machines* **2026**, *14*, 40. [\[CrossRef\]](#)
10. Sun, W.; Liu, X.; Hao, J.; Yao, Q.; Xi, H.; Wu, Y.; Xing, Z. AGS-YOLO: An Efficient Underwater Small-Object Detection Network for Low-Resource Environments. *J. Mar. Sci. Eng.* **2025**, *13*, 1465. [\[CrossRef\]](#)
11. Hasan, M.J.; Nguyen, C.K.; Boo, Y.L.; Jahani, H.; Ong, K.-L. Vehicle Damage Detection Using Artificial Intelligence: A Systematic Literature Review. *WIREs Data Min. Knowl. Discov.* **2025**, *15*, e70027. [\[CrossRef\]](#)
12. Edozie, E.; Shuaibu, A.N.; John, U.K.; Sadiq, B.O. Comprehensive review of recent developments in visual object detection based on deep learning. *Artif. Intell. Rev.* **2025**, *58*, 277. [\[CrossRef\]](#)
13. Wang, Q.; Wang, D.; Lu, J.; Xiao, G.; Liang, D.; Lu, G.; Shao, H. SAL-YOLO-DeepSeek: A lightweight real-time detection and LLM-driven decision framework for intelligent escalator safety monitoring. *Sci. Rep.* **2025**, *15*, 40600. [\[CrossRef\]](#)
14. Yu, H.; Liu, J.; Lin, M. A comprehensive literature review on YOLO-based small object detection: Methods, challenges, and future trends. *Comput. Mater. Contin.* **2026**, *87*, 7. [\[CrossRef\]](#)
15. Omodaratan, B.; Jamali, A.; Wiley, T.; Al-Saadi, Z.; Mallipeddi, R.; Asadi, E.; Asadi, H.; Sadeghian, R.; Sareh, S.; Khayyam, H. Advances in You Only Look Once (YOLO) algorithms for lane and object detection in autonomous vehicles. *Eng. Appl. Artif. Intell.* **2026**, *168*, 113893. [\[CrossRef\]](#)
16. Li, W.; Xie, H.; Lin, P.; Huang, L. HL-YOLO: Improving Vehicle Damage Detection with Heterogeneous Convolutions and Large-Kernel Attention. *World Electr. Veh. J.* **2025**, *16*, 640. [\[CrossRef\]](#)
17. Wang, X.; Li, W.; Wu, Z. CarDD: A New Dataset for Vision-based Car Damage Detection. *arXiv* **2022**, arXiv:2211.00945. [\[CrossRef\]](#)
18. Xu, F.; Wan, Y.; Ning, Z.; Wang, H. Comparative Study of Lightweight Target Detection Methods for Unmanned Aerial Vehicle-Based Road Distress Survey. *Sensors* **2024**, *24*, 6159. [\[CrossRef\]](#)
19. Xiao, Y.; Fan, X.; Xie, Z.; Lu, Y. Offline Knowledge Base and Attention-Driven Semantic Communication for Image-Based Applications in ITS Scenarios. *Big Data Cogn. Comput.* **2025**, *9*, 240. [\[CrossRef\]](#)

20. Yao, K.; Wang, Z.; Guo, F.; Li, F. Driver Abnormal Expression Detection Method Based on Improved Lightweight YOLOv5. *Electronics* **2024**, *13*, 1138. [\[CrossRef\]](#)
21. Shi, P.; Zhang, Y.; Cao, Y.; Sun, J.; Chen, D.; Kuang, L. DVCW-YOLO for Printed Circuit Board Surface Defect Detection. *Appl. Sci.* **2025**, *15*, 327. [\[CrossRef\]](#)
22. Yan, X.; Fan, X.; Yau, K.-L.A.; Xie, Z.; Men, R. Semantic communication based on bi-level routing attention in IoT environment. *J. Supercomput.* **2025**, *81*, 605. [\[CrossRef\]](#)
23. Wang, L.; Jiang, F.; Zhu, F.; Ren, L. Enhanced Multi-Target Detection in Complex Traffic Using an Improved YOLOv8 with SE Attention, DCN_C2f, and Siou. *World Electr. Veh. J.* **2024**, *15*, 586. [\[CrossRef\]](#)
24. Wang, J.; Qin, C.; Hou, B.; Yuan, Y.; Zhang, Y.; Feng, W. LCGSC-YOLO: A Lightweight Apple Leaf Diseases Detection Method Based on LCNet and GSConv Module under YOLO Framework. *Front. Plant Sci.* **2024**, *15*, 1398277. [\[CrossRef\]](#)
25. Huang, D.; Tu, Y.; Zhang, Z.; Ye, Z. A Lightweight Vehicle Detection Method Fusing GSConv and Coordinate Attention Mechanism. *Sensors* **2024**, *24*, 2394. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Yao, Z.-x.; Wang, H.; Meng, Z.-j.; Yang, L.-l.; Zhang, T.-h. Multi-scale Feature Alignment Network for 19-class Semantic Segmentation in Agricultural Environments. *Artif. Intell. Agric.* **2026**, *16*, 529–545. [\[CrossRef\]](#)
27. Hu, Z.; Kojima, R.; Han, X.-H. Multi-Scale Attention-Driven Hierarchical Learning for Fine-Grained Visual Categorization. *Electronics* **2025**, *14*, 2869. [\[CrossRef\]](#)
28. Ding, X.; Tang, J.; Li, Y.; Yang, G.; Lang, Y. Bi-Level Routing Attention and Enhanced Spatial-Temporal Inconsistency Learning for Deep VFI Video Detection. *ACM Trans. Multimed. Comput. Commun. Appl.* **2025**, *22*, 1–26. [\[CrossRef\]](#)
29. Cao, L.; Wang, Q.; Luo, Y.; Hou, Y.; Cao, J.; Zheng, W. YOLO-TSL: A Lightweight Target Detection Algorithm for UAV Infrared Images Based on Triplet Attention and Slim-neck. *Infrared Phys. Technol.* **2024**, *141*, 105487. [\[CrossRef\]](#)
30. Zhao, Z.; Huang, W.; Li, T.; Zhu, J. eCBAM and saSiou Co-Optimized YOLOv11 for Riverine Floating Garbage Classification Under Complex Aquatic Scenarios. *Appl. Sci.* **2026**, *16*, 651. [\[CrossRef\]](#)
31. 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. [\[CrossRef\]](#)
32. 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. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.