sustainability

MDPI

Article

SAW-YOLOv8l: An Enhanced Sewer Pipe Defect Detection Model for Sustainable Urban Drainage Infrastructure Management

Linna Hu [1,*], Hao Li [1], Jiahao Guo [1], Penghao Xue [1], Weixian Zha [1], Shihan Sun [1], Bin Guo [2] and Yanping Kang [3]

[1] School of Network and Communication Engineering, Jinling Institute of Technology, Nanjing 211169, China; 2216031053@stu.jit.edu.cn (H.L.); 2216031072@stu.jit.edu.cn (J.G.); 2407010018@stu.jit.edu.cn (P.X.); 2507010001@stu.jit.edu.cn (W.Z.); 2416901073@stu.jit.edu.cn (S.S.)
[2] College of Computer and Information Engineering, Xinjiang Agricultural University, Urumqi 830052, China; gb@xjau.edu.cn
[3] Jiangsu Hydraulic Research Institute, Nanjing 210017, China; kyp219@126.com
* Correspondence: hulnnj@jit.edu.cn

Abstract

Urban underground sewage pipelines often suffer from defects such as cracks, irregular joint misalignment, and stratified sedimentation blockages, which may lead to pipeline bursts, sewage overflow, and water pollution. Timely detection of abnormal defects in sewage pipelines is critical to ensuring public health and environmental sustainability. Vision-based sewage pipeline defect detection plays a crucial role in modern urban wastewater treatment systems. However, it still faces challenges such as limited feature extraction capabilities, insufficient multi-scale defect characterization, and poor positioning stability when dealing with low-contrast images and in environments with severe background interference. To address this issue, this study proposes an enhanced SAW-YOLOv8l model that integrates RT-DETR (real-time detection Transformer) with CNN (convolutional neural network) architecture. First, a C2f_SCA module improves the long-distance feature extraction capability and localization precision. Second, an AIFI-PRBN module enhances global feature correlation through attention-mechanism-based intra-scale feature interaction and reduces computational complexity using lightweight techniques. Finally, an adaptive dynamic weighted loss function based on Wise-IoU (weighted intersection over union) further improves training convergence and robustness by balancing the gradient distribution of samples. Experiments on a mixed dataset comprising Sewer-ML and industrial images demonstrate that the SAW-YOLOv8l model achieved mAP@0.5 of 86.2% and precision of 84.4%, which were improvements of 2.4% and 6.6% respectively over the baseline model, significantly enhancing the detection performance of abnormal defects in sewage pipelines.

Keywords: sewer pipe defect detection; YOLOv8l; RT-DETR; C2f_SCA; AIFI_PRBN; Wise-IoU; environmental sustainability

Check for updates

Academic Editor: Vasilis Kanakoudis

Received: 27 February 2026
Revised: 25 March 2026
Accepted: 3 April 2026
Published: 8 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The urban sewage pipeline system constitutes an indispensable component of modern urban infrastructure [1]. It plays a vital role in ensuring the safety and sustainability of urban environments by collecting, conveying, and discharging wastewater. However, over time and under the influence of factors such as corrosion, ground settlement, and hydraulic impacts, sewage pipelines gradually deteriorate and exhibit various defects, including pipe disconnections, sediment accumulation, and concealed connections of branch pipes [2]. To

better understand the causes of sewer pipeline defects and the importance of early-stage detection, it is essential to clarify the underlying deterioration mechanisms and their impacts on operation, economy, and safety. Sewer pipeline deterioration is not caused by a single factor; rather, it results from the long-term coupling of geotechnical and structural effects, hydraulic effects, corrosion processes, and ecological influences, and is further exacerbated by material aging and operating conditions. Differential settlement, void development, and traffic loading can induce joint opening, pipe deformation, and crack propagation, thereby increasing the probability of disconnection and leakage. Long-term overloading and wet–dry alternation can intensify leakage, sediment deposition, and scouring, while infiltration and exfiltration may cause soil loss and cavity enlargement. Corrosive gases and aggressive wastewater constituents weaken the pipe wall, and microbial action further accelerates material degradation. Tree-root intrusion through joints or micro-cracks can lead to blockage and even rupture. These mechanisms reinforce each other, driving small initial defects to evolve into significant structural and functional anomalies. Without timely detection and repair, such defects may reduce conveyance capacity, increase maintenance costs, raise the risk of service interruption, and pose hazards such as collapse and environmental contamination. However, pipeline defect detection still mainly relies on manual interpretation of CCTV inspection videos [3], which is time-consuming, labor-intensive, and subjective. Moreover, vision-based sewage pipeline defect detection still faces challenges such as limited long-distance feature extraction capabilities, insufficient multi-scale defect representation, and poor positioning stability. This paper proposes an accurate and efficient method for sewage pipeline defect detection to ensure the long-term safe and reliable operation of the sewage pipeline system.

The latest advancements in the field of deep learning, particularly convolutional neural networks (CNNs) and attention mechanism-based Transformer architectures, have driven steady progress in the detection of defects in sewage pipelines. These models demonstrate strong capabilities in feature representation, object localization, and real-time inference, enabling intelligent detection systems to replace or complement manual assessment. Convolutional neural networks (CNNs) extract local texture and edge cues through stacked convolutional operations to capture crack contours and corrosion morphologies. Their hierarchical feature maps can provide multi-scale descriptors for defect patterns [4,5]. Transformer-based architectures leverage self-attention to model long-range dependencies and contextual relations between defect regions and their surroundings, thereby improving robustness under complex illumination and background clutter [6,7]. Furthermore, lightweight models (e.g., MobileNetV2 and EfficientNet) combined with modern vision Transformer (ViT) variants enable real-time inference on embedded or edge hardware, meeting the throughput requirements for sewer pipe inspection [8,9].

However, there are still challenges in handling low-contrast images, severe background interference, and multi-scale feature extraction. For instance, when the environment inside the sewage pipeline undergoes changes in light, this can result in low image contrast or overexposure of the lens, and the details of the defect are easily obscured by noise; background interference factors such as water stains, deposits, and rust on the pipeline wall can conceal the true defect; dampness inside the pipeline can cause the lens to fog up, causing blurred vision and making it impossible to detect the defect normally, thereby increasing a model's false detection rate. Additionally, the size of sewage pipeline defects varies significantly, ranging from small pinhole corrosion to large structural cracks. How to effectively integrate and extract multi-scale features remains a current research challenge. Therefore, it is necessary to further study hybrid architectures that can effectively integrate global semantic reasoning and perceptions of local texture.

Recent improvements based on YOLOv8 are mostly local plug-in type optimizations, such as simply replacing the backbone network, inserting attention modules, or adjusting the detection head. Although these methods can enhance performance in specific scenarios, they fail to break through the inherent receptive field constraints of convolution operations and still rely on multiple layers of stacking to achieve long-distance feature correlation. This leads to two core problems: first, the feature correlation of small targets is poor, and there is a tendency to miss detections when observing at long distances; second, the robustness to low-quality samples (such as changes in lighting, occlusion, and blurriness) is insufficient. Existing RT-DETR hybrid schemes mostly consist of shallow layer concatenations at the architectural level, such as simply combining the CNN backbone with the Transformer detection head. These methods generally have a performance and efficiency contradiction that involves relying on LayerNorm during training and batch normalization during inference, making them difficult to deploy in real-time detection scenarios.

To address the aforementioned issues specifically, in view of the unique challenges encountered in the detection scenarios of sewer pipelines, such as long-distance observation, complex background interference, and low sample quality, this study constructed a hybrid detection architecture integrating convolutional neural network (CNN) and Transformer methods, and the SAW-YOLOv8l sewage pipeline defect detection model is proposed. This model simultaneously enhances the ability to perceive global semantic features and extract local detailed features. The core innovation points of this study are as follows.

1. The C2f_SCA module for long-distance context modeling was constructed. The high-dimensional nonlinear mapping capability of the Star Block module was introduced to enhance feature representation, while the context-aware attention (CAA) mechanism was leveraged to enable dynamic anchor point localization and cross-region contextual modeling. This addressed the limitation of the C2f module in YOLOv8, which failed to effectively capture the global context of long-distance targets (e.g., pipeline misalignment and interface separation), thereby significantly improving performance for feature extraction in sewage pipe networks.

2. An AIFI-PRBN module was designed to enhance global feature interaction via an attention-based same-scale feature interaction structure, while computational complexity was reduced using progressive re-parameterized batch normalization (PRBN). This module resolved two critical issues: the SPPF module in YOLOv8 lacking feature interaction ability, and the high computational cost and low real-time performance of the AIFI module. This achieved an efficient collaboration mechanism between local modeling ability and local perception ability and was adapted to the real-time requirements of sewer pipe defect detection.

3. An improved novel Wise-IoU loss optimization method automatically adjusts the gradient contribution of samples of different qualities through a learnable dynamic focusing coefficient, thereby improving the precision of bounding box regression. This method addresses the issues of gradient instability and reduced localization precision when processing low-quality samples and enhances the model's robustness and convergence performance in complex pipeline defect detection scenarios such as occlusion, blurriness, overexposure of the shot, and scale variation.

The core innovation of this study is the construction of a hybrid mechanism that integrates CNN and Transformer architecture. At the architectural level, the C2f_SCA and AIFI-PRBN modules are used to balance computational efficiency and training stability; at the learning theory level, the Wise-IoU loss function is adopted to achieve dynamic gradient adjustment of sample quality; at the design methodology level, the three improvements are coordinated with each other to form a progressive collaborative optimization

framework that includes local feature extraction, global semantic association, and adaptive loss weight adjustment.

2. Related Work

Recent advancements in the field of object detection have achieved remarkable progress driven by deep learning architectures. Representative convolutional neural network (CNN)-based detectors, such as two-stage detectors (e.g., R-CNN, Faster R-CNN and Mask R-CNN) [10–12] and single-stage frameworks of the YOLO series [13–15], have been widely applied to various visual tasks, including sewer pipe defect detection. Concurrently, end-to-end Transformer-based detectors, such as RT-DETR and its variants [16,17], have emerged, integrating global feature modeling and attention mechanisms to enable real-time detection with enhanced semantic understanding capabilities. These developments collectively form the technical foundation of modern intelligent detection systems, providing a viable pathway from local feature extraction to global contextual reasoning in complex visual environments.

2.1. Defect Detection of Sewer Pipes Based on Convolutional Neural Network

Kumar et al. employed the region-proposal-based detector Faster R-CNN variant proposed in ref. [18] for sewage pipeline detection. This method combined efficient object detection via region proposal with convolutional neural networks, achieving an improvement in average precision on a dataset containing four types of defects. However, owing to its intricate architecture and large number of parameters, the method demands substantial computing resources, which seriously affects the inference rate for live video processing. In contrast, single-stage object detection models like YOLO perform end-to-end bounding box regression and classification of objects categories without extracting candidate regions [19], thereby demonstrating superiority in processing real-time detection video streams of pipelines. In ref. [20], Jia et al. upgraded the YOLOv5 model by improving the backbone network, thereby enhancing the ability to identify small defects in the pipeline; however, the types of defects that can be detected are still limited. In ref. [21], Lu et al. employed StyleGAN2 for image preprocessing, integrated an enhanced feature fusion layer into YOLOX, and replaced the traditional loss function with CIoU to optimize the precision of bounding box regression, thereby achieving relatively high detection precision for defects such as deposition, leakage, and misalignment. However, this method is limited in terms of detectable categories and is unable to locate the specific position of defects. In ref. [22], Ren proposed an improved YOLOv8 algorithm for underground drainage pipeline defect detection, designed the SC_C2f module and LSKA_SPPF, and introduced WIoU (Wise-IoU) loss function to mitigate interference from irrelevant pipeline features during detection, so as to enhance discriminative capacity and locate pipeline defects. However, this model is complex, and the actual deployment cost is relatively high. Zhou et al. designed the DSK Block as a substitute for the Bottleneck block in C2f [23], improving the multi-scale feature extraction capability of the model for defects in drainage networks. This provides a computational complexity that meets the demands of real-time detection, but the precision of detection is not high. The YOLO series models satisfy the demands for the real-time detection in sewage pipelines, but their capability to capture the global features of sewage pipes under complex background conditions requires further improvement. These methods are compared in Table 1.

Table 1. Literature comparison.

Algorithm	Reference	Main Contribution	Detection Types	Advantages	Limitations
Faster R-CNN Variant	[18]	Combines region proposal and CNN for sewer defect detection	4 types	High precision (83% AP)	Complex structure; large parameters; low inference rate
Improved YOLOv5	[20]	Backbone modification for small-defect enhancement	5 types	Better small-object recognition	Limited defect categories
YOLOX+StyleGAN2+CIoU	[21]	Image preprocessing with GAN; enhanced feature fusion; CIoU loss	3 types (sediment, leakage, misalignment)	Improved regression precision	Cannot locate specific defect positions
Improved YOLOv8 (SC_C2f+LSKA_SPPF+WIoU)	[22]	Introduces Wise-IoU loss; enhances background discrimination	Multiple sewer defects	Strong generalization; robust to noise	Complex model; high deployment cost
YOLOv8+DSK Block (C2f substitute)	[23]	Replaces Bottleneck with DSK Block for multi-scale extraction	Multiple defects	Meets real-time detection demands	Precision not high enough
DefectTR (Transformer-based DETR)	[24]	Pure Transformer; uses Color Jitter, Leaky ReLU, CIoU	10 categories	Outperforms YOLOv4 and Faster R-CNN; mAP 60.2%	Weak for sub-mm defects; high complexity
RT-DETR	[25–28]	Hybrid encoder–decoder; auxiliary prediction heads	10+ types	Real-time performance; strong global modeling	Needs further tuning for small targets
YOLOv8+AIFI module	[29–32]	AIFI replaces SPPF for better feature interaction	Multiple	Enhanced feature fusion	Added computation overhead

Although traditional CNN-based object detectors have demonstrated excellent performance in numerous visual tasks, they exhibit significant limitations. On the one hand, the deficiency of CNNs in global feature extraction makes the effective fusion of global image semantics challenging. On the other hand, the inherent localized receptive field of CNN constrains its capacity to establish global dependencies between target pixels.

2.2. Transformer-Based Detectors and Training Enhancements

The Transformer architecture can dynamically model long-distance pixel dependencies through self-attention mechanisms, significantly enhancing global feature representation capabilities. This addresses the inherent limitations of CNN architectures and substantially improves the perception and understanding abilities of object detectors in complex scenarios.

Dang et al. proposed DefectTR, a detection framework employing the pure Transformer architecture based on DETR in [24]. This method optimizes the model performance by introducing ColorJitter data augmentation, the LeakyReLU activation function, and the CIoU loss function. On a dataset containing 47,100 annotated images covering 10 distinct defect categories, DefectTR achieved a mean average precision (mAP) of 60.2%, outperforming mainstream object detection architectures such as YOLOv4 and Faster R-CNN. However, DefectTR still has limitations in terms of performance for sub-millimeter-level defect detection, and it has a high computational complexity that requires further optimization. In ref. [25], the RT-DETR architecture was proposed, integrating three key components: an augmented Transformer decoder with auxiliary prediction heads, a hybrid encoder for efficient feature refinement, and a backbone network for preliminary feature extraction. By leveraging the Transformer, RT-DETR achieves global modeling capabilities while maintaining competitive precision and ensuring real-time performance in end-to-end detection. In ref. [26], a CMS-RTDETR based on RT-DETR was proposed. This method replaced the original ResNet18 network with a backbone network called CSP-MSEE, which improved the detection precision. It also introduced mechanisms such as multi-head self-attention modules to enhance robustness in complex backgrounds. In ref. [27], a steel pipe weld defect detection method based on XXX-DETR was proposed. By incorporating CSPMogaBlock, RepBN, and CSP-MSDC modules, the method enhances feature extraction and multi-scale defect modeling capabilities while maintaining computational efficiency. In ref. [28], an improved RT-DETR–based approach for real-time pipeline defect detection was reported, employing a lightweight backbone and incorporating enhanced adaptive feature fusion, dilated convolution, and structural re-parameterization to strengthen defect representation. The method shows notable improvements in both precision and efficiency compared with the original RT-DETR. The AIFI module is a core component of the efficient hybrid encoder in the RT-DETR model. The AIFI module is primarily applied to real-time object detection tasks and represents a key innovation for achieving a balance between real-time performance and high precision in the RT-DETR model. However, the self-attention mechanism of the AIFI module has a time complexity bottleneck. If applied directly to multi-scale feature scenarios, it leads to exponential growth in computational load, so external modules need to be relied upon for compression and optimization. The YOLOv8+AIFI framework has demonstrated significant effectiveness in small target detection scenarios, including unmanned aerial vehicle-based maritime small target detection, aero-engine component detection, citrus maturity detection, and wildlife target detection [29–32]. However, its potential in the field of pipeline defect detection remains to be thoroughly explored. These methods are compared in Table 1.

3. Methods

3.1. Framework Overview of YOLOv8

YOLOv8 crafted by Ultralytics evolves the YOLO lineage [33] through architectural refinements rather than a wholesale reinvention. Its blueprint remains tripartite; a detection head crowns a feature-pyramid neck, which in turn surmounts a backbone charged with initial representation learning. Model capacity is modulated across five tiers: extra-large (x), large (l), medium (m), small (s), and nano (n). Each variant differs in depth, layer count, and channel width. The C2f module in YOLOv8 is used to substitute the C3 block in YOLOv5 [34] in the backbone model design to optimize backpropagation dynamics and enhances its adaptability to complex backgrounds and noise. YOLOv8 significantly improves detection performance through a series of optimizations, including integrating modules such as C2f, SPPF (spatial pyramid pooling features), and MaxPool (max pooling), thus becoming an efficient and powerful object recognition tool. The network architecture of YOLOv8 is shown in Figure 1.

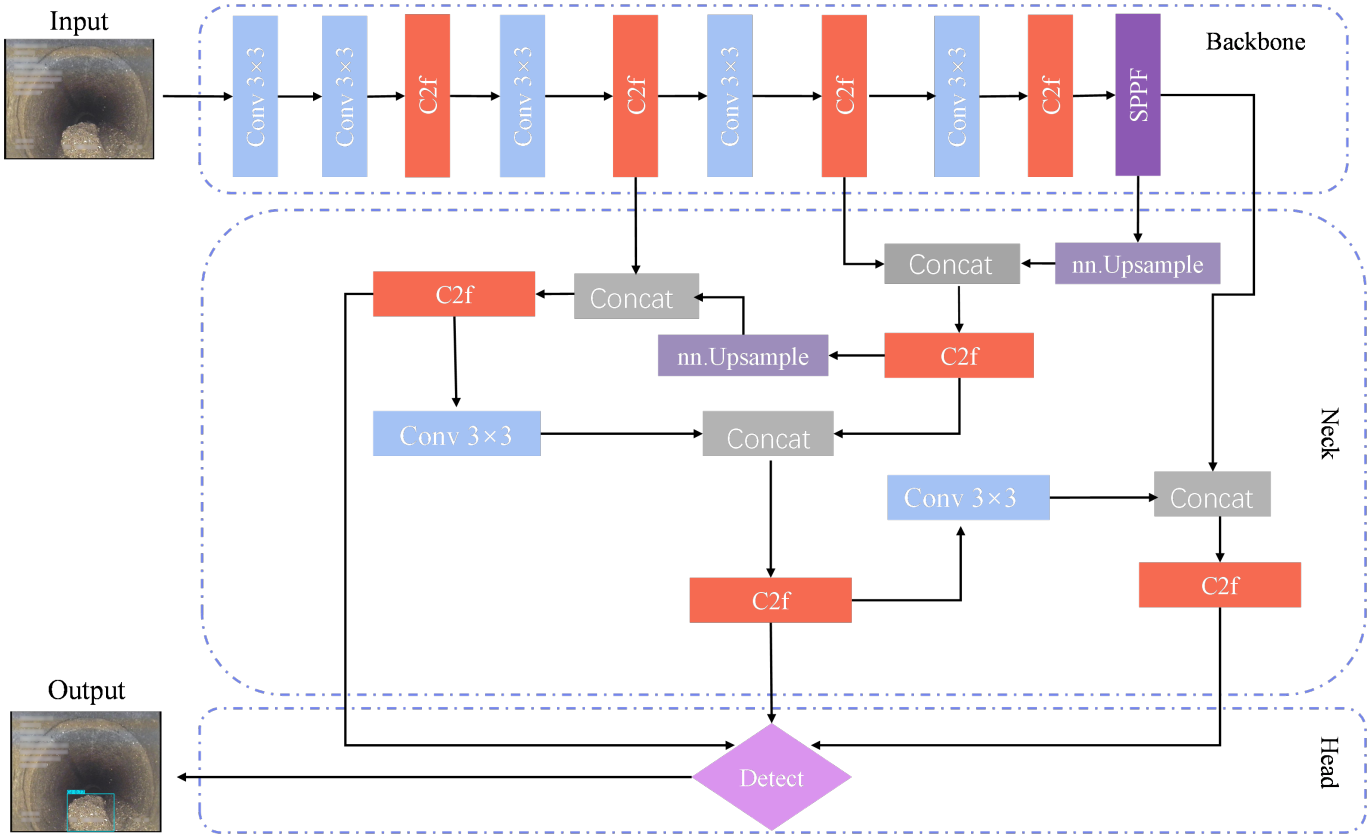


Figure 1. A Comprehensive Analysis of YOLOv8 Architecture.

3.2. The Proposed SAW-YOLOv8l Method

To address the challenges of visual similarity, image homogeneity, low-light conditions, and complex backgrounds in sewer inspection, this study builds upon YOLOv8l and proposes SAW-YOLOv8l through a progressive collaborative optimization strategy with three key upgrades. First, the original C2f module is upgraded to C2f_SCA to enhance fine-detail extraction, local feature discrimination, and long-distance contextual modeling for weak and distant defects. However, although this modification strengthens the convolutional representation path, explicit semantic interaction between defect features and their surrounding environment remains insufficient. Therefore, the original SPPF module is replaced with AIFI_PRBN to introduce same-scale attention-based feature interaction,

thereby improving global semantic correlation, enhancing discrimination under complex backgrounds, and reducing computational overhead. Meanwhile, because stronger attention interaction may amplify regression instability caused by sample quality differences, the original CIoU loss is further replaced with Wise-IoU based on dynamic focusing weights, to mitigate the influence of low-quality samples, stabilize localization training, and accelerate convergence. Through the coordinated optimization of feature representation, semantic interaction, and localization learning, SAW-YOLOv8l achieves more effective multi-scale defect recognition and more accurate localization in challenging sewer pipe environments. The overall architecture of SAW-YOLOv8l is illustrated in Figure 2.

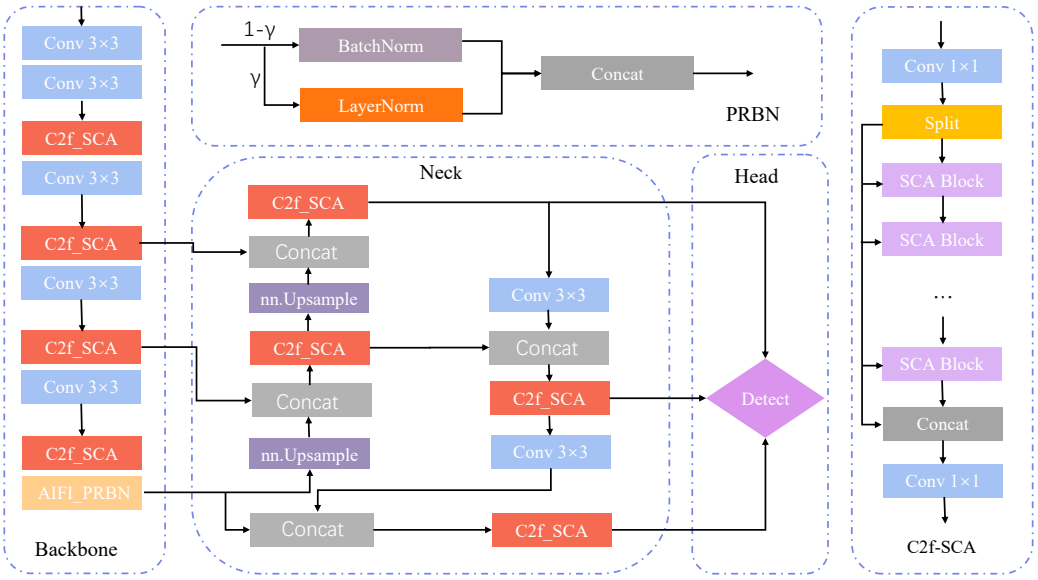


Figure 2. System architecture diagram for SAW-YOLOv8l. The model integrates the C2f_SCA module in the backbone and neck for enhanced long-distance context modeling, replaces the traditional SPPF with the AIFI_PRBN module to improve global feature interaction while maintaining computational efficiency, and utilizes the Wise-IoU loss in the detection head for accurate defect localization.

For the SAW-YOLOv8l model proposed in this study, a progressive collaborative optimization approach was adopted. Firstly, we improved C2f to C2f_SCA to enhance the ability of local feature discrimination and long-distance context modeling; however, this mainly strengthened the convolutional path, and the explicit semantic interaction between scale-specific defects and the environment was still insufficient. Therefore, the SPPF module was replaced with AIFI-PRBN to introduce attention-based intra-scale interaction. At the same time, attention enhancement may amplify the regression instability caused by differences in sample quality. Thus, we further added Wise-IoU dynamic re-balancing gradient to stabilize the localization training, enabling the representational and semantic advantages obtained in the previous two steps to be reliably transformed into more accurate defect localization.

3.3. The Improved C2f_SCA Module

Under long-distance observation conditions, reliable defect perception in pipeline inspection requires discriminative feature encoding and effective modeling of long-distance dependencies. Although the C2f in YOLOv8l alleviates gradient vanishing and explosion by repeatedly stacking bottleneck units, this strategy constitutes an inherent structural drawback, as excessive stacking tends to accumulate redundant and weakly relevant representations in hierarchical feature maps, ultimately impairing recognition precision. In sewer inspection, many defects such as misalignment, crack-like separation, and penetrating

objects are observed from long-distance viewpoints and are often surrounded by repetitive pipe-wall textures, making them difficult to distinguish using purely stacked convolutional bottlenecks. We therefore selected a design that combines implicit feature-space expansion and context-aware attention, rather than simply deepening or widening the convolutional block. To eliminate this limitation, a redesigned C2f architecture, termed C2f_SCA, has been constructed by replacing conventional bottleneck structures with the proposed SCA_Block, as illustrated in Figure 3. The SCA_Block fuses the context anchor attention (CAA) mechanism [35] with the Star_Block framework [36]. The reason for selecting this combination is twofold; the Star-based projection strengthens nonlinear feature transformation and cross-channel interaction without significantly increasing width, which is beneficial for weak defect texture representation; meanwhile, CAA introduces anchor-guided contextual aggregation, which helps the network preserve defect cues under long-distance observation and background interference. By jointly leveraging latent-space expansion and context-aware attention modeling, the proposed C2f_SCA module substantially enhances the model's capability for long-distance feature extraction under long-distance observation conditions, thereby improving detection robustness and overall recognition performance.

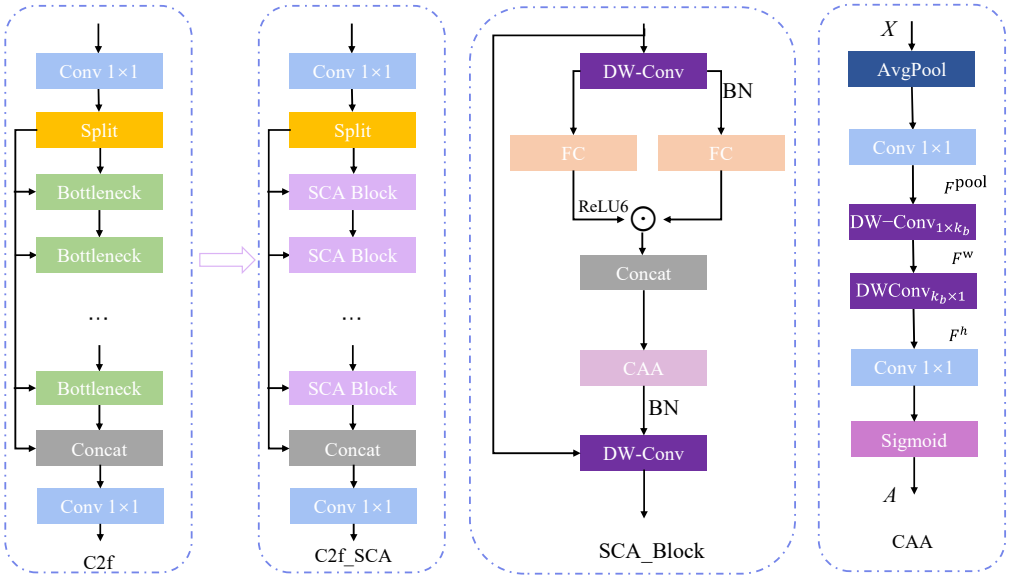


Figure 3. The network structure of C2f_SCA (⊙ Representative Star Operation).

The SCA_Block integrates star-based feature projection and CAA-based contextual aggregation. The star operation enhances latent feature interaction without significantly increasing width, whereas CAA strengthens long-distance contextual modeling for distant targets. The auxiliary operations, including depthwise convolution, normalization, and activation, follow standard implementations and mainly serve to support the proposed star-based projection and context-aware aggregation. As the core of the design, the star operation performs element-wise multiplication between features from two orthogonal subspaces, as shown in Equation (1):

$$\left(W_1^T X + B_1\right) \odot \left(W_2^T X + B_2\right) \tag{1}$$

where $W_1, W_2 \in \mathbb{R}^{(d+1)\times(d'+1)}$ are learnable projection matrices, and B_1, B_2 denote bias terms. The input feature is defined as $X \in \mathbb{R}^{(d+1)\times n}$, where d represents the channel dimension, n denotes the number of spatial locations (or tokens), and d' denotes the projected feature dimension after linear transformation. The additional dimension accounts

for the bias term by augmenting the input with a constant value of 1. Here, $\mathbb{R}$ denotes the set of real numbers.

The projection matrices map the input into two feature spaces of identical dimensionality, yielding $W_1^\top X$, $W_2^\top X \in \mathbb{R}^{d' \times n}$, which ensures dimensional compatibility for the element-wise multiplication operation $\odot$. This formulation follows the definition of the star operation [36].

The CAA module is introduced after the star operation to enhance feature discriminability for long-distance targets. The CAA module, originally proposed in PKINet [35], is a lightweight attention mechanism designed to capture long-distance contextual dependencies. The CAA mechanism first extracts global contextual information through the average pooling operation, then refines local regional features via 1×1 convolutional layer, with the final output F^{pool} formulated in Equation (2):

$$F_{l-1,n}^{pool} = Conv_{1\times1}\left(P_{avg}(X_{l-1,n})\right), n = 0, \ldots, N_l - 1 \tag{2}$$

where P_{avg} denotes the average pooling operation and $Conv_{1\times1}$ denotes the 1×1 convolutional layer.

The vertical and horizontal stripe depthwise convolutions are employed to simulate the effects of traditional large-kernel convolutions, achieving comparable feature extraction capabilities with significantly reduced parameter counts. These decomposed operations yield distinct outputs F^w and F^h as formulated in Equations (3) and (4), respectively:

$$F_{l-1,n}^{w} = DWConv_{1\times k_b}\left(F_{l-1,n}^{pool}\right) \tag{3}$$

$$F_{l-1,n}^{h} = DWConv_{k_b\times1}\left(F_{l-1,n}^{w}\right) \tag{4}$$

where $DWConv(\cdot)$ denotes depthwise convolution applied independently on each channel. The operation $DWConv_{1\times k_b}(\cdot)$ captures horizontal contextual dependencies, producing the intermediate feature $F_{l-1,n}^{w}$, while $DWConv_{k_b\times1}(\cdot)$ further aggregates vertical context, yielding the refined feature $F_{l-1,n}^{h}$.

Following the original PKINet design, the kernel size is defined as follows:

$$k_b = 11 + 2l \tag{5}$$

where l denotes the stage index. All depthwise convolutions are performed with stride 1 and padding $\lfloor k_b/2 \rfloor$, preserving spatial resolution. Through this decomposition, the effective receptive field approximately corresponds to that of a $k_b \times k_b$ depthwise convolution, while reducing the number of parameters from k_b^2 to $2k_b$. The use of depthwise convolution follows the standard formulation introduced in MobileNetV1 [37].

The CAA module dynamically adjusts the receptive field size to accommodate SCA_Blocks of different depths. The CAA adopts 1×1 convolutional layer and generates attention weight A by the function, as described in Equation (6):

$$A = \text{Sigmoid}\left(Conv_{1\times1}\left(F^h\right)\right) \tag{6}$$

The proposed C2f_SCA module is built upon the SCA_Block architecture, where the traditional bottleneck is replaced by the star operation to reduce redundant computation and enhance high-dimensional feature extraction from low-dimensional inputs. In addition, the integration of CAA strengthens long-distance contextual modeling and central feature representation. Together, these improvements enhance defect feature extraction and improve the efficiency and applicability of the module for sewer pipe defect detection.

3.4. The Proposed AIFI_PRBN Module

In sewer inspection images, many defects do not appear in isolation but are embedded in surrounding structures and contaminated backgrounds such as deposits, stains, corrosion textures, and illumination variation. Under these conditions, accurate recognition depends not only on local target texture, but also on whether the network can associate the target region with its surrounding semantic context at the same scale. However, the SPPF module in YOLOv8 lacks a feature interaction mechanism, making it difficult to effectively associate target information with contextual features, thereby negatively impacting detection precision. Based on this, instead of merely continuing to rely on SPPF to expand the receptive field, we introduced AIFI from RT-DETR to utilize the attention mechanism to achieve cross-feature interaction at the same scale, enabling the joint modeling of context evidence and defect evidence, thereby significantly improving the efficiency of feature extraction. However, the Transformer-based AIFI depends on LayerNorm, causing high computational overhead and reduced real-time performance. To address this limitation, this study further selected PRBN rather than directly retaining LayerNorm-based AIFI, because sewer defect detection requires not only stronger semantic interaction but also acceptable inference efficiency for engineering deployment. PRBN provides a transition from training-time multi-branch normalization to inference-time efficient BatchNorm, thereby preserving optimization stability while reducing inference cost. Therefore, the AIFI_PRBN module proposed in this paper incorporates progressive re-parameterized batch normalization (PRBN) for optimization [38]. Specifically, the LayerNorm layers in AIFI are replaced with PRBN, enabling multi-branch training to capture and fuse feature distribution characteristics. During inference, these branches are progressively merged into a single BatchNorm layer, thereby reducing computational complexity. By integrating global information modeling, the AIFI_PRBN module enhances detection precision while effectively balancing model speed and complexity, allowing the network to better utilize surrounding environmental pixels and reduce false positives.

The network architecture of the AIFI_PRBN module is shown in Figure 4. The AIFI_PRBN module primarily contains three major modules: the multi-head attention mechanism [39], the multi-layer perceptron (MLP) [40], and the progressive re-parameterized batch normalization (PRBN).

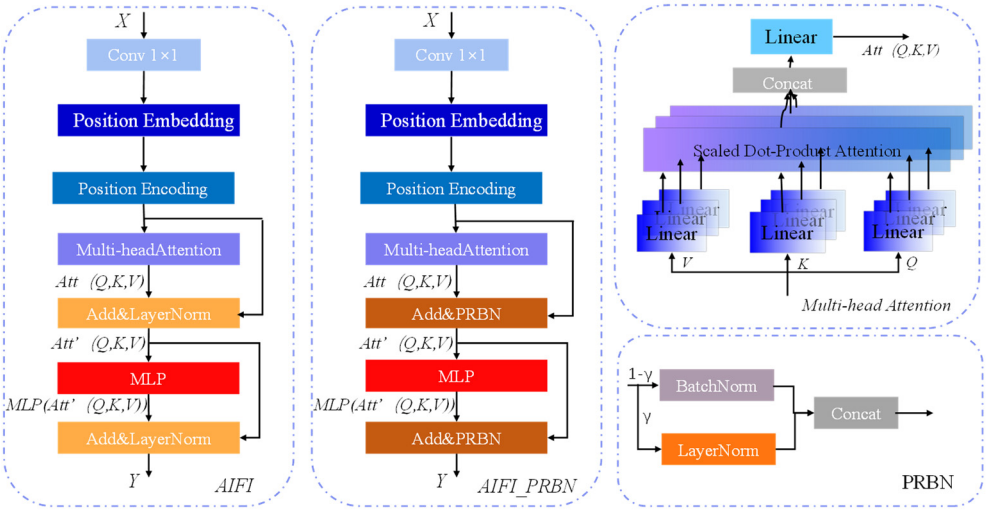


Figure 4. The network structure of AIFI_PRBN.

As shown in Figure 4, AIFI_PRBN mainly consists of a same-scale attention interaction branch, a nonlinear feature transformation branch, and a PRBN-based normalization mech-

anism. In this study, the attention branch is responsible for semantic association between defect regions and surrounding context, whereas PRBN is used to replace LayerNorm to reduce inference overhead while preserving optimization stability. The detailed formulations are given in Equations (6)–(14); standard components such as multi-head attention, residual connections and MLP are implemented in a standard manner. The mathematical formulation of the progressive re-parameterized batch normalization (PRBN) module is detailed as follows. Initially, the intermediate representation $RBN(X)$ is computed as follows:

$$RBN(X) = BN(X; \mu, \sigma, \alpha, \beta) + \eta XRBN(X) = BN(X; \mu, \sigma, \alpha, \beta) + \eta X \tag{7}$$

Subsequently, the final $PRBN(X)$ output is derived by dynamically weighting the $LayerNorm$ and the transitional state $RBN(X)$, as follows:

$$PRBN(X) = \gamma LayerNorm(X) + (1 - \gamma)RBN(X) \tag{8}$$

$$\gamma = 1 - \frac{Tcur}{T} \tag{9}$$

where μ and σ represent the mean and variance of the mini-batch, respectively, while α and β denote the learnable scale and shift parameters of the Batch Normalization layer. Furthermore, T indicates the total number of training steps utilizing $LayerNorm$, T_cur represents the current training step, and η is a learnable coefficient updated from end to end during the training process.

The multi-layer perceptron (MLP) is incorporated to introduce the capability of nonlinear transformation. The activation function GeLU has the property of continuous and differentiable gradients, enhancing the model's nonlinear expression ability while avoiding the issue of gradient mutation. The bottleneck structure design for dimension reduction and dimension increase is accomplished through the weight matrices of different layers ($W1$ and $W2$), further associating local features with global semantics. The specific internal calculation formulas of the AIFI_PRBN module are as follows:

$$Q_i = X \cdot W_i^q \cdot K_i = X \cdot W_i^k \cdot V_i = X \cdot W_i^v \tag{10}$$

$$Attention_i(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{11}$$

where W_i^q, W_i^k, and W_i^v represent the mapping matrices for queries, keys, and values, respectively, i is the sequence number of the subspace in this multi-head attention mechanism [6], $\frac{1}{\sqrt{d_k}}$ is the scaling factor, and softmax is the activation function.

$$Att(Q, K, V) = Concat \times (Attention_1, Attention_2, Attention_3, \ldots, Attention_i)\ W^o \tag{12}$$

Following the classical Transformer encoder architecture [6], residual connections and a multi-layer perceptron (MLP) are employed. The intermediate attention features are refined as follows:

$$Att'(Q, K, V) = PRBN(Att(Q, K, V) + X) \tag{13}$$

$$MLP\big(Att'(Q, K, V)\big) = W_2 \cdot \sigma\big(W_1 \cdot Att'(Q, K, V)\big) \tag{14}$$

$$Y = PRBN\big(MLP\big(Att'(Q, K, V)\big) + Att'(Q, K, V)\big) \tag{15}$$

where $W^O \in \mathbb{R}^{hd_v \times d_{model}}$, $W_1 \in \mathbb{R}^{d_{model} \times d_{ff}}$ and $W_2 \in \mathbb{R}^{d_{model} \times d_{ff}}$ are the weight matrices of the MLP layers, projecting the features from the model dimension d_{model} to a higher-

dimensional feed-forward space d_{ff} and back. The symbol σ represents the *GeLU* nonlinear activation function, which ensures continuous and differentiable gradients.

AIFI_PRBN introduces progressive re-parameterized batch normalization into the AIFI module to address the insufficient feature interaction of SPPF in YOLOv8 and the high computational cost caused by LayerNorm in the original AIFI. By improving semantic interaction between targets and surrounding context while reducing inference overhead, this module enhances defect perception and classification against complex sewer backgrounds, leading to more accurate and efficient detection.

3.5. Wise-IoU

The CIoU in YOLOv8 network has limitations when gauging the difference between the predicted box and the real box [41], which causes slow convergence and inaccurate positioning in the process of model optimization, and it is difficult to realize the high precision demands of sewage pipe defect detection when processing low-quality samples. Therefore, in the current proposal, the Wise-IoU [42,43] loss is adopted to replace the CIoU. By introducing improved geometric factors and adaptive adjustment mechanism, the Wise-IoU can better deal with low-quality samples, especially those with bad geometric characteristics (such as inconsistent aspect ratio, excessive distance, etc.), lower the effect of geometric errors on training, and thereby effectively upgrade the model's detection precision and training stability. The Wise-IoU loss is calculated as follows:

$$L_{WIoU} = rR_{WIoU}L_{IoU} \tag{16}$$

$$r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{17}$$

$$R_{WIoU} = exp\left(\frac{(x-x_{gt})^2 + (y-y_{gt})^2}{\left(W_g^2 + H_g^2\right)^*}\right) \tag{18}$$

$$L_{IoU} = 1 - IoU \tag{19}$$

where (x, y) and (x_{gt}, y_{gt}) denote the center coordinates of the predicted box and the ground-truth box, respectively. W_g and H_g represent the width and height of the smallest enclosing box covering both boxes. The superscript * indicates that these terms are detached from the computational graph. It should be noted that Equation (18) strictly follows the original WIoU v1 formulation [42] without modification.

Furthermore, to enhance robustness when dealing with samples of varying quality, a dynamic non-monotonic focusing mechanism is introduced via the gradient gain factor r, defined as follows:

$$r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{20}$$

where the dynamic variable β is defined as follows:

$$\beta = \frac{L_{IoU}^*}{\overline{L_{IoU}}} \tag{21}$$

where, β denotes the outlier degree of each sample; L_{IoU}^* is the IoU loss of the current sample and $\overline{L_{IoU}}$ is the running mean of the IoU loss. The parameters α and δ are hyperparameters that control the shape and peak position of the dynamic non-monotonic focusing function. By adaptively adjusting the gradient gain according to sample quality, WIoU suppresses the influence of low-quality samples while preventing excessive penalization of extremely hard examples, thereby improving training stability and localization precision.

4. Experiments

4.1. Dataset Preprocessing

There are various defects in urban drainage pipelines. In this study, we focus on six common types: AJ (hidden connection of branch pipes), CK (misalignment), CR (foreign objects penetrating), CQ (residual wall and dam roots), CJ (sedimentation), and TJ (disconnection of pipes). AJ denotes concealed or unrecorded branch connections that locally change the internal geometry and can easily induce turbulence and deposition. CK refers to misaligned joints where adjacent pipe segments are no longer concentric, which can cause leakage and localized sediment accumulation. CR describes foreign objects such as other pipes, cables, or steel bars that puncture or intrude into the drainage pipe and reduce the effective flow area. CQ includes residual construction walls, temporary dams that were not fully removed, and intrusive tree roots, all of which partially block the section and promote further sediment build-up. CJ corresponds to adherent deposits on the inner wall and settled sludge or sand on the invert, which reduce the drainage capacity under low-flow conditions. TJ denotes severe circumferential cracking, fracture, or loss of continuity between pipe segments, leading to serious leakage and potential structural failure. Examples of these defects are shown in Figure 5.



(**a**) AJ (**b**) CJ (**c**) CK (**d**) CQ (**e**) CR (**f**) TJ

Figure 5. Dataset defect image types. The Chinese characters in the picture represent the names of different Chinese sewerage companies or the shooting locations.

To enhance the diversity of inspection scenarios, the object detection dataset was constructed from both industrial and public sources. The industrial images were collected from CCTV inspections of underground drainage pipelines in Huai'an and Nanjing, Jiangsu Province, and Wuhan, Hubei Province, China, and were provided by multiple sewer inspection companies. In addition, public images were introduced from the Sewer-ML dataset [44,45], which contains sewer inspection images from eight European cities. The combination of multi-city industrial images and public European sewer inspection images increased the heterogeneity of the dataset in terms of inspection scenes, infrastructure conditions, and background appearance. Among them, the industrial images used in the experiment were extracted from the key frames of the CCTV inspection videos of sewer pipes, and only the frames containing objects were retained as the original data. The combined dataset includes images acquired under diverse inspection conditions, including artificial illumination, low-light environments, local overexposure or uneven illumination, different camera viewpoints and inspection positions, and varying degrees of pipe-wall contamination such as water stains, sediment deposition, and corrosion. Such a heterogeneous dataset composition, covering sewer pipeline scenes from different cities and regions together with diverse imaging conditions, provided a preliminary resource for evaluating the generalization performance of the proposed model under heterogeneous pipeline environments and imaging conditions. All images were saved in PNG format with resolutions ranging from 640 × 360 to 1280 × 720. All samples from different sources were annotated using the same defect taxonomy and the same LabelImg-based bounding-box protocol to reduce annotation inconsistency across data sources. The resulting sewer defect

dataset, which covered six defect categories, contained 5350 images (including 5082 industrial images and 268 public Sewer-ML images) and a total of 5374 defect bounding boxes (5101 from industrial images and 273 from public Sewer-ML images).We split the dataset using the conventional 8:1:1 ratio: 4280 images for the training set, 535 images for the validation set, and 535 images for the test set. The training and validation sets consisted entirely of industrial images, while the test set was a mixture of industrial images and public Sewer-ML images. The numbers of each defect are detailed in Table 2.

Table 2. The way defect samples are partitioned and distributed across each defect category.

Defect Category	AJ	CK	CR	CQ	CJ	TJ
Defect Quantity	1456	989	569	145	1701	514
Percentage (%)	27.1	18.4	10.5	2.7	31.7	9.6

The distribution of the defect boxes is summarized in Table 2. As shown in the table, this dataset had a problem of class imbalance: the CJ class (1701 boxes, accounting for 31.7%) and the AJ class (1456 boxes, accounting for 27.1%) were the main classes, while the CQ class contained only 145 boxes (accounting for 2.7%). This imbalance may have caused the detector to tend to encounter defect patterns that occurred more frequently. Therefore, to improve training robustness under limited samples and class imbalance, online data augmentation was applied during training using the Albumentations library, including horizontal and vertical flipping.

4.2. Experimental Setup

The CPU configuration used in the experiment was AMD Ryzen 9 9900X (Manufacturer: Advanced Micro Devices, City: Santa Clara, Country: United States of America), and the GPU configuration was NVIDIA GeForce RTX 4080S (16GB). The operating system environment for the experiment was Ubuntu 20.04, and the compilation environment consisted of PyTorch 2.2.2, Python 3.10, and CUDA Toolkit 12.1. All deep learning-based detectors in this study adopted the same hyperparameter settings, as can be seen in Table 3.

Table 3. Parameters setting.

Parameter Names	Value
Momentum	0.937
Batch size	8
Weight decay	0.0005
Initial learning rate	0.01
Final learning rate	0.01
Optimizer	SGD
Img Size	640 × 640
patience	30

4.3. Performance Evaluation

In this paper, a series of key indicators including precision, recall, F1, mAP, number of parameters, floating-point operations per second (FLOPS), and average confidence are used to assess the comprehensive execution of the proposed model.

Precision reflects prediction precision, while recall measures detection comprehensiveness. The F1 score balances these metrics, and mAP evaluates performance across varying IoU thresholds. The corresponding calculation formula is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{24}$$

$$AP = \int_0^1 P(R)dR \tag{25}$$

$$mAP = \frac{1}{n}\sum_{k=1}^{n} AP_k \tag{26}$$

where TP denotes the number of correctly detected positive samples, FP denotes the number of negative samples falsely classified as positive samples, TN stands for the count of correctly detected negative samples, FN represents the number of positive samples erroneously identified as negative samples, and the parameter n indicates the total number of classes.

5. Results and Discussion

5.1. Comparative Experiments on the C2f_SCA Module

The proposed C2f_SCA module improves the C2f module by embedding the CAA attention mechanism into the Star_Block module. To confirm the effectiveness of introducing CAA attention mechanism, this paper evaluates C2f_SCA module and C2f_Star module. In YOLOv8l+C2f_SCA, the SCA_Block with CAA attention mechanism is used to substitute C2f in the YOLOv8l model. The YOLOv8l+C2f_Star module uses the Star_Block structure to substitute the C2f module in YOLOv8l. Outcomes from the comparative experiments can be found in Table 4.

Table 4. Empirical comparisons of C2f_SCA.

Model	Precision	Recall	F1	mAP@0.5
YOLOv8l	77.8%	80.9%	0.79	83.8%
YOLOv8l+C2f_Star	79.8%	76.9%	0.78	84.3%
YOLOv8l+C2f_SCA	84.2%	77.1%	0.8	85.4%

As shown in Table 4, compared with YOLOv8l, YOLOv8L+C2f_Star improved the precision and mAP@0.5 by 2% and 0.5%, respectively, but decreased the recall and F1 by 4% and 0.01, respectively. The precision of YOLOv8l+C2f_SCA was enlarged by 6.4%, F1 increased by 0.01, while mAP@0.5 gained 1.6 percentage points. This shows that the performance of YOLOv8L+C2f_SCA was strongly improved compared with that of the YOLOv8l baseline model and YOLOv8L+C2f_Star model, which verifies that the incorporation of CAA can significantly enhance the capacity to understand local and global information.

The performance improvement brought by C2f_SCA is particularly significant for long-distance viewing or defects with relatively weak local appearance. Table 5 reports the detection precision of the C2f_SCA and C2f_Star modules across different types of pipeline defects. Compared with C2f_Star, the C2f_SCA module achieved consistent precision improvements for all four defect categories, with gains of 0.3%, 3.6%, 3.4%, and 0.2% for AJ, CK, CR, and TJ defects, respectively. This improvement indicates that this module is particularly effective for defects with blurred boundaries, small apparent sizes, or high dependence on surrounding structural cues. This is due to the projection based on star enhancing nonlinear feature transformation, while the CAA mechanism strengthens long-distance context aggregation. Therefore, the network no longer overly relies on local high-contrast textures and can identify weak textures or distant defects by jointly utilizing defect cues and their surrounding context.

Table 5. Comparison of detection precision of C2f_SCA in different defects.

Model	Classification					
	AJ	CJ	CK	CR	TJ	CQ
	AP@0.5	AP@0.5	AP@0.5	AP@0.5	AP@0.5	AP@0.5
C2f-Star	0.916	0.931	0.767	0.613	0.91	0.919
C2f_SCA	0.919	0.925	0.803	0.647	0.912	0.918

Due to the complex and uncertain internal structure of pipelines, defect detection is usually conducted at a long distance from the camera, which results in less detailed features of defects in the images. Figure 6 presents the defect detection visualization results of 10 images using the C2f_Star and C2f_SCA modules, respectively. These 10 images are divided into five pairs, where each pair consists of shots of the same target captured at different observation distances. Visualization results demonstrate that the C2f_SCA module equipped with the CAA attention mechanism achieved superior defect detection performance under long-distance observation conditions. By facilitating broader contextual modeling, the introduced CAA mechanism effectively enhances the extraction of long-distance features, thereby improving detection reliability for defects observed at extended distances.

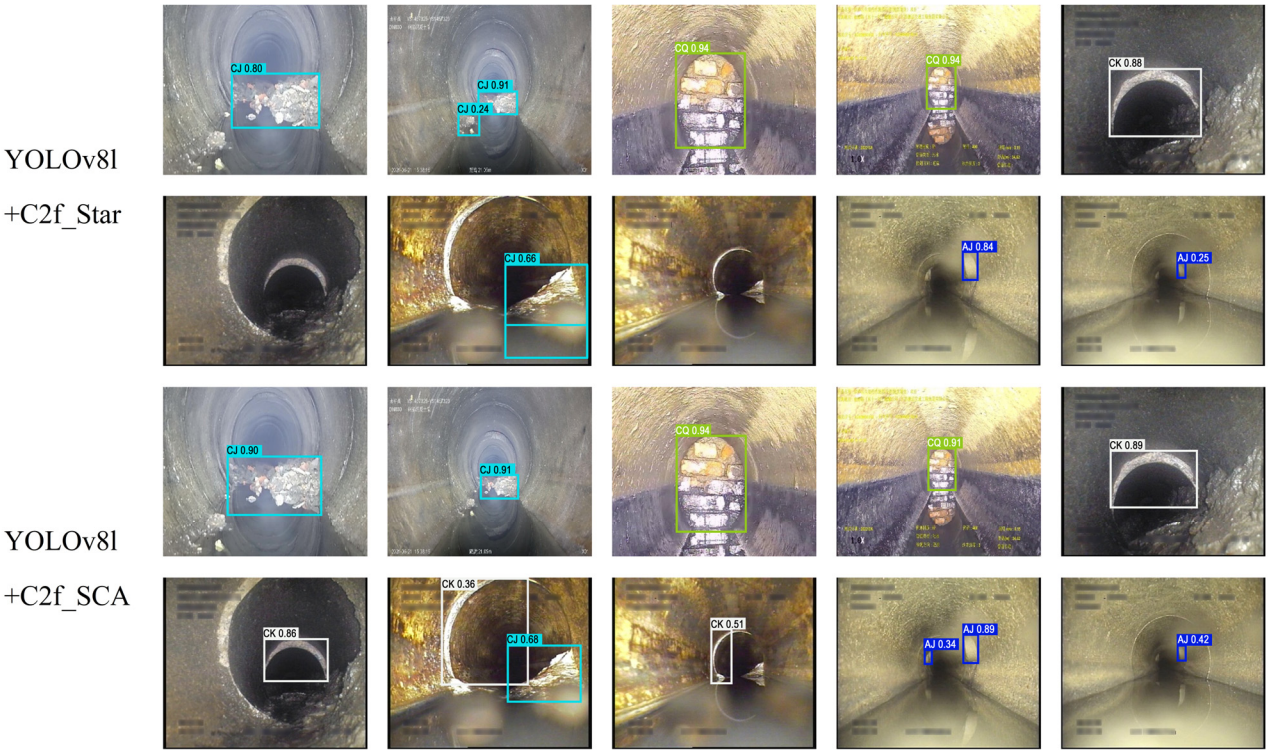


Figure 6. Visualization results comparison between the C2f_Star module and the proposed C2f_SCA module at different observation distances. Key observation: The inclusion of context anchor attention (CAA) in C2f_SCA significantly improves bounding box precision and prediction confidence, demonstrating superior long-distance feature extraction capabilities compared to the C2f_Star baseline, especially for small or distant defects.

5.2. Comparative Experiments on AIFI_PRBN Module

The proposed AIFI_PRBN module improves the AIFI module [46] by replacing layer norm with PRBN (progressive re-parameterized batch norm). To ascertain the effectiveness of AIFI_PRBN for reducing the amount of calculation and enhancing the stability of training, two sets of comparative experiments were carried out on YOLOv8l with the

C2f_SCA module added: YOLOv8l+C2f_SCA+AIFI and YOLOv8l+C2f_SCA+AIFI_PRBN. Figure 7 and Table 6 display the results obtained from the experiments.

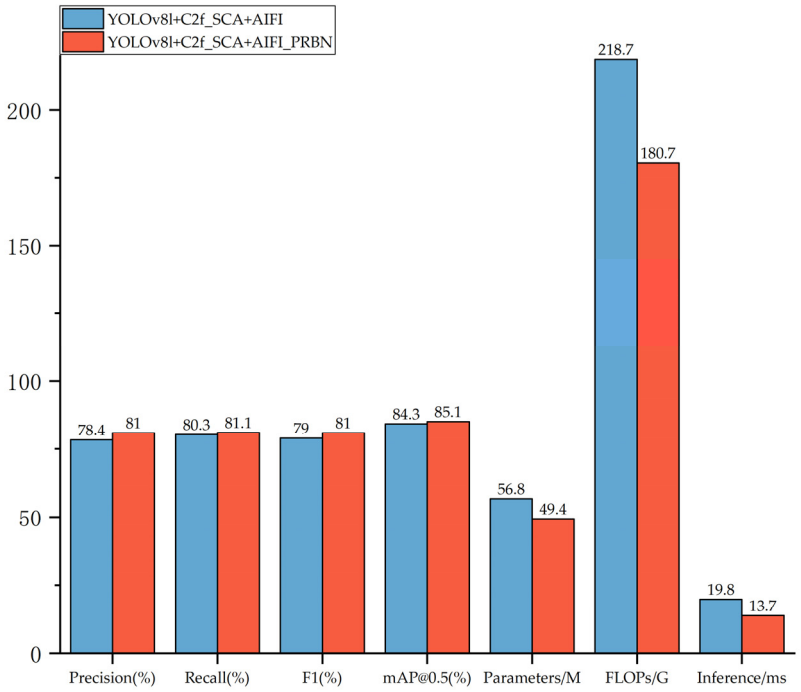


Figure 7. Comparative visualization of experimental results on AIFI_PRBN.

Table 6. Comparison of experimental results on AIFI_PRBN.

Model	Precision	Recall	F1	mAP@0.5	Parameters/M	FLOPS/G	Inference/ms	FPS
YOLOv8l+C2f_SCA+AIFI	78.4%	80.3%	0.79	84.3%	56.8	218.7	19.8	50.20
YOLOv8l+C2f_SCA+AIFI_PRBN	81%	81.1%	0.81	85.1%	49.4	180.7	17.1	58.4

As illustrated in Figure 7, the AIFI_PRBN module consistently outperformed the original AIFI module without re-parameterization across all evaluation metrics, demonstrating the effectiveness of the proposed strategy. The quantitative results reported in Table 6 further substantiate this observation. Specifically, compared with the original AIFI module, AIFI_PRBN reduced the number of parameters from 56.8 M to 49.4 M, decreased computational complexity from 218.7 G FLOPs to 180.7 G FLOPs, and shortened inference latency from 19.8 ms to 17.1 ms. In addition, the inference speed increased from 50.20 FPS to 58.4 FPS, corresponding to an improvement of 8.2 FPS, which indicates better real-time processing capability under the current experimental hardware setting. Meanwhile, AIFI_PRBN also achieves performance gains in precision, recall, F1 score, and mAP@0.5, with improvements of 2.6%, 0.8%, 0.02, and 0.8%, respectively. These results indicate that the re-parameterization strategy enhances the feature extraction capability of the AIFI module in complex background scenarios while partially reducing the computational burden introduced by the attention mechanism.

Table 7 shows the alterations in the precision of different kinds of defect detections before and after the improvement of the AIFI structure. By applying the AIFI-PRBN module, the detection precision for the four types of defects, CJ, CR, TJ, and CQ, increased by 1%, 0.9%, 2.7%, and 1.4%, respectively. The improved module enhances the correlation between the defect area and its surrounding context through scale-invariant semantic interaction, which is particularly important when the image contains deposits, water stains, corrosion textures, or uneven lighting. In these cases, relying solely on the local

defect appearance is often insufficient for reliable distinction. AIFI_PRBN reduces the computational cost of the original AIFI while enhancing semantic interaction, thus simultaneously improving environmental adaptability and practical application efficiency. Thus, the YOLOv8l+C2f_SCA+AIFI_PRBN strategy further improves environmental adaptability and precision while minimizing computational costs as much as possible.

Table 7. Comparison of experimental results on PRBN; detection precision.

Model	Classification					
	AJ	CJ	CK	CR	TJ	CQ
	AP@0.5	AP@0.5	AP@0.5	AP@0.5	AP@0.5	AP@0.5
AIFI	0.929	0.922	0.789	0.629	0.898	0.893
AIFI_PRBN	0.92	0.932	0.781	0.638	0.925	0.907

5.3. Comparative Experiments on Loss Functions

To evaluate the role of different box regression losses and their influence on performance metrics, we used YOLOv8l+C2f_SCA+AIFI_PRBN as the baseline and compared a variety of loss functions, as shown in Table 8.

Table 8. A Comparison study of loss function performance.

IoU	Precision	Recall	F1	mAP@0.5
CIoU	81%	81.1%	0.81	85.1%
EIoU	81.8%	80.5%	0.81	84.1%
Inner-CIoU	81.2%	76.8%	0.79	83.5%
Inner-MPDIoU	78.2%	79.5%	0.79	83.6%
Focaler-CIoU	79.2%	78.2%	0.79	82.9%
Focaler-EIoU	70.8%	78.4%	0.74	79.6%
Wise-CIoU	79.2%	82.5%	0.81	84.9%
Wise-IoU	84.4%	80%	0.82	86.2%

The experimental results showed that the precision of Wise-IoU achieved 84.4%, the F1 was 0.82, and the mAP@0.5 was 86.2%, which was superior to other loss functions. These three indicators were the highest among all bounding box regression loss functions, but its recall was 2.5% lower than that of Wise-CIoU. Compared with the original CIoU of YOLOv8l, Wise-IoU demonstrated significant improvement in three indicators; the precision increased by 3.4%, the F1 increased by 0.01, and the mAP@0.5 increased by 1.1%, but the recall slightly decreased by 1.1%. These results provide evidence that Wise-IoU outperformed other bounding box regression loss functions in terms of overall performance. By introducing improved geometric factors and adaptive adjustment mechanism, it can better deal with low-quality samples (such as samples with inconsistent aspect ratio and samples with excessive distance). The training loss curves in Figure 8 reinforce this advantage. Compared with the traditional CIoU, Wise-IoU maintained a lower box regression loss throughout training and converges more quickly. This shows that the improved geometric terms and adaptive adjustment mechanism not only boost precision, F1, and mAP@0.5, but also stabilize learning and make the model better able to handle low-quality samples, such as targets with large aspect-ratio discrepancies or very long distances.

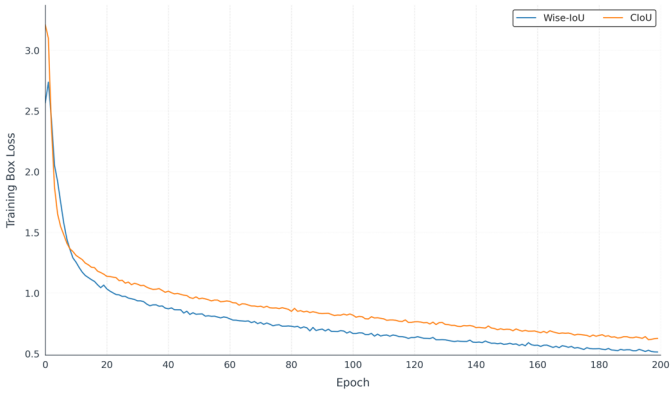


Figure 8. Iteration comparison of loss functions between CIoU and Wise-IoU.

Table 9 shows the changes in detection precision for different defect categories after replacing the loss function with Wise-IoU. The detection precision for AJ, CK, and CQ defects increased by 0.1%, 4.3%, and 4.4%, respectively. These results indicate that the adoption of Wise-IoU significantly enhances the capabilities of the model in multi-scale feature fusion, structural perception, and robustness under complex scenarios.

Table 9. Comparison of experimental results for detection precision for different defects; CIoU and Wise-IoU.

Module	Classification					
	AJ	CJ	CK	CR	TJ	CQ
	AP@0.5	AP@0.5	AP@0.5	AP@0.5	AP@0.5	AP@0.5
CIOU	0.92	0.932	0.781	0.638	0.925	0.907
WIOU	0.921	0.932	0.824	0.632	0.912	0.951

5.4. Ablation Experiment

To comprehensively and systematically evaluate the role of each improved module in SAW-YOLOv8l, ablation experiments were carried out for the three main innovations. Seven different combinations of modules were designed for the ablation experiments. Table 10 shows the experimental results of these combinations on the unseen dataset.

Table 10. Effectiveness of Individual Components.

YOLOv8	C2f_SCA	AIFI_PRBN	Wise-IoU	Precision	Recall	F1	mAP@0.5	mAP@0.5:0.95	Parameters/M
✓				0.778	0.809	0.79	0.838	0.695	43.6
✓	✓			0.842	0.771	0.8	0.854	0.699	49.4
✓		✓		0.771	0.837	0.8	0.853	0.699	43.6
✓			✓	0.796	0.805	0.8	0.854	0.708	43.6
✓	✓	✓		0.81	0.811	0.81	0.851	0.692	49.4
✓	✓		✓	0.829	0.786	0.81	0.861	0.697	49.4
✓		✓	✓	0.799	0.778	0.78	0.845	0.682	43.6
✓	✓	✓	✓	0.844	0.8	0.82	0.862	0.69	49.4

Note: All the training was conducted under the conditions of Img Size = 640 × 640, Batch Size = 8, and lr = 0.01. ✓ indicates that the module is included.

As shown in Table 10, when the C2f_SCA module was substituted for the C2f structure within the backbone and neck of YOLOv8l, precision increased to 84.2%, F1-score reached 0.8, and mAP@0.5 reached 85.4%, representing improvements of 6.4%, 0.01, and 1.6%, respectively, over the unoptimized model. However, the recall decreased by 3.8%. This indicates that the C2f_SCA module significantly enhances recognition precision for high-

confidence targets by optimizing feature learning, thereby improving evaluation metrics like precision, F1 score, and mAP@50. Nevertheless, this optimization strategy may also reduce sensitivity to some low-confidence targets, affecting recall. After evolving the backbone by replacing the SPPF of YOLOv8l and integrating the AIFI_PRBN module, although precision slightly decreased by 0.7%, the recall, F1 score, and mAP@0.5 all improved, with the most notable improvement being an increase in recall from 80.9% to 83.7%, a gain of 2.8%. The performance gain was the result of improved capability to handle complex backgrounds in sewer pipelines, improving environmental adaptability and enabling the model to prioritize more low-confidence targets. When CIoU in YOLOv8l was replaced by Wise-IoU, mAP@0.5 increased to 85.4%, constituting a 1.6% gain over the baseline model. This improvement mainly stems from the approach of Wise-IoU to mitigating the impact of low-quality samples, enhancing the model capability by calculating distance difference in coordinates between the target and predicted boxes and adjusting geometric factors, thereby enhancing generalization and achieving high localization precision in bounding box regression. Taken together, the ablation results suggest that the three improvements play different but complementary roles. C2f_SCA mainly strengthens local-context representation for weak and distant defects, AIFI_PRBN mainly improves semantic discrimination and environmental adaptability against cluttered backgrounds, and Wise-IoU mainly stabilizes localization learning for low-quality samples. Their combination therefore improves detection performance not only by increasing feature expressiveness, but also by making the model more robust to difficult imaging conditions and unstable regression cases.

By integrating the C2f_SCA, AIFI_PRBN, and Wise-IoU modules, the model achieved the highest values across three core metrics: precision (84.4%), F1 score (0.82), and mAP@0.5 (86.2%), exceeding the baseline by 6.6%, 0.03, and 2.4% respectively. It is worth noting that for mAP@0.5:0.95, the comprehensive model attained 69.0%. Although the precision was lower than that of Wise-IoU (0.708) and YOLOv8L (0.695) at baseline, we believe that for water utilities, the coarse precision of mAP@0.5 is sufficient to detect defects. In actual engineering applications, our SAW-YOLO model has significant advantages over other models in terms of improved precision and mAP@0.5.

In terms of parameter efficiency, SAW-YOLOv8l has 49.4M parameters, a 13% increase compared to the baseline model. However, this growth in parameters is accompanied by significant improvements in key performance metrics. From a practical deployment perspective, the parameter increase remains within a reasonable range, while the 2.4% gain in mAP@0.5 demonstrates a favorable trade-off between model complexity and performance, indicating good cost-effectiveness for engineering applications.

As illustrated in Figure 9, the training curves provide evidence that the proposed model consistently surpasses the former YOLOv8l across multiple metrics, particularly during the later training stages. This performance advantage is further confirmed by the comprehensive evaluation results in Table 10.

While the proposed architecture achieved substantial results including a 2.4% higher mAP@0.5, reaching 86.2%, the analysis of mAP@0.5:0.95 reveals a more nuanced picture. The complete model attained 69.0% in mAP@0.5:0.95, representing a minor decline compared to some individual module configurations, such as the Wise-IoU that achieved 70.8%. This indicates that while our model excels at detecting objects under standard IoU thresholds, there remains room for enhancement in localization precision under stricter matching criteria.

It is worth noting that these performance gains come with a 13% increase in parameter count compared to the baseline YOLOv8l, primarily due to the incorporation of the SCA_Block within the C2f structure. This parameter–performance trade-off proves favorable for applications where detection precision is prioritized, as the relative pa-

rameter increase remains moderate while delivering significant improvements in key detection metrics.

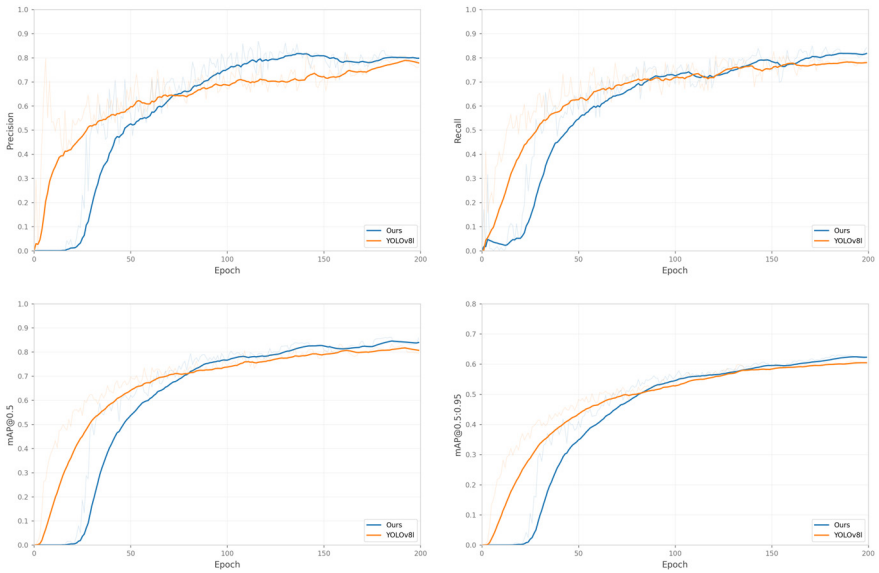


Figure 9. The proposed model is compared with the original YOLOv8l in terms of four indicators: precision, pecall, mAP@0.5, and mAP@0.5: 0.95 (the shaded area represents the fluctuations during training).

5.5. Comparative Experiments on Different Models

In this study, various object detection models including YOLOv5l, YOLOv8l, YOLOv11l, Faster R-CNN, Mamba-YOLO [47], D-FINE-m [48], RT-DETR-r18, and RT-DETR-R50 were tested. To ensure a rigorous and fair comparison, all models were trained under identical experimental configurations: a consistent dataset, an input resolution of 640 × 640, a batch size of eight, and a uniform data augmentation strategy. The initial learning rate was set to 0.01, supplemented by an early stopping strategy with a patience of 30 epochs to prevent over-fitting. Notably, all models were trained from scratch without the use of pretrained weights to maintain experimental consistency. For model-specific parameters, the official default settings were adopted where applicable.

Table 11 and Figure 10 summarize the comparative results, demonstrating that SAW-YOLOv8l achieved overall superior performance among the evaluated architectures. As shown in Table 11, SAW-YOLOv8l attained an F1 score of 0.82 and an mAP@0.5 of 86.2%, outperforming the other mainstream approaches. In addition to the precision improvement, the proposed model also maintained favorable inference efficiency under the current experimental setting. Although the enhanced architecture introduces a moderate increase in model parameter quantity compared with the baseline, the final model still achieved 58.4 FPS, indicating that it can satisfy real-time inspection requirements on a GPU-accelerated platform. These results suggest that the proposed method achieved a reasonable balance between detection precision and computational cost, making it suitable for practical inspection scenarios with adequate computing resources. Regarding recall, SAW-YOLOv8l was 2.3% lower than YOLOv5l and differed by only 0.9% compared to the baseline YOLOv8l. Although this might slightly increase the risk of missed detections, the precision of the proposed model and the overall improvement in mAP@0.5 precisely indicate that this method can achieve a better balance in practical applications. In Figure 10, different shapes represent the positions of different models in the mAP@0.5, precision and recall three-dimensional space. From the figure, it can be seen that our model is located at the deepest part of the three-dimensional graph, indicating that itachieved the

best performance when compared comprehensively on these three indicators. Overall, these enhancements enable SAW-YOLOv8l to maintain competitive inference efficiency while substantially improving detection precision and robustness. These findings validate the effectiveness and practical applicability of SAW-YOLOv8l for sewage pipe defect detection tasks.

Table 11. Experimental results of object detection models.

Model	Precision	Recall	F1	mAP@0.5	Parameters/M	FPS
YOLOv5l	77.9%	82.3%	0.8	82.6%	46.6	111.6
YOLOv8l	77.8%	80.9%	0.79	83.8%	43.6	121.7
YOLOv11l	84.6%	76.3%	0.8	84.8%	25.3	235.3
Faster R-CNN	58.1%	78.8%	0.661	76%	28.3	36.3
Mamba-YOLO	84.6%	76.9%	0.803	83.8%	21.8	91.37
D-FINE-m	84.2%	71.6%	0.77	79.9%	19.2	81.6
RT-DETRr18	85.5%	79%	0.82	82.9%	19.88	130.3
RT-DETRr50	83.6%	75.2%	0.79	78%	42	108.0
SAW-YOLOv8l	84.4%	80%	0.82	86.2%	49.4	58.4

Note: All the training was conducted under the conditions of Img Size = 640 × 640, Batch Size = 8, and lr = 0.01.

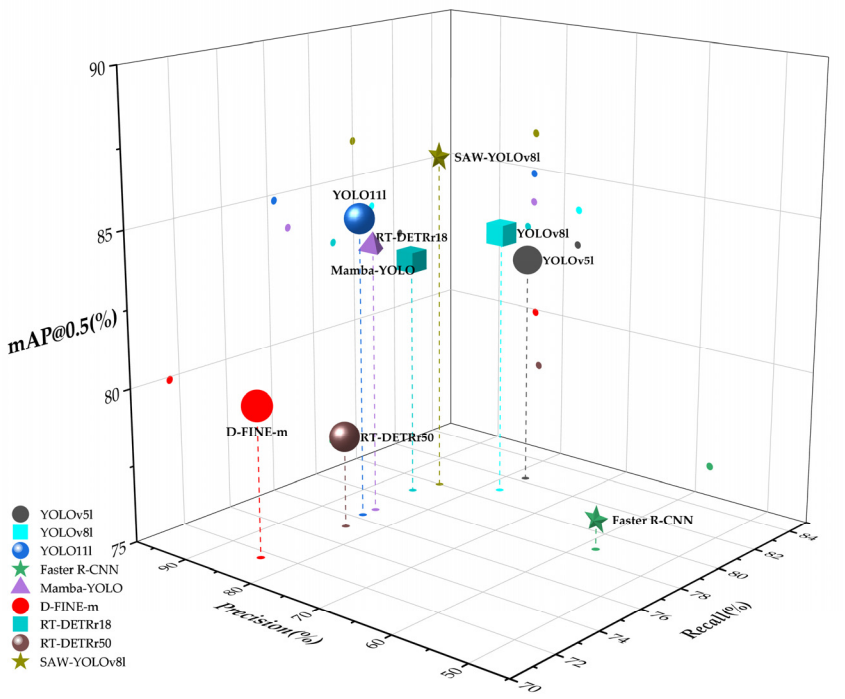


Figure 10. 3D visualization of experimental results across different models. (Different colored dots represent the projections of each model on each plane.)

5.6. Visualization Analysis

To comprehensively evaluate the performance of SAW-YOLOv8l, images were randomly sampled from the dataset for visual analysis. The visualization experiments were organized into two parts. The first part of this section presents qualitative comparisons between SAW-YOLOv8l and several mainstream object detection models, including RT-DETRr50, Faster R-CNN, and representative YOLO-based methods (YOLOv11l, YOLOv8l, and YOLOv5l). The second part focuses on a robustness evaluation under challenging imaging conditions, where detection results are compared across models on images affected by blur and exposure variations.

As shown in Figure 11, the prediction results of Faster R-CNN (Figure 11(b1–b6)) exhibited frequent false detections and redundant bounding boxes, indicating limited

localization reliability. Owing to the insufficient sensitivity of the Transformer architecture to local features, RT-DETRr50 produced prediction results with noticeably lower confidence than SAW-YOLOv8l, and false positives were observed (Figure 11(c3)). Regarding the YOLO-series models, both YOLOv11l and YOLOv5l suffered from missed detections (Figure 11(a3)), and the YOLO-based methods failed to reliably identify TJ defects located near background boundaries (Figure 11(a4)). In contrast, the proposed SAW-YOLOv8l effectively resolved the missed detection issues of the baseline YOLOv8l (Figure 11(d1,d4)) and achieved complete defect recognition with higher prediction confidence.

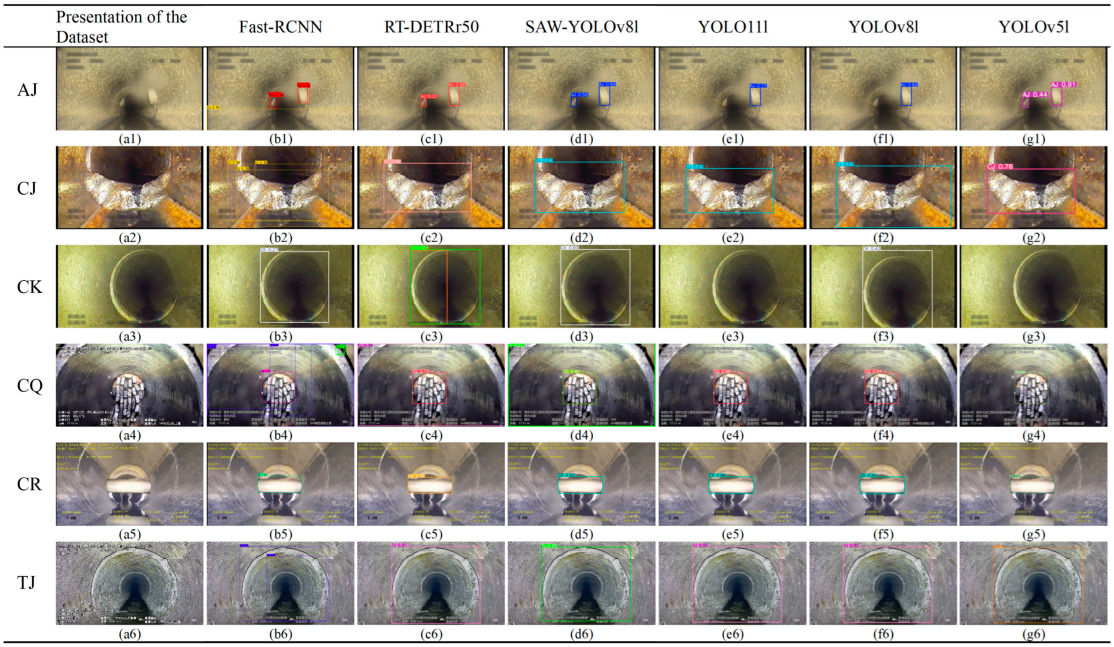


Figure 11. Qualitative visualization results comparing SAW-YOLOv8l against mainstream object detection models (Faster R-CNN, RT-DETRr50, YOLOv11l, YOLOv8l, and YOLOv5l). Key observation: the proposed SAW-YOLOv8l successfully identifies complex defects, such as concealed branch connections (AJ) and misalignments (CK), with tighter bounding boxes and higher confidence scores. It effectively mitigates the redundant false positives observed in Faster R-CNN and the missed detections commonly seen in other YOLO variants. (**a1**–**a6**) Original input images of the dataset (ground truth, without detection boxes); (**b1**–**b6**) Detection results of the Fast-RCNN model; (**c1**–**c6**) Detection results of the RT-DETRr50 model; (**d1**–**d6**) Detection results of the SAW-YOLOv8l mode; (**e1**–**e6**) Detection results of the YOLO11l model; (**f1**–**f6**) Detection results of the YOLOv8l model; (**g1**–**g6**) Detection results of the YOLOv5l model. Rows 1–6 correspond to different sewer pipe scenarios.

Figures 12 and 13 present the detection results of SAW-YOLOv8l and YOLOv8l under exposure variation and imaging blur conditions, respectively. The visualization results demonstrate that, compared with YOLOv8l, the proposed SAW-YOLOv8l maintained more stable detection performance and achieved higher detection precision under challenging imaging conditions. Even in the presence of severe exposure variation and imaging blur, SAW-YOLOv8l consistently identified pipeline defects, which indicates that the C2f_SCA and AIFI_PRBN modules proposed in this paper can still maintain the discriminative features under conditions of weak visual cues or degraded imaging, further demonstrating the strong robustness and effectiveness of the proposed model in addressing common practical challenges in sewage pipeline inspection.

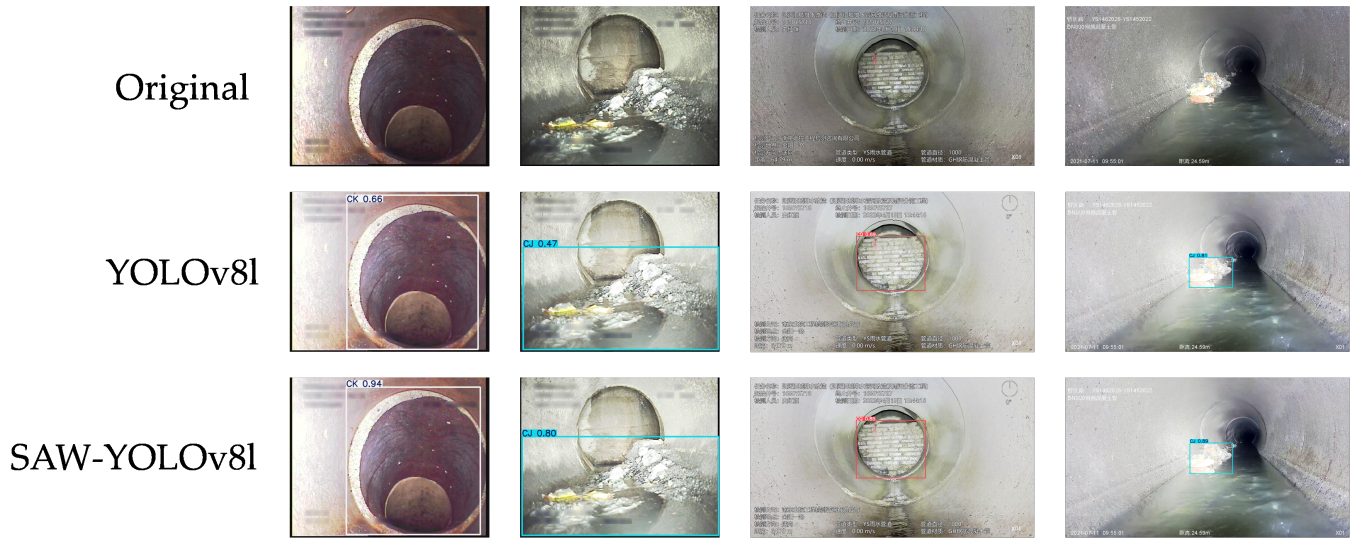


Figure 12. Visualization of experimental results under exposure conditions. The first row shows the original images. The second and third rows present the detection results of the corresponding models on the original images processed with exposure adjustment.

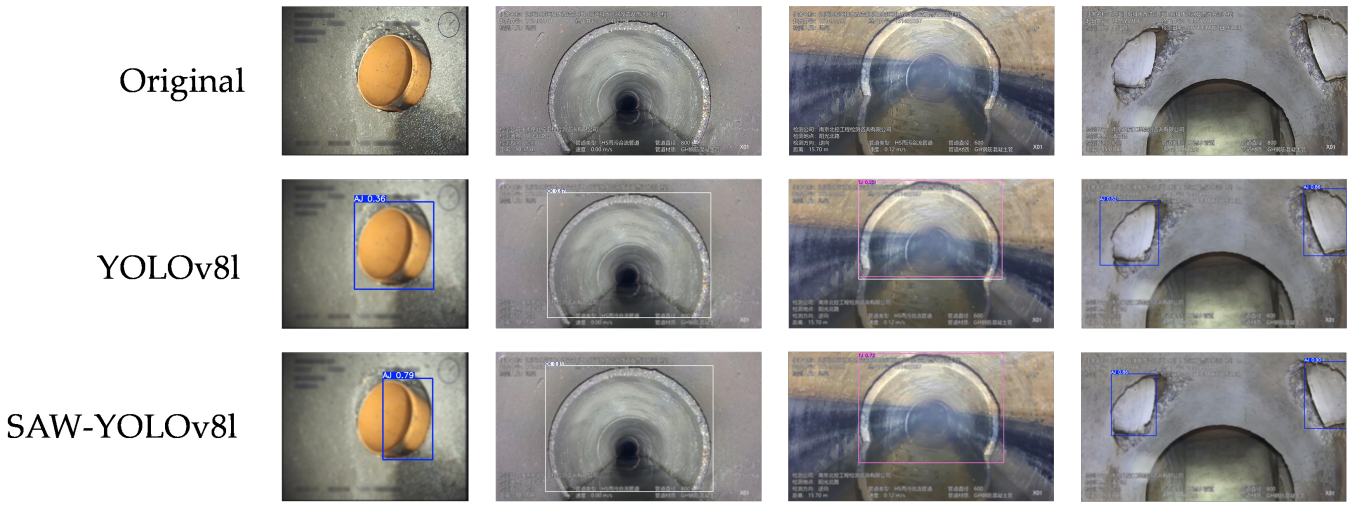


Figure 13. Visualization of experimental results under blurring conditions. The first row shows the original images. The second and third rows present the detection results of the corresponding models on the original images processed with blurring.

6. Discussion

The SAW-YOLO model showed improvements in precision, F1 score, and mAP@0.5 to varying degrees. From a practical perspective, the performance improvements demonstrate clear value for real-world sewer pipeline inspection systems. The 6.6% increase in precision effectively reduces false positives, thereby lowering the workload of manual verification in CCTV inspection videos. Meanwhile, the 2.4% improvement in mAP@0.5 enhances the detection capability for small-scale and low-contrast defects, reducing the risk of missed detections and potential pipeline failures. In addition, the proposed model achieved an inference speed of 58.4 FPS, meeting the real-time requirements of mainstream CCTV inspection systems (25–30 FPS), enabling practical deployment for real-time detection tasks. However, it should be noted that the current real-time performance has only been validated on a high-performance hardware platform. Therefore, although the proposed model shows promising deployment potential, its feasibility on resource-constrained edge devices still

requires further investigation. Future work will focus on lightweight optimization and hardware-aware deployment to improve its applicability in practical inspection systems. In practical sewer inspection scenarios, model performance may also be affected by environmental and internal factors. Imaging conditions can vary significantly due to illumination fluctuation, camera pose changes, and pipeline contamination such as biofilm, grease, and sediment accumulation. These factors may lead to domain shifts between training and deployment environments, potentially degrading the model's generalization performance. From a system-level perspective, such variability not only affects detection precision but also influences the reliability and continuity of long-term infrastructure monitoring. For instance, inconsistent imaging conditions may result in unstable detection outputs across inspection cycles, thereby impacting maintenance decision-making and risk assessment.

7. Conclusions

To address the challenge of complex background interference, multi-scale defect representation, and long-distance target feature extraction in the closed space of underground pipelines, an enhanced SAW-YOLOv8l detection model combining RT-DETR and convolutional neural network architecture is innovatively proposed. The improved C2f_SCA model and AIFI_PRBN model were constructed, the loss function and dynamic attention weight have been optimized, and the global and local feature extraction of semantic information in sewage pipe images are thereby enhanced. Experimental verification showed that the mAP@0.5, F1, and precision of the proposed model on the sewage pipeline defect dataset were increased by 2.4%, 0.03, and 6.6% respectively. At the same time, considering possible situations of overexposure and fogging of the lens in actual detection, visual simulation experiments were conducted for corresponding scenarios, and combined with visual analysis, the robustness of the model was further proved. However, while detection precision was improved, the number of parameters and model complexity also increased. Considering practical deployment issues, our future research will strive to significantly reduce redundant parameters while maintaining detection performance. This will be achieved by using pruning and other methods to reduce the model size, making it easier to deploy on devices. Furthermore, to more comprehensively detect defects in different scenarios, we will expand the model to include more scenarios to improve its generalization performance.

Author Contributions: L.H., H.L. and J.G. wrote the main manuscript text and conceived the simulations, H.L., W.Z. and P.X. conducted the simulations, B.G., S.S. and Y.K. analyzed the results. All authors have read and agreed to the published version of the manuscript.

Funding: This research was by the Jiangsu Provincial Department of Water Resources (No. 2023045), the Jiangsu hydraulic research institute (No. 2024Z009), the High-level Talent Project by the Jinling Institute of Technology (No. jit-b-202302) and the Jiangsu Province Industry-Academia-Research Collaboration Projects (No. BY20230562).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used during the current study are available from the corresponding author upon reasonable request.

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

References

1. Alqahtani, F.K.; Alsharef, A.; Hommadi, G.M.; Alammari, M.A. Assessment Framework for the Maintainability of Sewer Pipeline Systems. *Appl. Sci.* **2023**, *13*, 11828. [CrossRef]
2. U.S. EPA. *Condition Assessment of Underground Pipes*; U.S. Environmental Protection Agency: Washington, DC, USA, 2015.

3. Hawari, A.; Alamin, M.; Alkadour, F.; Elmasry, M.; Zayed, T. Automated Defect Detection Tool for CCTV-Inspected Sewer Pipelines. *Autom. Constr.* **2018**, *89*, 99–109. [CrossRef]
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Lake Tahoe, NV, USA, 3–8 December 2012*; Neural Information Processing Systems Foundation: San Diego, CA, USA, 2012; pp. 1097–1105.
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017*; Neural Information Processing Systems Foundation: San Diego, CA, USA, 2017; pp. 5998–6008.
7. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
8. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 4510–4520.
9. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019*; PMLR: Cambridge, MA, USA, 2019; pp. 6105–6114.
10. Virasova, A.; Klimov, D.; Khromov, O.; Gubaidullin, I.; Oreshko, V. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Radio Eng.* **2021**, *85*, 115–126. [CrossRef]
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
12. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2017; pp. 2961–2969.
13. Situ, Z.; Teng, S.; Feng, W.; Zhong, Q.; Chen, G.; Su, J.; Zhou, Q. A Transfer Learning-Based YOLO Network for Sewer Defect Detection. *Dev. Built Environ.* **2023**, *15*, 100191. [CrossRef]
14. Zhao, X.; Xiao, N.; Cai, Z.; Xin, S. YOLOv5-Sewer: Lightweight Sewer Defect Detection Model. *Appl. Sci.* **2024**, *14*, 1869. [CrossRef]
15. Oh, C.; Dang, L.M.; Han, D.; Moon, H. Robust Sewer Defect Detection with Text Analysis Based on Deep Learning. *IEEE Access* **2022**, *10*, 46224–46237. [CrossRef]
16. Lv, Z.; Dong, S.; Xia, Z.; He, J.; Zhang, J. Enhanced RT-DETR for Robotic Inspection of Underwater Bridge Pier Cracks. *Autom. Constr.* **2025**, *170*, 105921. [CrossRef]
17. Cui, Y.; Han, Y.; Guo, D. TS-DETR: Multi-Scale DETR for Traffic Sign Detection and Recognition. *Pattern Recognit. Lett.* **2025**, *190*, 147–152. [CrossRef]
18. Kumar, S.S.; Wang, M.; Abraham, D.M.; Jahanshahi, M.R.; Iseley, T.; Cheng, J.C.P. Deep Learning-Based Automated Detection of Sewer Defects in CCTV Videos. *J. Comput. Civ. Eng.* **2020**, *34*, 04019047. [CrossRef]
19. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A Survey and Performance Evaluation of Deep Learning Methods for Small Object Detection. *Expert Syst. Appl.* **2021**, *172*, 114602. [CrossRef]
20. Jia, P.T.; Guo, T.; Guo, F.J.; Wang, B.; Xiong, Q. Detection Model of Drainage Pipe Defect Based on Improved YOLOv5. In *Proceedings of the International Conference on Intelligent Computing and Signal Processing (ICSP)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1950–1955.
21. Lu, Q.; Ding, X.; Liang, Y. An Algorithm for Identifying Defects in Underground Drainage Pipes Based on an Improved YOLOX. *Electron. Meas. Technol.* **2022**, *45*, 161–168.
22. Ren, H.; Liu, F.F.; He, X.F. Research on the Defect Detection Algorithm for Underground Drainage Pipes Using Improved YOLOv8. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*; IEEE: Piscataway, NJ, USA, 2024; pp. 272–278.
23. Zhou, M.; Zhang, X.; Zeng, P.; Jiang, Y. Drainage Pipe Network Defect Detection Based on Improved YOLOv8. *Electron. Meas. Technol.* **2025**, *48*, 128–137.
24. Dang, L.M.; Wang, H.; Li, Y.; Nguyen, T.N.; Moon, H. DefectTR: End-to-End Defect Detection for Sewage Networks Using a Transformer. *Constr. Build. Mater.* **2022**, *325*, 126584. [CrossRef]
25. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. *arXiv* **2023**, arXiv:2304.08069.
26. Xiao, L.; Zhang, H. CMS-RTDETR: Multi-scale enhanced real-time detection for underground pipeline defect. *Meas. Sci. Technol.* **2025**, *36*, 076015. [CrossRef]
27. Lv, Z.; Feng, Y.; Chen, C.; Xu, P.; Zhou, C.; Luo, J.; Yin, L. A Defect Detection Method for Steel Pipe Welds Based on XXX-DETR Network. In *Proceedings of the 2025 4th International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5.

28. Wei, G.; Yu, Z.; Li, D. WEHD-DETR: A real-time defect detection algorithm for sewer pipelines based on improved RT-DETR. *Digit. Signal Process.* **2025**, *168*, 105585. [CrossRef]
29. Ling, P.; Zhang, Y.; Ma, S. Maritime Small Object Detection Algorithm in Drone Aerial Images Based on Improved YOLOv8. *IEEE Access* **2024**, *12*, 176527–176538. [CrossRef]
30. Liu, J.; Wei, Y.; Yue, Y.; Liu, J.; Liu, Q. SEA-YOLOv8: An Enhanced Method for Detecting Small Targets in Aero-Engine Components Based on YOLOv8n. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 8522–8533. [CrossRef]
31. He, Y.; Li, Y.; Li, Z.; Song, R.; Xu, C. An improved YOLOv8-based lightweight approach for orange maturity detection. *J. Food Meas. Charact.* **2025**, *19*, 4740–4754. [CrossRef]
32. Chen, L.; Li, G.; Zhang, S.; Mao, W.; Zhang, M. YOLO-SAG: An improved wildlife object detection algorithm based on YOLOv8n. *Ecol. Inform.* **2024**, *83*, 102791. [CrossRef]
33. Lv, Z.; Dong, S.; He, J.; Hu, B.; Liu, Q.; Wang, H. Lightweight Sewer Pipe Crack Detection Method Based on Amphibious Robot and Improved YOLOv8n. *Sensors* **2024**, *24*, 6112. [CrossRef]
34. Wang, T.; Li, Y.; Zhai, Y.; Wang, W.; Huang, R. A Sewer Pipeline Defect Detection Method Based on Improved YOLOv5. *Processes* **2023**, *11*, 2508. [CrossRef]
35. Cai, X.; Lai, Q.; Wang, Y.; Wang, W.; Sun, Z.; Yao, Y. Poly Kernel Inception Network for Remote Sensing Detection. In *Proceedings of the IEEE/CVF Conference Computer Vision and Pattern Recognition (CVPR) 2024, Seattle, WA, USA*; IEEE: Piscataway, NJ, USA, 2024; pp. 27706–27716.
36. Ma, X.; Dai, X.; Bai, Y. Rewrite the Stars. In *Proceedings of the IEEE/CVF Conference Computer Vision and Pattern Recognition (CVPR) 2024, Seattle, WA, USA, 17–21 June 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 5694–5703.
37. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861. [CrossRef]
38. Guo, J.; Chen, X.; Tang, Y.; Wang, Y. SLAB: Efficient Transformers with Simplified Linear Attention and Progressive Re-parameterized BatchNormalization. In *Proceedings 41st International Conference Machine Learning (ICML'24)*; JMLR: Norfolk, MA, USA, 2024; Volume 235, pp. 16802–16812.
39. Cordonnier, J.-B.; Loukas, A.; Jaggi, M. Multi-Head Attention: Collaborate Instead of Concatenate. *arXiv* **2020**, arXiv:2006.16362.
40. Taud, H.; Mas, J. Multilayer Perceptron (MLP). In *Geomatic Approaches for Modeling Land Change Scenarios*; Camacho Olmedo, M., Paegelow, M., Mas, J.F., Escobar, F., Eds.; Springer: Cham, Switzerland, 2018; pp. 451–455.
41. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation (CIoU). *IEEE Trans. Cybern.* **2021**, *52*, 8574–8586. [CrossRef]
42. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
43. Li, Q.; Guan, Q.; Yu, K.; Gao, S.; Wang, H.; Jiang, J. Detection Algorithm for Safety Helmet Wearing of Chemical Plant Personnel Based on Improved YOLOv5m (Wise-IoU). In Proceedings of the International Conference on Computer Design (ICCD), Washington, DC, USA, 6–8 November 2023.
44. Haurum, J.B.; Moeslund, T.B. Sewer-ML: A Multi-Label Sewer Defect Classification Dataset and Benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
45. Lu, J.; Song, W.; Zhang, Y.; Yin, X.; Zhao, S. Real-time defect detection in underground sewage pipelines using an improved YOLOv5 model. *Autom. Constr.* **2025**, *173*, 106068. [CrossRef]
46. Tang, H.; Chen, F.; Feng, Z.; He, L.; Wu, W. Application Effect of YOLOv8 Algorithm Combined with AIFI Technology in Insulator Detection of Power System. *Res. Sq.* **2025**. [CrossRef]
47. Wang, Z.; Li, C.; Xu, H.; Zhu, X.; Li, H. Mamba yolo: A simple baseline for object detection with state space model. In *Proceedings of the AAAI Conference on Artificial Intelligence 2025*; AAAI Press: Washington, DC, USA, 2025; Volume 39, pp. 8205–8213.
48. Peng, Y.; Li, H.; Wu, P.; Zhang, Y.; Sun, X.; Wu, F. D-FINE: Redefine regression task in DETRs as fine-grained distribution refinement. *arXiv* **2024**, arXiv:2410.13842.

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.