processes

MDPI

Article

A Sewer Pipeline Defect Detection Method Based on Improved YOLOv5

Tong Wang [1,2,3], Yuhang Li [1,2], Yidi Zhai [1,2], Weihua Wang [4,*] and Rongjie Huang [1,2,3]

[1] Henan Key Laboratory of Intelligent Manufacturing of Mechanical Equipment, Zhengzhou University of Light Industry, Zhengzhou 450002, China; 2009039@zzuli.edu.cn (T.W.); 15137334826@163.com (Y.L.); 18134494997@163.com (Y.Z.); nysyhrj@163.com (R.H.)
[2] College of Mechanical and Electrical Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China
[3] Food Laboratory of Zhongyuan, Luohe 462300, China
[4] China Special Equipment Inspection and Research Institute, Beijing 100029, China
* Correspondence: whlt1982@163.com

Abstract: To address the issues of strong subjectivity, low efficiency, and difficulty in on-site model deployment encountered in existing CCTV defect detection of pipelines, this article proposes an object detection model based on an improved YOLOv5s algorithm. Firstly, involution modules and GSConv simplified models are introduced into the backbone network and feature fusion network, respectively, to enhance the detection accuracy. Secondly, a CBAM attention mechanism is integrated to improve the detection accuracy of overlapping targets in complex backgrounds. Finally, knowledge distillation is performed on the improved model to further enhance its accuracy. Experimental results demonstrate that the improved YOLOv5s achieved an mAP@0.5 of 80.5%, which is a 2.4% increase over the baseline, and reduces the parameter and computation volume by 30.1% and 29.4%, respectively, with a detection speed of 75 FPS. This method offers good detection accuracy and robustness while ensuring real-time detection and can be employed in the on-site detection process of sewer pipeline defects.

Keywords: detection of sewer defects; improved YOLOv5; involution; GSConv; attention mechanism; knowledge distillation

Citation: Wang, T.; Li, Y.; Zhai, Y.; Wang, W.; Huang, R. A Sewer Pipeline Defect Detection Method Based on Improved YOLOv5. *Processes* **2023**, *11*, 2508. https://doi.org/10.3390/pr11082508

Academic Editor: Anna Wołowicz

Received: 12 June 2023
Revised: 25 July 2023
Accepted: 25 July 2023
Published: 21 August 2023

1. Introduction

The sewer pipeline system is one of the important components of urban infrastructure construction and an important guarantee for maintaining the cleanliness and hygiene of cities [1]. With the increase in the service time of pipelines, some sections may develop defects such as misalignment and rupture. In order to ensure the normal operation of the pipeline network, the municipal government invests a lot of manpower and resources every year to carry out daily inspections and maintenance work on it.

Currently, closed-circuit television (CCTV) inspection is the most extensively employed method for pipeline inspection globally [2]. The process of CCTV inspection comprises two stages, namely, on-site video information collection and off-site evaluation. On-site video is collected using either pipeline cameras or robots and is submitted to experts for evaluation. Finally, manual inspection reports are provided by the experts [3]. However, in the process of manual evaluation, there are many factors that can lead to inaccurate results and low efficiency, such as different levels of technical expertise among personnel, varying video quality, and excessive workload.

The development of deep learning technology has brought about many remarkable object detection algorithms in the field, such as Fast R-CNN [4], SSD [5], YOLO [6–9], etc. These algorithms have surpassed traditional computer vision detection techniques in terms of detection accuracy and speed in various application scenarios. The YOLO

models have demonstrated their excellent detection performance in a variety of fields. Sergio and Abdussalam [10] conducted an investigation into the correlation between image size, training time, and the performance of YOLO series models, resulting in the successful detection of a vehicle dataset. Furthermore, Yang et al. [11] employed YOLOv5 for defect detection in steel pipe welds. In recent years, some scholars have applied deep learning technology to the field of pipeline defect detection. Srinath et al. [12] used YOLOv3 as the detection network to locate defects such as tree roots and sediment in pipelines, and compared it with other detection models. Tan et al. [13] utilized Mosaic data augmentation on top of YOLOv3, introduced generalized intersection over union (GIOU), and employed adaptive anchor boxes. Chanmi et al. [14] utilized YOLOv5 as the architecture and incorporated a small object detection layer while introducing attention mechanisms to enhance the detection accuracy of small objects.

Despite the outstanding achievements of the aforementioned models, it is important to note that they are predominantly proposed based on non-field evaluations, overlooking the challenges linked to on-site video collection. Limitations in device computational power, among other factors, can hinder the deployment of these models. Additionally, video quality holds significant importance as a contributing factor to detection performance. Moreover, persisting challenges such as weak lighting conditions and complex background structures within the pipelines pose difficulties in identifying specific defects. Hence, further research is warranted to tackle concerns such as model lightweighting and enhancing detection accuracy.

The aim of the paper is to present a novel enhanced detection model, built upon YOLOv5, aimed at mitigating the challenges prevalent in existing object detection models. The proposed model specifically tackles issues such as suboptimal accuracy, high parameter and computational complexity, excessive memory consumption due to large model weights, and constraints associated with deploying on mobile devices.

In this study, the YOLOv5s model is selected as the foundational detection network. To facilitate efficient deployment, the main network incorporates the Involution [15] operator, while the feature fusion network adopts the GSConv [16] technique to construct a lightweight network structure. Furthermore, to mitigate the effects of challenging factors such as weak lighting and complex backgrounds in pipeline environments on detection accuracy, the CBAM [17] attention mechanism is introduced in the feature fusion network to effectively integrate semantic features across different network layers, thereby augmenting the detection accuracy. Lastly, knowledge distillation [18] is employed, utilizing YOLOv5m as the teacher network, to distill the improved model and further enhance its accuracy and generalization capabilities.

The contributions of this paper can be summarized as follows:

(1) In light of the limitations associated with current pipeline defect detection methods, an enhanced YOLOv5 model is presented in this study. This model strikes a fine balance between lightweight architecture and detection accuracy, thereby facilitating its effective deployment for on-site sewer pipeline defect detection tasks.
(2) Based on the model's detection accuracy, a comparative analysis is conducted to evaluate the effects of three distinct attention mechanisms, namely SE [19], CA [20], and CBAM, on the precision of the improved model. Heatmaps are employed to visually illustrate the regions of interest captured by each attention mechanism. Furthermore, ablation experiments are carried out to examine the impact of different enhancement modules on the detection performance of the network.
(3) The experimental findings unequivocally establish the superior detection accuracy of the enhanced YOLOv5 model in comparison to its original counterpart. Moreover, the improved model showcases reduced parameter and computational complexity, thereby satisfying the real-time detection prerequisites essential for on-site scenarios.

2. Network Architecture

2.1. YOLOv5

YOLOv5 is a widely used one-stage object detection model in engineering projects, with five models of YOLO5n, YOLO5s, YOLO5m, YOLO5l, and YOLO5x based on the depth and width of the network. The complexity and detection accuracy of a model are influenced by different network widths and depths. In the context of sewer pipeline defect detection, where both accuracy and model complexity are crucial, YOLOv5s was selected as the fundamental network architecture for a series of improvements in this study. The architecture of the YOLOv5s network is illustrated in Figure 1.

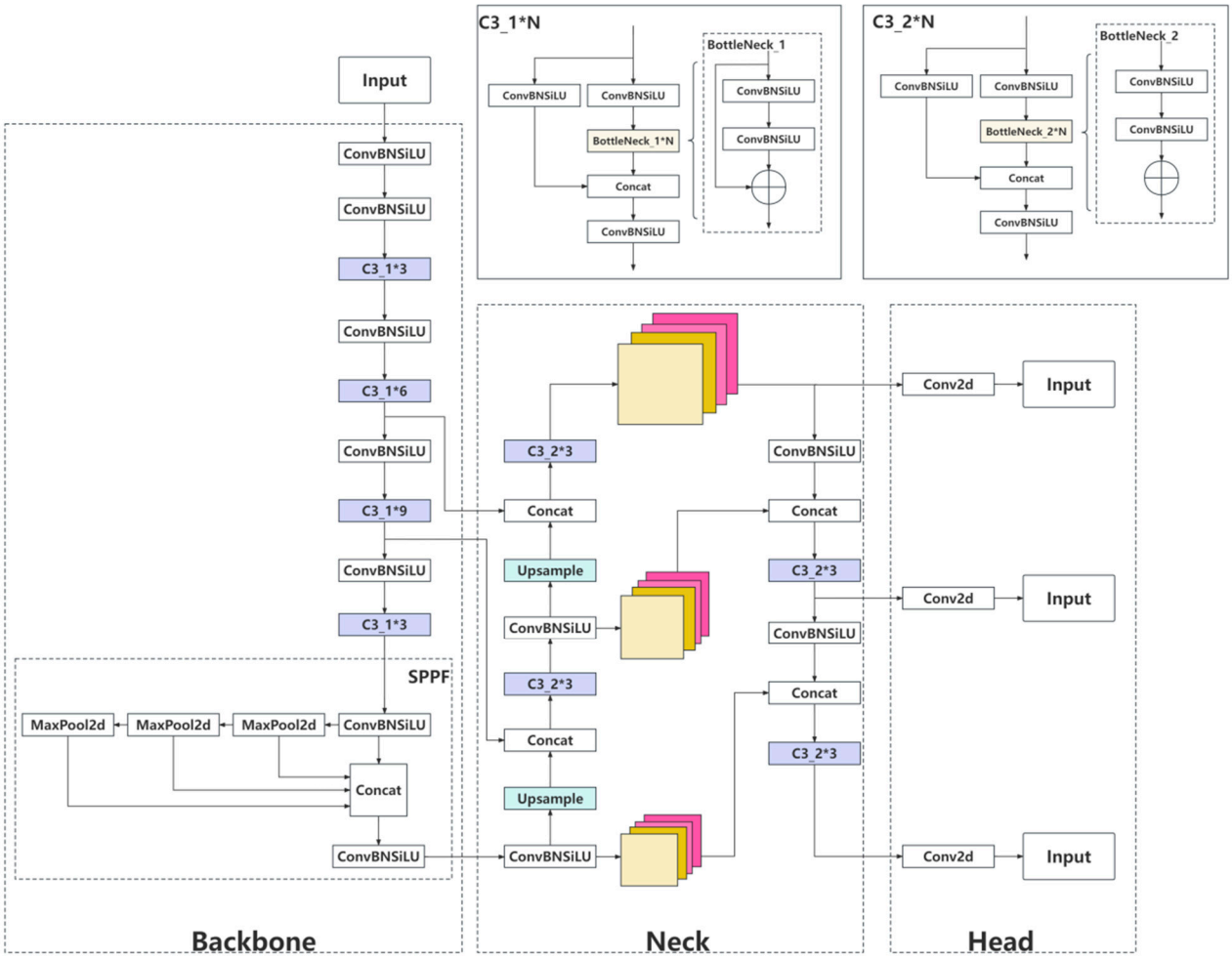

Figure 1. YOLOv5s network structure diagram.

The YOLOv5 network architecture can be divided into four parts: input, backbone, neck, and head. The input part applies data augmentation, adaptive anchor box calculation, and adaptive image scaling to the input images. The backbone of YOLOv5 is a feature extraction network consisting of convolution (Conv) modules, cross stage partial network with 3 convolutions (C3) modules, and spatial pyramid pooling fusion (SPPF) modules. The Conv modules extract features and organize the feature maps, while the C3 modules mainly increase the depth of the network and enhance its feature extraction capabilities. "Spatial Pyramid Pooling Fast (SPPF)" is an improvement based on "SPP" with faster speed. The goal of SPPF is to concatenate the feature representations at different scales of the same feature map. In the YOLOv5 network architecture, "Neck" is a feature fusion network composed of feature pyramid network (FPN) and path aggregation network (PAN). The shallow visual features are combined with the deep semantic features to obtain a more comprehensive feature representation. The "Head" section is composed of three detection

layers of different sizes, which output the final detection results by computing the loss function and performing non-maximum suppression.

2.2. Improved YOLOv5

This paper presents a modified model structure based on YOLOv5s, which is illustrated in Figure 2 (the improved modules are highlighted in green and red in the figure). To reduce the model size, involution and GSConv were introduced into Backbone and Neck, respectively. The detection accuracy of the model is enhanced by incorporating the CBAM attention module into Neck and performing knowledge distillation on the improved YOLOv5s. The aforementioned improvement measures can achieve a balance between detection accuracy and speed by reducing the number of parameters and computational complexity while ensuring detection accuracy. This facilitates deployment on terminals.

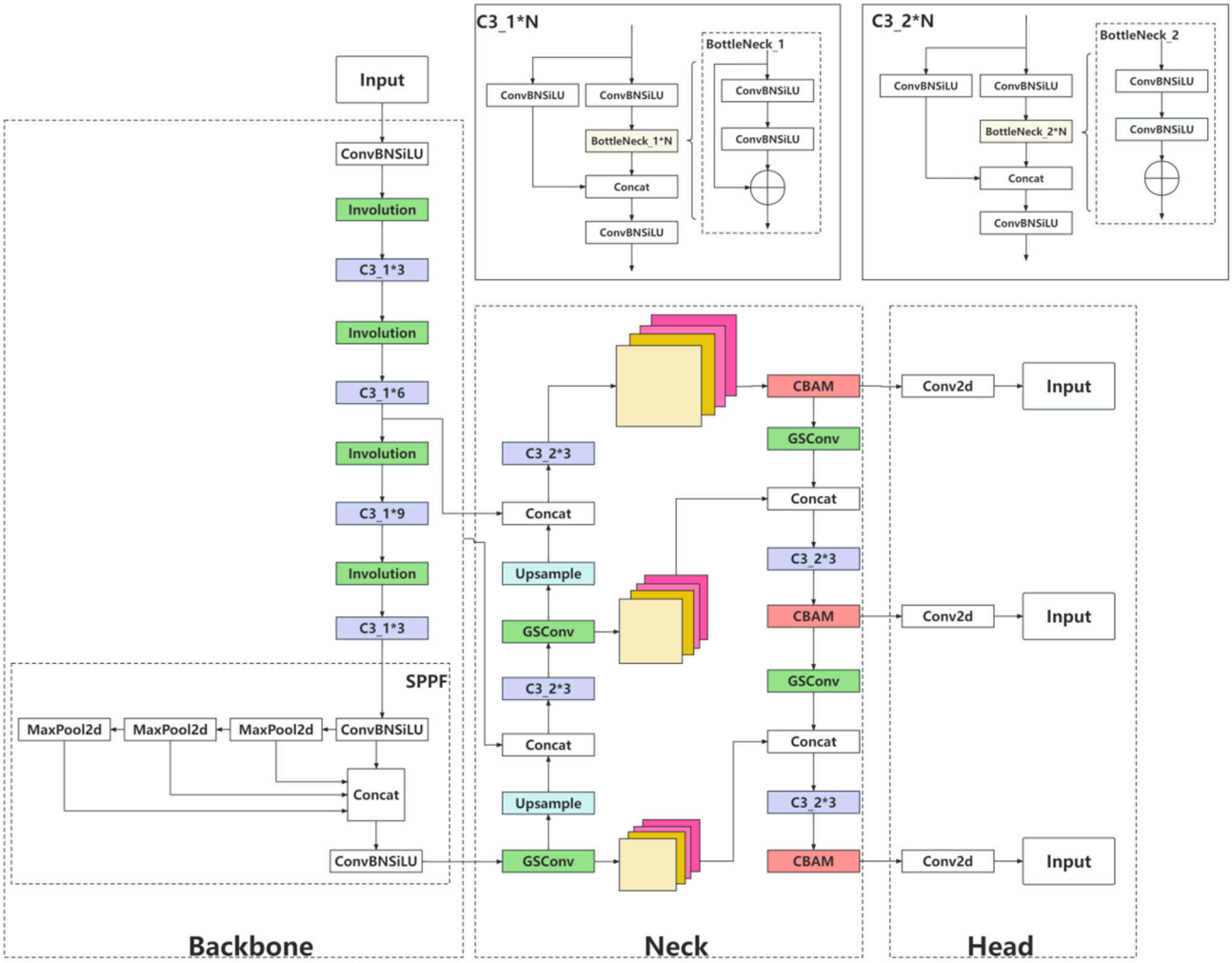

Figure 2. Improved YOLOv5s network structure diagram.

3. Methods

3.1. Involution

Convolution exhibits two fundamental properties: spatial invariance and channel specificity, which enable it to fully exploit the translational equivariance of visual features and the modeling information between channels. Nonetheless, these properties can also restrict the modeling capability of convolution kernels in various spatial positions and result in a substantial computational and parameter cost due to the non-sharing of parameters between channels. Thus, the involution operator is proposed as having opposite characteristics to the convolution operator, specifically spatial specificity and channel invariance. By

sharing parameters between channels, the involution operator can reduce the number of parameters and computational complexity.

The involution operator partitions the number of feature channels into G groups, where each group shares one kernel and different kernels are used for different spatial coordinates. The size of the involution kernel can be represented as $H \times W \times K \times K \times G$, and the output feature map of the involution operator can be denoted by Equation (1):

$$Y_{i,j,k} = \sum_{(u,v)\in\Delta K} \eta_{i,j,u+\lfloor K/2 \rfloor, v+\lfloor K/2 \rfloor, \lceil kG/C \rceil} X_{i+u,j+v,k}, \tag{1}$$

In this equation, X represents the input feature map, Y represents the output feature map, $\eta \in R$ is the kernel vector of the involution operator, R represents the entire pixel coordinate space, $\lceil kG/C \rceil$ denotes the number of shared groups within a channel, and ΔK represents the set of offset values for the neighborhood of the central pixel convolution. Unlike convolution, the involution kernel is dynamically generated based on the input features. Specifically, the input feature map X is mapped to form a dynamic convolution kernel, which can be expressed in a general form using Equation (2):

$$\begin{gathered} \eta_{i,j} = \phi(X_{i,j}) = W_1 \sigma(W_0 X_{i,j}), \\ W_0 \in R^{\frac{C}{r} \times C}, W_1 \in R^{(K \times K \times G) \times \frac{C}{r}}, \end{gathered} \tag{2}$$

$X_{i,j}$ represents the input feature map at pixel point i, j, ϕ denotes the kernel generation function of the involution operator, W_0 and W_1 represent two linear transformations, and the inter-channel dimension is controlled by the downsampling ratio r for efficient processing. σ denotes the nonlinear activation function processed on the above linear transformations after batch normalization. The principle and operation of involution are shown in Figure 3. Each group of pixel coordinates is mapped by the ϕ function to obtain a $1 \times 1 \times K^2$ feature map, which is then restructured by the reshape function into the shape of the involution operator's kernel. Finally, a multiplication and addition operation is performed with the feature vector of the neighborhood of this coordinate point to obtain the output feature map.

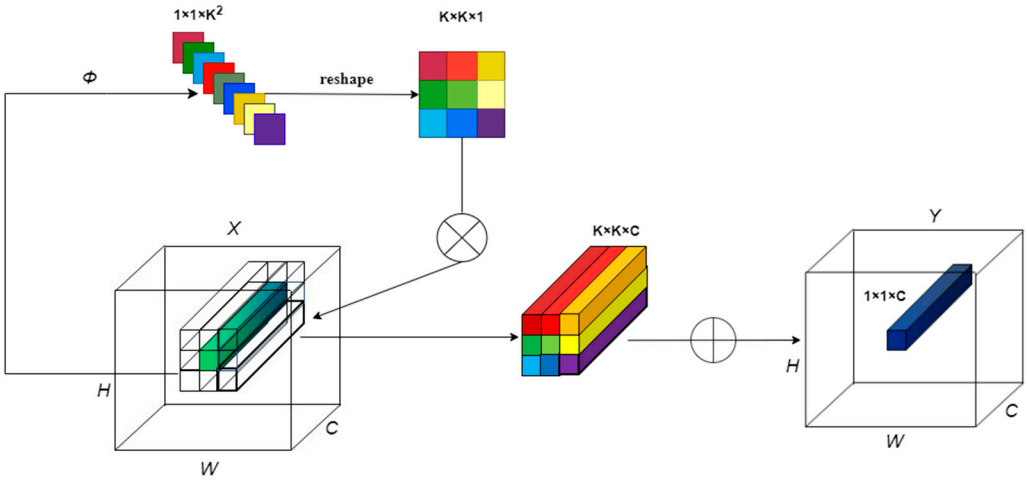

Figure 3. Schematic diagram of involution.

Compared to traditional convolution, the involution operator enhances spatial modeling information while weakening channel modeling information. If all ordinary convolution operators are replaced with the involution operator, it will cause a significant drop in accuracy. Considering that downsampling with a stride of 2 in traditional convolution will cause the phenomenon of spatial information loss, this paper replaces the downsampling convolution block in the backbone network with the involution block to balance the model size and accuracy.

3.2. GSConv

When designing lightweight networks, depth-wise separable convolution (DSConv) is often utilized as a replacement for standard convolution (SConv) to reduce computational costs. However, DSConv separates the channel information of the input feature map, which leads to lower feature extraction and fusion capabilities when compared to SConv. This article presents the introduction of GSConv to the feature fusion network, as illustrated in Figure 4. By concatenating and shuffling the feature tensors output by SConv and DSConv, the model's nonlinear expression capability is enhanced. This leads to a balance between the lightweight structure of the network and detection accuracy.

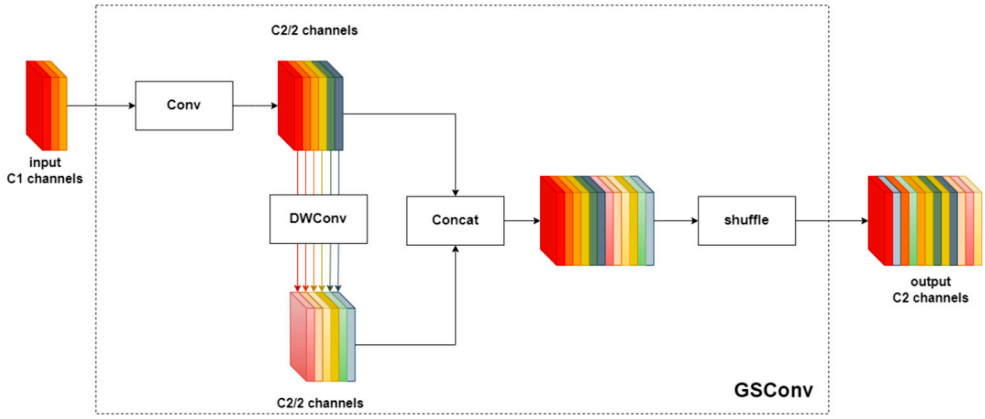

Figure 4. The structure of the GSConv. Here, DWConv indicates the DSC operation.

3.3. CBAM

Incorporating attention mechanisms in the construction of neural networks can effectively suppress irrelevant information and enhance network efficiency. One of the widely used attention modules is the CBAM, which is simple yet effective, as illustrated in Figure 5.

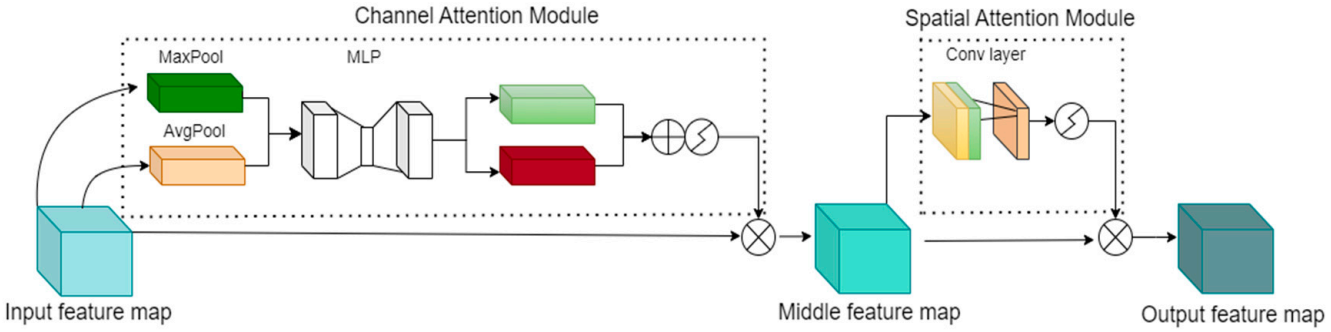

Figure 5. CBAM module structure.

CBAM consists of two independent sub-modules: the channel attention module (CAM) and the spatial attention module (SAM). To begin, the channel attention mechanism is applied to the input feature map, which generates two two-dimensional feature maps via parallel max-pooling and average-pooling operations. These feature maps are then fed into a multi-layer perceptron (MLP) with shared weights. The resulting features are summed and passed through a sigmoid activation function. Finally, the output of the activation function is multiplied with the input feature map, resulting in an intermediate feature map. The intermediate feature map undergoes parallel max-pooling and average-pooling operations in the spatial attention module, producing two two-dimensional feature maps that are concatenated. After passing through a 7×7 convolutional layer and a sigmoid activation function, the final output feature map is obtained by multiplying it with the intermediate feature map.

3.4. Knowledge Distillation

The main concept of knowledge distillation is to transfer knowledge from a large and accurate network (teacher network) to a lightweight network (student network). This paper conducts offline distillation of the improved model using the response-based distillation strategy proposed in Rakesh et al. [21]. The distillation process is shown in Figure 6. The output of the teacher network is heated to obtain soft labels, and one branch of the student network is also heated during output to obtain soft predictions. The distillation loss is calculated by computing the loss function between the soft predictions and the soft labels generated by the teacher network. The other branch calculates the student loss by computing the loss function between the unheated output and the true label.

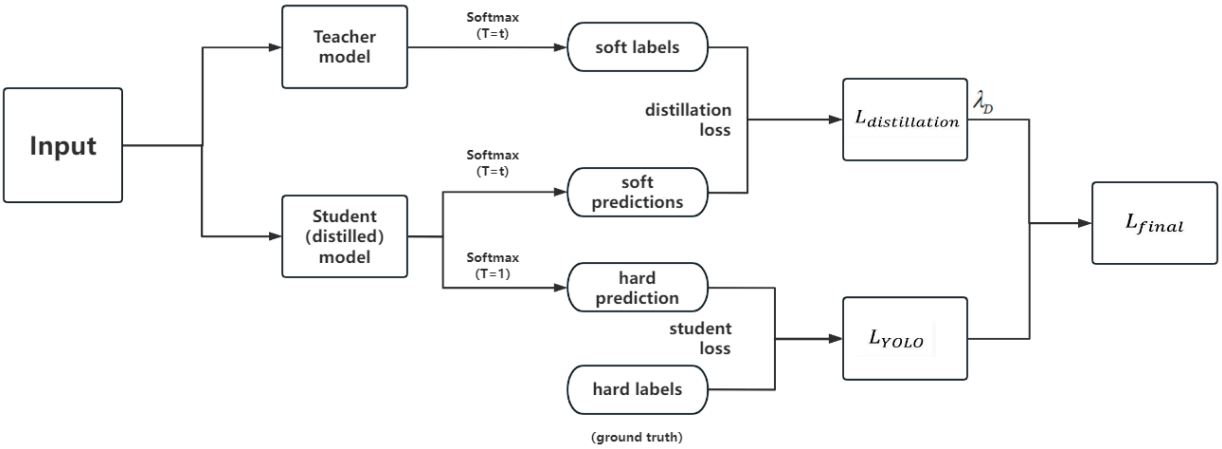

Figure 6. Knowledge distillation process.

The loss function of the YOLO algorithm is represented by Equation (3):

$$L_{YOLO} = f_{obj}(o_i^{gt}, \hat{o}_i) + f_{cl}(p_i^{gt}, \hat{p}_i) + f_{bb}(b_i^{gt}, \hat{b}_i), \tag{3}$$

In Equation (3) of the YOLO algorithm, f_{obj}, f_{cl}, f_{bb} represent the losses of objectness, class probability, and bounding box coordinates, respectively. $\hat{o}_i$, $\hat{p}_i$, and $\hat{b}_i$ represent the target objectness, class probability, and coordinate information of the predicted bounding box by the student network, while $o_i^{gt}, p_i^{gt}, b_i^{gt}$ represent the target objectness, class probability, and coordinate information of the ground truth bounding box. Moreover, building upon this foundation, the concept of distillation loss is introduced, accompanied by its corresponding loss function presented in Equation (4):

$$L_{Distillation} = f_{obj}(o_i^T, \hat{o}_i) + \hat{o}_i{}^T \bullet f_{cl}(p_i^T, \hat{p}_i) + \hat{o}_i{}^T \bullet f_{bb}(b_i^T, \hat{b}_i), \tag{4}$$

The variables o_i^T, p_i^T, b_i^T denote the objectness, class probability, and coordinate information of the bounding box predicted by the teacher network. The output o_i^T after applying the sigmoid function is denoted as $\hat{o}_i{}^T$, which is used as the coefficient for both the classification and localization losses to prevent the student network from learning incorrect background box information. The total loss function comprises both distillation loss and student loss, and the parameter λ_D is introduced to balance the object detection loss and distillation loss of the student network. Thus, the total loss function can be expressed as Equation (5):

$$L_{final} = L_{YOLO} + \lambda_D \bullet L_{Distillation}. \tag{5}$$

The network structure of the teacher model is usually more complex than that of the student model. However, the difference between the teacher model and the student model should not be too large, otherwise the student model will have difficulty fitting the predictions of the teacher model, resulting in poor knowledge distillation. Therefore, in

this paper, the YOLOv5m network with the same structure as the improved YOLOv5s is used as the teacher network to perform knowledge distillation on the improved YOLOv5s to improve the performance of the model.

4. Experimental Results and Analysis

4.1. Dataset and Preprocessing

This paper utilizes a dataset from SewerML [22], which was originally designed for multi-label image classification. Therefore, only six commonly occurring defects, namely break (PL), displaced joint (CK), roots (SG), intruding sealing material (TL), branch pipe (AJ), and obstacle (ZW), are selected for analysis in this study. In this study, a total of 2122 defect images were utilized, and 3490 annotation boxes were manually labeled using the labeling tool.

The dataset was divided into training, testing, and validation sets in a ratio of 7:2:1. To enhance the sample size and improve the generalization ability and robustness of the model, online data augmentation techniques, such as HSV enhancement, random rotation, random scaling, random translation, and Mosaic4, were applied to the training set. Partial defect images are shown in Figure 7.

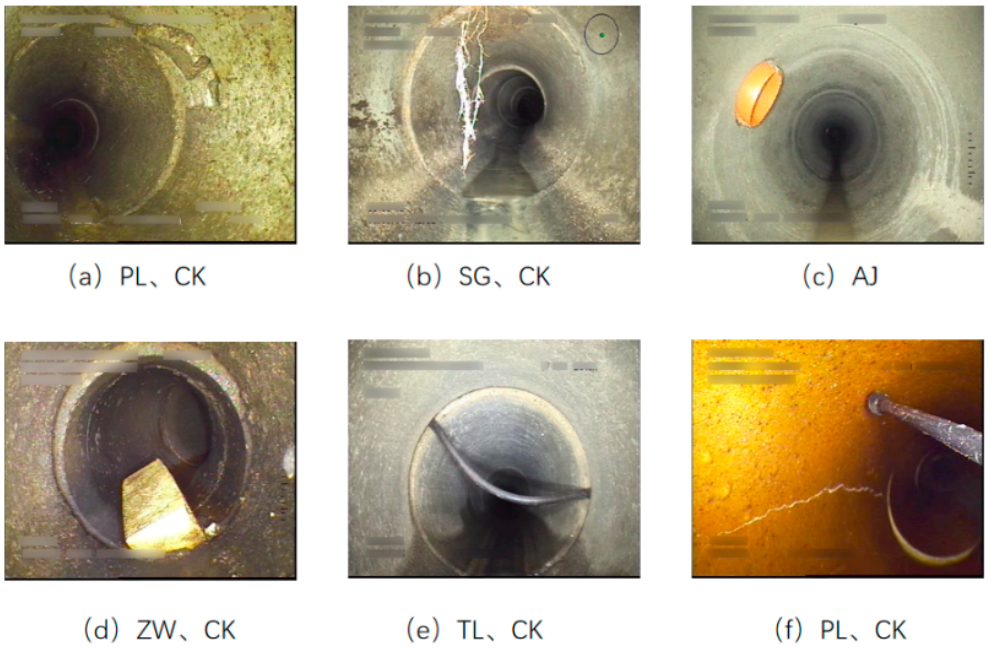

Figure 7. Example images of the sewer defect.

4.2. Experimental Environment and Hyperparameter Settings

The experiments were carried out using a Windows 11 operating system, an NVIDIA GeForce RTX 3050 graphics card, an Intel Core i5-11260H2 CPU, and 16 GB of memory. The model was constructed, trained, and validated using the PyCharm 2018 editor and the PyTorch 1.12.1 deep learning framework, respectively, to achieve the research objectives.

During training, the epoch was set to 300, and the initial learning rate was set to 0.01. The learning rate decay employed the cosine annealing method, with a final decay of 0.0001. The optimizer used was SGD, with a momentum of 0.937. The batch size was set to 16, and the input image size was set to 480 × 480. For the knowledge distillation experiment, the value of λD was set to 0.5.

4.3. Evaluation Index

To compare the accuracy of models with different structures for detecting defects in sewer pipelines, mean average precision $mAP@0.5$ and $mAP@0.5:0.95$ were used as evaluation metrics. The formulas for the evaluation metrics are as shown in Equations

(6)–(9): TP represents the number of true positive samples detected, FP represents the number of false positive samples detected, and FN represents the number of undetected positive samples. AP is the average precision for a given category, N is the total number of categories, and P and R represent precision and recall, respectively.

$$P = \frac{TP}{TP + FP}, \tag{6}$$

$$R = \frac{TP}{TP + FN}, \tag{7}$$

$$AP = \int_0^1 P(R)dR, \tag{8}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i. \tag{9}$$

To reflect the speed and complexity of model detection, the evaluation metrics included frames per second (FPS) and the number of parameters (Params), floating-point operations per second (FLOPs), and the size of model weight files (Weights).

4.4. Different Attention Mechanism Comparative Experiments

The CBAM attention mechanism was introduced in the feature fusion network of this paper. In order to verify its effectiveness and investigate the impact of different attention mechanisms on the improved model, SE and CA attention mechanisms were introduced at the same position for comparative experiments, as shown in Table 1 for the results.

Table 1. Comparative experimental results of attention mechanisms.

Model	mAP@0.5/%	mAP@0.5:0.95/%	Parameters	GFLOPs
Involution + GSConv	78.1	45.2	4,875,367	11.3
Involution + GSConv + SE	78.2	45.3	4,918,375	11.3
Involution + GSConv + CA	78.3	45.9	4,911,047	11.3
Involution + GSConv + CBAM	79.3	46.3	4,918,669	11.3

To further analyze the impact of different attention mechanisms on the improved network's prediction results, GradCAM++ [23] was employed to visualize the feature maps of the last layer of the neck network and generate heat maps, as presented in Figure 8. The heat map indicates that the attention of the three different attention mechanisms has been enhanced to varying degrees compared to the original model, with darker colors representing higher attention in the corresponding areas. Notably, among the three types of defects, namely break, displaced joint, and intruding sealing material, the CBAM mechanism shows higher coverage of the target region. Thus, based on the experimental findings and heat map analysis, this paper selects the CBAM attention mechanism, which exhibits more pronounced improvement in the model's performance.

4.5. Ablation Experiments

To further analyze the influence of various improvement modules on network detection performance, ablation experiments were conducted on the test set, and various performance indicators are presented in Table 2. The symbol "√" denotes the inclusion of the corresponding improvement module.

Based on the observation of Table 2, it is apparent that the integration of the involution module into the original network leads to a substantial decrease in both the number of parameters and the computational complexity. GSConv adds shuffling operation on the basis of DSConv to enhance the network's nonlinear expression ability, enabling the network to

achieve higher accuracy while slightly reducing the number of parameters and computational complexity. While the introduction of CBAM reduces detection speed, it compensates for the accuracy loss incurred by the lightweight process. Table 3 presents the selection of YOLOv5m as the teacher network, with mAP@0.5 and mAP@0.5:0.95 reaching 79.3% and 46.3%, respectively, prior to knowledge distillation. Post distillation, the final model demonstrated a 2.4% and 2.6% improvement in mAP@0.5 and mAP@0.5:0.95, respectively, when compared to the baseline model. The number of parameters and computational cost decreased by 30.1% and 29.4%, respectively, and the detection speed reached 75 FPS. The above data demonstrates the effectiveness of the proposed improvement method, which achieves lightweight network and improved prediction accuracy while ensuring real-time on-site detection.

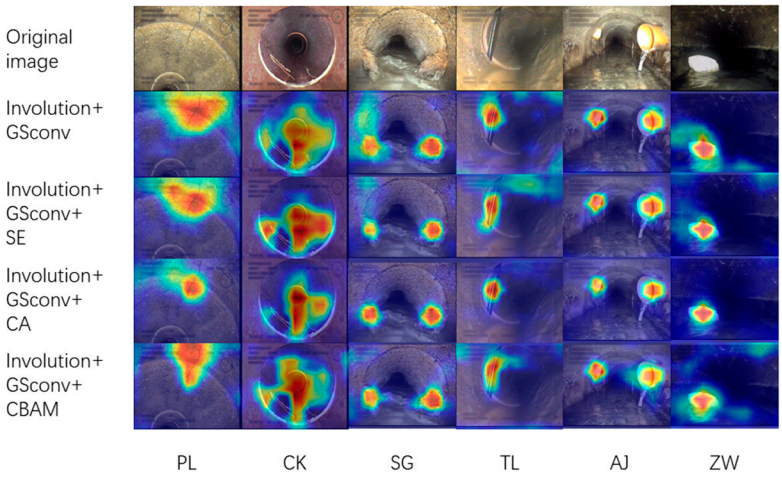

Figure 8. Comparison of heat maps before and after adding different attention modules.

Table 2. Results of the ablation experiment.

Method	Involution	GSConv	CBAM	KD	mAP@0.5/%	mAP@0.5:0.95/%	Parameters	GFLOPs	FPS
YOLOv5s					78.1	46.1	7,035,811	16	98
A	√				77.6	45	5,316,327	11.8	95
B		√			78.9	46.7	6,594,851	15.4	93
C			√		79.2	46.6	7,079,113	16.1	85
D	√	√			78.1	45.2	4,875,367	11.3	90
E	√	√	√		79.3	46.3	4,918,669	11.3	75
F	√	√	√	√	80.5	48.7	4,918,669	11.3	75

Table 3. Results of knowledge distillation experiments.

Model	λ_D	mAP@0.5/%	mAP@0.5:0.95/%
Involution + GSConv + CBAM		79.3	46.3
YOLOv5m(Teacher)		80.8	51.2
Involution + GSConv + CBAM + KD	0.5	80.5	48.7

4.6. Comparison Experiment

To objectively demonstrate the effectiveness of the improved YOLOv5 model proposed in this paper for detecting defects in sewer pipes, some mainstream object detection models, including SSD, Faster R-CNN, and the YOLO series, were trained and tested on the same dataset. Moreover, the enhancement strategy proposed by Chanmi et al. [14] (YOLOv5LC, micro-scale detection layer + CBAM) for the detection of sewer pipeline defects using the YOLOv5s model has been successfully reproduced in this study. Subsequently, a comparative analysis was conducted between the improved model and theirs. The evaluation metrics used were mean average precision (mAP), frames per second (FPS), and model weight size. The experimental results are presented in Table 4.

Table 4. Comparison experimental results of mainstream algorithms.

Model	mAP@0.5/%	mAP@0.5:0.95/%	Weight Size/MB	FPS
SSD	76.2	41.7	93.1	26
Faster-RCNN	78.5	44	108	8
YOLOv3	78.4	45.7	117	41
YOLOv4	80.9	52	100	44
YOLOv5s	78.1	46.1	13.7	98
YOLOv5(MobileNetV3)	76.4	45.3	9.96	72
YOLOv5(ShuffleNetV2)	71.2	37.7	6.36	103
YOLOv7tiny	77.1	44.5	11.7	95
YOLOv7	82.3	53.5	71.3	46
YOLOv5sLC [14]	79.4	48.3	14.5	64
Improved model	80.5	48.7	9.7	75

According to Table 4, the improved algorithm achieved a detection speed of 75 FPS and has good real-time performance. The precision metric mAP@0.5 reached 80.5%, which is 2.0% and 4.3% higher than the classic object detection algorithms Faster R-CNN and SSD, respectively. It is also significantly better than the same type of algorithms YOLOv3, and YOLOv5s. Compared to other improved lightweight algorithms such as YOLOv5 (backbone based on ShuffleNetV2 [24] and MobileNetV3 [25]) and the newer YOLOv7tiny, the improved model achieves the highest detection accuracy while satisfying the real-time detection requirements. Compared to YOLOv5sLC, the proposed approach demonstrates superior suitability for sewer pipeline defect detection tasks in terms of accuracy, speed, and lightweight design. Although its accuracy is slightly lower than large models such as YOLOv4 and YOLOv7, due to its smaller model size and faster detection speed, it is more suitable for field deployment in sewer pipe defect detection.

4.7. Detection Results

In order to further validate the efficacy of the final improved model, a comprehensive comparison was conducted between the improved model and YOLOv5s on the test dataset, aiming to assess their performance and effectiveness. Table 5 demonstrates that the improved model demonstrates diverse degrees of performance enhancement across different defect detection scenarios. Figure 9 presents the PR curve plots for the original and the improved models. As shown in the figure, it is evident that the PR curve of the improved model achieves a larger area under the curve, indicating its superior performance compared to the original model.

Table 5. Comparison of precision, recall, mAP between the original and improved models.

Method	Indicator	PL	CK	SG	TL	AJ	ZW	ALL
YOLOv5s	P/%	74.3	70.1	79.6	83.1	86.3	82.7	79.4
	R/%	64.6	64.5	79.7	74.8	82.1	70	72.6
	mAP@0.5/%	70.4	68.5	81.5	83.2	85.4	79.8	78.1
	mAP@0.5:0.95/%	31.7	44.9	39.9	47.6	60.4	52.2	46.1
Improved model	P/%	78.2	70	83	84.7	88.5	84.8	81.5
	R/%	64.6	71.5	83.2	78.3	80.6	77.3	75.9
	mAP@0.5/%	71.9	73.5	82.8	84.3	89.1	81.7	80.5
	mAP@0.5:0.95/%	31.6	51.1	44	51.1	62.6	52	48.7

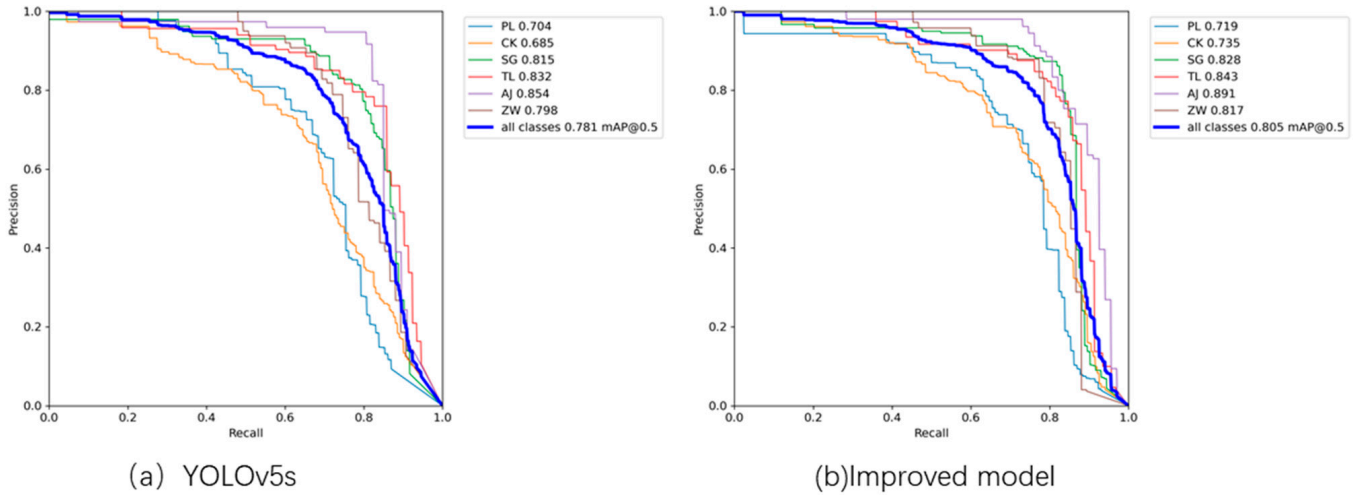

Figure 9. PR curve plots for the original and improved models.

In the task of detecting sewer pipeline defects, missed detections often occur due to factors such as low lighting conditions and occlusions. Analysis of Table 5 and Figure 9 reveals a notable enhancement in the recall rate of the improved model compared to YOLOv5s. A higher recall rate indicates a reduced number of missed defects by the model, which holds significant importance for sewer defect detection tasks. Figure 10 presents the visual results of the original and improved models, providing visual evidence of the enhanced performance of the improved model in detecting sewer pipeline defects. In Figure 10, a comparison between (a) and (g), as well as (e) and (k), reveals that the improved model, compared to the original model, successfully detects break (PL) under low-light conditions. Comparing (d) and (j), the improved model demonstrates excellent detection capabilities for smaller roots (SG) defects as well.

In conclusion, the model proposed in this study demonstrates substantial advancements over YOLOv5s, facilitating enhanced detection of diverse defect types and precise localization within sewer pipelines.

It is important to acknowledge that while the improved model demonstrates the capability to identify the majority of diverse sewer pipeline defects, there might be specific scenarios where certain limitations or oversights in the improved model's performance could arise. In Figure 10j, the model failed to detect the roots (SG) extending from the lateral branch. This can be attributed to the similarity in height between the roots and the background sediment, which leads to less distinct features and subsequently results in missed detections. In Figure 10k, a break (PL) at a specific location was not detected, possibly due to the small size of the fracture, leading to a missed detection (missed defects are indicated by green arrows in Figure 10).

Typically, in pipeline engineering regulations, fractures are classified into four distinct levels. Hence, in this study, both cracks and substantial wall damages are categorized as break (PL) defects. Moreover, existing object detection methods lack the capability to evaluate the severity level of individual defects, thereby requiring the incorporation of segmentation techniques. This presents a promising avenue for future research in the domain of sewer pipeline inspection tasks.

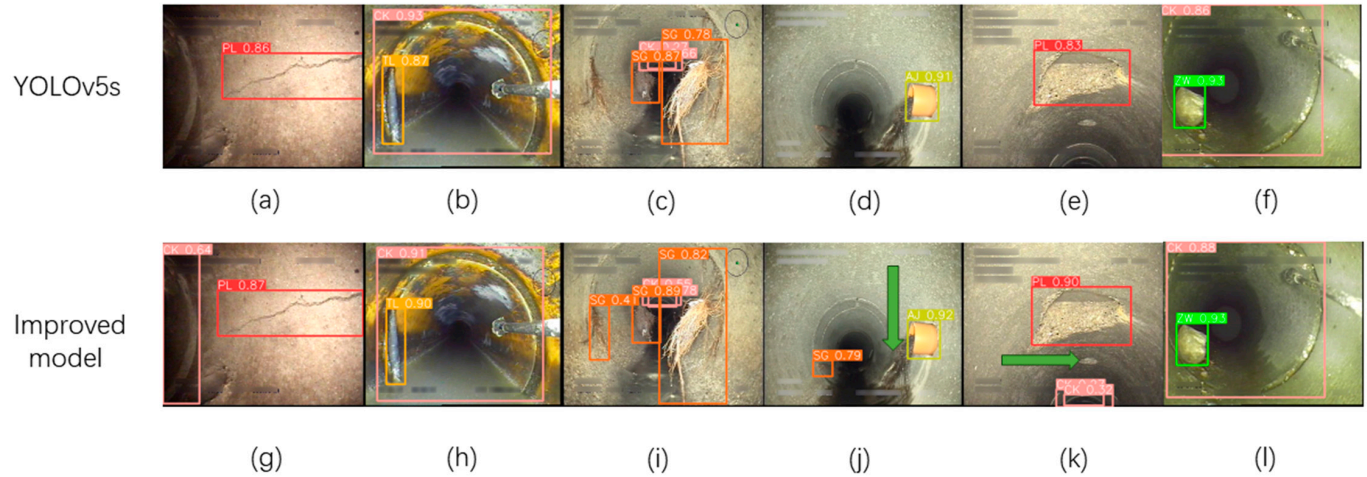

Figure 10. Detection results of YOLOv5s (**a**–**f**) and the improved model (**g**–**l**).

5. Conclusions

This paper presents an improved algorithm for sewer pipeline defect detection based on YOLOv5s, effectively tackling the subjectivity, low efficiency, and on-site model deployment challenges associated with existing CCTV-based defect detection methods. The model presented in this paper possesses the following advantages:

(1) The proposed model exhibits a reduced number of parameters and computational complexity. In this study, a lightweight network architecture is constructed by incorporating involution and GSconv, thereby mitigating the dependence on computational power of the device. Compared to the YOLOv5s model, the proposed model exhibits a reduction of 30.1% in the number of parameters and a 29.4% decrease in computational complexity.

(2) The model proposed in this study exhibits a high level of detection performance. The incorporation of the CBAM attention mechanism enhances the detection capability of the model, particularly in complex backgrounds. Furthermore, the utilization of knowledge distillation is employed to enhance the model's generalization performance. Ultimately, the improved model successfully attained an mAP@0.5 score of 80.5% and an mAP@0.5:0.95 score of 48.7% on the test dataset. Moreover, the detection speed demonstrated remarkable performance, achieving a rate of 75 frames per second (FPS), effectively meeting the stringent real-time demands for on-site detection.

Through comparative experiments, the proposed model demonstrated superior performance compared to well-known models such as SSD and Faster R-CNN. Furthermore, it surpassed its counterparts in the YOLOv3 and YOLOv5s series. When compared to other lightweight enhancement algorithms, such as YOLOv5s (ShuffleNetV2 and MobileNetV3) and YOLOv7tiny, the proposed model achieves the highest level of detection accuracy while satisfying the real-time detection demands. Although the proposed model's mAP is slightly lower compared to YOLOv4 and YOLOv7, it offers notable advantages in terms of model size and detection speed. Moreover, in comparison to the YOLOv5sLC model, a target detection model for the same task, the proposed model exhibits even greater advantages on the dataset employed in this study.

Based on the aforementioned data, the model proposed in this study fulfills the real-time requirements for on-site sewer pipeline defect detection. It demonstrates low computational overhead and achieves enhanced accuracy compared to mainstream algorithms at the current stage. Thus, it is well-suited for deployment on mobile devices. While current object detection models demonstrate the ability to accurately detect defects, they encounter difficulties in assessing the precise severity level of individual defects. In future studies, semantic segmentation will be carried out on the detected defect images, followed by the evaluation of their respective severity levels, considering their geometric characteristics.

Author Contributions: Supervision, T.W.; project administration, T.W.; resources, T.W., W.W. and R.H.; conceptualization, Y.L., W.W. and R.H.; methodology, Y.L.; writing—original draft, Y.L.; validation, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was sponsored by the Natural Science Foundation of Henan (NO.232300420091) with a funding amount of 20,000 RMB and the Henan Provincial Department of Science and Technology Research Project (NO.222102210270) with a funding amount of 15,000 RMB.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request. The Sewer-ML is available online at http://vap.aau.dk/sewer-ml (accessed on 1 December 2022).

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

References

1. Haurum, J.B.; Moeslund, T.B. A Survey on Image-Based Automation of CCTV and SSET Sewer Inspections. *Autom. Constr.* **2020**, *111*, 103061. [CrossRef]
2. Li, Y.; Wang, H.; Dang, L.M.; Song, H.K.; Moon, H. Vision-based defect inspection and condition assessment for sewer pipes: A comprehensive survey. *Sensors* **2022**, *22*, 2722. [CrossRef]
3. Yin, X.; Chen, Y.; Bouferguene, A.; Zaman, H.; Al-Hussein, M.; Kurach, L. A deep learning-based framework for an automated defect detection system for sewer pipes. *Autom. Constr.* **2020**, *109*, 102967. [CrossRef]
4. 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.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
5. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Volume 9905, pp. 21–37.
6. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
7. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
8. Ultralytics: Yolov5. Available online: https://github.com/ultralytics/yolov5 (accessed on 1 October 2022).
9. Wang, C.Y.; Bochkovskiy, A.; Liao, H. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
10. Saponara, S.; Elhanashi, A. Impact of image resizing on deep learning detectors for training time and model performance. In Proceedings of the International Conference on Applications in Electronics Pervading Industry, Environment and Society, Pisa, Italy, 21–22 September 2021; pp. 10–17.
11. Yang, D.; Cui, Y.; Yu, Z.; Yuan, H. Deep learning based steel pipe weld defect detection. *Appl. Artif. Intell.* **2021**, *35*, 1237–1249. [CrossRef]
12. 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]
13. Tan, Y.; Cai, R.; Li, J.; Chen, P.; Wang, M. Automatic detection of sewer defects based on improved you only look once algorithm. *Autom. Constr.* **2021**, *131*, 103912. [CrossRef]
14. 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]
15. Li, D.; Hu, J.; Wang, C.; Li, X.; She, Q.; Zhu, L.; Zhang, T.; Chen, Q. Involution: Inverting the Inherence of Convolution for Visual Recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12316–12325.
16. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles. *arXiv* **2022**, arXiv:2206.02424.
17. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *Eur. Conf. Comput. Vis.* **2018**, *11211*, 3–19.
18. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531.
19. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [PubMed]
20. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
21. Mehta, R.; Ozturk, C. Object Detection at 200 Frames Per Second. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018; Part V 15. pp. 659–675.
22. Bruslund Haurum, J.; 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, Nashville, TN, USA, 20–25 June 2021; pp. 13456–13467.

23. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018.
24. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical Guidelines for Efficient CNN Architecture Design. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 11218, pp. 122–138.
25. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.