This is a Just Accepted version, keep updated with https://doi.org/10.26599/LLES.2026.9660019

Research Article

Lightweight segmentation model for real-time detection of multi-type façade defects using enhanced YOLOv12

Yi-Si Lu[1], Kang Zhou[1], Ao Zhou[2], Tuo Liao[3], Ming-Ming Wang[3], Jia-Le Shi[4, ✉]

[1] *Research Center for Wind Engineering and Engineering Vibration, Guangzhou University, Guangzhou, China*

[2] *School of Civil and Marine Engineering, Harbin Institute of Technology, Shenzhen, China*

[3] *Guangdong Province Academy of Building Research Group Co., Ltd., Guangzhou, China*

[4] *College of Civil Engineering, Hefei University of Technology, Hefei, China*

✉ Address correspondence to Jia-Le Shi, shijl@mail.hfut.edu.cn

Received: 08 May 2026; Revised: 10 June 2026; Accepted: 02 August 2026

ABSTRACT

Accurate segmentation of façade defects is essential for quantifying damage extent and supporting maintenance decisions in urban infrastructure management. Accurate segmentation of façade defects is essential for extracting visible damage regions and providing auxiliary information for urban infrastructure maintenance. However, existing detection methods based on bounding boxes often fail to delineate defect boundaries precisely, limiting their practical applicability. To address this challenge, this study proposes a lightweight instance segmentation model for real-time multi-

type façade defect detection, built upon an enhanced YOLOv12n architecture. Three task-oriented modules are incorporated and adapted for façade defect segmentation: a Global Edge Information Transfer (GEIT) module that enhances boundary representation by integrating multi-scale edge features into shallow layers; a Hierarchical Attention Fusion Block (HAFB) that captures both local details and global semantics via parallel attention and re-parameterizable convolutions; and a Lightweight Shared Convolutional Segmentation Head (LSCSH) that improves efficiency and robustness through parameter sharing, scale adaptation, and group normalization. Experimental results show that the proposed model achieves 91.5% precision, 84.7% recall, 87.2% mAP@0.5, and 64.8% mAP@0.5:0.95, with a compact size of 5.1 MB and real-time speed of 201 FPS, demonstrating strong segmentation performance and deployment potential.

KEYWORDS

Building façade defects, Real-time segmentation, YOLOv12, Edge-aware feature extraction, Multi-scale attention mechanism

1. Introduction

The building façade serves as a critical envelope system, shielding interior environments from external conditions and ensuring the normal operation of architectural functions. Over long-term service periods, façades are subjected to a combination of environmental degradation, mechanical stresses, and material aging, making them highly susceptible to various forms of damage such as cracking, peeling, and detachment. These defects not only compromise the durability of the structure but also pose significant risks to public safety. In recent years, a series of façade-related accidents have underscored the severity of this issue. For instance, in December 2019, a terracotta piece fell

from a building near Times Square in New York, resulting in the death of a 60-year-old pedestrian. In July 2024, the collapse of a hospital façade in Shandong, China, caused one fatality and injured two others. Similarly, in September 2024, a brick detachment incident in Tokyo, Japan, led to the death of a 67-year-old. In February 2025, falling concrete fragments from a commercial building on Queen Street, Auckland, New Zealand, caused head injuries to an elderly woman and eye trauma to another pedestrian. These incidents highlight the urgent need for early detection of façade defects to ensure urban public safety.

Traditional façade inspection methods predominantly rely on manual visual surveys, which suffer from low efficiency, high subjectivity, and limited coverage (Lourenco et al. 2017, Zhang et al. 2020, Wang et al. 2024). Such approaches are increasingly inadequate for the growing demands of intelligent maintenance in modern cities. With the rapid development of deep learning and computer vision technologies, vision-based automated damage detection methods have emerged as promising alternatives to manual inspections (Zhao et al. 2018, Chai et al. 2021). Among these, object detection algorithms such as Faster R-CNN and the YOLO series have demonstrated strong performance in identifying and localizing common façade defects (Ren et al. 2016, Linyi et al. 2023). For example, Perez and Tah (2019) evaluated transfer learning-based convolutional neural networks for detecting and localizing building defects such as mould, stain, and paint deterioration, demonstrating the feasibility of deep learning for building-condition assessment. Perez et al. (2021) further developed a smartphone-based deep learning application for real-time detection of building defects, highlighting the potential of lightweight vision systems for practical inspection scenarios. Interlando et al. (2024) used ensembles of deep neural networks to classify building façade defects from images, showing the value of recent architectures for façade-condition assessment. Lee et al.

(2020) employed Faster R-CNN to detect cracks, delamination, and detachment, achieving an mAP@0.5 of 62.7%. Li et al. (2024) applied YOLOv5 for crack detection, reporting an mAP@0.5 of 70.3%. Luo et al. (2025) incorporated a Coordinate Attention mechanism and a decoupled detection head, which enhanced detection accuracy for surface damage in concrete structures, reaching an mAP@0.5 of 89.1%. Idjaton et al. (2022) integrated a Transformer module into YOLOv5 for peeling detection, attaining an mAP@0.5 of 79.0%. Huang et al. (2023) combined Ghost Convolution, Squeeze-and-Excitation blocks, and a Focal-EIoU loss function within YOLOv8 to effectively detect exposed reinforcement and tile loss, achieving an mAP@0.5 of 83.4%. Zhou et al. (2025) proposed an improved YOLOv10-based model that integrates Vanillanet, BiFPN, Inner-CIoU, and a P2 detection head, demonstrating superior performance in defect detection for glass curtain walls of high-rise buildings.

Although these object detection approaches have achieved considerable progress in the identification and localization of façade defects, their inherent reliance on bounding boxes imposes significant limitations. These methods typically enclose target regions with coarse rectangular frames, making it difficult to accurately delineate the true boundaries of defects. For elongated or irregularly shaped damage types such as cracks, peeling, and detachment, bounding boxes often contain excessive background noise and tend to overlook critical geometric details. As a result, precise quantification of defect properties (e.g., length, width, and area) becomes challenging. This lack of granularity adversely affects downstream tasks, including damage progression analysis and maintenance decision-making, thereby constraining the practical value of these methods in real-world urban infrastructure management.

The deep learning-basedinstance segmentation models offer pixel-level precision, enabling

detailed identification and accurate quantification of façade damage (Wang et al. 2022, Hussain et al. 2025, Kyem et al. 2025). Recent studies have begun incorporatinginstance segmentation into structural defect detection. For instance, Cao et al. (2023) developed the YOLOM model by integrating YOLOv7 with BlendMask for tile detachment segmentation in aerial imagery. Despite achieving high accuracy, the method operated at only 4 FPS, limiting its applicability in real-time scenarios. Guo et al. (2021) employed a Mask R-CNN model to segment façade images from Singapore, attaining an mAP of 58.4%. Dais et al. (2021) utilized U-Net and Feature Pyramid Network (FPN) architectures, combined with various pre-trained CNN backbones, for pixel-level segmentation of masonry cracks, with U-Net-MobileNet and FPN-InceptionV3 achieving the highest accuracies of 79.9% and 81.3%, respectively. Chen et al. (2021) proposed a two-stage deep learning method for close-range building façade crack inspection using UAV images, combining patch-level crack classification and U-Net-based pixel-level segmentation to separate crack regions from complex façade backgrounds. Junior et al. (2021) developed a ceramic crack segmentation method and a dedicated ceramic crack database, showing that surface texture, grout lines, illumination variation, and material-specific appearance are important challenges in façade-related segmentation tasks. While these approaches demonstrate promising results in specific contexts, they generally suffer from two critical drawbacks: (a) most studies target only a single defect type, limiting generalizability to scenarios with coexisting heterogeneous damage; and (b) many models feature complex structures and redundant parameters, leading to poor inference efficiency and hindering deployment in real-time urban maintenance systems.

To address these challenges, a lightweightinstance segmentation model is proposed in this study to enable accurate and efficient detection of several common façade defects, including

concrete cracks, exposed reinforcement, tile loss, peeling, and delamination. Built upon the YOLOv12 architecture, the proposed model integrates an edge-aware enhancement module, a hierarchical attention fusion mechanism, and a shared feature head to improve fine-grained feature extraction in heterogeneous background textures while substantially reducing parameter overhead. The model achieves a balance between segmentation accuracy and inference speed, making it suitable for real-world deployment. The remainder of this paper is organized as follows: Section 2 details the proposed segmentation architecture based on the YOLOv12; Section 3 presents the dataset construction and experimental settings; Section 4 provides results and discussions on ablation and comparative experiments; and Section 5 concludes the paper with key findings.

2. Methodology

This section begins with an overview of the baseline model YOLOv12, followed by a series of architectural enhancements tailored to the task of façade defect segmentation. An improved segmentation framework is then proposed to meet the demands of accuracy and efficiency in complex urban environments.

2.1 Overview of the YOLOv12 algorithm

YOLOv12 is a state-of-the-art real-time object detection framework that introduces attention-oriented designs to improve the balance between accuracy and inference speed (Alif et al. 2025, Tian et al. 2025). Compared with previous YOLO versions that mainly rely on convolutional operations, YOLOv12 enhances feature extraction and computational efficiency through several compact structural improvements.

Specifically, the Area Attention module in A2C2f reduces computational complexity by partitioning the feature space and works with Flash Attention to improve memory access efficiency.

The Residual Efficient Layer Aggregation Network (R-ELAN) strengthens feature fusion and training stability, while the adjusted multilayer perceptron (MLP) ratio and large-kernel depth-wise separable convolutions help maintain spatial awareness with fewer redundant computations.

The YOLOv12 framework is available in several configurations, namely YOLOv12n, YOLOv12s, YOLOv12m, YOLOv12l, and YOLOv12x, each offering a different balance between model complexity and detection performance. Considering the requirements of real-time inference and the constraints associated with large-scale inspection platforms such as unmanned aerial vehicles, the lightweight version YOLOv12n is selected as the foundational architecture in this study, and its network structure is presented in Figure 1.

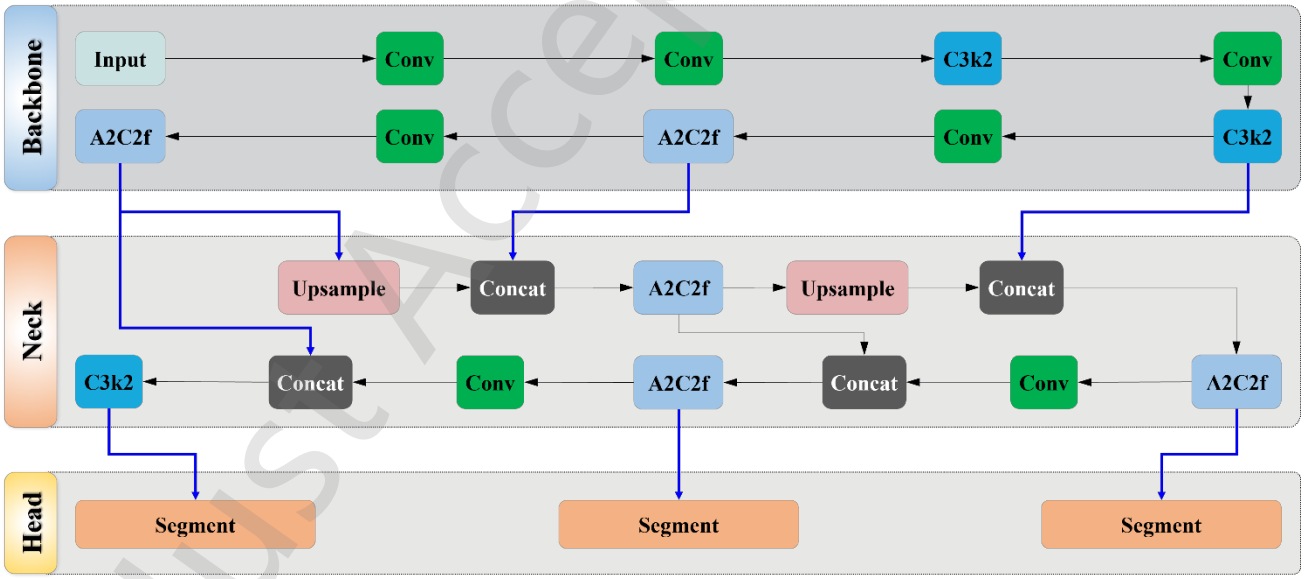


Figure 1. The network structure of YOLOv12n

2.2 Proposed façade defect segmentation algorithm

To meet the practical demands of façade defect segmentation, this paper introduces a set of structural enhancements to the YOLOv12n architecture, aiming to improve its capability in fine-grained damage perception, segmentation accuracy, and inference efficiency. The overall architecture of the improved model is shown in Figure 2. The proposed modifications include the

following key components:

(a) Global Edge Information Transfer (GEIT): This edge-aware enhancement module incorporates multi-scale edge information into shallow feature representations and fuses it with backbone semantic features. The integration significantly enhances the model's ability to capture the boundaries of clearly contoured defects such as cracks and peeling.

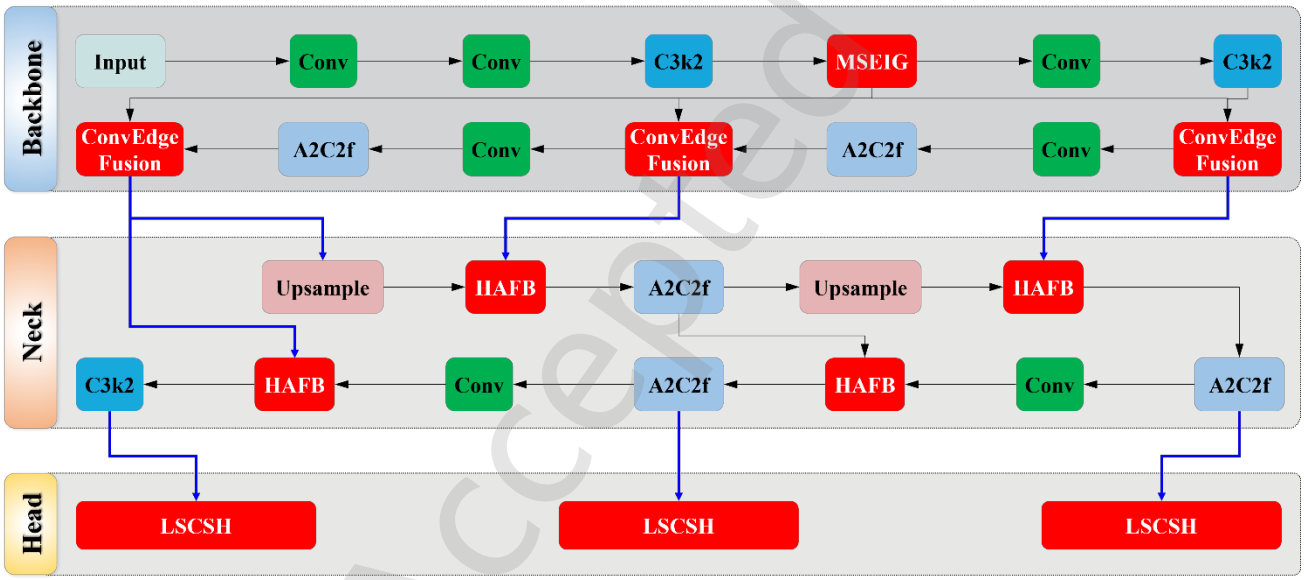


Figure 2. Architecture of the proposed algorithm

(b) Hierarchical Attention Fusion Block (HAFB): This module combines parallel local-global attention mechanisms with a re-parameterizable convolutional structure to enable comprehensive multi-scale and multi-semantic information integration. It strengthens contextual awareness and detail retention in structurally complex regions.

(c) Lightweight Shared Convolutional Segmentation Head (LSCSH): By leveraging parameter sharing, scale-adaptive mechanisms, and group normalization, this module effectively reduces the model's complexity while improving the robustness and efficiency of multi-scale segmentation.

2.2.1 Global edge information transfer module

The original YOLOv12n architecture shows limitations ininstance segmentation tasks,

particularly when dealing with façade defects that exhibit fine structures and distinct boundaries. It tends to perform poorly in edge perception and contour fitting due to its backbone network prioritizing high-level semantic abstraction while neglecting edge-specific detail modeling. As a result, predictions often suffer from blurred or fragmented boundaries. To address this issue, a module named global edge information transfer (GEIT) is introduced, as illustrated in Figure 3. The design aims to improve boundary modeling for structural damages with clear outlines, such as cracks and peeling.

Instance segmentation requires precise pixel-level delineation, making edge information especially critical in damage scenarios with strong boundary features. GEIT enhances this capability by incorporating a Multi-Scale Edge Information Generator (MSEIG) into the shallow layers of the backbone network. This generator applies multi-scale convolutions combined with max-pooling operations to extract edge features across different resolutions. Additionally, a Sobel operator is used to compute gradients along horizontal and vertical directions within the shallow feature maps, effectively capturing regions with sharp grayscale transitions and generating initial edge responses (Kanopoulos et al. 1988, Vairalkar et al. 2012). Compared to edge extraction directly from raw images, this approach reduces background noise while preserving salient edge contours. The use of multi-scale convolution and pooling further improves the model's ability to perceive edges of varying scales. These edge features are then injected into multiple levels of the backbone network (such as P3, P4, and P5), enabling a joint representation of shallow detail and deep semantic information.

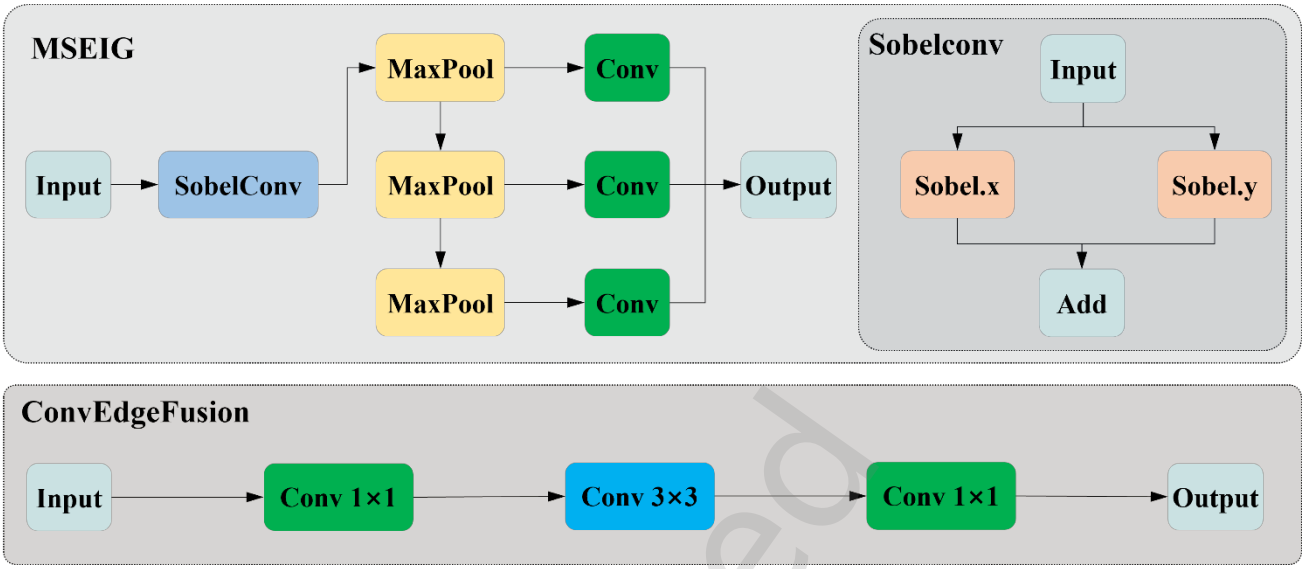


Figure 3. Architecture of GEIT module

To facilitate efficient integration, GEIT employs a dedicated fusion module referred to as ConvEdgeFusion. After edge generation, the edge features are transferred to the backbone feature layers corresponding to P3, P4, and P5. For each target layer, the edge feature is resized to the same spatial resolution as the corresponding backbone feature map and adjusted by a 1×1 convolution for channel alignment. The aligned edge feature and semantic feature are then concatenated along the channel dimension and fused within ConvEdgeFusion. A 3×3 convolution is applied to strengthen local boundary and texture responses, and a final 1×1 convolution restores the output channel dimension so that the enhanced feature can be passed to the subsequent network layers. This transfer path enables shallow edge cues to guide multi-scale semantic features at different depths, while the scale alignment strategy avoids mismatches between edge maps and backbone feature resolutions. In this way, GEIT strengthens boundary localization and contour continuity for façade defects without introducing a heavy additional branch.

2.2.2 Hierarchical attention fusion block module

Façade defects exhibit diverse spatial scales and rich semantic characteristics, ranging from

fine-grained features such as microcracks to large-scale areas of peeling. This heterogeneity poses significant challenges forinstance segmentation, where single-scale features often fail to capture both local details and global structural context. In particular, YOLOv12n and similar architectures suffer from limited multi-scale feature fusion capability, resulting in blurred boundaries and missing structural details. These limitations hinder the model's ability to achieve fine-grained and robust defect segmentation.

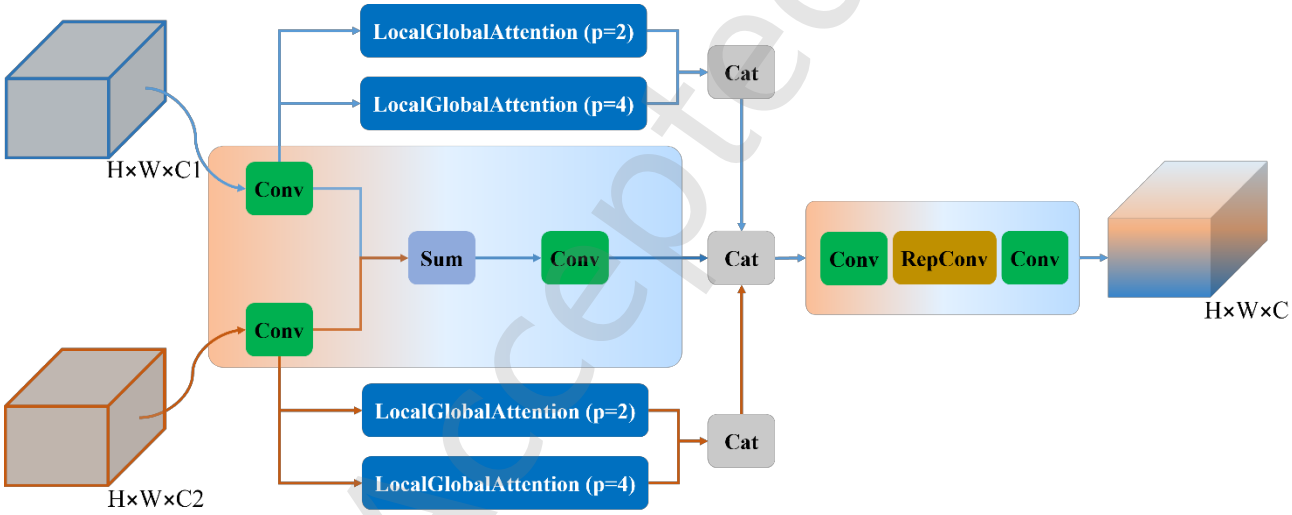


Figure 4. Architecture of HAFB module

In recent computer vision research, multi-scale feature fusion and attention mechanisms have been widely adopted to enhance perceptual capacity in complex scenes. However, many existing methods still struggle to simultaneously preserve local textures and model global semantic structures, limiting their effectiveness in precise multi-scale representation. To address these challenges, a Hierarchical Attention Fusion Block (HAFB) is proposed, as illustrated in Figure 4. This module is designed to adaptively model and integrate both local and global features, thereby enhancing the model's expressiveness and discrimination capability across multiple scales.

The HAFB module consists of three main stages: feature preprocessing, local-global attention fusion, and feature reorganization. In the preprocessing stage, the input feature map is first processed

by a 1×1 convolution to adjust the channel dimension and reduce unnecessary computation. A subsequent 3×3 convolution is then used to aggregate neighboring spatial information and generate a baseline feature representation. During the attention fusion stage, HAFB adopts a dual-branch Local-Global Attention structure (Shao et al. 2024). The local branch extracts fine-grained responses from nearby pixels and small receptive fields, which is useful for preserving crack edges, peeling boundaries, and texture discontinuities. The global branch aggregates broader contextual information from the feature map to capture the overall façade region and suppress background interference. Attention weights are generated from these two branches and are used to recalibrate the corresponding local and global feature responses. The weighted local and global features are then combined with the baseline representation, allowing the module to retain detailed defect textures while incorporating higher-level semantic context. This hierarchical fusion is beneficial for façade defect segmentation because visible defects often differ greatly in scale, morphology, and surface texture: small cracks require local boundary sensitivity, whereas larger peeling or delamination regions require broader contextual consistency. Finally, the fused features are reorganized through standard convolution, re-parameterizable convolution (RepConv), and a concluding convolution. RepConv uses a multi-branch structure during training to enrich feature representation and is merged into a single convolutional kernel during inference, thereby improving feature interaction without increasing deployment complexity (Soudy et al. 2022).

In summary, HAFB effectively improves the coordination between multi-scale details and global semantics through its hierarchical attention mechanism and efficient convolutional restructuring. This design significantly boosts segmentation performance in complex façade damage scenarios.

2.2.3 Lightweight shared convolutional segmentation head

The segmentation head in YOLOv12n adopts a decoupled structure, where separate convolutional branches are used for class prediction and spatial localization. While this design provides a certain degree of representational capacity, it suffers from several notable limitations. First, the use of multiple independent convolutional layers results in substantial parameter redundancy and reduced inference efficiency. Second, the architecture exhibits limited capability in modeling multi-scale objects, making it less effective when dealing with targets that vary significantly in size across complex scenes. Additionally, the decoupled structure constrains information exchange between branches, which weakens overall feature coordination and adversely affects segmentation accuracy.

To address these issues, a task-oriented Lightweight Shared Convolutional Segmentation Head (LSCSH) is incorporated, as illustrated in Figure 5. In the improved YOLOv12n-based model, LSCSH replaces the original multi-branch segmentation head used after the multi-scale neck outputs, while the backbone and feature-fusion neck remain unchanged. Instead of assigning an independent convolutional prediction branch to each feature layer, LSCSH applies a shared convolutional block to the multi-scale feature maps, so that the same set of convolutional parameters is reused across different prediction scales. Before prediction, Group Normalization (GN) is first used to stabilize feature distributions, and the scale-adaptive module then adjusts the responses of features from different resolutions so that shallow high-resolution details and deep semantic features can be jointly used for mask prediction.

Through this design, LSCSH reduces redundant convolutional parameters and repeated computation in the segmentation head while preserving the ability to process multi-scale façade

defects. The scale-adaptive operation provides lightweight feature adjustment before the final prediction layers, which helps maintain segmentation capability for defects with different sizes and shapes without introducing separate heavy heads for each scale.

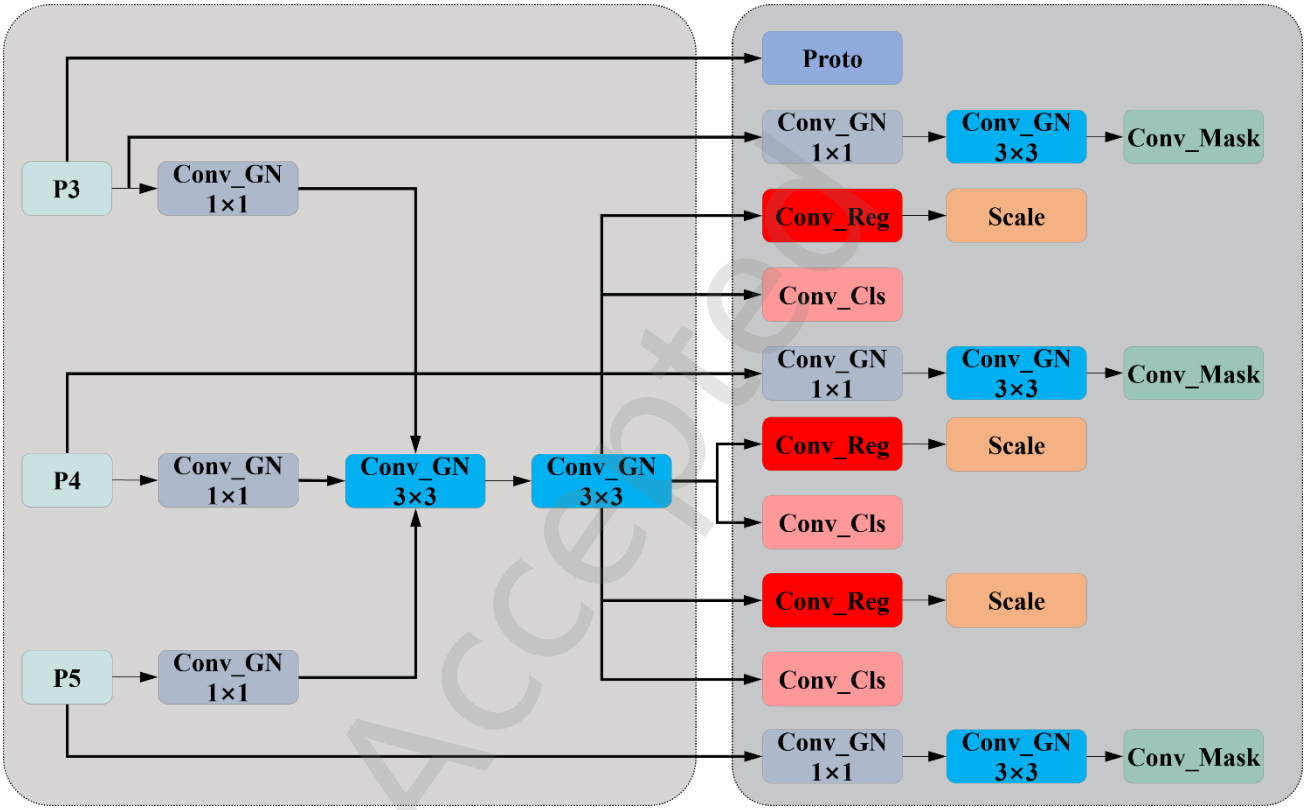


Figure 5. Architecture of the LSCSH module

3. Dataset and Experimental Setup

3.1 Building façades dataset

To support pixel-level segmentation of façade defects, a high-qualityinstance segmentation dataset was developed with a focus on precise boundary identification and detailed semantic region parsing. The dataset comprises 6156 images and their corresponding pixel-wise annotated masks, covering five common defect types: concrete cracks, exposed reinforcement, tile loss, peeling, and delamination, as illustrated in Figure 6.

The five defect categories were defined according to their visible appearance and engineering meaning during annotation. Concrete cracks refer to narrow linear or branching discontinuities on concrete surfaces; the mask boundary follows the visible crack trace and includes connected dark or open crack pixels, while isolated stains or shadows are excluded. Exposed reinforcement denotes regions where the concrete cover has detached and steel bars or corroded reinforcement are visibly exposed; the annotated mask covers the exposed steel and immediately damaged surrounding concrete, but excludes intact surface areas. Tile loss refers to missing façade tiles or cladding units with clearly visible substrate or empty tile positions; the mask follows the missing-tile contour rather than the entire tile grid. Peeling indicates surface coating or finishing layers that are lifted, flaked, or detached from the substrate; only the visibly peeled region is annotated, and discoloration without material separation is not included. Delamination refers to larger areas of separation, bulging, hollowing, or detachment of façade material layers; the mask is drawn along the visible damaged region, and ambiguous transition zones are annotated according to the most evident boundary of material separation. For all categories, annotations follow a closed-contour rule, and overlapping or adjacent defects are labeled according to the dominant visible defect type.

Among them, 4298 images were collected from urban environments in Guangzhou and Hefei, China. These samples encompass a wide variety of building façade types. Particular attention was paid to the complexity of defect boundaries, the continuity of damage morphology, and interference from background textures, in order to improve the model's robustness in handling fuzzy edges and small-scale anomalies. In addition, 1858 high-quality samples were introduced from several publicly availableinstance segmentation datasets to enhance diversity and generalization capability.

(a) Concrete crack

(b) Exposed reinforcement

(c) Tile loss

(d) Delamination

(e) Peeling

Figure 6. Examples of building façade defects

Table 1. Sample composition of the dataset

Category	Concrete crack	Exposed reinforcement	Tile loss	Peeling	Delamination
Number of training objects	2389	2031	2275	2017	2409
Number of validation objects	562	437	682	587	572
Number of test objects	417	362	539	506	491

All defect regions were manually annotated using the open-source tool Labelme, following a strict closed-contour principle to ensure accurate boundary definition. The annotations were reviewed by professionals with a structural engineering background to ensure semantic consistency and labeling precision. The complete dataset was divided into training, validation, and testing subsets in a ratio of 8:1:1. Each subset maintains a balanced distribution of defect types and façade categories. Detailed statistics are presented in Table 1.

3.2 Experimental configuration

The training and evaluation processes were conducted on a high-performance workstation equipped with an Intel i9-13900K CPU, an NVIDIA RTX 4090 GPU, and 24 GB of RAM, operating under Windows 10. The model was implemented using Python 3.10.0 and PyTorch 1.12, with CUDA 11.7 as the GPU acceleration backend. All input images and their corresponding masks were resized to 640 × 640 pixels before being fed into the network. Training was performed over 900 epochs using the stochastic gradient descent (SGD) optimizer. The learning rate was initialized at 0.01 and decayed progressively to 0.0001 using a linear schedule. Weight decay was set to 0.0005, and a momentum factor of 0.937 was used to stabilize convergence. A batch size of 32 was chosen to maintain a balance between memory efficiency and optimization dynamics. No additional data

augmentation operations, such as Mosaic augmentation, random rotation, flipping, color jittering, or random scaling, were applied in this study. In addition, no pretrained weights were used; all models were trained from scratch on the constructed façade defect dataset.

To comprehensively evaluate the model's performance, a combination of quantitative metrics was employed. Precision and recall were used to assess the effectiveness of defect identification, while segmentation quality was evaluated through mean Average Precision at two thresholds, namely mAP@0.5 and mAP@0.5:0.95. In addition, model efficiency was quantified using frames per second (FPS) and total model size, providing insight into real-time applicability and deployment feasibility. It should be noted that FPS was measured on the workstation described above and therefore reflects computational efficiency under this specific hardware configuration (Zhou et al. 2025).

4. Results and discussions

This section presents the experimental results and analysis. First, a series of ablation studies are conducted to quantitatively assess the individual contributions of each proposed module to the overall model performance, thereby validating the effectiveness of the architectural improvements. Subsequently, the proposed method is compared with several commonly used segmentation models in the field of structural damage segmentation, including YOLOv5n-seg, YOLOv8n-seg, and the two-stage approach Mask R-CNN. This comparative analysis demonstrates the advantages of the proposed model in terms of segmentation accuracy, inference speed, and model compactness.

4.1 Ablation experiments

To systematically evaluate the effectiveness of the proposed structural enhancements to the

YOLOv12n framework for façade defect segmentation, six ablation experiments were conducted. These experiments were designed to assess the individual and combined contributions of the three proposed modules: Global Edge Information Transfer (GEIT), Hierarchical Attention Fusion Block (HAFB), and Lightweight Shared Convolutional Segmentation Head (LSCSH). The performance metrics examined include model size, inference speed (FPS), precision, recall, and mean Average Precision at both 0.5 (mAP@0.5) and 0.5:0.95 (mAP@0.5:0.95) thresholds, as summarized in Table 2.

Table 2. Ablation experiment results

YOLOv12n	GEIT	HAFB	LSCSH	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Size/MB	FPS
√				78.8%	78.5%	83.5%	60.6%	6.4	151
√	√			86.5%	81.9%	85.3%	62.9%	6.6	157
√		√		86.8%	82.3%	86.1%	63.4%	5.8	178
√			√	88.8%	79.9%	86.3%	64.1%	5.5	194
√	√	√		87.3%	83.6%	86.7%	63.5%	6.1	189
√	√	√	√	**91.5%**	**84.7%**	**87.2%**	**64.8%**	**5.1**	**201**

The first set of results demonstrates that incorporating the GEIT module improves the model's ability to perceive fine-grained structural boundaries by injecting multi-scale edge information. Compared with the baseline YOLOv12n, GEIT produces clear improvements in both mAP@0.5 and mAP@0.5:0.95, indicating that edge-aware feature transfer is beneficial not only for coarse defect localization but also for stricter mask-quality evaluation. The accompanying increases in precision and recall suggest that the module helps reduce both incorrect defect predictions and missed boundary regions. Meanwhile, the inference speed is not negatively affected, showing that the edge-enhancement strategy improves boundary representation with limited computational overhead.

The second experiment evaluates the effect of HAFB on feature interaction and contextual representation. After introducing the local-global attention fusion block, the model shows simultaneous improvements in precision, recall, and mAP, which suggests that the module strengthens the interaction between local defect textures and broader façade context. Compared with the baseline, HAFB also maintains a compact model size and a high inference speed, indicating that the re-parameterizable convolutional design provides additional representational capacity during training without causing excessive inference burden. This result supports the effectiveness of HAFB in balancing segmentation accuracy and computational efficiency.

The third experiment focuses on the LSCSH module, which optimizes the segmentation head through parameter sharing, scale adaptation, and group normalization. Compared with the original segmentation head, LSCSH achieves a more compact model and faster inference while still improving mAP values. This comparison indicates that the shared-convolution design reduces redundant prediction-head computation, whereas the scale-adaptive operation helps preserve multi-scale mask prediction capability. Therefore, LSCSH contributes most directly to the lightweight and high-speed characteristics of the final model.

In addition to evaluating each module individually, the experiments also examined various combinations to explore their synergistic effects. When GEIT and HAFB were integrated jointly, the model achieved a recall of 83.6% and an mAP@0.5:0.95 of 63.5%, confirming that the modules complement each other in capturing edge details, contextual dependencies, and multi-scale representations.

The full integration of all three modules yielded the best overall performance. The final model maintained a compact size of 5.1 MB and reached an inference speed of 201 FPS. Precision

increased to 91.5%, recall to 84.7%, mAP@0.5 to 87.2%, and mAP@0.5:0.95 to 64.8%. These results demonstrate that the proposed improvements not only preserve lightweight architecture and high inference efficiency but also significantly enhance segmentation accuracy and robustness in complex façade damage scenarios. It is worth noting that the reduction in model size and the increase in FPS do not result from simply adding modules to the baseline network. The efficiency improvement is mainly attributed to replacing the original segmentation head with LSCSH, whose parameter-sharing and lightweight design reduce redundant convolutional computation. HAFB also provides an auxiliary contribution through compact feature fusion and re-parameterizable convolution, whereas GEIT mainly improves boundary perception and segmentation accuracy. Therefore, the smaller model size and higher inference speed of the final model should be understood as the combined effect of the lightweight LSCSH design, the compact HAFB structure, and their coordinated integration with GEIT.

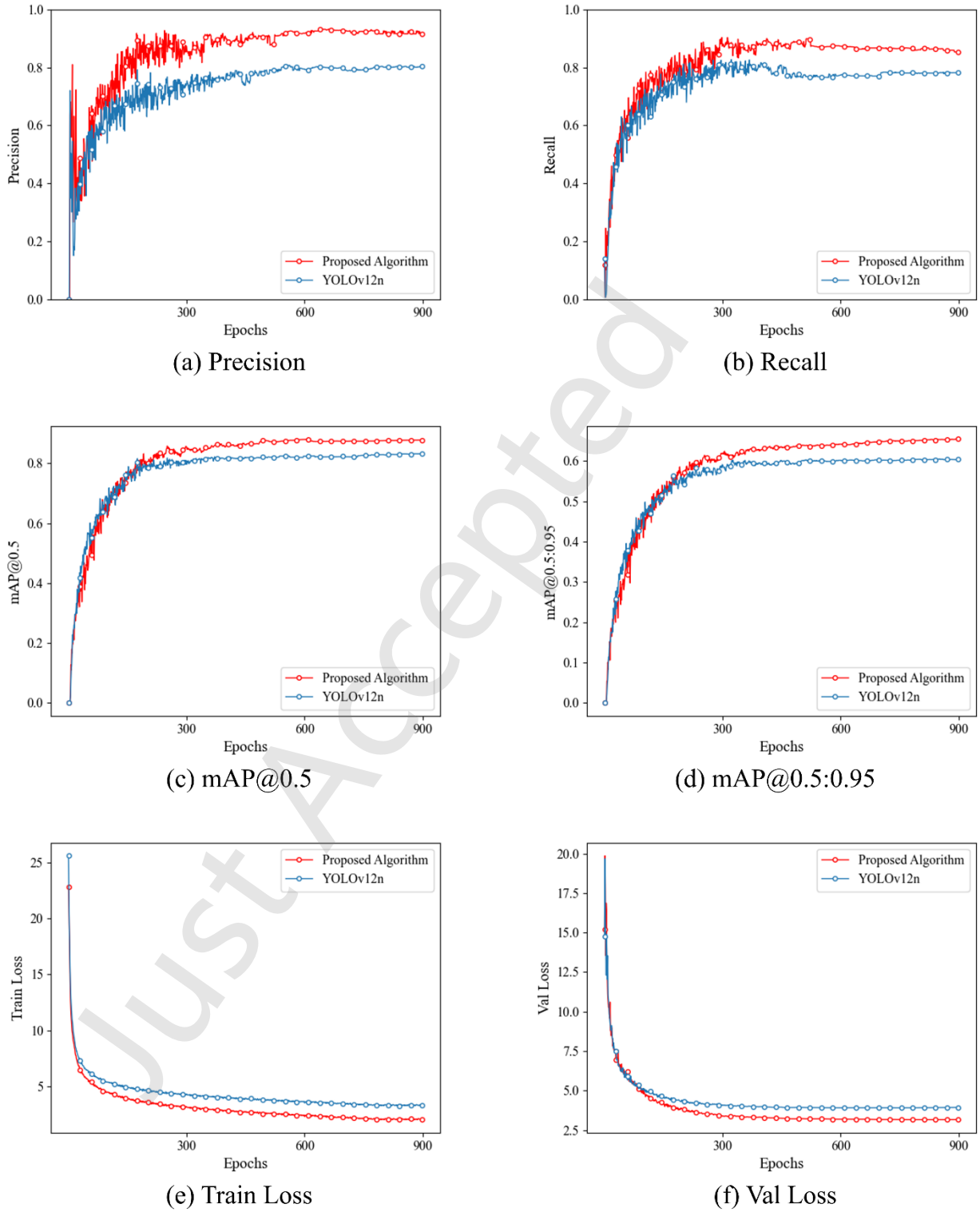


Figure 7. Metrics evolution over 900 training epochs

The training performance of the proposed algorithm compared to the baseline YOLOv11n is illustrated in Figure 7. As shown in Figures 7(a) and 7(b), the proposed model consistently

outperformed YOLOv12n in both precision and recall throughout the 900 training epochs, especially in the later stages, indicating superior classification capacity and reduced false positives and false negatives. Figures 7(c) and 7(d) further show that the proposed model maintains higher mAP@0.5 and mAP@0.5:0.95 values, suggesting that the introduced edge-enhancement, attention-fusion, and lightweight segmentation-head designs improve mask prediction quality under different IoU thresholds. In terms of optimization behavior, Figures 7(e) and 7(f) show that both the training and validation losses gradually decrease and tend to converge within 900 epochs. Compared with the baseline, the proposed model presents lower loss values and smoother convergence trends, indicating more stable training and better feature representation.

In summary, the proposed architecture consistently outperforms YOLOv11n across multiple key performance metrics, offering superior classification accuracy, segmentation quality, and model stability.

4.2 Comparative experiments

To comprehensively evaluate the performance advantages of the proposed improvements, a comparative analysis was conducted between the enhanced YOLOv12n model and several mainstream instance segmentation algorithms. The benchmark models include the two-stage method Mask R-CNN and four lightweight single-stage segmentation networks: YOLOv5n-seg, YOLOv8n-seg, YOLOv10n-seg, and YOLOv11n-seg. To ensure a fair comparison, all compared models were trained and evaluated using the same training, validation, and testing subsets. The input resolution, training epochs, batch size, optimizer settings, data augmentation strategy, and pretrained-weight strategy were kept consistent with those described in Section 3.2. The comparison is based on key evaluation metrics, including precision, recall, mAP@0.5, mAP@0.5:0.95, model

size, and inference speed. The experimental results are summarized in Table 3.

Table 3. Comparative experiment results

Algorithm	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Size/MB	FPS
YOLOv5n-seg	72.9%	70.8%	80.3%	53.4%	5.3	185
YOLOv8n-seg	76.4%	72.3%	81.2%	59.9%	6.5	128
YOLOv10n-seg	76.7%	73.5%	81.1%	58.3%	6.0	139
YOLOv11n-seg	77.2%	77.3%	83.6%	60.1%	5.6	162
Mask R-CNN	49.6%	40.2%	50.1%	30.8%	332.0	91
Proposed algorithm	**91.5%**	**84.7%**	**87.2%**	**64.8%**	**5.1**	**201**

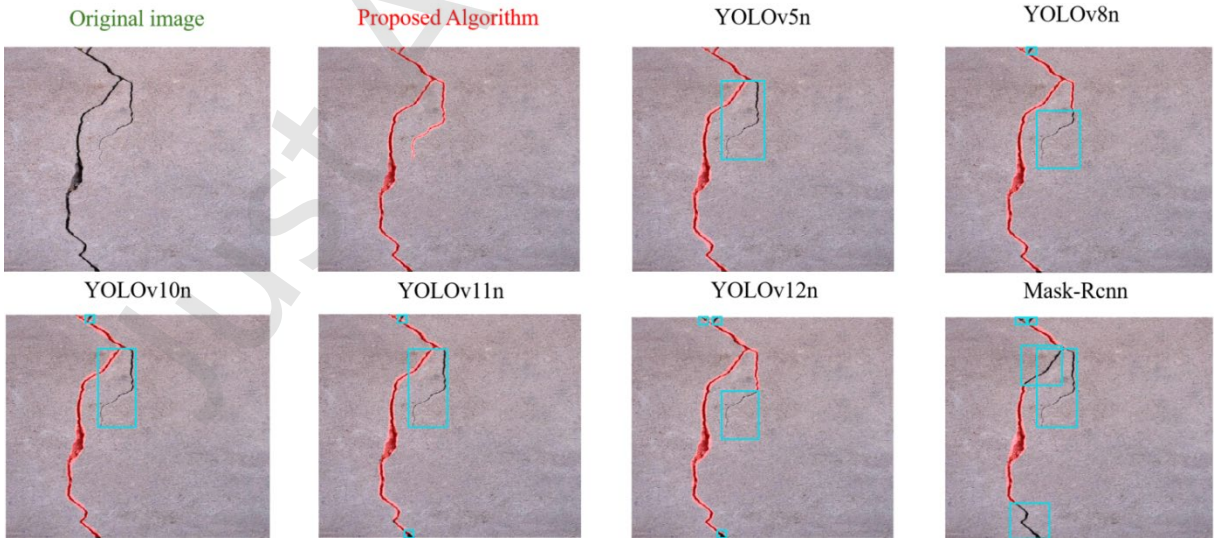
The results show that the proposed algorithm achieves the best overall performance across multiple indicators, with a precision of 91.5%, recall of 84.7%, mAP@0.5 of 87.2%, and mAP@0.5:0.95 of 64.8%. In particular, the model demonstrates a clear advantage in recall and mean average precision, reflecting its superior capability in identifying various types of façade defects comprehensively and accurately.

Regarding model complexity, the improved YOLOv12n has a parameter size of only 5.1 MB, which is the smallest among all single-stage models evaluated. It is slightly smaller than YOLOv5n-seg (5.3 MB) and significantly more compact than YOLOv8n-seg (6.5 MB) and YOLOv10n-seg (6.0 MB). Furthermore, it is vastly more lightweight than Mask R-CNN, which requires 332 MB. In terms of inference speed, YOLOv12n achieves 201 frames per second, making it the fastest among all models. This speed is approximately 1.57 times that of YOLOv8n-seg (128 FPS) and 2.2 times that of Mask R-CNN (91 FPS), demonstrating its strong capability for real-time processing.

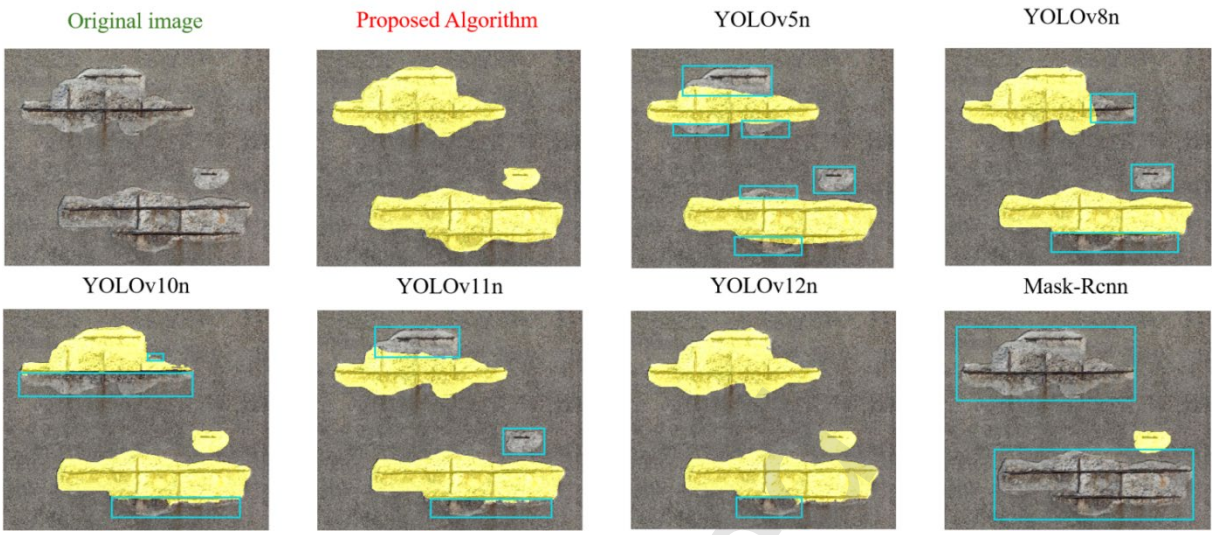
Considering the trade-offs between accuracy, computational efficiency, and deployment readiness, YOLOv12n strikes a favorable balance that makes it well-suited for real-world applications requiring both high detection accuracy and rapid response. Compared to the high-overhead, slower two-stage models such as Mask R-CNN and the less accurate lightweight models

such as YOLOv5n-seg, YOLOv12n delivers both efficiency and precision, offering greater practical value and deployment potential in engineering scenarios.

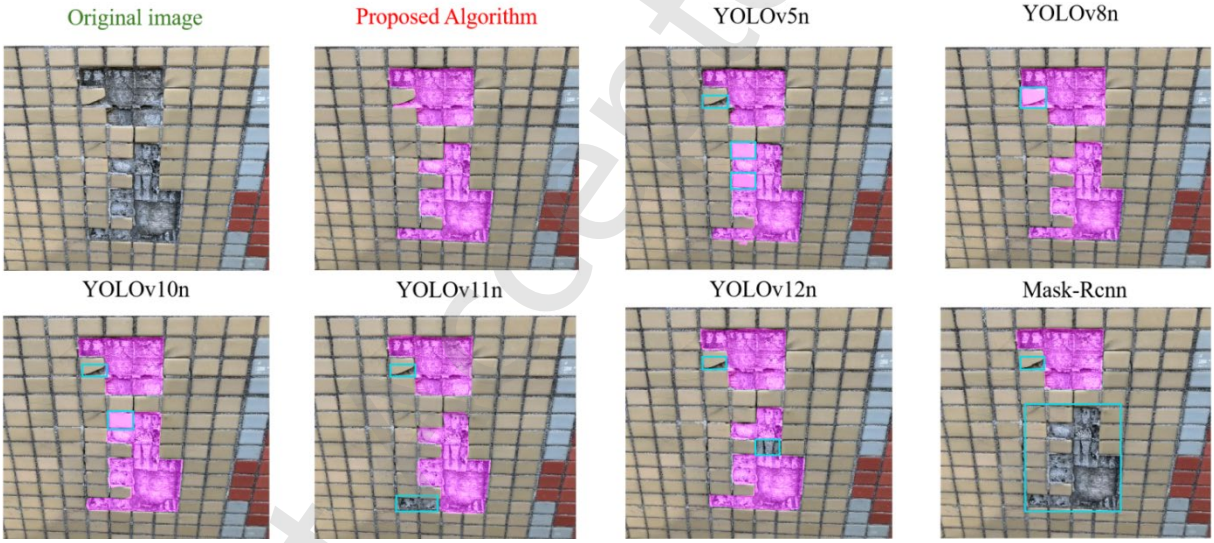
In addition, a qualitative comparison was conducted to evaluate segmentation performance across typical façade defect scenarios, as illustrated in Figure 8. The results reveal that most existing models struggle with challenging cases such as blurred boundaries, weak textures, and densely distributed multiple defect types. Common issues observed include imprecise contours, broken boundaries, missing fine details, and under-segmentation near edges. These problems primarily stem from interference introduced by large background regions during training, which suppresses the model's ability to learn discriminative features for small-scale defects such as concrete cracks and low-texture damage. Moreover, many baseline models lack effective mechanisms for multi-scale feature modeling and integration, limiting their segmentation accuracy and robustness in complex structural environments.



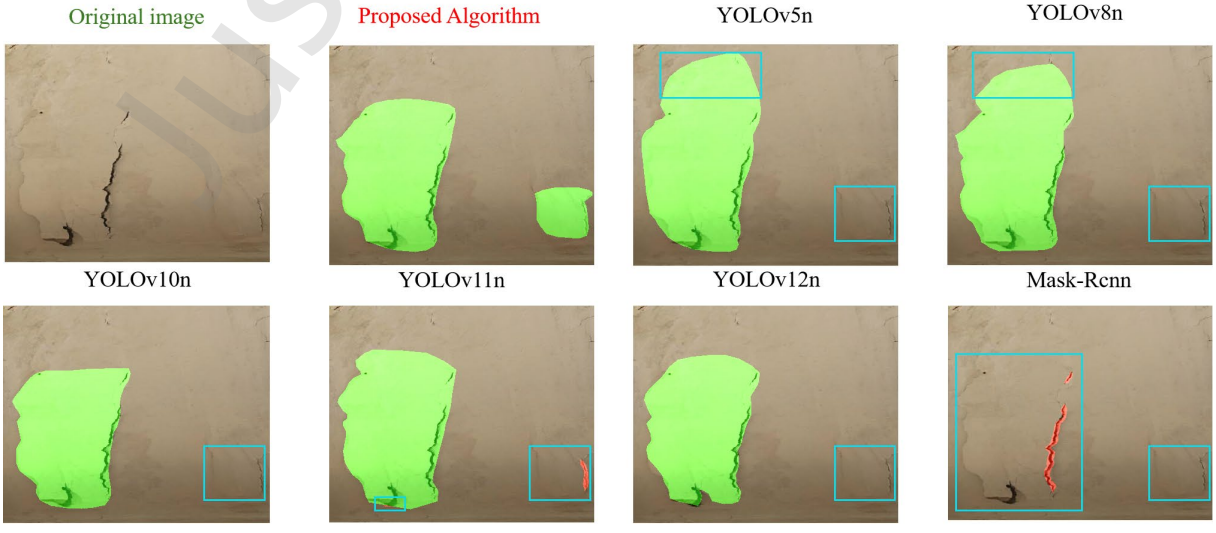
(a) Concrete cracks



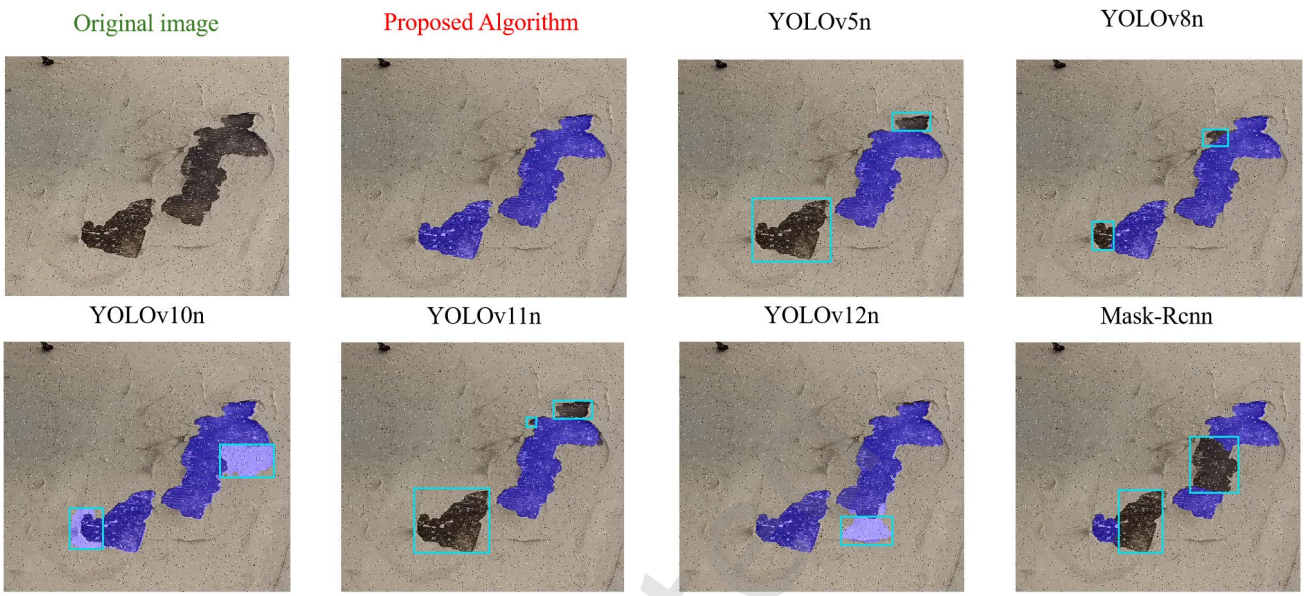
(b) Exposed reinforcement



(c) Tile loss



(d) Delamination



(e) Peeling

Figure 8. Segmentation performance of different algorithms (the cyan boxes show the false results of the mask compared with the the original image)

By contrast, the proposed algorithm addresses these challenges through several structural innovations. The integration of the edge-aware GEIT module improves boundary localization and enhances detail reconstruction. The hierarchical attention mechanism incorporated in HAFB boosts the model's responsiveness to diverse defect types and reduces class confusion. Furthermore, the shared segmentation head facilitates effective fusion of shallow texture and deep semantic information. These improvements collectively enable the model to maintain stable performance under diverse material textures and mixed defect scenarios. Additionally, multi-scale fusion and spatial refinement strategies embedded in the segmentation head further strengthen pixel-level spatial representation and detail continuity, resulting in more complete and accurate delineation of defect regions.

4.3 Limitation and Future Work

The main limitation of this study lies in the fact that the current model is developed and

evaluated using two-dimensional visible-light images with pixel-wise masks. Alth ough the proposed method can extract visible façade defect regions at the pixel level, the segmented masks cannot be directly converted into physical defect areas, crack widths, or severity grades without camera calibration, scale conversion, and engineering assessment procedures. Future work will therefore integrate scale calibration and field measurement information to support more reliable quantitative damage assessment.

Another limitation is that the current validation mainly covers images collected from Guangzhou and Hefei, together with selected public samples. Although these data include multiple façade materials and defect appearances, the model has not yet been systematically evaluated across unseen cities, building types, camera viewpoints, and inspection distances. Future work will involve expanding the field dataset to include more diverse façade scenes and difficult imaging conditions, such as strong reflection, occlusion, moisture, low illumination, and motion blur, thereby improving robustness and generalization in real inspection scenarios.

Future work will also prioritize the practical deployment of the proposed algorithm. The FPS values reported in this study were obtained under the workstation hardware configuration, and therefore indicate real-time potential under the reported hardware setting rather than verified embedded UAV-side performance. Further optimization and deployment tests will be conducted on drone-based or field inspection platforms, with consideration of image transmission, onboard computing capacity, power consumption, and integration with maintenance decision workflows.

5. Conclusions

This paper proposed a lightweightinstance segmentation algorithm for real-time detection of

multiple-class façade defects, built upon the YOLOv12n framework with targeted architectural enhancements. Extensive ablation and comparative experiments were conducted to validate the effectiveness, efficiency, and deployability of the proposed model. The key conclusions are summarized as follows:

(1) A high-quality façade defect segmentation dataset was constructed, comprising 6156 pixel-level annotated images. The dataset covers five typical types of structural damage commonly found on building exteriors, including concrete cracks, exposed reinforcement, tile loss, peeling, and delamination. The images were sourced from urban environments in Guangzhou and Hefei, supplemented by publicly available datasets, with all annotations verified by domain experts.

(2) The proposed algorithm incorporates three task-oriented lightweight modules: the GEIT module for enhanced edge perception, the HAFB module for hierarchical attention-based multi-scale fusion, and the LSCSH module for compact and adaptive segmentation. These modules adapt and integrate existing edge extraction, attention, re-parameterization, normalization, and shared-convolution mechanisms to improve segmentation accuracy, detail recovery, and computational efficiency for façade defect scenarios.

(3) Ablation studies demonstrated that each module independently contributes to performance gains, while their combined use further enhances overall effectiveness. The final integrated model achieved a precision of 91.5%, recall of 84.7%, mAP@0.5 of 87.2%, and mAP@0.5:0.95 of 64.8%, with a model size of only 5.1 MB and an inference speed of 201 FPS, striking an optimal balance among accuracy, speed, and model compactness.

(4) Comparative experiments with five state-of-the-art models (i.e., Mask R-CNN, YOLOv5n-seg, YOLOv8n-seg, YOLOv10n-seg, and YOLOv11n-seg) confirmed the superiority of the

proposed method in terms of segmentation performance, boundary precision, and inference efficiency, especially under complex urban conditions involving mixed defect types and textured backgrounds.

The proposed segmentation framework offers a practical solution for the automatic detection of visible façade defects in buildings. It enables pixel-level extraction of defect regions and provides useful spatial information for subsequent defect review, inspection prioritization, and maintenance planning. The framework demonstrates strong potential for real-time façade inspection under the reported hardware configuration and can serve as a useful technical basis for future urban infrastructure maintenance systems.

Acknowledgment

This study was supported by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2026A1515010049), the Young Science and Technology Talent Cultivation Program of the Guangdong Association for Science and Technology (Grant No. SKXRC2026620), and the Guangdong Provincial Key Laboratory of Intelligent Disaster Prevention and Emergency Technologies for Urban Lifeline Engineering (Grant No. 2022B1212010016).

References

Alif M A R, Hussain M. Yolov12: A breakdown of the key architectural features. arXiv preprint arXiv:2502.14740, 2025.

Cao M T. Drone-assisted segmentation of tile peeling on building façades using a deep learning model. Journal of Building Engineering, 2023, 80: 108063.

Chai J, Zeng H, Li A, Nagi E. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications, 2021, 6: 100134.

Chen K, Reichard G, Xu X, Akanmu A. Automated crack segmentation in close-range building façade inspection images using deep learning techniques. Journal of Building Engineering, 2021, 43: 102913.

Dais D, Bal I E, Smyrou E, Sarhosis V. Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning. Automation in Construction, 2021, 125: 103606.

Guo J, Wang Q, Li Y. Evaluation-oriented façade defects detection using rule-based deep learning method. Automation in Construction, 2021, 131: 103910.

Huang Q, Wan F, Lei G, Xu L. Detection of Building Façade Damage Based on Improved YOLOv8. 2023 5th International Conference on Frontiers Technology of Information and Computer (ICFTIC). IEEE, 2023: 1065-1070.

Hussain T, Li Y, Ren M, Li J. Pixel-level crack segmentation and quantification enabled by multi-modality cross-fusion of RGB and depth images. Construction and Building Materials, 2025, 487: 141961.

Idjaton K, Desquesnes X, Treuillet S, Brunetaud X. Transformers with yolo network for damage detection in limestone wall images. International conference on image analysis and processing. Cham: Springer International Publishing, 2022: 302-313.

Interlando M, Pacifico MG, Novellino A, Pastore VP. Ensembles of deep neural networks for the automatic detection of building facade defects from images. IEEE Access, 2024, 12: 164953-164965.

Junior G S, Ferreira J, Millán-Arias C, Daniel R, Junior A C, Fernandes B J T . Ceramic cracks segmentation with deep learning. Applied Sciences, 2021, 11(13): 6017.

Kanopoulos N, Vasanthavada N, Baker R L. Design of an image edge detection filter using the Sobel operator. IEEE Journal of solid-state circuits, 1988, 23(2): 358-367.

Kyem B A, Asamoah J K, Aboah A. Context-cracknet: A context-aware framework for precise segmentation of tiny cracks in pavement images. Construction and Building Materials, 2025, 484: 141583.

Lee K, Hong G, Sael L, Lee S, Kim H Y. MultiDefectNet: Multi-class defect detection of building façade based on deep convolutional neural network. Sustainability, 2020, 12(22): 9785.

Li F X, Lai J A, Liu Y R, Ling Y, Pan R Y, Wang L S, Wan J L ,Pang Y. Crack detection method for building exterior walls based on YOLOv5. Practical Electronics,2024,32(23):64-67.

Linyi W, Jing B, Wenjing L, Jinzhe J. Research Progress of YOLO Series Target Detection Algorithms. Journal of Computer Engineering & Applications, 2023, 59(14).

Lourenço T, Matias L, Faria P. Anomalies detection in adhesive wall tiling systems by infrared thermography. Construction and Building Materials, 2017, 148: 419-428.

Luo D M, Xie J K, Li F,Liu H T.Concrete structural apparent damage detection method based on improved YOLOv5 algorithm.Journal of Xi'an University of Architecture & Technology,2025,57(02):159-166.

Perez H, Tah J H M, Mosavi A. Deep learning for detecting building defects using convolutional neural networks. Sensors, 2019, 19(16): 3556.

Perez H, Tah J H M. Deep learning smartphone application for real-time detection of defects in buildings. Structural Control and Health Monitoring, 2021, 28(7): e2751.

Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.

Shao Y. Local-Global Attention: An Adaptive Mechanism for Multi-Scale Feature Integration. arXiv preprint arXiv:2411.09604, 2024.

Soudy M, Afify Y, Badr N. RepConv: A novel architecture for image scene classification on Intel scenes dataset. International Journal of Intelligent Computing and Information Sciences, 2022, 22(2): 63-73.

Tian Y, Ye Q, Doermann D. Yolov12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524, 2025.

Vairalkar M K, Nimbhorkar S U. Edge detection of images using Sobel operator. Int. J. Emerg. Technol. Adv. Eng, 2012, 2(1): 291-293.

Wang P, Xiao J, Qiang X, Xiao R, Liu Y, Sun C, Hu J, Liu S. An automatic building façade deterioration detection system using infrared-visible image fusion and deep learning. Journal of Building Engineering, 2024, 95: 110122.

Wang Y, Ahsan U, Li H, Hagen M. A comprehensive review of modern object segmentation approaches. Foundations and Trends in Computer Graphics and Vision, 2022, 13(2-3): 111-283.

Zhang X, Ye P, Leung H, Gong K, Xiao G. Object fusion tracking based on visible and infrared images: A comprehensive review. Information Fusion, 2020, 63: 166-187.

Zhao M, Yan M, Li T. Vision-based positioning: Related technologies, applications, and research challenges. 2018 IEEE 9th International Conference on Software Engineering and Service

Science (ICSESS). IEEE, 2018: 531-535.

Zhou K, Shi J L, Fu J Y, Zhang S X, Liao T, Yang C Q, Wu J R, He Y C. An improved YOLOv10 algorithm for automated damage detection of glass curtain-walls in high-rise buildings. Journal of Building Engineering, 2025, 101: 111812.