Measurement xxx (xxxx) 120548

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Boosting Low-Light sewer defect detection via Illumination-Aware lightweight YOLO framework

Yanfen Li [a], Hanxiang Wang [a,*], Yuanke Zhang [a], Junliang Shang [a], Guangshun Li [a], L. Minh Dang [b,c,**]

[a] School of Computer Science, Qufu Normal University, Rizhao 276826, China
[b] Institute of Research and Development, Duy Tan University, Da Nang 550000, Viet Nam
[c] Faculty of Information Technology, Duy Tan University, Da Nang 550000, Viet Nam

ARTICLE INFO

Keywords:
Multi-scale features
Transformer
Attention mechanism
Defect detection

ABSTRACT

Sewer defect detection is imperative for urban infrastructure management, as it directly impacts environmental safety and public health. Traditional methods for detecting defects in sewer systems are inefficient, costly, and reliant on expert knowledge. This study introduces an advanced illumination-guided You Only Look Once (YOLO) framework to automatically address the challenges of detecting subtle and imperceptible defects in complex backgrounds. The framework includes a Retinex-inspired feature extractor for brightness enhancement, which integrates convolutional neural networks (CNNs) and transformers to leverage global information and local details. Additionally, it features a Cross Stage Partial with fewer-parameter block (CFBlock) for reduced computational complexity, a pyramid structured Mixed Path Aggregation Network (MPANet) for multi-scale feature fusion, and a dynamic attention module in the YOLO head for improved defect localization. This cutting-edge model architecture not only achieves superior accuracy and efficiency but also boasts competitive parameter count and rapid inference speed, setting a new benchmark in real-time sewer defect detection. Our work represents a significant leap forward in the field of automation in construction, offering a practical and powerful solution for urban infrastructure maintenance and the prevention of environmental issues stemming from undetected sewer defects.

1. Introduction

As an important component of public infrastructure, the operation status of sewer directly affects the urban environmental sanitation and the quality life of residents. With the acceleration of urbanization, the construction and maintenance of sewer systems have become increasingly notable tasks of city management. However, the sewer system is susceptible to defects caused by various factors such as pipeline aging, corrosion, and damage, when exposed to the external environment for a long time. If these defects are not detected and repaired in a timely manner, they may lead to sewer leakage and pollution, which is apt to cause serious environmental pollution, urban traffic paralysis, as well as bring inconvenience and even harm to urban operation and residents' lives. In recent years, massive defect detection methods based on computer vision have gradually developed into the mainstream and achieved good performance. However, how to perfectly distinguish complex backgrounds, subtle or imperceptible defects, and even pseudo defects with similar appearances remains a thorny issue for many researchers.

In the exploration history of defect detection, the earlier methods mainly relied on manual inspection or professional equipment for detection. This kind of traditional sewer detection methods exited problems, such as low efficiency, high cost, and dependence on expert personnel [1]. With the development of machine learning, some detection methods based on hand-crafted features and prior knowledge have been applied to simple defect detection. However, such methods exhibit extreme lack of generalization ability to handle defects in different scenarios [2]. Subsequently, the emergence of CNNs satisfied the high pursuit of precision and efficiency by researchers in this field for the first time. Since the truth that the actual receptive field in CNNs is smaller than the theoretical receptive field, numerous approaches have been proposed to boost the contextual awareness of CNNs. For example, en-

* Corresponding author.
** Corresponding author at: Institute of Research and Development, Duy Tan University, Da Nang 550000, Viet Nam.
E-mail addresses: hanxiang@qfnu.edu.cn (H. Wang), danglienminh@duytan.edu.vn (L.M. Dang).

https://doi.org/10.1016/j.measurement.2026.120548
Received 27 April 2025; Received in revised form 19 January 2026; Accepted 21 January 2026

larging the convolution kernel [3], using atrous convolutions [4], pyramid pooling [5], or incorporating attention mechanisms [6]. Although relevant experimental results have proven the effectiveness of such methods, the inherent locality of two-dimensional (2D) convolution still has limitations on the overall performance of the algorithm. Recently, visual transformers have gained popularity across various applications, because of their capacity for extensive contextual analysis over extended sequences. However, the pure transformer's model property limits its ability to learn local details in pixel-wise fine defect localization tasks [7]. Through the above discussion, it is necessary to construct an algorithm that can simultaneously consider both global information and local details. From now on, the newest effort is to integrate the advantages of CNN and transformers into a unified model architecture.

Based on the above analysis, this paper proposes an interpretable fusion network architecture for sewer defect detection. This fusion network combines CNNs and transformers to address their respective limitations and enhance their respective advantages.

This study's key contributions can be encapsulated in the following points:

(1) A feature extractor that utilizes Retinex theory and Transformer mechanisms is designed to obtain the features with enhanced brightness.
(2) Integrating the newly designed CFBlock into Retinex-Inspired CFBlock (RCFNet) reduces computational complexity compared to traditional convolutional blocks, making the network more suitable for real-time applications.
(3) A proposed pyramid structure named MPANet effectively fuses features at different scales, capturing both global and local details essential for accurate defect detection.
(4) The enhancement of the YOLO head with a novel attention module that dynamically integrates features, leading to improved accuracy in identifying and localizing defects in sewer systems.

The structure of the remainder of this paper is as follows. Section 2 offers an overview of the existing literature on defect detection utilizing a variety of algorithms. Section 3 provides an in-depth explanation of the methods and strategies proposed in this study. Section 4 showcases the unique experimental outcomes along with a thorough analysis. Conclusively, Section 5 presents the final summary of the research.

2. Related work

Since the manual inspection based on the hand-crafted features and knowledge is a time-consuming and error-prone task [2]. Recently, the CNN-based algorithms have achieved impressive results in defect detection field, ascribing to the excellent capability in learning the features with different levels. Although numerous CNNs acquire large receptive field (RF) using the analogous structures with two dimensional convolutions, such strategies cannot present an ideal performance. Because the empirical RFs of CNNs are actually smaller than the notional one. Thus, various approaches were applied in CNN models to obtain the global dependency. For example, the enlarged kernel was used to explore the contextual correlation within a larger image patch [3]. But this kind of solutions are followed by huge computational burden. Alternative method attempts to gain the global information without raising the parameters by involving the atrous convolution [8]. Chollet and François presented pyramid pooling to extract objective features [9]. In addition, some studies introduce attention mechanism to deal with the problem under the channel and spatial dimension [10,11]. However, such designs in CNNs need complex architecture and considerable computation cost.

Different from the CNN-based methods, latterly emerging transformer-based methods can link each part of the data via the self-attention mechanism. The transformer architecture is initially constructed for natural language processing, and later it is introduced to object detection in recent advances owing to its inherent capture ability of the contextual relations. Zhou et al. proposed a lightweight vision transformer (ViT) in the defect detection network [12]. In another work, ViT was firstly applied as an encoder in the segmentation model [13]. But such transformers produce columnar outputs, which brings poor results. In order to solve the above problem, the pyramid structure was added into the transformer to create the multiple outputs [14].

Although both CNNs and transformers achieved good results, there still exits room for improvement, especially for some complicated cases. Hence, the current research makes the effort to combine the transformers and CNNs into an integrated whole, which can retain their individual advantages of distinct methods. The incipient research adopted transformers on the features learned from CNNs, yet it is restricted to inferior feature representation [15]. More recently, the convolution and pooling layers of CNNs were put into the sections of transformer [16,17].

By reviewing the related literature, it is demonstrated that CNNs and transformer are widely applied to particular sewer defect detection task. A systematic summary comparing the existing methods is shown in Table 1. Tan, Yi, et al. proposed an improved CNN-based YOLO algorithm in order to detect sewage pipe defects [18]. Dang et al. presented a sewer defect detection model by using the convolution-free transformer [19]. In [20], an enhanced transformer-based architecture was put forward to classify and localize the sewer defects in small-sample closed circuit television (CCTV) images. However, the existing CNN-Transformer models failed to strike a balance between the computation and precision, especially under challenging conditions. Also, few studies apply CNN and transformer to a unite model for quality sewer defect detection.

3. Methodology

In this study, the whole flow-process diagram of the presented framework for the proposed sewer defect detection approach is illustrated in Fig. 1. In the data collection step, the data was collected by a specific robot with high-resolution cameras. The obtained data was enhanced by using several preprocessing approaches. In the training step, firstly, the split sewer training images are input into a proposed extractor, RCFNet. The specific design details of the network are described in Section 3.1. Next, the features from the initial feature extraction stage are fed into feature pyramid, MPANet (see details in Section 3.2), for further process. And then, the defects are recognized and localized via the proposed head architecture, which adopted a customized dynamic attention module, see details in Section 3.3. Finally, the model's weight file was saved and used for the testing step.

Table 1
Key challenges in the existing sewer defect detection methods.

Method Type	Key Studies	Approach	Advantages	Limitations
CNN-Only	Tan et al. [18]	Improved YOLO for pipe defect detection.	High precision in controlled conditions.	Struggles with global context; high computation.
Transformer-Only	Dang et al. [19]	Convolution-free Transformer (DefectTR).	Global defect context capture.	Columnar output reduces spatial accuracy.
CNN-Transformer Hybrid	Gao et al. [15], Guo et al. [17]	CNNs + Transformers fused (e.g., UTNet, Cmt).	Balances local/global features.	Computation-precision trade-off; less study applies a unite CNN-transformer model for sewer defect detection

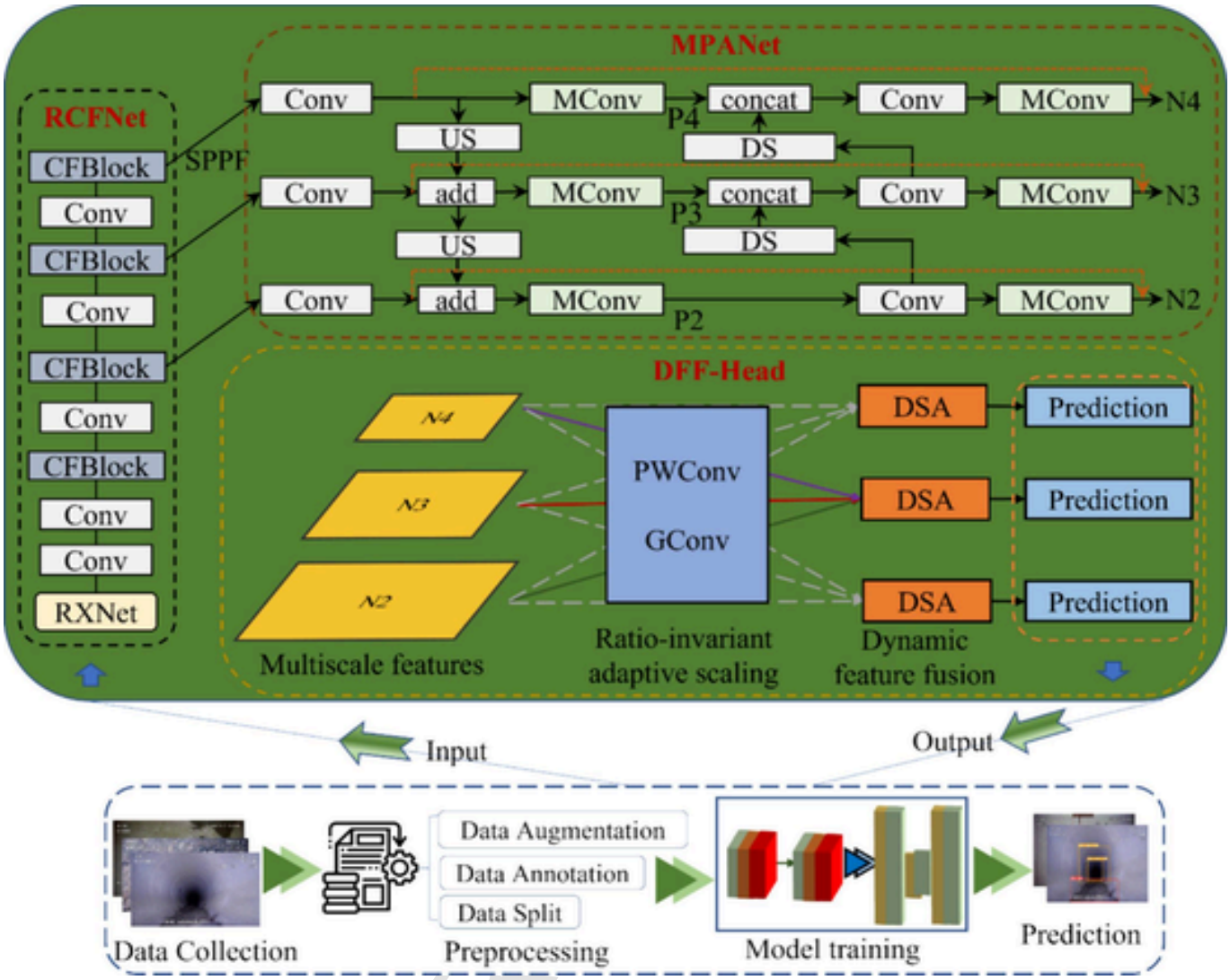


Fig. 1. The overall flowchart of the presented method, which includes data collection step, preprocessing step, model training step, postprocessing step, and the final prediction step.

3.1. *Efficient Retinex-Inspired feature extractor*

The performance of sewer defect detection is significantly influenced by lighting conditions. To address the issue of low light, we have proposed a Retinex-Inspired Feature Extractor named RCFNet. The complete configuration of the model is depicted in Fig. 2 (b), while Fig. 2 (a) illustrates the feature extraction backbone of YOLOv8 [21], C2f-DarkNet. The primary advantage of C2f-DarkNet lies in its ability to maintain high detection accuracy with minimal computational resources. However, the model's performance in low-light conditions is less than optimal, necessitating improvements to the network structure or the introduction of algorithms specifically designed for dimly lit environments. Our proposed RCFNet achieves high-speed target recognition in low-light conditions by incorporating two innovative modules—Retinex-Inspired Network (RXNet) and Cross Stage Partial with fewer-parameter block (CFBlock).

3.1.1. *RXNet*

Since its inception in 1963, Retinex theory [22] has made significant strides in addressing issues of uneven lighting and color bias. This article introduces RXNet, an innovative application of Retinex theory within the framework of modern convolutional neural networks, and it ingeniously incorporates a self-attention mechanism (depicted in Fig. 3).

RXNet, with its lightweight architecture, is strategically positioned at the beginning of the feature extractor and seamlessly integrated into the entire network. This integration enables end-to-end feature enhancement and detection. RXNet consists of two core modules: the lighting analyzer and SmoothNet. The lighting analyzer is responsible for deeply analyzing the lighting information within the input image, capturing and extracting global lighting features. The design of this module is greatly inspired by Retinex theory. By closely integrating the processes of lighting processing and feature extraction, it effectively improves the model's capacity to enhance images under low-light conditions. The essence of Retinex theory lies in its unique insight into the image formation mechanism, where the pixel intensity of an image can be seen as the product of the object's reflectance and the scene's lighting (Equation (1).

$$I(x,y) = R(x,y) \bullet L(x,y) \tag{1}$$

Here, $I(x,y)$, $R(x,y)$, and $L(x,y)$ represent the intensity, reflectance, and illumination at pixel (x,y), respectively. Light analysis involves ex-

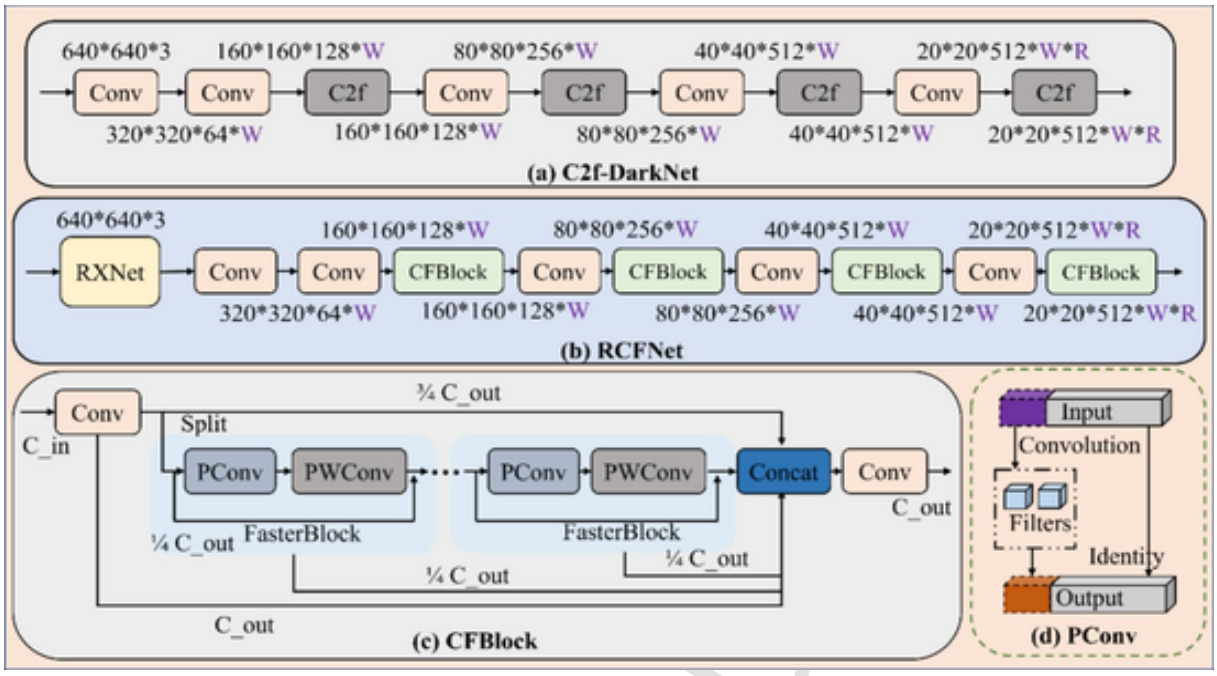


Fig. 2. The architecture of C2f-DarkNet (a) and the proposed RCFNet (b). RCFNet improves upon C2f-DarkNet by incorporating specialized modules for low-light performance: RXNet and CFBlock. (c) details the CFBlock module, which utilizes PConv (d) to reduce computation by processing only a fraction of input channels (C_in), followed by PWConv to fuse features, ultimately outputting the full channel dimension (C_out).

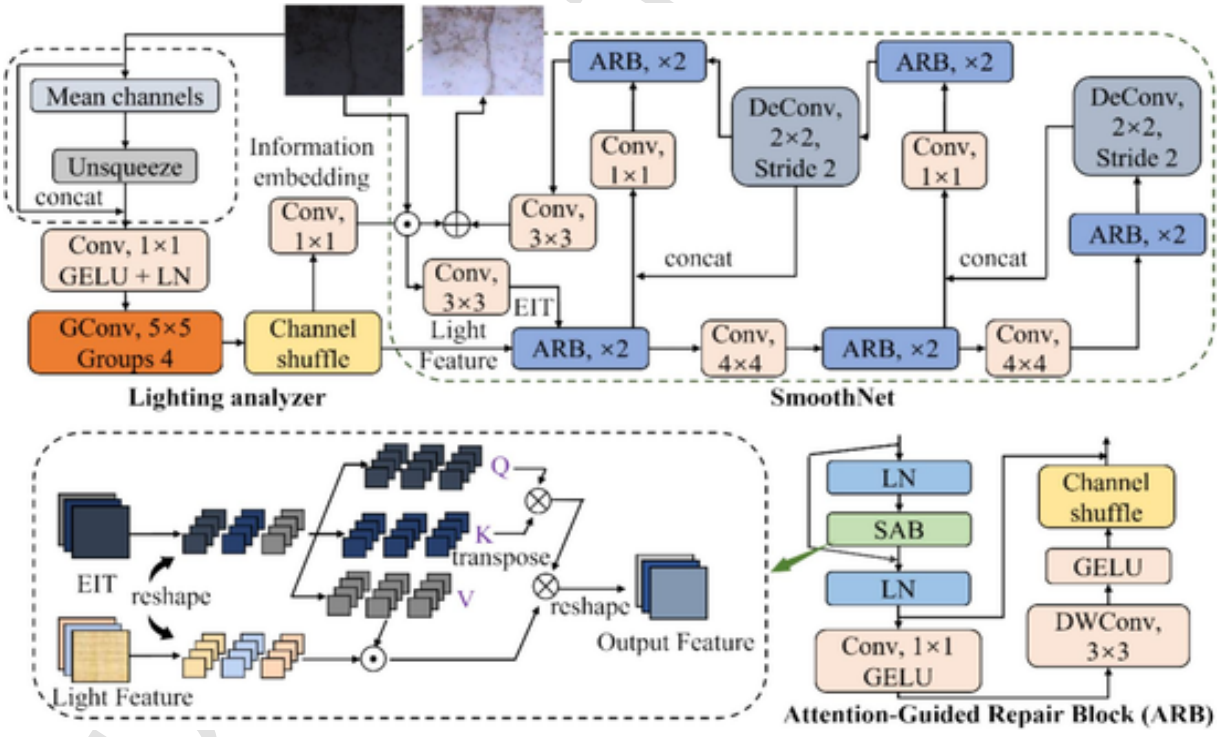


Fig. 3. Core components of RXNet for low-light enhancement and noise reduction (Lighting Analyzer & SmoothNet). **Note:** DeConv (Transposed Convolution), GConv (Grouped Convolution, groups = 4), DwConv (Depth-Wise Convolution), LN (Layer Normalization), GELU (Gaussian Error Linear Unit), SAB (Multi-Head Self-Attention Block), EIT (Enhanced Input Tensor).

tracting $L(x,y)$ from the input image. It begins by calculating the global light estimation $M(x,y)$ as the mean across the image channels. This mean map $M(x,y)$ is then concatenated with the original image $I(x,y)$ to form the input feature map $I'(x,y)$ for further processing.

$$M(x,y) = \frac{1}{C}\sum_{i=1}^{C} I_i(x,y) \tag{2}$$

$$I'(x,y) = Concat(I(x,y) \bullet M(x,y)) \tag{3}$$

The concatenated input undergoes feature extraction through a 1×1 convolution (denoted as C_1), followed by an activation function ($GELU$) and Layer Normalization (LN). This sequence of operations yields preliminary lighting features. Subsequently, a group convolution operation (GC) is employed to refine the feature information further, which can be construed as modeling the local variations in illumination. Since GC (such as DWConv) processes each group of channels independently, it limits inter-channel interaction. To address this issue, a Channel Shuffle (CS) operation is applied to rearrange the channel positions, enabling information flow across groups and enhancing the expressive power of the features. Finally, a second 1×1 convolution (C_2) produces the definitive illumination feature map $L(x,y)$.

$$F(x,y) = LN(GELU\left(C_1\left(I'(x,y)\right)\right)) \tag{4}$$

$$F'(x,y) = CS(GC(F(x,y))) \tag{5}$$

$$L(x,y) = C_2(F'(x,y)) \tag{6}$$

SmoothNet is designed to eliminate noise and overexposure in the features extracted by upper layers, adeptly capturing and restoring the fine details within the image through a multi-stage encoding and decoding architecture. The feature integration and up/down-sampling between layers allow the network to process and reconstruct image information at various resolutions, culminating in the output of a denoised result.

The core component of SmoothNet is the Attention with Residual Blocks (ARB), which takes as input the illumination features along with the enhanced input vectors. As depicted in Fig. 3, input features are subjected to LN before being directed into the Multi-Head Self-Attention Block (SAB). Within the SAB, the input features (X) are transformed linearly to yield the query vectors (Q), key vectors (K), and value vectors (V). Subsequently, an attention weight matrix (A) is computed, a process that necessitates scaling the dot product by the dimensionality of the key vectors (d_k) to regulate the influence of each component within the mechanism.

$$Q = W_Q \bullet X, K = W_K \bullet X, V = W_V \bullet X \tag{7}$$

$$A = softmax(\frac{Q \bullet K^T}{\sqrt{d_k}}) \tag{8}$$

The attention weight matrix (A) is utilized to weight the value vector (V), thereby generating a novel result of attention-weighted features. These features are then element-wise multiplied with the Illumination Attributes (IA) for further weighting, enhancing feature expressiveness and adaptability. The multi-head self-attention output is formed by concatenating the results from all attention heads ($head_1, head_2, \cdots, head_n$), amalgamating information from different subspaces. This concatenated output undergoes a linear transformation (ω) to refine and output the final features, ensuring flexible and deep feature transformation:

$$V' = A \cdot V \odot IA \tag{9}$$

$$MultiHead(Q,K,V) = \omega \cdot concat(head_1, head_2, \cdots, head_n) \tag{10}$$

To enhance the model's stability and information transmission capability, ARB employs multiple residual connections. Additionally, the layer LN and CS operations maintain the original information of the input while strengthening the expression of features related to brightness, thereby enabling effective extraction of key information even in low-light environments.

3.1.2. CFBlock

In the field of neural network research, the efficiency of feature extraction is a key metric for measuring model performance. The feature extraction of CNNs primarily relies on convolutional operations, the design of which should be inspired by the mechanisms of the biological visual system. In this study, we propose a novel structure, CFBlock, to replace the C2f module in C2f-DarkNet. As shown in Fig. 2 (c), CFBlock employs a multi-path feature fusion strategy, effectively simulating the integration process of coarse-grained and fine-grained features in the biological visual system, thereby enhancing the efficiency and accuracy of feature extraction.

To further reduce computational complexity, we adopted a feature extraction approach in CFBlock that combines Partial Convolution (PConv) [23] and Pointwise Convolution (PWConv) [24]. The fundamental concept of PConv involves performing convolutional processes on a limited range of channel features from the input map (as illustrated in Fig. 2(d)), leaving the other channels unchanged. This reduces the computational load and enhances processing speed. PConv relies on a binary mask to distinguish valid data points, where valid points are marked as 1, and missing points are marked as 0. During the convolution operation, only positions where the mask value is 1 are convolved, and the resulting convolution outputs are normalized to ensure comparability across different regions. Following the PConv, we employ a PWConv that performs 1x1 convolutions across all channels, thereby integrating the channel information not processed by PConv. This ensures that the network can capture a more comprehensive feature representation. With this combination, CFBlock not only enhances the efficiency of feature fusion but also reduces computational costs.

3.2. Mixed path Aggregation network (MPANet)

The PANet [25] is an innovative architecture specifically designed to enhance feature extraction in object detection by effectively aggregating information across multiple levels of a feature pyramid. Building upon the Feature Pyramid Network (FPN) [26], PANet introduces additional pathways for feature fusion, ensuring that both high-level semantic details and low-level fine-grained characteristics are fully utilized. With its unique bottom-up enhancement pathway from N2 to N4, PANet strengthens the construction of the feature pyramid, optimizes the information flow mechanism, and achieves accurate detection of targets across various scales.

Although PANet significantly improves feature representation and detection accuracy, its complex feature fusion process also incurs a relatively high computational cost. To address this issue, our study has designed a hybrid convolutional module (as shown in Fig. 4 (b)), which replaces the traditional 3x3 convolutions within the Top-Down and Bottom-Up paths of PANet. This designed module named MConv integrates Partial Convolution (PConv), Pointwise Convolution (PWConv), and channel shuffle (CS) operations to achieve an efficient fusion of feature maps. Here, this approach alleviates the computational burden on the model without compromising its precision. Specifically, PConv is configured at ratios of 1/4, 1/2, and 3/4 to downsample the input features, so it performs convolution on only 25%, 50%, or 75% of the input channels while leaving the rest untouched. This enables a tunable trade-off between computational saving and feature fidelity. PWConv further refines feature extraction, and the final shuffle operation effectively integrates the results of the two convolutions.

Additionally, to address the issue of feature attenuation in neural networks, where the feature signals gradually weaken as the network depth increases, making it difficult for deep networks to capture important detailed information, we have introduced residual structures in

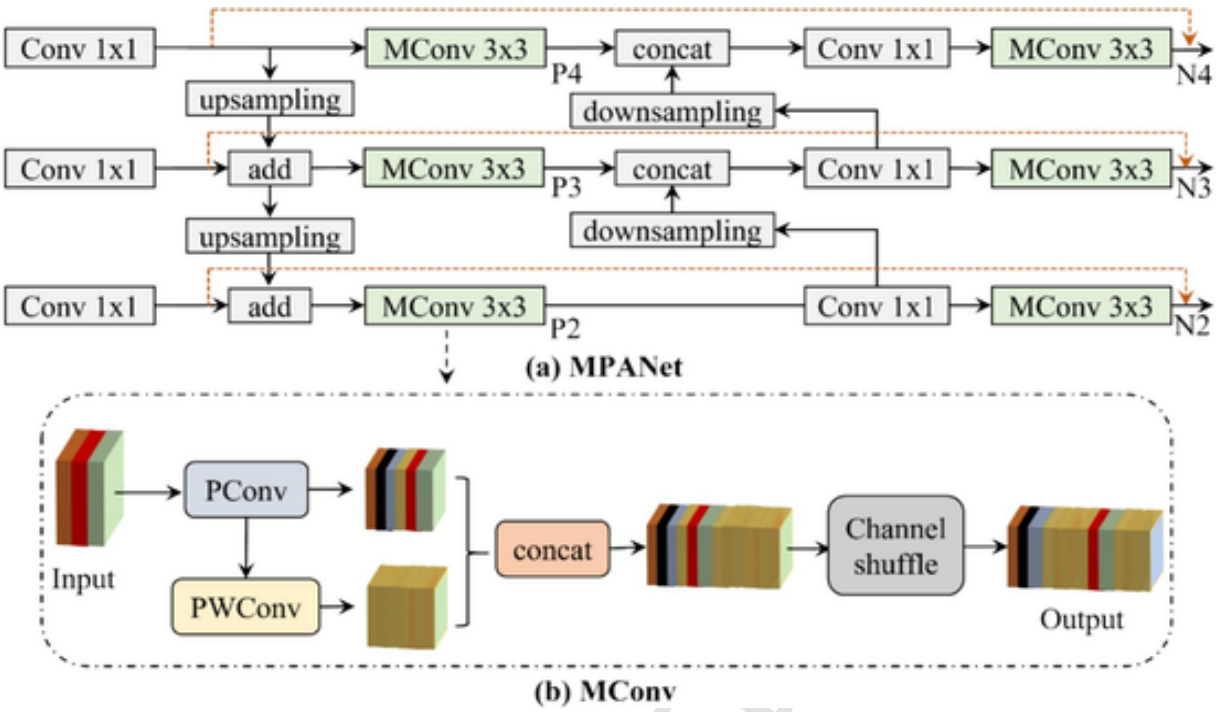


Fig. 4. The structure of MPANet (a) and its combined module MConv (b), where each fusion node in the bi-directional feature pyramid employs MConv—an efficient three-step operator consisting of PConv that performs 3 × 3 depth-wise convolution on only a selectable quarter, half, or three-quarters of the channels, followed by PWConv to fuse the partial features and restore the full channel dimension, and a final Channel Shuffle to reorder channels and enhance cross-group information flow.

every feature fusion path. This enables the module to maintain the original detailed features through multiple convolutional operations.

3.3. Dynamic feature fusion head (DFF-Head)

In multi-scale object detection algorithms, model performance heavily relies on the effective extraction of contextual information. Therefore, we introduce an innovative head architecture for the integration of features (as shown in Fig. 5), allowing the model to efficiently utilize both local features and global context. First, feature maps of three different scales undergo ratio-invariant adaptive scaling using Pointwise Convolution (PWConv) or Grouped Convolution (GConv). Next, the Dynamic Spatial Attention (DSA) module extracts attention maps from the three input feature maps and computes the corresponding attention weights. This weight matrix dynamically re-weights important discriminative features across the input feature maps, ensuring that the most relevant features from each scale are emphasized and effectively fused.

Finally, the model performs convolution operations and loss computation to predict both detection and classification tasks. Within the DSA module (brown dashed box in Figure), input features $I \in \mathbb{R}^{H\times W\times C}$ undergo channel encoding along both spatial dimensions. This generates direction-sensitive feature vectors by averaging features horizontally and vertically:

$$F_H = W^{-1}\sum_{x=1}^{W} I(H,x) \tag{11}$$

$$F_W = H^{-1}\sum_{y=1}^{H} I(y,W) \tag{12}$$

yielding the orientation-aware representations $[F_H, F_W]$ (Equations (11) and (12).

The generated vectors pass through the CBS convolutional block to capture inter-channel correlations. After that, two 1 × 1 convolutions restore channel dimensions, producing directional attention maps $[M_i^H, M_i^W]$. Softmax activation is applied to these maps across three feature levels, generating attention weights $[T_i^H, T_i^W]$ that dynamically weight the input feature maps. The final output feature map O_i for level i is computed through element-wise multiplication, as defined by Equations 13–15.

$$T_i^H = \frac{\exp(\omega M_i^H)}{\sum_{j=1}^{3}\exp(\omega Z_j^H)} \tag{13}$$

$$T_i^W = \frac{\exp(\omega Z_i^W)}{\sum_{j=1}^{3}\exp(\omega M_j^W)} \tag{14}$$

$$O_i = I_i \times T_i^H \times T_i^W \tag{15}$$

4. Sewer defect dataset

The experimental dataset utilized in the study was developed through the collection and meticulous validation of defect images from diverse locations within the Korean sewer system, captured via CCTV inspection videos acquired by the Civil Engineering and Building Technology institute in Seoul, Korea. The dataset encompasses 4,383 original images, identifying 5,385 defects. To enhance the dataset's performance, augmentation techniques such as cutout, Gaussian blur, and channel shuffle were implemented. After enlarging the dataset, there are 3506 and 877 images for the training and testing sets, respectively. In the data annotation process, a total of 8 classes frequent sewer defects are labelled via the LabelMe tool. In order to better describe the

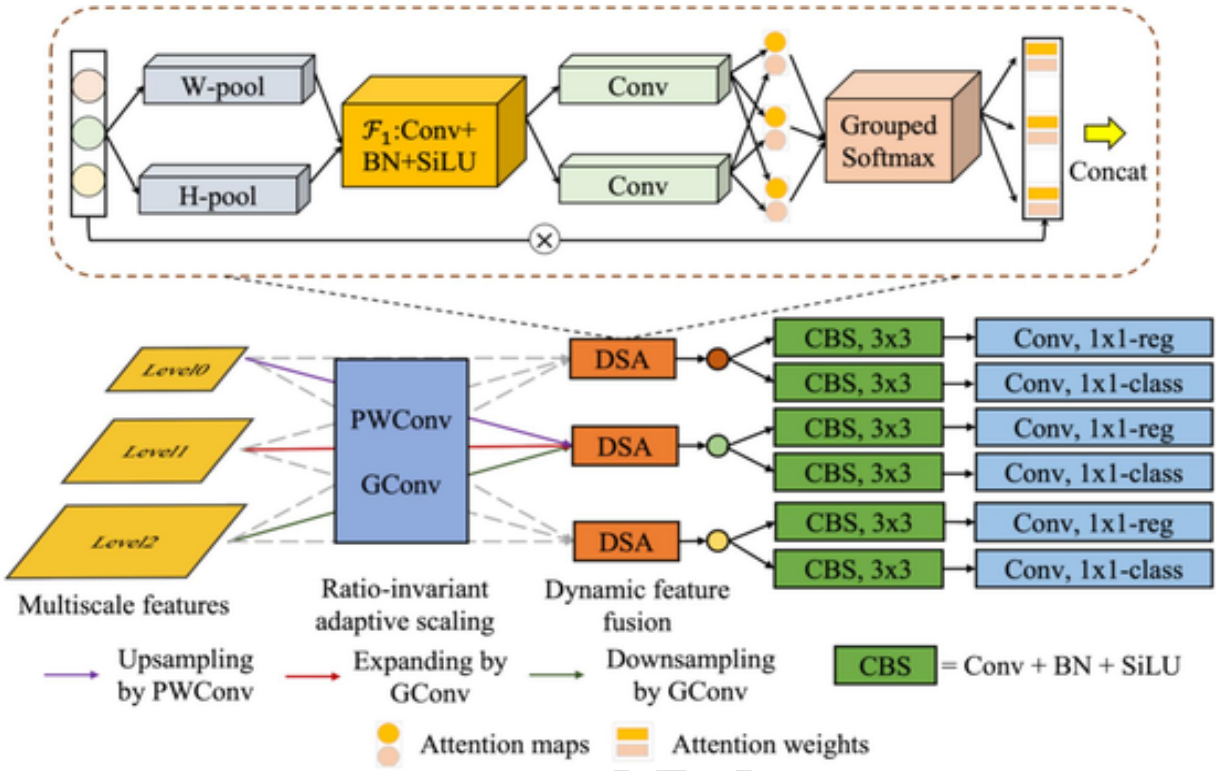


Fig. 5. The architecture of the proposed DFF-Head. The features from the backbone are first processed by the DSA to achieve dynamic feature fusion, then passes them through two CBS blocks followed by separate 1 × 1 convolutions for class (cls) and regression (reg) prediction, while GConv/PWConv and W/H-pooling branches supply complementary spatial context. Level 0, Level 1 and Level 2 denote the three detection scales.

defects, each class is named by the specific code. The details of the experimental dataset are given in Table 2.

5. Experimental results

In the section dedicated to the experimental setup of our proposed sewer defect detection research, we provide a comprehensive overview of the hardware and software environment that was utilized to conduct our experiments. To be specific, we harnessed the capabilities of a high-performance server equipped with two RTX 4090 GPUs, a configuration that has become standard for deep learning applications. Furthermore, our server was decked out with an Inter(R) Core (TM) CPU i9-14900 k running at 6.0 GHz, coupled with a substantial 256 GB of RAM. This setup offered ample computational power for both the training and evaluation phases of our deep learning models. For the development of our presented framework, we opted to employ PyTorch as our deep learning library, driven by its widespread adoption and user-friendly interface. The experimental procedures were carried out utilizing the Ubuntu 18.04 OS. In order to validate the dependability of our findings, we executed all experiments under identical hardware and software conditions. This methodological choice allows for an equitable assessment of outcomes, mitigating any variances that might arise from differences in technological setups.

Table 2
The details of the proposed sewer defect dataset.

Defect name	Defect code	Before data augmentation	After data augmentation	Training set	Testing set
Debris silty	DS	395	1,031	825	206
Horizontal crack	HC	711	1,843	1,474	369
Vertical crack	VC	709	1,852	1,482	370
Joint faulty	JF	770	1,999	1,599	400
Joint open	JO	652	1,698	1,358	340
Lateral protruding	LP	1,020	2,652	2,122	530
Pipe broken	PB	657	1,711	1,369	342
Surface damage	SD	471	1,224	979	245
Total		5,385	14,010	11,208	2,802

5.1. Feature extractor

The purpose of this experiment is to substantiate the efficacy of the proposed RXNet module in the feature extractor (RCFNet). As shown in Fig. 6, the first row lists several input images under low-light conditions, while the second row details the targets present in each input. The third row showcases the enhancement results achieved by RXNet, significantly improving visibility and clarity, which is crucial for accurate detection. The final row presents the predicted detection results, highlighting increases in confidence scores due to RXNet's integration. For example, the confidence score for 'VC' increased by 0.16. Thus, RXNet not only enhances low-light images but also boosts detection accuracy and confidence.

Moreover, the comparative results between the RXNet-equipped models and their baseline counterparts across eight distinct defect categories, along with the mean Average Precision (mAP), are graphically represented in Fig. 7. As illustrated in Fig. 1, the incorporation of the RXNet module yields an improvement over the baseline model in six out of eight evaluated categories. This increase underscores the efficacy of the RXNet in seizing nuanced information that contribute to the accurate identification of defects. Even in categories where the baseline model already performs well, such as 'DS', the RXNet module still manages to further improve precision, albeit marginally. This suggests that

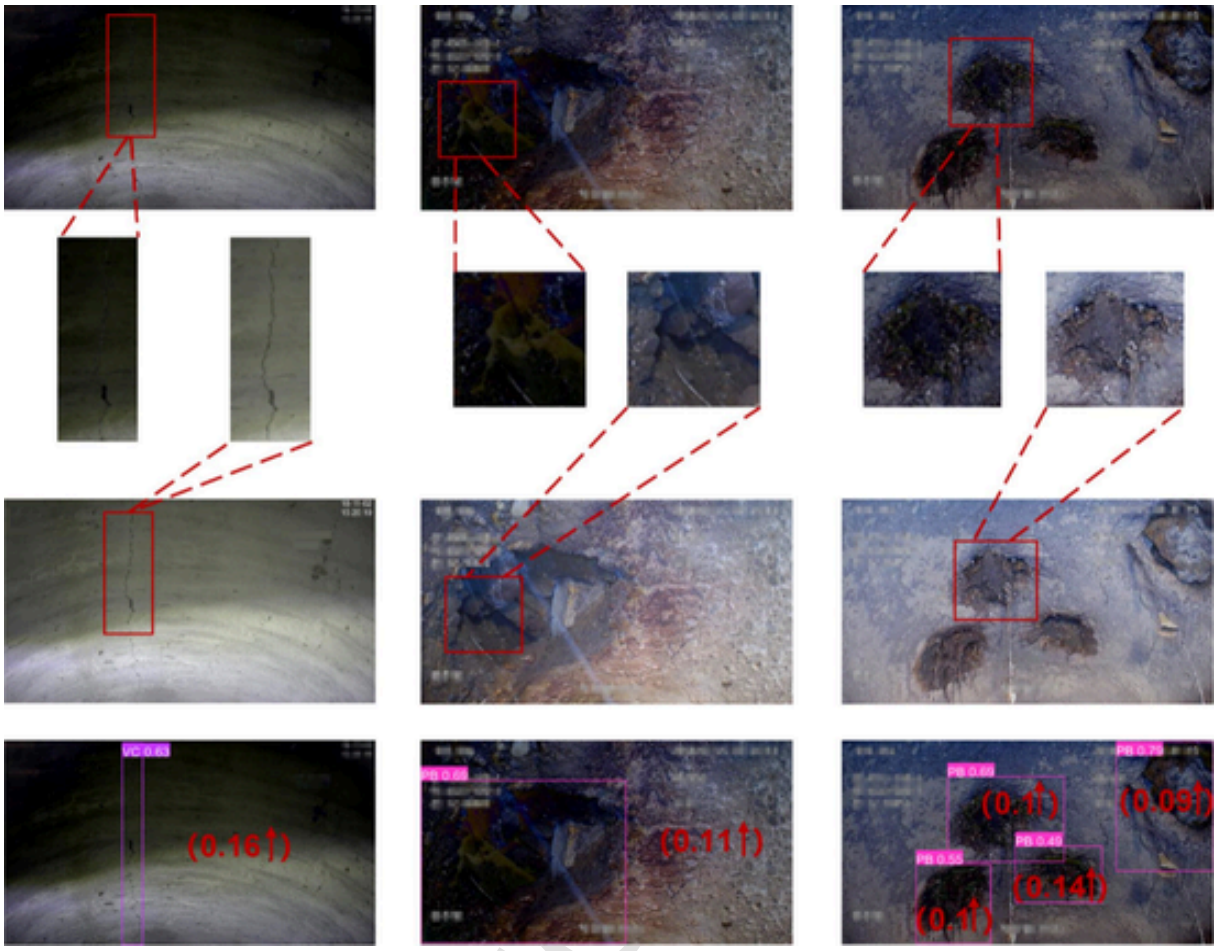
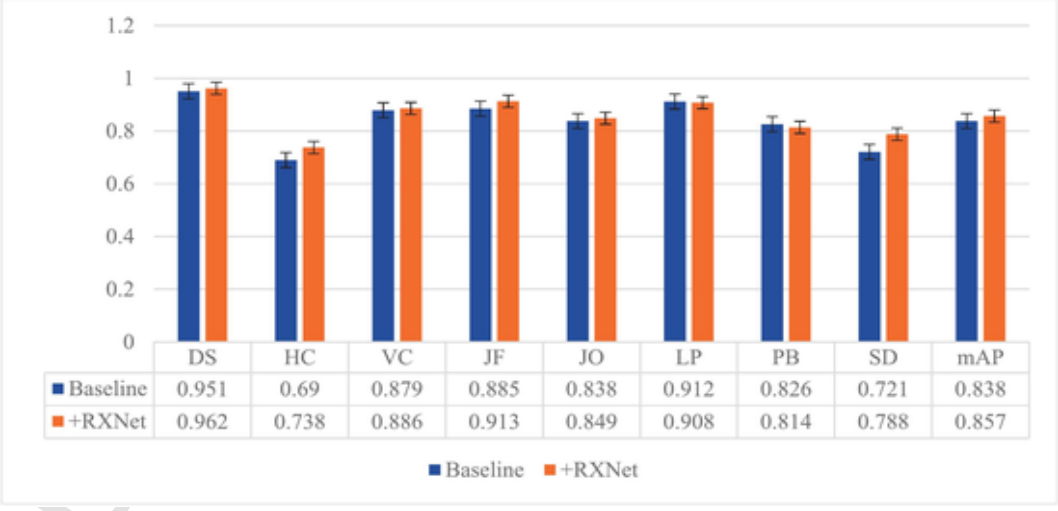


Fig. 6. The effectiveness of the designed RXNet. The initial row consists of input images, while the subsequent row displays the specifics of the targets. The third row low-light enhancement results of RXNet, and the result improvements are in the final row.



	DS	HC	VC	JF	JO	LP	PB	SD	mAP
Baseline	0.951	0.69	0.879	0.885	0.838	0.912	0.826	0.721	0.838
+RXNet	0.962	0.738	0.886	0.913	0.849	0.908	0.814	0.788	0.857

Fig. 7. The performance comparison between the baseline feature extractor and our proposed feature extractor in terms of the average precision (AP) and mean average precision (mAP).

the RXNet module is universally beneficial in enhancing feature extraction for a variety of defect types.

In addition, Table 3 presents the detection performance of YOLOv8 models with and without RXNet enhancement across three low-light illumination levels (High, Middle, Low). The evaluation metrics include mAP and Precision. As demonstrated, integrating RXNet consistently improves performance across all lighting conditions, with more substantial gains observed in darker environments. Specifically, YOLOv8 with RXNet achieves mAP improvements of 4.8, 3.5, and 1.9 percentage points under High, Middle, and Low light conditions, respectively, validating the effectiveness of RXNet for robust low-light object detection.

Apart from the RXNet module, another proposed module was incorporated into the feature extractor. Herein, the results of CFBlock were tested and compared in Table 4 via multiple evaluation metrics, which include the number of parameters (Param.), computational load measured in floating-point operations (FLOPs), mean Average Precision at IoU = 0.5 (mAP_{50}), mAP, and inference time per image. The model is consistent in terms of neck components and the used detection framework. However, there are differences in the backbone for feature extraction. The experiment demonstrates that the RCFNet, which integrates both RXNet and CFBlock modules, achieved superior performance across all evaluated metrics. C2F-CSPDarkNet(n) exhibited mAP scores of 61.5%, with corresponding mAP_{50} scores of 83.8%. The introduction of the RXNet module to the C2F-CSPDarkNet(n) backbone led to a modest improvement in mAP to 63.8% and a more significant increase in mAP_{50} to 85.7%, while also slightly increasing the parameter count and FLOPs. RCFNet has two configurations. The RCFNet (n) has a reduced parameter count of 2.69 M and lower FLOPs of 7.0G. The RCFNet (s) achieved a mAP of 67.9% and an impressive mAP50 of 90.7%, outperforming all the baseline models.

5.2. Feature pyramid network

In this section, an ablation study was presented to evaluate the impact of the proposed modifications to the MPANet Neck structure. The study compares the baseline PANet model with various configurations of the MPANet, assessing the overall performance based on the number of parameters, precision, recall, and F1-score (Table 5). Starting with the baseline PANet (Precision 0.781, Recall 0.863, F1 0.820, 2.69 M params), each row represents incremental modifications and their impact on performance and model size. Replacing PANet's 3 × 3 convolutions with MConv blocks at ratio 1/4 improves the F1 to 0.832 with a 0.43 M reduction in parameters. Upping the MConv ratio to 1/2 adds only 0.05 M parameters while further boosting F1 to 0.840. The full MPANet at ratio 3/4 delivers 0.854 F1 at 2.31 M params, showing diminishing returns beyond half. Finally, introducing shortcut connections increases F1 to 0.869 without any parameter cost. The MPANet Neck thus achieves a 6.0% F1 improvement over PANet while reducing the parameter count by 0.37 M.

In addition, Fig. 8 compares the intermediate feature maps produced by PANet and the proposed MPANet at three different pyramid levels. Across all scales, MPANet yields activation maps that are denser on the object extent and sharper at boundaries, while suppressing background clutter more effectively. This qualitative improvement suggests that MPANet enriches multi-scale representation without relying on heavier computation, providing a better foundation for subsequent detection heads.

Fig. 9 presents a visual comparison of detection results and associated saliency attention maps for PANet (top row) and our proposed MPANet (bottom row). The first two columns illustrate successful detections under standard conditions. In contrast, columns three and four highlight challenging scenarios where both models exhibit failure cases, as marked by yellow boxes. These failures manifest as missed detections and localization offsets, demonstrating that even our enhanced MPANet architecture faces difficulties with particularly challenging defect patterns. While MPANet generally produces more focused attention maps, these examples underscore the persistent challenges in low-light defect detection.

5.3. Dynamic feature fusion module

Table 6 presents an ablation study of four variants, each with different head configurations built upon the RCFNet (n) and MPANet backbone. The baseline RCFNet (n) + MPANet + original head achieves precision of 83.4% and mAP_{50} of 88.2%. The introduction of adaptive scaling results in a modest increase in performance, with Precision rising to 84.6% and mAP_{50} to 89.6%, indicating a positive yet incremental effect. Adding the SA module further lifts Precision to 85.5% and mAP_{50} to 89.9%. SA enhances the model by recalibrating spatial responses, suppressing background noise while highlighting object regions. Replacing SA with SK attention yields 86.3% Precision and 90.6% mAP_{50}. SK dynamically aggregates multi-scale features via an adaptive kernel selection mechanism, providing richer context for objects of varying sizes. For small object detection, mAP_{small} shows consistent gains, rising from 61.5% at baseline to 62.4% with adaptive scaling, 62.6% with SA, 63.1% with SK, and finally reaching 63.7% with DSA. Finally, the model equipped with the DSA achieves the highest scores, with Precision at 86.9% and mAP_{50} at 91.1%. This result underscores the significant impact of the DSA attention module on enhancing the model's precision and reliability in the object detection task.

Fig. 10 presents the class-activation maps (CAM) produced by the model with and without the proposed DSA module. It can be clearly observed that, in the absence of DSA, the saliency map is diffusely distributed: numerous background regions are erroneously activated, whereas the truly discriminative target details (circled in red) exhibit extremely low saliency and are even submerged. After embedding DSA, the dynamic channel–spatial attention recalibrates the feature weights: background clutter is drastically suppressed, and the saliency values are precisely focused on the key textures and edge structures of

Table 3
Effect of RXNet on YOLOv8 performance across different low-light levels.

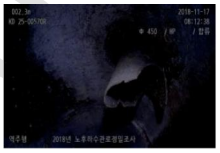
Models	Low-light levels of testing dataset		
	High	Middle	Low
			
YOLOv8 (w/o RXNet)	mAP:70.9 Precision:61.1	mAP:78.8 Precision:68.2	mAP:83.8 Precision:75.2
YOLOv8 (w/ RXNet)	mAP:75.7 Precision:65.7	mAP:82.3 Precision:72.5	mAP:85.7 Precision:77.8

Table 4
Comparative Analysis of Object Detection Models with Different Backbones. **Note:** Metrics comprise the count of Parameters (Param.), Floating Point Operations (FLOPs), mean Average Precision (mAP, mAP_{50}), and Inference Time per Image.

Backbones	Neck	Framework	Param.	FLOPs	mAP (%)	mAP_{50} (%)	Inference (ms)
C2F-CSPDarkNet (n)	PAFPN	YOLOv8	3.08 M	8.4G	61.5	83.8	32
C2F-CSPDarkNet (s)			11.41 M	29.4G	65.2	88.3	64
FasterNet (T2)			15.26 M	37.3G	63.9	86.1	80
C2F-CSPDarkNet (n) + RXNet (Proposed)			3.16 M	9.2G	63.8	85.7	46
RCFNet (n) (Proposed)			**2.69 M**	**7.0G**	64.5	86.8	**30**
RCFNet (s) (Proposed)			8.93 M	23.4G	**67.9**	**90.7**	63

the target. This comparison intuitively verifies that the DSA module guides the model to attend to the most discriminative features.

5.4. Model performance

In this section, we present a comprehensive comparison of various state-of-the-art methods on the proposed dataset, focusing on the detection of eight distinct defect categories. All methods were retrained on the same dataset to ensure a fair comparison. The selection of benchmarks was carefully considered to include a diverse range of state-of-the-art models. Pipe-VFNet [27], based on the MMDet framework, represents advanced anchor-based detection methods. DEIM [28], the current best-performing model within the DETR framework, showcases the potential of transformer-based approaches. Other compared methods such as YOLO-FAS [29], FBRT-YOLO [30], RT-DETRv3 [31] and YOLOv12 [32] are also recent structures that have demonstrated significant performance in object detection tasks.

Table 7 shows that RX-YOLO surpasses other methods in multiple metrics, making it highly suitable for real-time defect detection in challenging conditions like low light. Notably, RX-YOLO integrates all strategies within the model itself, enabling end-to-end detection without requiring any preprocessing or postprocessing steps. This streamlined approach, combined with its lightweight design, significantly en-

Table 5
PANet is the baseline. MConv is a customed convolution block. CS refers to the channel shuffle.

Feature Extractor	Neck						Param.	Precision	Recall	F1-Score
	PANet	Conv → MConv				ShortCut				
		1/4	1/2	3/4	CS					
C2F-CSPDarkNet (n)	√		√		√	√	2.72 M	0.786	0.849	0.816
RCFNet	√						2.69 M	0.781	0.863	0.820
	√	√					2.26 M	0.795	0.873	0.832
	√		√				2.31 M	0.803	0.880	0.840
	√			√			2.35 M	0.807	0.883	0.843
	√		√		√		**2.31 M**	0.819	0.892	0.854
	√		√		√	√	2.32 M	**0.834**	**0.906**	**0.869**

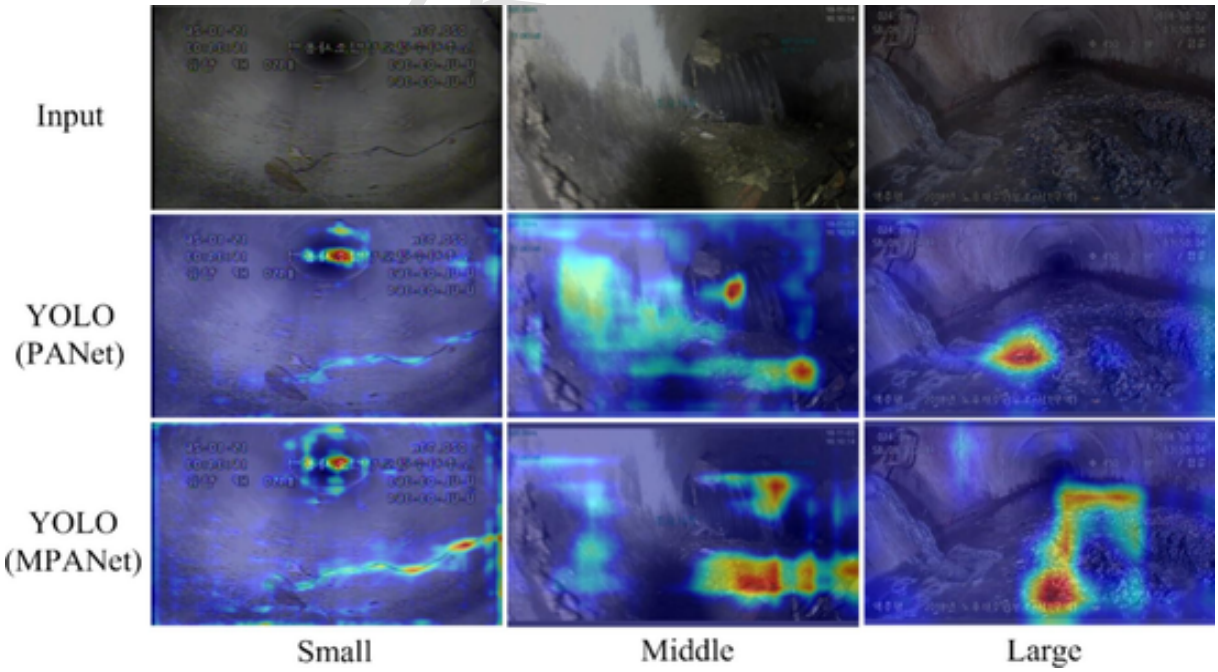


Fig. 8. The saliency attention maps of the PANet and our proposed MPANet under different scales.

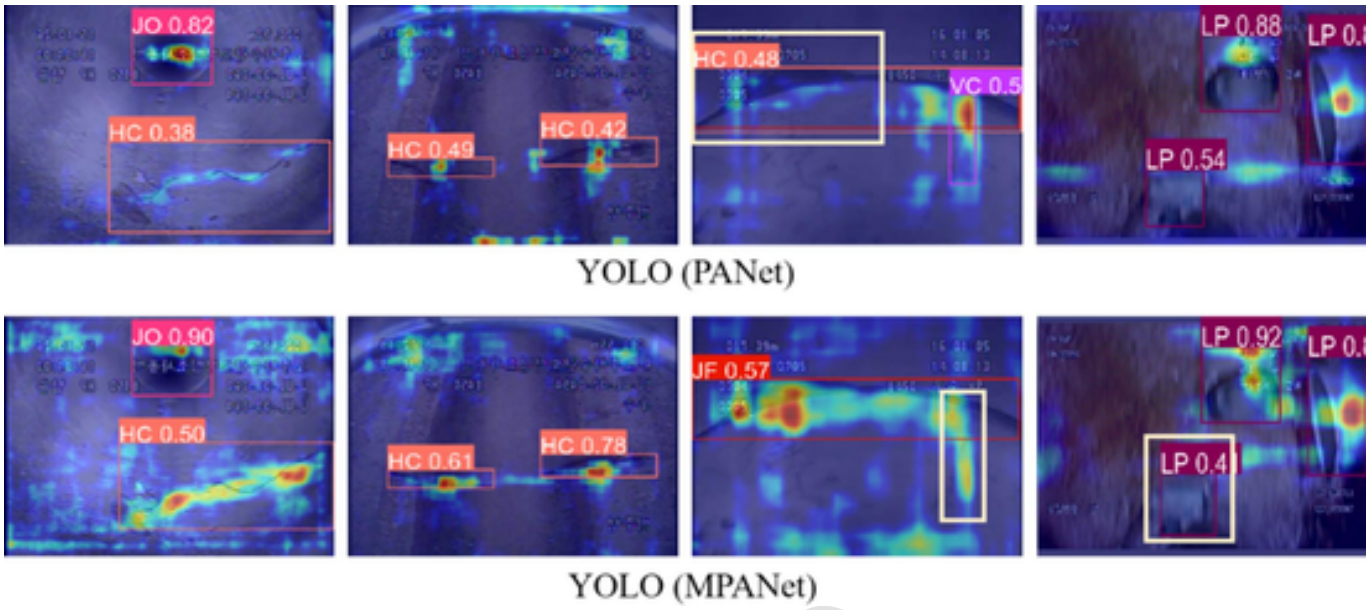


Fig. 9. Visual comparison of detection performance and saliency attention maps between PANet and MPANet. The yellow boxes indicate failure cases.

Table 6
Based on the RCFNet and MPANet, four variants are configured and their respective results are compared. SA, SK, and DSA refer to three different attention modules.

Backbone	Variants				Precision (↑) (%)	mAP_{50} (↑) (%)	mAP_{small} (↑) (%)
	Adaptive Scaling	SA [34]	SK [35]	DSA			
C2F-CSPDarkNet (n) + PANet	√			√	77.6	85.1	59.7
RCFNet (n) + MPANet					83.4	88.2	61.5
	√				84.6 (1.2)	89.6 (1.4)	62.4 (0.9)
	√	√			85.5 (2.1)	89.9 (1.7)	62.6 (1.1)
	√		√		86.3 (2.9)	90.6 (2.4)	63.1 (1.6)
	√			√	**86.9 (3.5)**	**91.1 (2.9)**	**63.7 (2.2)**

hances detection efficiency and achieves a remarkable 40.5 FPS. The model's 2.64 M parameters further exemplify its balance between efficiency and accuracy, rendering it highly effective for real-time applications. RX-YOLO's advanced features, including the RXNet backbone and CFBlock module, are specifically designed for low-light environments, thereby improving feature extraction and detection accuracy. These attributes collectively position RX-YOLO as a strong contender for real-time inspections in demanding settings. Its high FPS rate also ensures rapid processing of video streams or image sequences, which is critical for real-time monitoring systems. RX-YOLO's superior performance across eight defect categories, with the highest mAP scores in five classes, underscores its robustness and adaptability.

Moreover, Table 7 indicates varying performance gains across the different defect categories. Herein, a detailed discussion is provided to analyze the potential reasons for these disparities. To be more specific, although the defects named 'DS' have irregular shapes, they occupy a certain area and have noticeable texture and color differences from the background, making them relatively easy to detect. The excellent performance of RX-YOLO (0.98) indicates that its backbone network can effectively capture this texture change during the feature extraction stage. On the contrary, FCOS (0.81) as an anchor-free method may not be accurate enough for regression of dense targets with irregular shapes. The 'HC' class is one of the most challenging defects, its small characteristics make it a typical "small object detection" problem. RFLA (0.89) performs excellently because its design was intended to address small targets and complex background issues. RFLA may have specifically optimized the feature pyramid, enhancing the feature representation of small targets. However, RX-YOLO (0.78) and other YOLO models performed averagely here, indicating that standard CNN backbone networks and FPNs are prone to information loss in deep networks when extracting such subtle and low contrast linear features. In sharp contrast to HC, almost all models perform much better on VC than HC. This is because the visual features of vertical cracks are more prominent in the collected dataset. The outstanding performance of RX-YOLO proves that it can almost perfectly capture such "relatively easy" linear features. The large size and obvious structural abnormalities make defects such as 'JF','JO', and 'LP' easy targets. All advanced models perform well. The slight differences in performance may be more due to the overall regression accuracy of the model. There is a big difference in 'PB' performance, and the success of YOLOv12 may be due to its model's ability to better understand the global context to determine the damaged state, rather than just detecting local features. As for the 'SD' class, it requires the model to be able to capture subtle texture changes. The leading performance of RX-YOLO is due to its use of more effective attention mechanisms, focusing on local texture differences.

To validate the generalization capabilities of the proposed model, experiments were conducted on the RDD2022 dataset [33], a multinational road-damage benchmark that includes a large proportion of low-illumination images. Figure 11 illustrates a performance comparison of different detection methods based on the mAP_{50} metric. It clearly shows that the proposed RX-YOLO model achieves the highest mAP_{50} value of 59.80, surpassing other state-of-the-art methods such as YOLOv12 and DEIM. This demonstrates the effectiveness of RX-YOLO in detecting objects within the context of the RDD2022 dataset and its potential applicability to other datasets.

In this experiment, multiple defect images captured under diverse conditions are tested and presented in Fig. 12, in order to access the robustness of the detection system. The images depict eight types of sewer defects, which are annotated with bounding boxes and labels for clarity. The first image in the second row illustrates the detection of defects

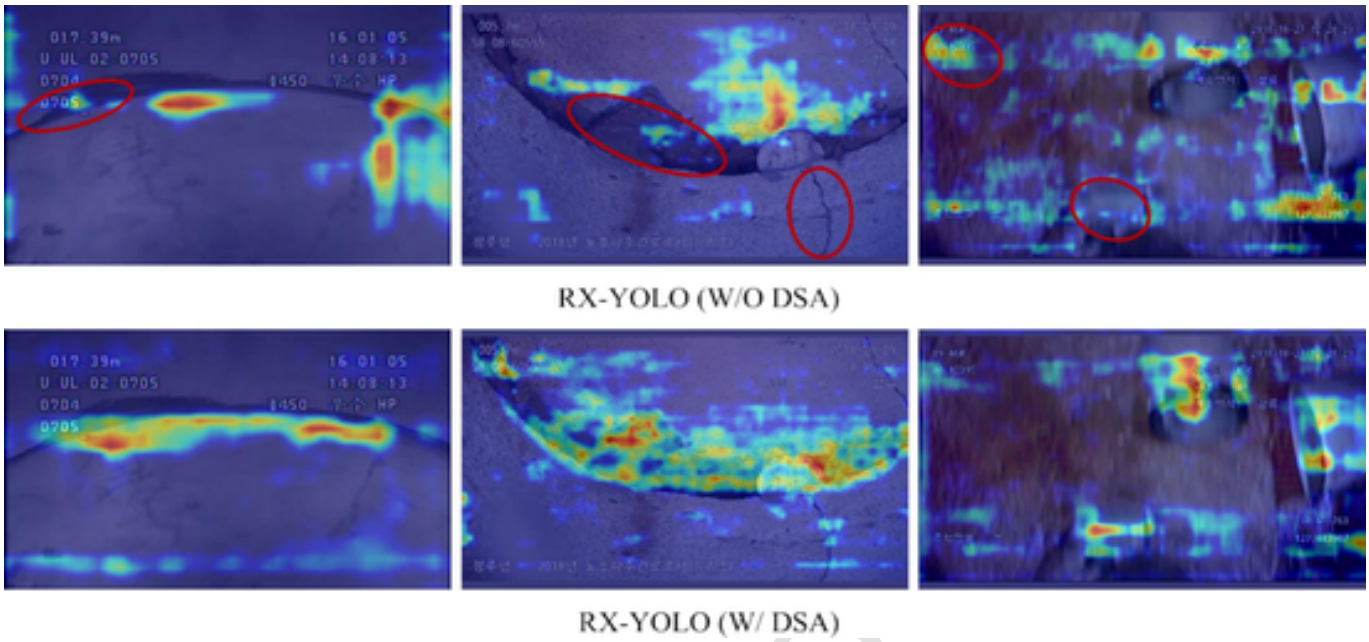


Fig. 10. Saliency-map comparison with vs. without DSA module. Red circles mark unattended regions.

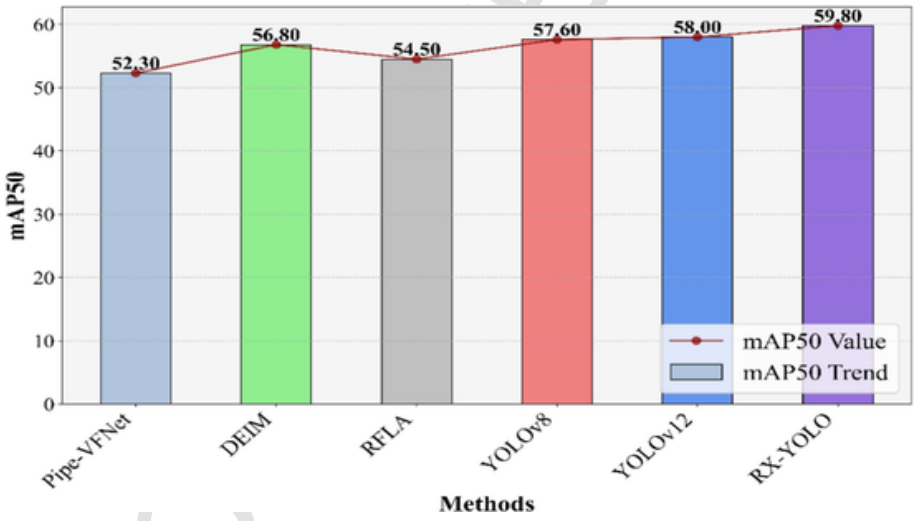


Fig. 11. Performance Comparison of mAP50 Across Different Methods on the RDD2022 Dataset.

Table 7
The comparison with other work based on the proposed dataset.

Methods	DS	HC	VC	JF	JO	LP	PB	SD	Param.	mAP_{50}	FPS
FCOS [36]	0.81	0.49	0.72	0.85	0.74	0.76	0.63	0.68	31.86 M	0.71	16.5
Pipe-VFNet [27]	0.88	0.57	0.70	0.85	0.77	0.79	0.54	0.77	38.62 M	0.73	12.4
HDETR [37]	0.92	0.84	0.71	0.88	0.73	0.82	0.70	0.75	36.84 M	0.79	7.6
DEIM [28]	0.95	0.86	0.82	0.9	0.77	0.82	0.84	0.78	10.19 M	0.84	23.7
RFLA [38]	0.90	**0.89**	0.77	0.93	0.71	0.86	0.79	0.77	41.23 M	0.83	7.2
YOLO-FAS [29]	0.96	0.67	0.89	0.93	0.85	0.9	0.84	0.81	4.32 M	0.86	27.7
YOLOv8 [21]	0.95	0.69	0.88	0.89	0.84	0.91	0.83	0.72	3.08 M	0.84	31.3
YOLOv12 [32]	0.96	0.81	0.90	**0.94**	0.82	0.87	**0.85**	0.76	**2.61 M**	0.86	36.3
FBRT-YOLO[30]	0.93	0.85	0.76	0.89	0.76	0.88	0.81	0.78	2.93 M	0.83	25.5
RT-DETRv3 [31]	0.91	0.86	0.82	0.88	0.81	0.85	0.83	0.75	20 M	0.84	32.1
RX-YOLO (Proposed)	**0.98**	0.78	**0.96**	0.92	**0.86**	**0.91**	0.82	**0.84**	2.64 M	**0.88**	**40.5**

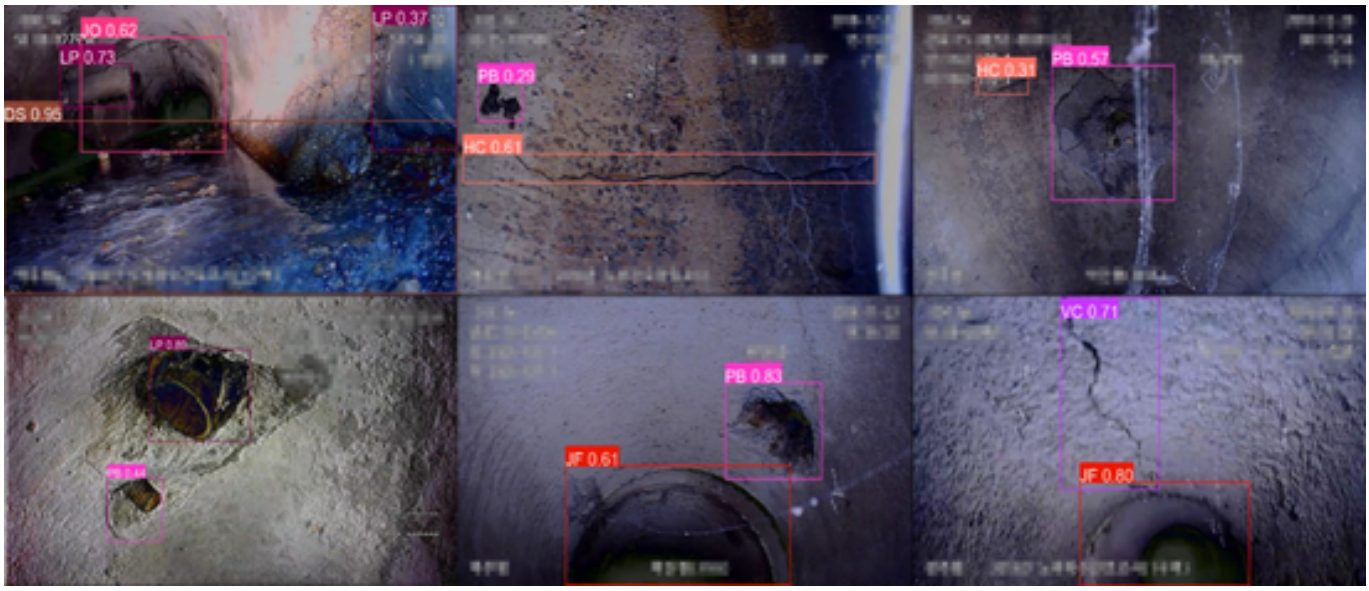


Fig. 12. The visualized prediction results under different conditions.

in a controlled environment, while the first image in the first row presents the system's performance in more challenging scenarios, such as low visibility, densely distributed defects, and complex backgrounds. The consistent identification of defects across these conditions highlights the system's ability to maintain high accuracy and reliability, even in less-than-ideal circumstances. This robust performance is crucial for practical applications where the detection system must operate effectively in a wide range of real-world conditions.

6. Conclusion

In this study, an illumination-guided YOLO framework was presented for sewer defect detection, combining the strengths of Convolutional Neural Networks (CNNs) and Transformers to effectively capture both global information and local details. The backbone of RX-YOLO integrates a lightweight Retinex-based feature extractor with Transformer mechanisms to enhance brightness and extract features, and employs a CFBlock to reduce computational complexity and improve real-time processing efficiency. The neck section features a novel pyramid structure, MPANet, designed for multi-scale feature fusion to capture rich contextual details. Additionally, a dynamic attention module was incorporated into the YOLO head to enhance defect identification and localization.

The proposed RX-YOLO model outperformed state-of-the-art methods across multiple metrics, achieving the highest mAP in five out of eight classes, with a low parameter count and high FPS, demonstrating suitability for real-time applications. The RXNet module notably improved image clarity under low-light conditions, boosting detection accuracy, while MPANet optimized feature integration through hybrid convolution and residuals. The DFF-Head structure further improved multi-scale detection via dynamic spatial attention. This work advances the precision and efficiency of sewer defect detection, contributing significantly to automation in construction.

Despite the achievements of this study, there are some limitations. The presented method was primarily tested on a Korean sewer dataset, so its generalization to other regions and environments needs validation. Computationally, though optimized, it still demands significant resources, especially with large datasets or high-resolution images, limiting scalability. Future work could explore deeper networks for better accuracy and robustness. Techniques like transfer learning from related large-scale datasets or leveraging pre-trained models on similar defect detection tasks could provide a stronger foundation for feature learning. Additionally, semi-supervised or unsupervised learning techniques could significantly reduce the dependency on large annotated datasets. For instance, implementing a self-supervised pre-training phase where the model learns useful representations from unlabeled data through contrastive learning or auto-encoding tasks before fine-tuning on a smaller labeled dataset could be beneficial.

CRediT authorship contribution statement

Yanfen Li: Writing – original draft, Methodology. **Hanxiang Wang:** Writing – review & editing, Supervision, Methodology. **Yuanke Zhang:** Formal analysis. **Junliang Shang:** Validation, Investigation. **Guangshun Li:** Investigation, Formal analysis. **L.Minh Dang:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 62502271 and 62472250), the Natural Science Foundation of Shandong Province (No. ZR2025QC630), and the Natural Science Foundation of Rizhao City (Nos. RZ2024ZR33 and RZ2024ZR34).

Data availability

Data will be made available on request.

References

[1] S.I. Hassan, L.M. Dang, I. Mehmood, S. Im, C. Choi, J. Kang, H. Moon, Underground sewer pipe condition assessment based on convolutional neural networks, Autom. Constr. 106 (2019) 102849.
[2] J. Wang, G. Xu, C. Li, Z. Wang, F. Yan, Surface defects detection using non-convex total variation regularized RPCA with kernelization, IEEE Trans. Instrum. Meas. 70 (2021) 1–13.

[3] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, Q. Meng, PGA-Net: Pyramid feature fusion and global context attention network for automated surface defect detection, IEEE Trans. Ind. Inf. 16 (12) (2019) 7448–7458.
[4] G. Song, K. Song, Y. Yan, EDRNet: Encoder–decoder residual network for salient object detection of strip steel surface defects, IEEE Trans. Instrum. Meas. 69 (12) (2020) 9709–9719.
[5] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.
[6] T. Ding, G. Li, Z. Liu, Y. Wang, Cross-Scale Edge Purification Network for salient object detection of steel defect images, Measurement 199 (2022) 111429.
[7] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: Transformer for semantic segmentation, in: In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7262–7272.
[8] Gao, Y., Lin, J., **e, J., & Ning, Z. (2020). A real-time defect detection method for digital signal processing of industrial inspection applications. IEEE Transactions on Industrial Informatics, 17(5), 3450-3459.
[9] F. Chollet, Xception: deep learning with depthwise separable convolutions, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[10] J. Cao, G. Yang, X. Yang, A pixel-level segmentation convolutional neural network based on deep feature fusion for surface defect detection, IEEE Trans. Instrum. Meas. 70 (2020) 1–12.
[11] D. Zhang, K. Song, J. Xu, Y. He, M. Niu, Y. Yan, MCnet: Multiple context information segmentation network of no-service rail surface defects, IEEE Trans. Instrum. Meas. 70 (2020) 1–9.
[12] H. Zhou, R. Yang, R. Hu, C. Shu, X. Tang, X. Li, Etdnet: efficient transformer-based detection network for surface defect detection, IEEE Trans. Instrum. Meas. (2023).
[13] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, L. Zhang, Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.
[14] Wang, W., **e, E., Li, X., Fan, D. P., Song, K., Liang, D., ... & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 568-578).
[15] Gao, Y., Zhou, M., & Metaxas, D. N. (2021). UTNet: a hybrid transformer architecture for medical image segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24 (pp. 61-71). Springer International Publishing.
[16] Q. Zhang, Y.B. Yang, Rest: an efficient transformer for visual recognition, Adv. Neural Inf. Proces. Syst. 34 (2021) 15475–15485.
[17] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, C. Xu, Cmt: Convolutional neural networks meet vision transformers, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12175–12185.
[18] Y.i. Tan, et al., Automatic detection of sewer defects based on improved you only look once algorithm, Autom. Constr. 131 (2021) 103912.
[19] L.M. Dang, H. Wang, Y. Li, T.N. Nguyen, H. Moon, DefectTR: End-to-end defect detection for sewage networks using a transformer, Constr. Build. Mater. 325 (2022) 126584.
[20] Q. Ren, et al., Automated classification and localization of sewer pipe defects in small-sample CCTV imagery: an enhanced transformer-based framework, J. Civ. Struct. Heal. Monit. (2025) 1–18.
[21] Reis D, Kupec J, Hong J, Daoudi A. Real-time flying object detection with YOLOv8. arXiv preprint arXiv:2305.09972. 2023 May 17.
[22] Land, E. H., & McCann, J. J. (1971). Lightness and retinex theory. Josa, 61(1), 1-11.
[23] J. Chen, S.H. Kao, H. He, W. Zhuo, S. Wen, C.H. Lee, S.H.G. Chan, Run, don't walk: chasing higher FLOPS for faster neural networks, in: In proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12021–12031.
[24] B.S. Hua, M.K. Tran, S.K. Yeung, Pointwise convolutional neural networks, in: In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 984–993.
[25] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.
[26] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[27] Y. Li, H. Wang, L.M. Dang, H.K. Song, H. Moon, Attention-guided multiscale neural network for defect detection in sewer pipelines, Comput. Aided Civ. Inf. Eng. 38 (15) (2023 Oct) 2163–2179.
[28] Huang S, Lu Z, Cun X, Yu Y, Zhou X, Shen X. DEIM: DETR with Improved Matching for Fast Convergence. arXiv preprint arXiv:2412.04234. 2024 Dec 5.
[29] H. Duan, M. Yu, T. Ai, M. Zhu, H. Jiang, S. Guo, YOLO-FAS: a lightweight model for detecting rebar intersections location and tying status, Neurocomputing 1 (624) (2025 Apr) 129485.
[30] Xiao Y, Xu T, Xin Y, Li J. FBRT-YOLO: Faster and Better for Real-Time Aerial Image Detection. InProceedings of the AAAI Conference on Artificial Intelligence 2025 Apr 11 (Vol. 39, No. 8, pp. 8673-8681).
[31] Wang S, Xia C, Lv F, Shi Y. RT-DETRv3: Real-time end-to-end object detection with hierarchical dense positive supervision. In2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2025 Feb 26 (pp. 1628-1636). IEEE.
[32] Tian Y, Ye Q, Doermann D. Yolov12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524. 2025 Feb 18.
[33] D. Arya, H. Maeda, S.K. Ghosh, D. Toshniwal, Y. Sekimoto, RDD2022: a multi-national image dataset for automatic road damage detection, Geosci. Data J. 11 (4) (2024 Oct) 846–862.
[34] Zhang H, Wu C, Zhang Z, Zhu Y, Lin H, Zhang Z, Sun Y, He T, Mueller J, Manmatha R, Li M. Resnest: Split-attention networks. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022 (pp. 2736-2746).
[35] Li X, Wang W, Hu X, Yang J. Selective kernel networks. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 510-519).
[36] Z. Tian, C. Shen, H. Chen, T. He, FCOS: a simple and strong anchor-free object detector, IEEE Trans. Pattern Anal. Mach. Intell. 44 (4) (2020 Oct 19) 1922–1933.
[37] Jia D, Yuan Y, He H, Wu X, Yu H, Lin W, Sun L, Zhang C, Hu H. Detrs with hybrid matching. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2023 (pp. 19702-19712).
[38] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "RFLA: Gaussian receptive based label assignment for tiny object detection," in Proc. Eur. Conf. Comput. Vis., 2022, pp. 526–543.