IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2017.DOI*

# Robust Sewer Crack Detection with Text Analysis based on Deep Learning

**CHANMI OH[1], L. MINH DANG[2], DONGIL HAN[1], AND HYEONJOON MOON[1]**
[1]Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea)
[2]Department of Information Technology, FPT University, Ho Chi Minh city 70000, Viet Nam)

Corresponding author: Hyeonjoon Moon (e-mail: hmoon@sejong.ac.kr).

**ABSTRACT** Sewerage systems play a vital role in building modern cities, providing appropriate ways to release liquid wastes. Due to the rapid expansion of cities, the deterioration of sewage pipes are increasing. Hence, systematic maintenance methods are require to overcome this problem. In most cases, sewer inspection is done by human inspectors, which is error-prone, time-consuming, costly, and lacking appropriate survey evaluations. In this paper, we introduce a new automated framework for detecting sewage pipe cracks based on the attention mechanism, improved YOLOv5 architecture, and location information recognition from CCTV videos. The main contributions include (1) the addition of a micro-scale detection feature in the layers to improve the crack detection mechanism; (2) the application of a convolutional block attention module for better channel/spatial features; (3) construction of a larger crack-detection dataset for the 12 most common crack types; and (4) implementation of the TPS-ResNet-BiLSTM-Attn (TRBA) model for the text-information recognition mechanism from CCTV videos. The experimental results show that the proposed real-time sewer crack detection model achieved the mean average precision (mAP) of 75.9% on the proposed dataset, outperforming state-of-the-art models, such as YOLO and SSD.

**INDEX TERMS** deep learning, sewer crack detection, crack classification, text recognition, attention mechanism.

## I. INTRODUCTION

SEWER pipe infrastructure is one of the most important components for building a modern society because it is designed to collect information about domestic, industrial-toxic, and storm sewerage. Although, existing well-developed sewerage networks are developed by complicated pipelines, operating such a big network and identifying the breakage into pipelines are extremely hard work because of the widespread sewer infrastructure [1]. Additionally, the state of affairs becomes even worse when sewer structures suddenly deteriorate due to various faults such as infiltration, permanent obstruction, lining cracks, and deformation. Long-term, sewer systems leads to a significant financial burden, labor costs, and time shortages regarding construction [2]. As a result, sewer investigation must be performed regularly to detect cracks at their early stage to reduce the fast deterioration rate of sewer pipelines. Currently, sewer investigation is comprised of three main processes: collection of functional data on pipelines, data collection using automated devices (e.g., robot), and manual assessment of cracks by experts.

Surveying the entire sewer network and collecting manual data are extremely time-consuming, requiring extensive resources, proper management, and expert human interaction at each level. In video-based sewer approaches, the usage of high-resolution cameras, is normally used to obtain the interior situation of sewer pipelines by capturing videos. While performing inspection, experts require manual checking of defaults in sewer pipes and proper recording of data, including: location, pipe number, type, and crack codes. After recording videos, experts need to investigate the crack again in order to evaluate the sewer pipeline condition. This process is quite cumbersome and the timeframe for finishing the project being unpredictable, leaving room for a lot of errors. Sometimes, the records can be inconsistent because of different interpretations of the inspector, i.e., the crack belongs to a particular class but is reported in a different class.

The recorded sewer videos are available on a massive scale; it is extremely critical to train a deep CNN and achieve superior results to deal with a fault in the sewer and extract the textual information presented on each frame [3]. Conventional ML algorithms, which rely on the manual feature engineering process, work well on a specific crack dataset [4]. However, they showed poor performance on images where the crack region that occupied over 70% [5]. The main reason is because of the dataset-specific crack features that are extracted manually [4].

Compared to conventional ML algorithms, deep learning-based models learn to extract abstract features automatically and achieve state-of-the-art results given sufficient dataset [6]. Moreover, deep learning-based systems can be deployed in real-world applications, because they are robust against noise, challenging environments, such as the pitch-dark sewer environment, and can perform crack detection in real-time [7]. Although many methods are proposed to achieve the structural assessment, the number of classes is limited [8]. Substantial needs and efforts are required in order to develop a new sewer framework that can automatically recognize the different cracks and is enough robust to adopt in a real-world application.

This paper introduces a CNN and attention-based approach that puts forward existing sewer frameworks, especially a design to detect 12 types of faults, including buckling, broken pipe, longitudinal crack, multiple crack, silty deposit, displaced joint, separated joint, lining crack, protruding lateral, sealing lateral, sags, and temporary obstruction crack as presented in Fig. 2 to extract rich features and achieve top-performing classification results [8], [9]. Firstly, this research proposes an improved YOLOv5 architecture with an attention module that contributes to automated crack detection in videos. Secondly, the research focuses on the text recognition mechanism that captures valuable information from sewer images extracted from robot-recorded CCTV videos. Thirdly, this research collects a new crack dataset that derives from high-resolution CCTV videos. Based on sewer videos, extraction of crack images from recorded videos ensure that the crack image belongs to an exact class and then performs a labeling task. The entire dataset construction of pipelines is done manually. Overall, the whole framework automatically detects sewer cracks and classifies by the specific crack. Meanwhile, the text detection model is integrated to recognize the entire information present in an image in term of name, date, position-time inspection id number. The proposed system is inspired by the prior and exiting studies of the CNN architectures in various fields that promised to capture relevant information from the training set. Further, the text recognition model is utilized to extract textual information present in the crack image. Various experiments were conducted on the collected dataset to show the robustness and effectiveness of our sewer crack framework. Finally, three models are trained from scratch in order to evaluate the accuracy of our improved YOLOv5 models and original YOLOv3 [10], v4, and retinanet models.

The paper is divided as follows: Section II extensively covers the previous and existing approaches in terms of sewer crack classification, detection, and datasets. In Section III, the CNN and attention-based framework is presented. Next, the statistics of new collected pure crack dataset. In Section IV, various experimented are performed to evaluate our sewer framework on the collected dataset. Finally, the research highlights the strength of our method which are summarized in Section V.

## II. RELATED WORK

### A. SEWER CRACK DETECTION

Initially, the inspection of widespread sewer pipelines was investigated manually by sending experts into the pipeline to note all the crack information, a slow, time-consuming, and dangerous process. To overcome inspection problems, various devices have been proposed, including scanners, CCTVs, laser devices, etc. Liu et al. [11] reviewed state-of-the-arts technology and highlighted several limitations in CCTVs, including how the camera had to stop at every location to recognize a crack. All data collection devices have certain limitations [12]. For example, sonar systems provide high resolution on high frequency, but the scanning capability is not good on low frequency. Even though CCTV cameras are appropriate for capturing sewer cracks, they cannot extract detailed information about the cracks, such as the location where the crack appears and crack depth. Even though various sensors can be utilized to effectively collect additional information, they are costly to deploy and are computational complex during the data analysis [13].

To solve manual problems mentioned above and to reduce inspection time, traditional methods were developed to detect cracks in a sewer image, including edge detection, hand-crafted features, and small classifiers. For instance, Mahmoud and Jantira [14] utilized various two image preprocessing approaches, SVM and histogram, to recognize faults in pipelines, using a sobel filter to identify the edge frame and applying special filters to detect cracks. However, this approach highly emphasized traditional computer vision techniques and took a long time to perform an operation on an image. Iraky et al. [15] developed a Markov-based model to detect crack and used the canny edge algorithm to recognize faults. However, this study was limited to several classes, and the procedure to detect cracks was time-consuming.

With the rapid development of deep learning, DL-based models have shown state-of-the-art performances in various computer vision fields, such as image classification [6], object detection [5], and segmentation [2], [16]. Particularly, CNN architectures automatically extract abstract features from the training set, outperforming conventional ML approaches, which rely heavily on the manual feature engineering process [4]. Therefore, deep CNNs have been increasingly investigated to solve sewer crack classification and detection problems. The image-based sewer crack classification problem has been extensively investigated in recent years. Qian et al. [17] proposed a fully automated practical approach to classify sewer cracks by developing hierarchical CNNs based on two-level to address data imbalanced problems and solve the limited data problems. The model obtained 94.96% classification accuracy on the testing set.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3168660, IEEE Access

However, the authors claimed that the raw images varied in resolution. Thus, all images were resized to 256 × 256 before training. In addition, this study only classifies six crack classes that commonly appear in the dataset. Srinath et al. [18] developed an automated sewer framework by using multiple CNNs to detect specific faults in pipelines. The model can be easily classified as a crack when passed through the ensemble of CNNs. Further, specific CNN were developed to extract sewer crack features, particularly in fog and grease situations. However, the proposed CNNs could not distinguish between medium and fine roots, or longitudinal and spiral cracks. In another research, Cheng and Wang [1] utilized the Faster R-CNN approach to automatically recognize cracks in pipelines such as longitudinal crack, multi crack and etc. Various hyper-parameters were employed to achieve superior accuracy on the model considering the highest performance and computational cost. The final accuracy on the image-based test set was 83% mAP. However, this technique only worked on static images and poorly classified crack images due to the colour intensity.

CNNs have achieved top-performing results and obtained state-of-the-art performance compared to traditional machine learning approaches especially solving sewer crack detection problems [19]. Compared with conventional techniques, CNN-based methods were more robust to learning image features and extracting pixel-level information without the requirement of developing a specific feature extractor. Yi et al. [20] extended YOLOv3 and presented an improved YOLOv3 version for automatic sewer crack detection that was primarily focused on the enhancement of data augmentation techniques, involving bounding box prediction, loss function, and network architecture. However, this technique for fault detection did not have particular information on geographical area and actual information of fault remained inconsistent. Li et al. [21] introduced a new deep CNN to tackle imbalanced data problems in sewers by utilizing hierarchical softmax. These hierarchical function categorized the faults at a high-level (normal images) to a low level (detecting cracks). Additionally, an 18 layer-based ResNet architecture was utilized as a backbone network to preserve knowledge. The detection performance on high-level faults were significantly increased from 78.4% to 83.2%. However, the proposed technique showed imbalance of data and some classes were labelled improperly. Zuo et al. [22] contributed a novel framework that integrated various CV and ML techniques and developed a feature-based method to automatically identify common crack types in sewer images. The pixel differential approach categorized views and identified pipes jointly. Then, edge detection techniques were utilized to classify cracks. However, low-resolution sewer images were used in the approach and only focused on cracks and was more limited to spiral and hinge cracks.

The main drawback of previous models is that they just utilized the original CNN structures [23], without any modification to adapt to the sewer crack images. Another aspect is that most datasets on the sewer crack detection topic are private. Although currently some large-scale sewer crack datasets are introduced, such as SewerML [24], most of them are proposed for performing the crack classification. Finally, the number of images and classes in the dataset is imbalanced [8], which discouraged researchers from working on the problem [4].

### B. TEXT DETECTION AND RECOGNITION

Most studies have proposed novel CNN architectures to overcome sewer classification and detection problem. However, the text recognition part in sewer images was still not addressed and few studies were available on such topics. Dang et al. [25] presented a new text recognition approach to extract contextual information from sewer images based on Korean subtitles applying a multi-scale approach. The text detection component was focused on single-line text rather than text region consisting of multiple lines. This approach contributed a great deal in recognition. Jeonghun et al. [26] developed a novel STR model to detect and identify text in all direction considering low and high-resolution images. The model was validated on seven benchmark datasets, achieving the highest accuracy among the existing models. Chen et al. [27] presented a framework to generate synthetic Chinese data to enhance exiting models. They have collected $7,000$ labelled data and 20M synthetic data and achieved 89.64% on a pre-trained model. Recently, Rowel et al. [28] proposed a transformer-based method ViTSTR, single stage and parameter efficient approach. His approach obtained a competitive performance of 82.6% and 84.2% using several data augmentation techniques.

Most of the studies have been proposed to perform sewer crack classification and detection problems [5], [7], they did not investigate the text printed on the CCTV videos, which can be used to obtain useful information regarding the detected cracks.

To address those problems, this research presents an improved YOLO with Convolution Block Attention Module (CBAM) [29] to detect cracks in pipelines. This paper also solves the aforementioned difficulties in the sewer by constructing a first 12-class crack dataset, analyzing the text part in order to save time and provide more relevant information of detected cracks.

## III. METHODOLOGY

Fig. 1 illustrates a detailed architecture of the real-time crack detection framework with five main components. (1) extract frames from CCTV inspection videos; (2) pre-process and annotate images with augmentation; (3) an improved YOLOv5 [30] architecture with attention block and training; (4) crack detection; and (5) text detection and recognition for frame containing cracks. Two different improved YOLOv5 models are trained for performing two different tasks, including crack detection and text detection.

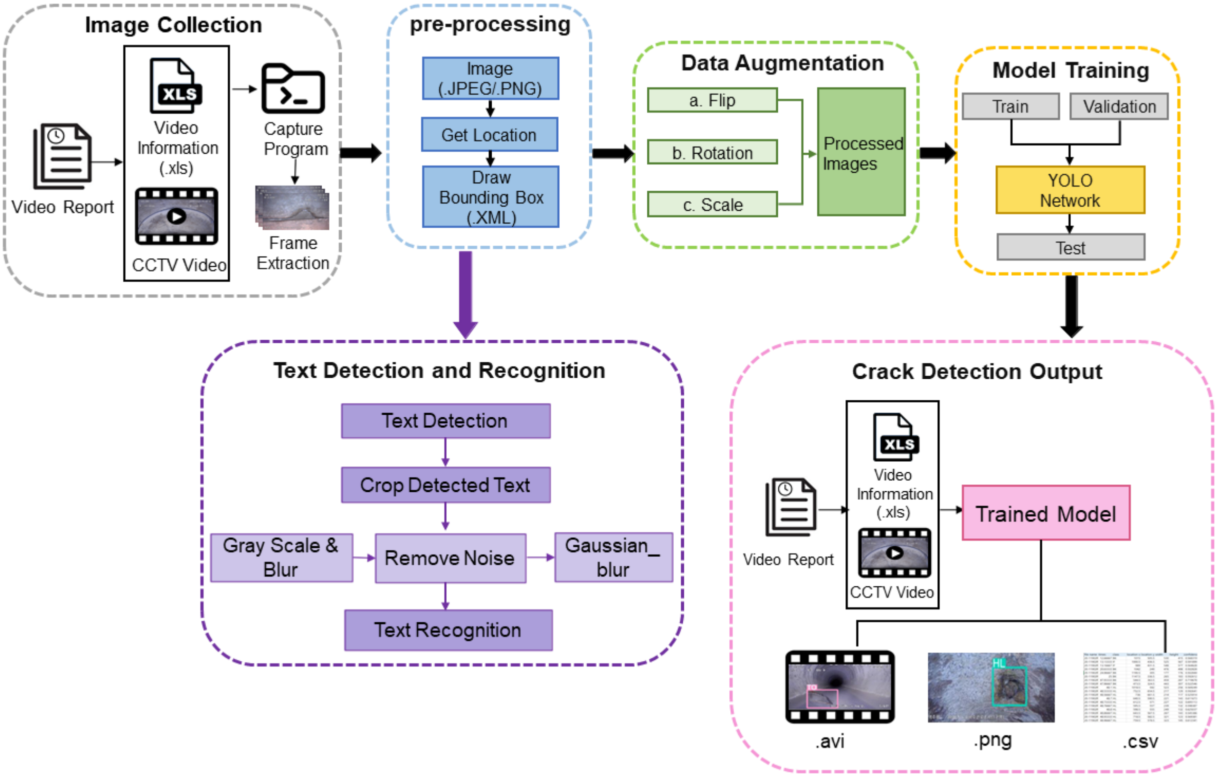


**FIGURE 1.** The flow chart of the deep learning-based automatic fault detection system framework

### A. DATASET PREPARATION AND PRE-PROCESSING

#### 1) Crack Dataset Preparation

The inside of the sewerage pipeline could not be evaluated directly by humans due to concerns, such as harmful gases and continuous water flow. As a result, the CCTV inspection robot-based approach is usually deployed because the inspectors can inspect the sewer pipeline remotely. The inner condition of the sewage pipe can be monitored using a small screen and operation equipment on the ground. A filming device (module) integrated at the head of the robot, which is waterproof and dust-proof, is used to investigate the sewer pipes while the inspection robot runs inside a sewer pipeline. The device can capture 360° consecutive rotations in the side vision, and 90° rotations in the upper and lower left and right, reaching up to 2K resolution (2048 × 1024). In addition, the robot is equipped with six 1000 lumens (lm) halogens to enable it to record videos in varying lighting conditions. The video duration is between one and fifteen minutes. In addition, recorded CCTV videos also contain the necessary information as video subtitles, such as pipe type, robot moving distance, and diameter, which can be used for further inspection.

At the end of the data collection phase, a total of $2,086$ raw CCTV videos are collected to facilitate the crack detection in this study. After that, each CCTV is manually reviewed by experts. When a crack appears in a frame, it is manually annotated with a bounding box and labels using the open-source labelImg tool[1]. At the end of the labeling process, $4,456$ images for $12$ types of sewer cracks were manually annotated. Sample images for each crack type are demonstrated in Fig. 2.

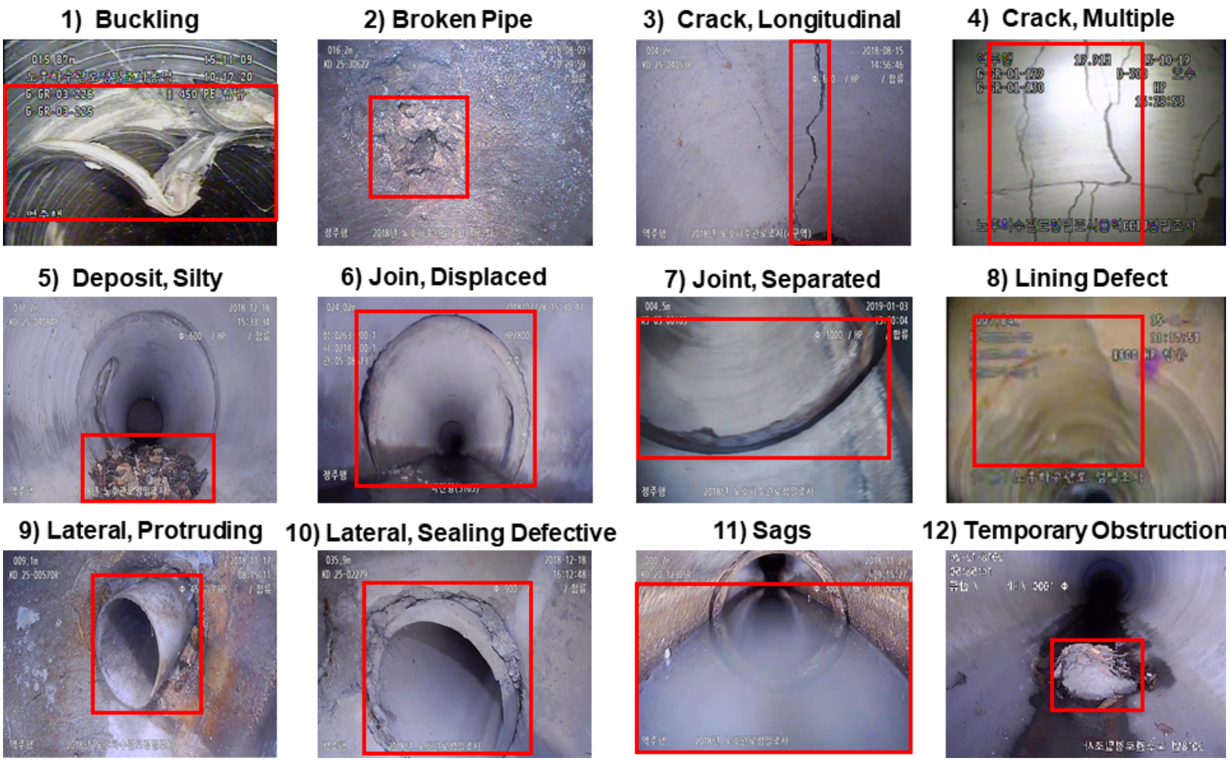
#### 2) Crack Data Pre-processing

Data augmentation is a common pre-processing approach that has been proved to reduce over-fit and increase the number of training data [5]. Some common data augmentation techniques are flipping, rotation, and translation. As a result, flipping, rotation, scaling augmentation are implemented in this study to increase the number of images for each class. After the augmentation process, the number of images is increased by three times to a total of $12,456$ from the original $4,152$ images.

#### 3) Crack Data Description

The dataset is comprised of training, validation, and testing sets. Each dataset was randomly divided into 8:1:1. Table 1 describes information on sewage crack datasets. The training set provides more capabilities to train the robust model and improve accuracy on the test set. Because the size of the frame extracted from various images varies, all images were resized to $640 \times 640$.

[1] https://github.com/tzutalin/labelImg



1) Buckling
2) Broken Pipe
3) Crack, Longitudinal
4) Crack, Multiple
5) Deposit, Silty
6) Join, Displaced
7) Joint, Separated
8) Lining Defect
9) Lateral, Protruding
10) Lateral, Sealing Defective
11) Sags
12) Temporary Obstruction

FIGURE 2. Example images for each class of the sewer crack detection dataset.

TABLE 1. Detailed description of the number of images for each type of crack in the proposed sewer crack dataset. Note: indicates the number of images

| Crack Name | Code | # of Training Set | # of Testing Set |
|---|---|---|---|
| Buckling | BC | 1,080 | 120 |
| Broken Pipe | BK | 1,011 | 144 |
| Crack, Longitudinal | CL | 1,335 | 150 |
| Crack, Multiple | CM | 1,383 | 156 |
| Deposit, Silty | DS | 1,665 | 186 |
| Join, Displaced | JD | 1,059 | 117 |
| Joint, Separated | JS | 1,227 | 138 |
| Lining Defect | LD | 909 | 102 |
| Lateral, Protruding | LP | 981 | 108 |
| Lateral, Sealing | LS | 1,389 | 156 |
| Sags | SG | 1,113 | 126 |
| Temporary Obstruction | TO | 1,554 | 171 |

### 4) Text Recognition Dataset

The caption information in the CCTV image provides important information to investigate the sewer pipe. Travel distance and sewage pipe ID information were detected and recognized as shown in Fig. 3.

The travel distance was crucial information during the inspection because it allowed the inspectors to know the exact location of a crack based on the robot's referenced travel distance. Sewer pipe ID was also important and useful information that assisted the investigators in distinguishing the recorded videos when numerous amounts of videos were processed by the system. The character information extraction dataset used CCTV images such as crack datasets on top. The research extracted only one image per 30 frames for the recognition of text information in the image. A total of $1,910$ text images and labeled them as containing text for text detection. The dataset was divided into 8:1:1 (train, validate and test), with $1,528$, $191$ and $191$, respectively. Then, the images were labeled and the images and labeling text files were converted into lmdb format.

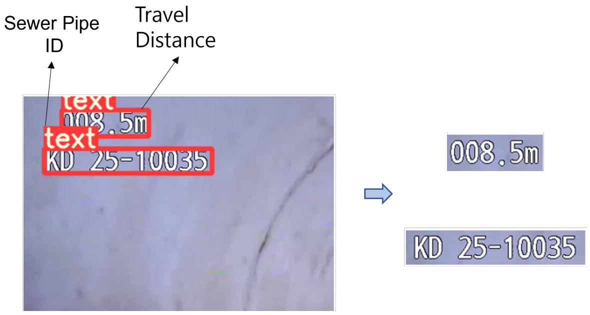


FIGURE 3. An example of a detected character

### B. CRACK DETECTION

#### 1) YOLOv5 Network Structure

Object detection is a technique to detect object and draw a bounding box that provides the object's position and categorical probability for each object. As the latest product in the YOLO architecture series, YOLOv5 [31] has high detection accuracy (mAP) of network models, a fast inference speed (FPS), and the fastest detection speed is up to 140 frames per second. The YOLOv5 architecture relies on three modules: backbone network (focus on image features), neck (number of mixed layers) and head (bounding box and prediction). When inputting, YOLOv5 uses an adaptive anchor frame calculation function in which the values of the optimal anchor frames in different training sets are adaptively calculated during each training to update the optimal anchor box values in the dataset.

Backbone is a part of an extracting feature map from images, and YOLOv5 uses the cross stage partial network (CSPNet) [32]. In the backbone, the original 640 × 640 × 3 image is fed into the main part of the architecture, and the slicing work by the focus module initially 320 × 320 × 12 feature map, and a convolutional operation consisting of 32 convolutional kernels produces an output feature map of 320 × 320 × 32. The BottleneckCSP module is designed to perform feature extraction on feature maps to richly extract deep features from images. The bottleneckCSP consists mostly of bottleneck modules and is capable of reducing gradient information redundancy in the process of optimizing convolutional neural networks compared to that of other large CNNs. YOLOv5 is divided into 4 different versions, including Small, Medium, Large, and Extra-Large based on the width and depth of the BottleneckCSP module, whereas the Backbone, Neck, and Head are kept the same. The spatial pyramid pooling (SPP) module [33] of the backbone network is designed to convert feature maps of all sizes into fixed-size feature vectors, increasing the acceptance field of the network and capturing functions of various sizes. In the C3 block, input separates from two half. As shown in the Fig. 5, one passes through the conv, bottleneck block, the other passes through the conv layer, and two are concatenated and another conv layer is connected.

The neck of YOLOv5 is based on Feature Pyramid Network (FPN) [34] and Path Aggregation Network (PAN) structures [35], such as YOLOv4. It uses PANet as Neck and aggregates functions and creates a multi-scale feature maps with the help of FPN. This operation improves the upward path and enhances the network's capability to recognize target objects on different scales.

#### 2) Improved YOLOv5 with Attention Mechanism Module

In many research areas, the detection accuracy and position accuracy of the detection model has been improved after adding the attention mechanism [23]. Previously, Hu et al. proposed a compact Squeeze-and-Excitation (SE) module to obtain the CNN structure's inter-channel relationship [36]. Global average-pooled features are extracted in the Squeeze-and-Excitation module to calculate channel-wise attention. However, this module ignored spatial attention, which is imperative in deciding which part of the image, the model should focus on. Therefore, the CBAM is implemented to enhance the performance of the proposed model because it overcomes the drawback of the SE module [29]. The CBAM structure is shown in Fig. 4

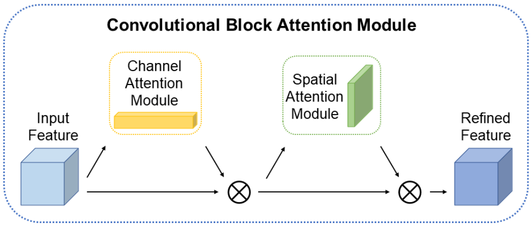


FIGURE 4. Convolutional block attention module (CBAM) module network structure.

CBAM uses the sequential application method, first reducing the input dimension of the tensor and then using only one convolution to calculate spatial attention and use location information. As displayed in Fig. 4, CBAM consists of channel attention and spatial attention modules. The channel attention focuses on specific channel using the relationship between the channel of the feature map. When the feature map in the middle of the model is given as an input like $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, where C is the input channel, W is the width, and H is the height feature map. The channel attention module initially generates a channel attention map by utilizing the internal relationship of the input feature F, and for effective calculation, the spatial dimension of the input feature map is flattened to become $C \times 1 \times 1$. That is, a 1D channel attention map $\mathbf{M_c} \in \mathbb{R}^{C \times 1 \times 1}$is generated. In addition, average pooling and max pooling are applied to integrate spatial information. Using the two pooling operations together improves performance. It can be expressed as a formula as [29]:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(\mathrm{MaxPool}(F))) \\ &= \sigma\left(\mathbf{W_1}\left(\mathbf{W_0}\left(F^C_{\text{avg}}\right)\right) + \mathbf{W_1}\left(\mathbf{W_0}\left(F^C_{\text{max}}\right)\right)\right) \end{aligned} \tag{1}$$

The working principle is to integrate spatial information from feature maps using both avg pooling and max pooling to generate descriptors $\mathrm{F}^{\mathrm{c}}_{\mathrm{avg}}$ and $\mathbf{F}^{\mathrm{c}}_{\mathrm{max}}$. In this formula, MLP stands for multi-layer perceptron. Each of the $\mathrm{F}^{\mathrm{c}}_{\mathrm{avg}}$ and $\mathbf{F}^{\mathrm{c}}_{\mathrm{max}}$ is passed to the MLP to generate each attention map, and then the two are added to create the final channel attention map. In the above equation, $\sigma$ is the sigmoid activation function, and $W_0$ and $W_1$ represent the weights of MLP.

$$M_s(F) = \sigma\left(f^{7\times7}([\mathrm{AvgPool}(F); \mathrm{MaxPool}(F)])\right) \tag{2}$$

$$= \sigma\left(f^{7\times7}\left(F^S_{\text{avg}}; F^S_{\text{max}}\right)\right) \tag{3}$$

The spatial attention module focuses on where the important information is located. From the channel attention feature map, the spatial attention module is retrieved. The 2D spatial attention map $\mathbf{M_c} \in \mathbb{R}^{C\times1\times1}$ concatenates two values of $\mathrm{F}^{\mathrm{c}}_{\mathrm{avg}}$ and $\mathbf{F}^{\mathrm{c}}_{\max}$ of $1 \times H \times W$ generated by applying Maxpool and Avgpool to channel information by multiplying the channel attention map and the input feature map. This process is calculated as equ. 2. Then, as in equ. 3, the attention weight is calculated with one filter with a size of $7 \times 7$. Woo et al. [29] investigated the effect of 3 and 7 kernel sizes on the convolutional layer and found that adopting a larger kernel size produces better accuracy because a large receptive area was needed to determine spatially significant regions. Therefore, $7 \times 7$ convolution kernel size is used.

Additionally, the research proposes an improved YOLOv5 using attention module to detect sewage pipe cracks in real time. As shown in the Fig. 5, in our model structure, CBAM attention modules were added to the neck part because the CBAM attention module helps the model resist confusing information and accurately extract important features. This study added modules to both the backbone and the neck, as well as adding additional modules to the part of the neck. However, the results showed that the accuracy was the highest when adding only the neck part, as shown in the Fig. 5.

### 3) Micro-scale Detection Layer

Cracks of various sizes were detected on different scales. However, when detecting cracks in a sewer pipe in a video, it was difficult to detect micro cracks by various angles without stopping the video at all angles. Micro-scale crack detection was difficult because YOLOv5 detected three scales and was given by down-sample input dimensions of 32, 16, and 8, respectively [37]. Therefore, this study additionally down-sampled the input image dimension to 4 and added a new detection head to create a wider, more detailed detection network structure for detecting small cracks. The new multi-scale fusion can see $160 \times 160$, $80 \times 80$, $40 \times 40$, $20 \times 20$ four feature scales as shown in Figure 5. The additional unsampled fused feature maps are linked to the feature map of $160 \times 160$ pixels from the backbone network to create a new layer and C3 modules and concat modules are used in this process.

## C. TEXT IMAGE PRE-PROCESSING

The most important and critical task was to detect text information of the sewer image about Travel Distance and Sewer Pipe ID, and the images were extracted and cut to a $142 \times 270$ size for better recognition. For the text recognition of high accuracy of the sewer CCTV image, the preprocessing process was applied because noise had to be removed. At first, gray scales and blurry images were performed as shown in Fig. 6. Noise was removed by using a median filter that replaces each filter value with the median value of neighboring pixels. Finally, the image was smoothed using a gaussian filter that replaces the current pixel value using the weighted average of the current filter value and neighboring pixel values.

## D. TEXT DETECTION AND RECOGNITION

After preprocessing the text detection dataset, it was fed to the proposed improved YOLOv5 model for performing text detection. As mentioned previously in the text recognition dataset, the model was trained to detect two crucial subtitle information, including pipe ID and travel distance information. The detected text information can be recognized to provide additional useful information for the inspectors.

After the detection process, text recognition is implemented. This study applied the Pre-trained TPS-ResNet-BiLSTM-Attn (TRBA) framework to perform the subtitle recognition. The four main stages of TRBA are transformation, feature extraction, sequence modeling, and prediction. In the transformation stage, the input text image is normalized using the Spatial Transformer Network (STN) [38] in order to ease other stage's tasks. After that, feature extraction is carried out to map the normalized image to a representation that concentrates on the features that are crucial for text recognition. Moreover, irrelevant features such as font, color, size, and background are also suppressed at this stage. Sequence modeling is then applied to extract the contextual information within a sequence of characters for the following stage to robustly predict each character, rather than doing it independently. Finally, the prediction stage evaluates the output character sequence from the identified features of the processed image. There are two options for the prediction phase: Connectionist temporal classification (CTC) [39] can predict the characteristics of variable numbers, and CTC's key method is to predict characters in each column and delete repeated characters and spaces to modify full character sequences into unfixed character flows.Attention-based sequence prediction (Attn) [40], [41] allows the STR model to learn a character-level language model that represents output class dependency. Attention is used in this paper with the highest accuracy in STR.

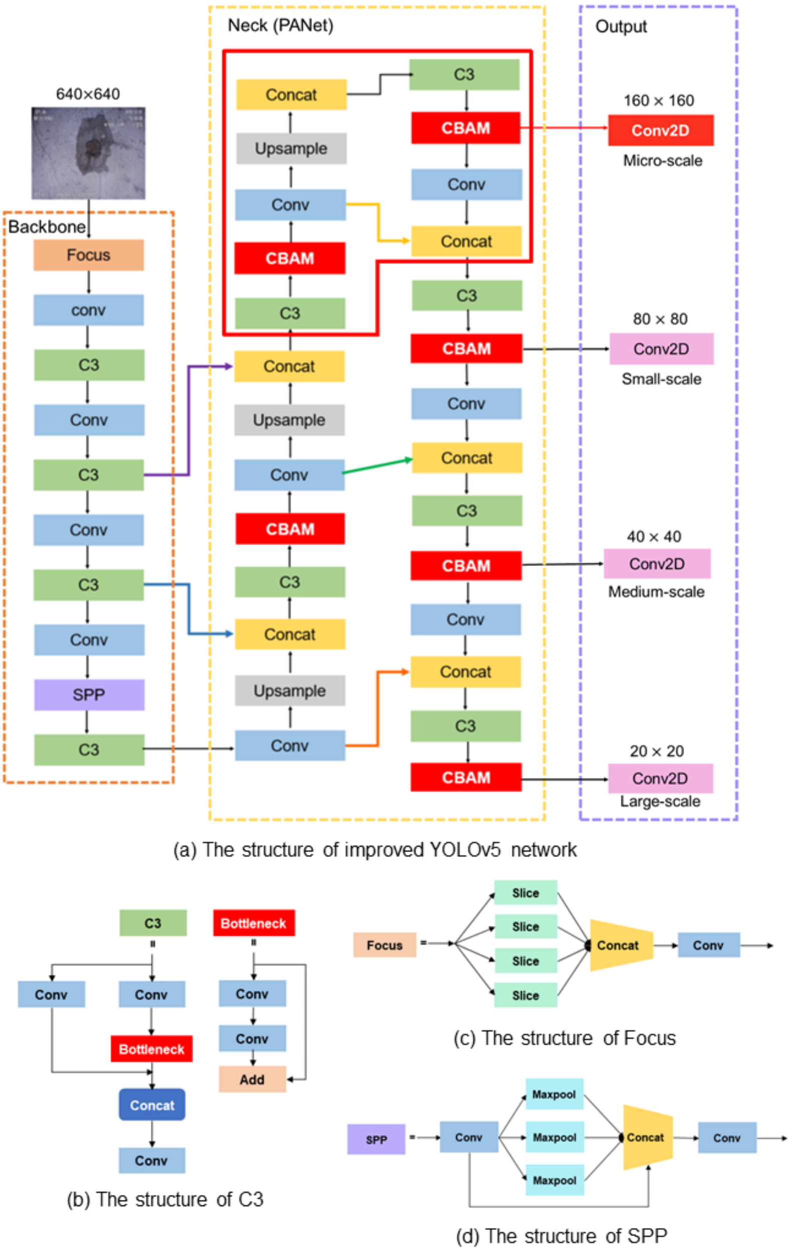
In summary, TPS transformation is implemented in the transformation stage. ResNet is utilized as the feature extractor to improve the expressiveness, in case of severe background or chaos fonts. In the third stage, the BiLSTM model is deployed to extract sequence modeling and reduce truncated characters. Finally, Attn prediction finds missing information or characters [26].

# IV. EXPERIMENTAL RESULTS

All experiments were developed using Pytorch 1.7.0. The programming environment was Ubuntu 20.04 with an NVIDIA Tesla V100 (32GB) GPU.

## A. EVALUATION METRICS

Evaluation metric is a crucial tool used for evaluating model performance. This research applied several evaluation metrics, including Precision (1), Recall (2), mAP (3), and number of frames per second (FPS), to evaluate the sewer crack detection model.



(a) The structure of improved YOLOv5 network

(b) The structure of C3

(c) The structure of Focus

(d) The structure of SPP

**FIGURE 5.** Structure of the improved YOLOv5 network where CBAM and micro-scale are added. Note: the red rectangular box indicates the added layers for micro-scale.

Equ. 4 and Equ. 5 describe the precision and recall.

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{Falsepositive}} \tag{4}$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \tag{5}$$

In Table 2, TP is the detection of precisely identified objects that originally existed in an image, and FP is the detection of the wrong object; FN is an object that actually exists in the picture but is not detected in the classification model.

**TABLE 2.** Confusion matrix

| | | **Prediction** | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Ground Truth** | **Positive** | True Positive (TP) | False Positive (FP) |
| | **Negative** | False Negative (FN) | True Negative (TN) |

The detection model could be evaluated based solely on Precision and Recall. In order to comprehensively evaluate the model performance, the interpolated AP presented by

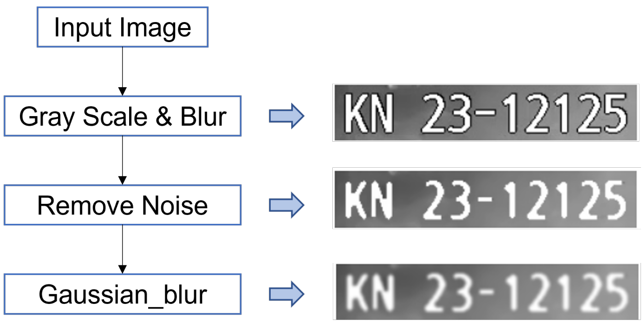


**FIGURE 6.** Character preprocessing process

[42] to the examined object detector model, as in equ. 6, where P is Precision, R is Recall, and AP is calculated as the area under the graph line in the precision-recall graph. mAP is to evaluate the performance of the algorithm by calculating the AP for each class, summing them all, and dividing them by the number of object classes equ. 7, where $N$ is the total number of classes.

$$\mathrm{AveragePrecision(AP)} = \int_0^1 \mathrm{P(R)}d\mathrm{R} \tag{6}$$

$$\mathrm{MeanAveragePrecision(mAP)} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{AP_i} \tag{7}$$

## B. PARAMETER SETTINGS

### 1) Sewer crack detection model

Since YOLOv5 is not a model developed for sewerage maintenance, the architecture and parameters of YOLOv5 could not satisfy the requirements for real-time sewer crack detection. Therefore, the research modified the parameter settings to suit the characteristics of the sewer pipe crack. To enhance performance of model, the technique of increasing resolution was used, and the color and brightness of the image were adjusted and learned among the augmentation used in YOLOv5. The output dimension was $(classes + 5) \times 3$, where 3 represents the box of each scale, and 5 represents the coordinates (x-axis, y-axis, width, height) and confidence of each prediction box. The coco dataset had 80 classes. The output dimensions were $3 \times (5 + 80) = 255$. Since 12 classes were predicted, The classifier part of YOLOv5 needs to be fixed. Our output dimension became $3 \times (5 + 12) = 51$.

The improved YOLOv5l network use the momentum factor of 0.937. SGD optimizer is used as the main optimizer with the learning rate of 0.001 and the weight decay rate of 0.0005. The IoU threshold is set at 0.6. The total number of epochs is 100 with the batch size of 16.

### 2) Text detection and text recognition model

To train model for the sewer image dataset, the hyper parameters values were set as layout size=16, momentum=0.1, learning rate = 0.01, Leaky ReLU activation function, optimizer=Adam, loss function=binary cross entropy to create a model for detecting sewer information. In this study, the TPS-ResNet-BiLSTM-Attn model implemented in PyTorch was selected as the basic model, and the number of fiducial points of TPS was set to 20, the number of ResNet output channels was set to 512, and the size of the BiLSTM hidden state was set to 256. The batch size was 192, and the model was verified after every 2,000 iterations, and a maximum of 300,000 iterations was set. For the optimization of the learning model, the Lr = 1.0, the value of the Adam optimizer was set to 0.9, the decay rate of the Adadelta was set to 0.95, and the eps was set to 1e-8.

## C. CRACK DETECTION RESULT

### 1) Proposed our detection model performance

The most important thing in this study was to consider both the accuracy and speed of crack detection. Our improved model achieved an mAP@0.5% of 75.9, Precision% of 75.8, Recall of 73.1, and FPS of 31. The detection speed and accuracy of our model satisfy the real-time performance in terms of computational complexity.

**TABLE 3.** Ablation experiment of the model

| Data Aug | Micro-scale Detection Layer | CBAM | mAP @.5 | Prec | Recall | Epoch |
|---|---|---|---|---|---|---|
| | | | 74.5 | 73 | 70.1 | 100 |
| √ | | | 75.2 | 74.4 | 72.2 | |
| √ | √ | | 75.3 | 74.3 | 71.9 | |
| √ | √ | √ | **75.9** | **75.8** | **73.1** | |

In Table 3, using the data augmentation method, the mAP improved from 74.5 to 75.2. And prec increased by 1.4 and recall by 2.1. After adding the micro-scale detection layer, the mAP reached 75.3. Finally, our model obtained the highest mAP of 75.9 with the addition of cbam and compared with the mAP 74.5 of the original YOLOv5 model.

The improved YOLOv5 detected smaller cracks and was superior to other models because it adopted a learning method that added CBAM modules to focus more on important parts. The loss function shows the difference between the prediction and the ground truth. The optimization function uses the loss function to minimize the error in the algorithm. It is used to evaluate how good or bad the model performs. Fig. 7. shows the validation loss curve of the original YOLOv5 and improved YOLOv5 models. The validation loss of YOLOv5 and the improved YOLOv5 decreases significantly until epoch 20th. After that, the validation loss of YOLOv5 and improved YOLOv5 declines gradually and stops at 0.03 and 0.02, respectively, at epoch 100th. As a result, it can be seen that the validation loss of the improved YOLOv5 is better than that of the original YOLOv5 model.

Table 4 shows the mAP for each class tested on the same test dataset using the original YOLOv5 and enhanced YOLOv5 models. In general, more than 80% of mAP works well in Buckling, Multiple Crack, Deposit Silty, and Lateral Sealing. The highest mAP of 90.1% was obtained in Buckling however, this model showed poor performance with

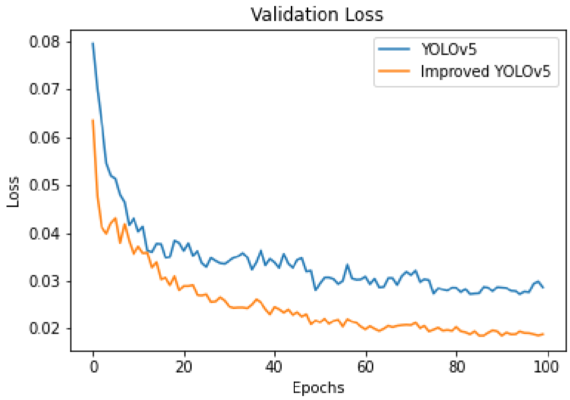


FIGURE 7. Validation loss graph compared to YOLOv5 and Ours

a low mAP value of 55.6% in each of the Sags specific classes. The reason why Sag class shows poor performance is that sometimes flooding in the image is severe and flooding throughout the screen often occurs. It can be seen that the proposed model has improved mAP by 1.4%, precision by 4.6%, and recall by 2% compared to the original YOLOv5.

TABLE 4. Comparison between the original model and the our improved model. Note: Prec indicate precision and Rec means recall

| Category | YOLOv5 | | | Ours | | |
|---|---|---|---|---|---|---|
| | mAP | Prec | Rec | mAP | Prec | Rec |
| Buckling | 89.9 | 77.2 | 90.8 | **90.1** | 86.9 | 95.5 |
| Broken Pipe | 76.9 | 76.7 | 80 | 77.8 | 74.4 | 80.0 |
| Crack, Longitudinal | 67.5 | 75.6 | 71.4 | 72.9 | 81.1 | 65.8 |
| Crack, Multiple | 78.1 | 71.8 | 80.9 | **82.5** | 92.0 | 72.3 |
| Deposit, Silty | 80.3 | 87.5 | 71.4 | **81.6** | 79.1 | 74.7 |
| Join, Displaced | 76.5 | 79.6 | 67.4 | 77.4 | 79.6 | 72.1 |
| Joint, Separated | 86.4 | 82.4 | 68.4 | 75.8 | 76.1 | 86.0 |
| Lining Defect | 69.3 | 88.8 | 67.9 | 67.5 | 86.2 | 64.3 |
| Lateral, Protruding | 76.1 | 74.4 | 61.5 | 78.8 | 69.0 | 67.2 |
| Lateral, Sealing | 82.6 | 71.1 | 83.9 | **84.8** | 73.1 | 90.3 |
| Sags | 48.5 | 40.7 | 38.7 | 55.6 | 59.6 | 38.7 |
| Temporary Obstruction | 62.3 | 50.8 | 57.1 | 66.6 | 74.1 | 57.1 |
| Average | 74.5 | 73.0 | 70.0 | 75.9 | 77.6 | 72.0 |

The precision-recall curve is computed to demonstrate the precision and recall value as shown in Fig. 8. The average mAP for all classes is 75.9%. The SG(Sags) class has the lowest mAP at 56.6%, whereas the BC(Bucking) class has the highest mAP at 90.1%.

### 2) Comparison with other models

In this section, several standard detection models were compared for state-of-the-art performance using our same dataset. Table 5 shows the comparison of detection accuracy (mAP) and FPS tested with the evaluation models, SSD [43], Retinanet [44], YOLOv3 [10], YOLOv4 [45], YOLOv5, and enhanced YOLOv5, based on the validation dataset. The proposed model shows the mAP@0.5% of the modified model (75.9) was improved by 1.4 more than that of the original YOLOv5 (74.5), the precision% of our model (75.8) was 2.8 higher than that of the original (73), and the recall% was higher than the original (73). (72), our model (73.1) improved by 1.1.

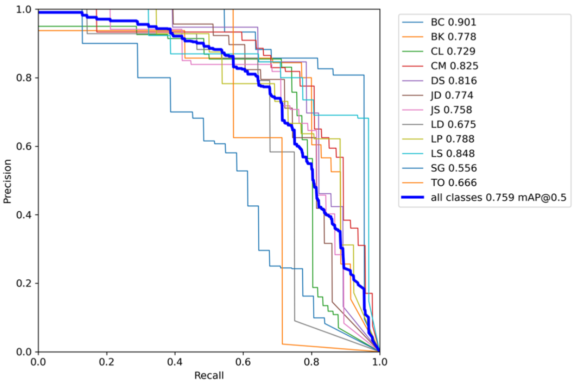


FIGURE 8. Precision-recall curve of the proposed model on the testing set.

TABLE 5. Performance comparison of five object detection networks including SSD, Retinanet, YOLOv3, YOLOv4, YOLOv5

| Method | mAP | FPS |
|---|---|---|
| SSD [43] | 53.2 | 20 |
| RetinaNet [44] | 58.8 | 22 |
| YOLOv3 [10] | 60.3 | 35 |
| YOLOv4 [45] | 67.1 | 31 |
| YOLOv5 | 74.6 | **34** |
| **Ours** | **75.9** | 31 |

Fig. 9. shows the training loss curve of the different models.

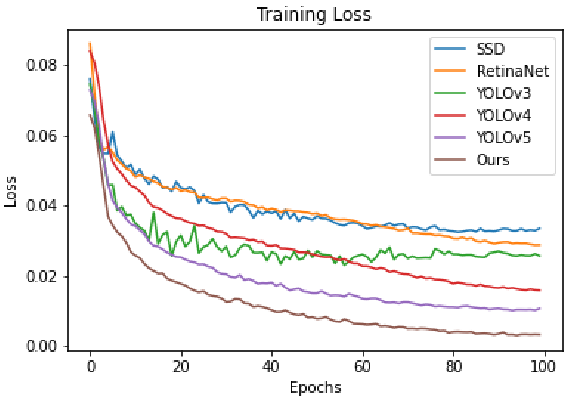


FIGURE 9. Training loss graph comparing with other models (SSD, RetinaNet, YOLOv3, YOLOv4, YOLOv5, Ours).

As the number of repetitions increased, the improved YOLOv5 algorithm curve gradually converged more than the original algorithm curve, and the loss value decreased more rapidly. Our improved model is illustrated in the Fig. 9, where loss value is stable even with just 100 epochs. Compared to the original YOLOv5, our model demonstrates that the regression is faster and more accurate, improving the accuracy of the improved YOLOv5.

It can be seen from the results in Table 6 that the proposed model achieves the highest mAP of 68.1% on the COCO benchmark dataset, which improves the mAP by 0.8 compared to the original YOLOv5 model and outperform all other models. In addition, the proposed model shows that it can process five more frames compared to the original YOLOv5 model.

TABLE 6. Performance comparison of the related methods using COCO benchmark dataset.

| Method | mAP | FPS |
|---|---|---|
| SSD [43] | 48.5 | 22 |
| RetinaNet [44] | 57.5 | **5** |
| YOLOv3 [10] | 57.9 | 20 |
| YOLOv4 [45] | 65.7 | 23 |
| YOLOv5 | 67.3 | 35 |
| **Ours** | **68.1** | 30 |

### D. QUALITATIVE EVALUATIONS

In Fig. 10, it is shown that the same image is captured at the same time. Through the comparative photos, it can be confirmed that the small crack detection ability was improved as targets that were not detected before were recognized. Even when crack detection was difficult because of light, crack detection was good and the confidence score of other cracks increased. Testing shows that our improved model can reach 0.004s per image processing speed, which is consistent with real-time detection, even though the model structure is more complex.

In addition, Fig. 11 show the qualitative evaluation results of the proposed network in the test dataset, whereas Fig. 12 shows the detection results of the proposed model on challenging cases.

The Fig. 11 shows that many cracks can be recognized in each different scene. Also, even if the image is blurry and the screen condition is poor, the performance was still highly efficient. However, in Fig. 12, there were a few bad outcomes where the cracks that were not clear were not accurately detected or recognized. Our dataset has other images with higher image quality, but problems occurred in the test process due to low image quality images and poor brightness conditions. In the first picture, no joint cracks were detected because the cracks were fainter and narrower than the training datasets. Temporary obstruction (TO) was incorrectly recognized as desposits/silty (DS) because of branches and some soils. In the second image, the crack is blurred and the entire part of the crack was not detected. Finally, the third picture shows that the box range of the sags class is incorrect, because our model is trained only with clean image datasets without bubbles as in the picture.

### E. TEXT DETECTION AND RECOGNITION RESULT

The text detection results using the YOLOv5 model on 120 test images are as followed: Precision is at 0.994, Recall is at 1, and mAP is 0.993. The results show that the model effectively detected all text with a high detection rate. In addition, with the implementation of image preprocessing, the accuracy was improved to 97.3%. As shown in Fig. 13, the text detection model works well in complex background and simple background scenarios.

Three different pre-trained text recognition models, including ViTSTR [28], TextOCR [46], and Azure [47], are implemented in order to show the effectiveness of the text recognition method (TPS-ResNet-BiLSTM-Attn) used in this study. The hyperparameters of those models are set like those recommended by the original papers. Table 7 shows the text recognition results of the four models on the test dataset, and it can be seen that TRBA has the highest accuracy of 97.3%, whereas other models, such as Azure and TextOCR achieve slightly lower accuracy of 96.4% and 96.9%, respectively.

TABLE 7. The performance of various models for text recognition.

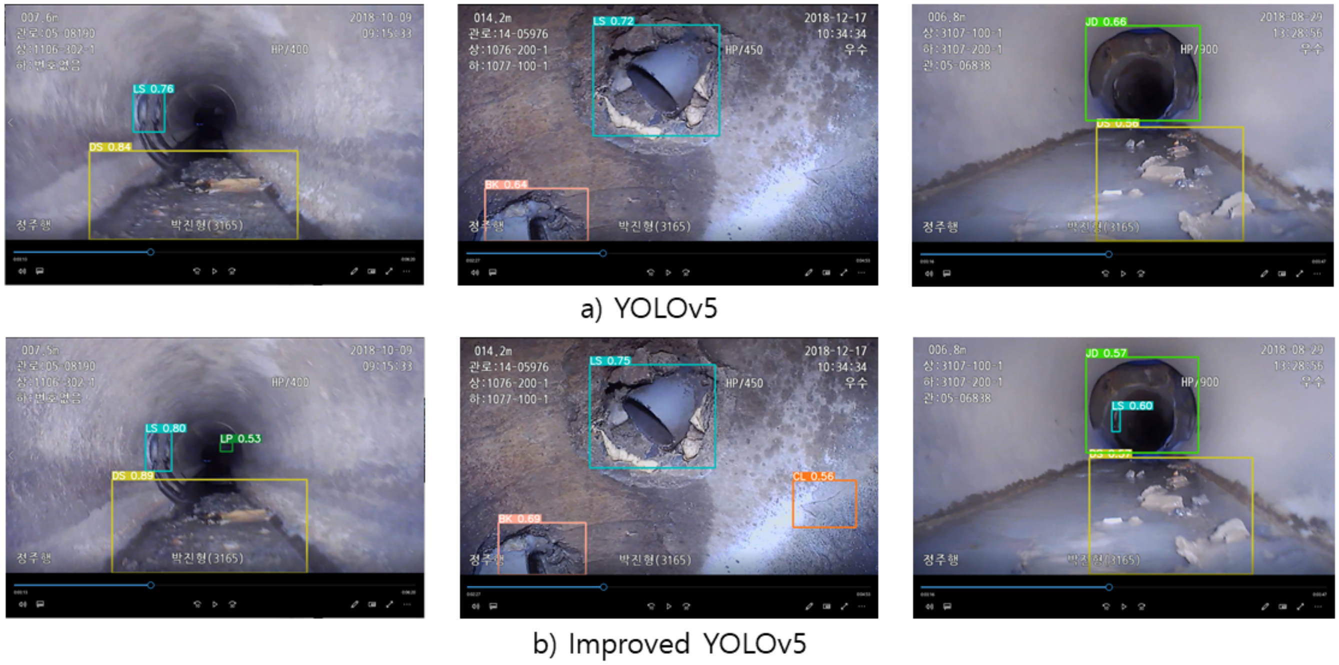
| **Method** | **ViTSTR** [28] | **TextOCR** [46] | **Azure** [47] | **Ours(TRBA)** |
|---|---|---|---|---|
| Val acc | 95.8 | 96.4 | 96.9 | 97.3 |

## V. CONCLUSION

In this paper, we introduced a new crack detection approach and text detection and recognition method for underground sewers photographed from CCTV. We first created a dataset from the video that was photographed in the sewer pipe, compared with five superior architectures, and confirmed the model accuracy on test dataset. Among them, the performance of our YOLOv5 model was the highest at mAP@0.5/% 75.9. To effectively improve detection accuracy, a CBAM attention module was added into the network and the model was modified by adding a micro-scale for higher position accuracy in small crack detection. The current model obtains a 1.4% higher in mAP@0.5/% compared to that of the original model, and even smaller cracks were well detected. In order to recognize important information in the image such as Sewer Pipe ID and Travel Distance, the text information detection and recognition module first detected important information efficiently, and applied an appropriate pre-trained TRBA (TPS-ResNet-BiLSTM-Attn) model. In the future study, we plan to train a large dataset to improve performance, seeking to find a way to control the speed and accuracy while making the model lightweight.

## ACKNOWLEDGMENT

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540) and National Research Foundation of Korea (NRF) grant funded by the Korea government, Ministry of Science and ICT (MSIT) (2021R1F1A1046339) and by a grant(20212020900150) from "Development and Demonstration of Technology for Customers Bigdata-based



a) YOLOv5

b) Improved YOLOv5

**FIGURE 10.** Comparison between the cracks detected by the original model and the improved yolo model.

Energy Management in the Field of Heat Supply Chain" funded by Ministry of Trade, Industry and Energy of Korean government.

## REFERENCES

[1] J. C. Cheng and M. Wang, “Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques,” Automation in Construction, vol. 95, pp. 155–171, 2018.

[2] G. Pan, Y. Zheng, S. Guo, and Y. Lv, “Automatic sewer pipe defect semantic segmentation based on improved u-net,” Automation in Construction, vol. 119, p. 103383, 2020.

[3] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: Review of methods and applications,” Expert Systems with Applications, vol. 73, pp. 220–239, 2017.

[4] J. B. Haurum and T. B. Moeslund, “A survey on image-based automation of cctv and sset sewer inspections,” Automation in Construction, vol. 111, p. 103061, 2020.

[5] Y. Li, H. Wang, L. M. Dang, M. J. Piran, and H. Moon, “A robust instance segmentation framework for underground sewer defect detection,” Measurement, p. 110727, 2022.

[6] S. I. Hassan, L. M. Dang, I. Mehmood, S. Im, C. Choi, J. Kang, Y.-S. Park, and H. Moon, “Underground sewer pipe condition assessment based on convolutional neural networks,” Automation in Construction, vol. 106, p. 102849, 2019.

[7] L. M. Dang, H. Wang, Y. Li, T. N. Nguyen, and H. Moon, “Defecttr: End-to-end defect detection for sewage networks using a transformer,” Construction and Building Materials, vol. 325, p. 126584, 2022.

[8] D. Meijer, L. Scholten, F. Clemens, and A. Knobbe, “A defect classification methodology for sewer image sets with convolutional neural networks,” Automation in Construction, vol. 104, pp. 281–298, 2019.

[9] L. M. Dang, S. Kyeong, Y. Li, H. Wang, T. N. Nguyen, and H. Moon, “Deep learning-based sewer defect classification for highly imbalanced dataset,” Computers & Industrial Engineering, vol. 161, p. 107630, 2021.

[10] A. Farhadi and J. Redmon, “Yolov3: An incremental improvement,” in Computer Vision and Pattern Recognition. Springer Berlin/Heidelberg, Germany, 2018, pp. 1804–2767.

[11] Z. Liu and Y. Kleiner, “State of the art review of inspection technologies for condition assessment of water pipes,” Measurement, vol. 46, no. 1, pp. 1–15, 2013.

[12] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, “Sensor-based and vision-based human activity recognition: A comprehensive survey,” Pattern Recognition, vol. 108, p. 107561, 2020.

[13] S. Costello, D. Chapman, C. Rogers, and N. Metje, “Underground asset location and condition assessment technologies,” Tunnelling and Underground Space Technology, vol. 22, no. 5-6, pp. 524–542, 2007.

[14] M. R. Halfawy and J. Hengmeechai, “Automated defect detection in sewer closed circuit television images using histograms of oriented gradients and support vector machine,” Automation in Construction, vol. 38, pp. 1–13, 2014.

[15] I. Khalifa, A. E. Aboutabl, and G. S. A. Aziz, “A new image model for predicting cracks in sewer pipes based on time,” International Journal of Computer Applications, vol. 87, no. 9, 2014.

[16] L. M. Dang, H. Wang, Y. Li, Y. Park, C. Oh, T. N. Nguyen, and H. Moon, “Automatic tunnel lining crack evaluation and measurement using deep learning,” Tunnelling and Underground Space Technology, vol. 124, p. 104472, 2022.

[17] Q. Xie, D. Li, J. Xu, Z. Yu, and J. Wang, “Automatic detection and classification of sewer defects via hierarchical deep learning,” IEEE Transactions on Automation Science and Engineering, vol. 16, no. 4, pp. 1836–1847, 2019.

[18] S. S. Kumar, D. M. Abraham, M. R. Jahanshahi, T. Iseley, and J. Starr, “Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks,” Automation in Construction, vol. 91, pp. 273–283, 2018.

[19] X. Fang, W. Guo, Q. Li, J. Zhu, Z. Chen, J. Yu, B. Zhou, and H. Yang, “Sewer pipeline fault identification using anomaly detection algorithms on video sequences,” IEEE Access, vol. 8, pp. 39 574–39 586, 2020.

[20] Y. Tan, R. Cai, J. Li, P. Chen, and M. Wang, “Automatic detection of sewer defects based on improved you only look once algorithm,” Automation in Construction, vol. 131, p. 103912, 2021.

[21] D. Li, A. Cong, and S. Guo, “Sewer damage detection from imbalanced cctv inspection data using deep convolutional neural networks with hierarchical classification,” Automation in Construction, vol. 101, pp. 199–208, 2019.

[22] X. Zuo, B. Dai, Y. Shan, J. Shen, C. Hu, and S. Huang, “Classifying cracks at sub-class level in closed circuit television sewer inspection videos,” Automation in Construction, vol. 118, p. 103289, 2020.

[23] P. Wang, H. Huang, M. Wang, and B. Li, “Yolov5s-fcg: An improved yolov5 method for inspecting riders’ helmet wearing,” in Journal of

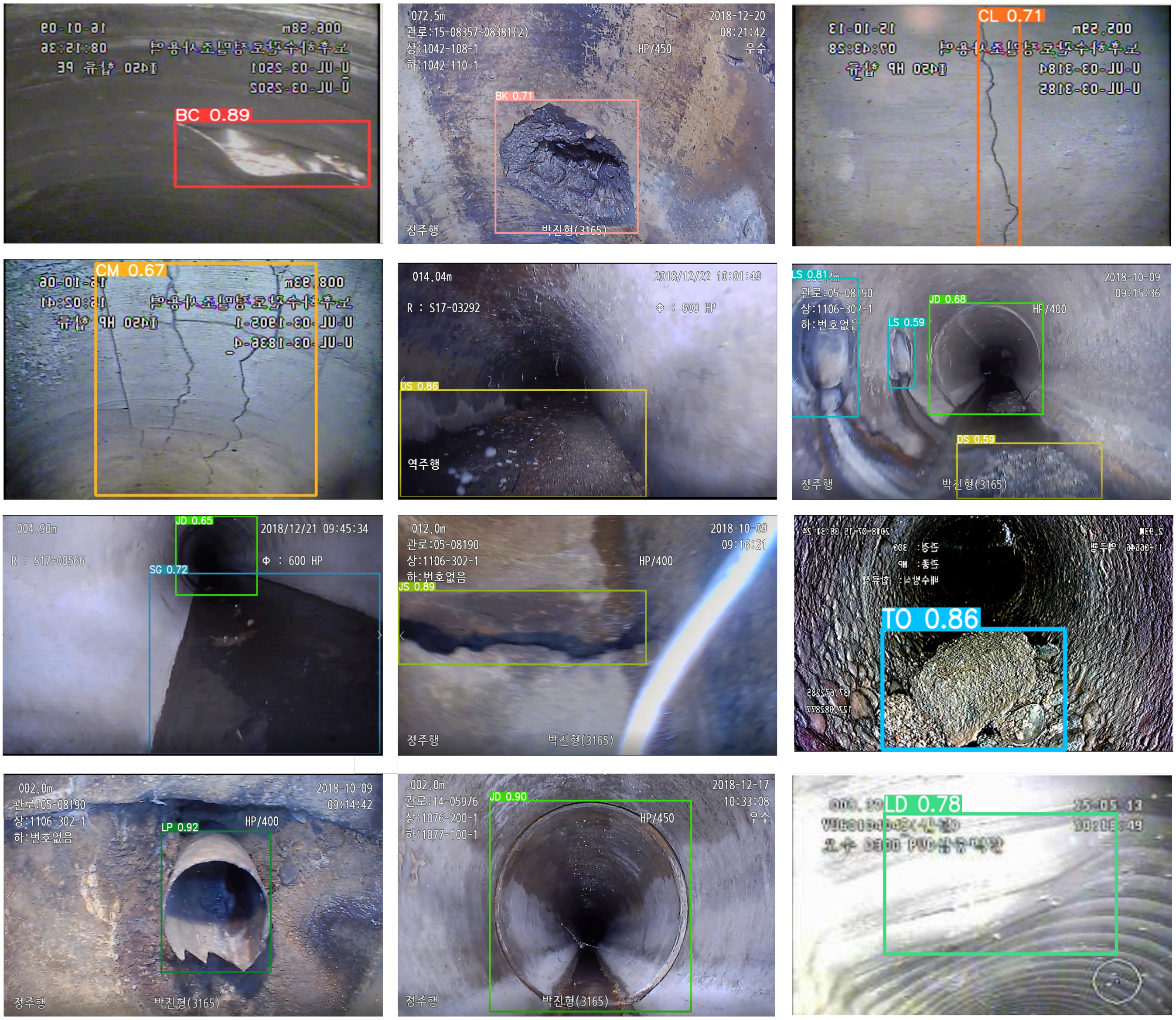


**FIGURE 11.** Detection results for various types of sewer cracks (Correct cases)

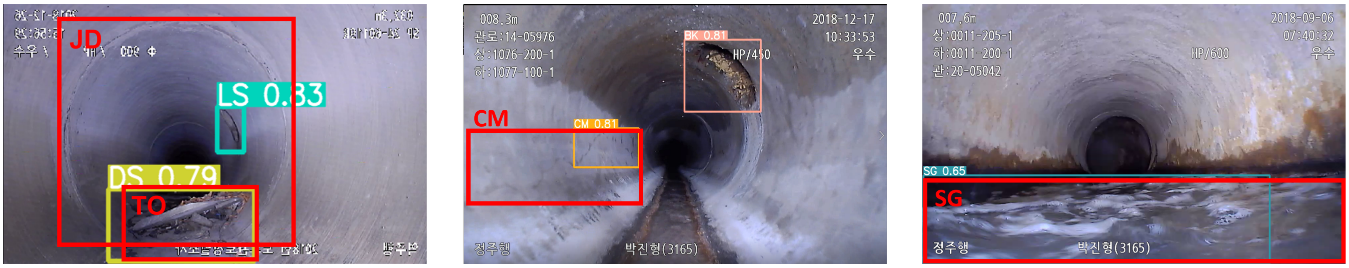


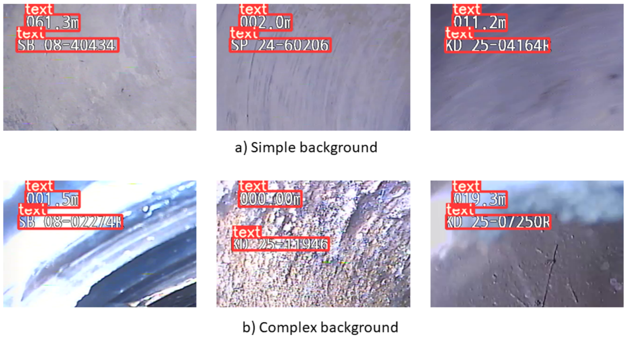
**FIGURE 12.** Detection results for various types of sewer cracks (Failed cases)

Physics: Conference Series, vol. 2024, no. 1. IOP Publishing, 2021, p. 012059.

[24] J. B. Haurum and T. B. Moeslund, “Sewer-ml: A multi-label sewer defect classification dataset and benchmark,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13 456–13 467.

[25] L. M. Dang, S. I. Hassan, S. Im, I. Mehmood, and H. Moon, “Utilizing text recognition for the defects extraction in sewers cctv inspection videos,” Computers in Industry, vol. 99, pp. 96–109, 2018.

[26] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, “What is wrong with scene text recognition model comparisons? dataset and model analysis,” in Proceedings of the IEEE/CVF International Con-



a) Simple background

b) Complex background

**FIGURE 13.** Text detection results on simple background and complex background scenarios

ference on Computer Vision, 2019, pp. 4715–4723.

[27] Y.-C. Chen, Y.-C. Chang, Y.-C. Chang, and Y.-R. Yeh, "Traditional chinese synthetic datasets verified with labeled data for scene text recognition," arXiv preprint arXiv:2111.13327, 2021.

[28] R. Atienza, "Vision transformer for fast and efficient scene text recognition," arXiv preprint arXiv:2105.08582, 2021.

[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[30] Y. Liu, B. Lu, J. Peng, and Z. Zhang, "Research on the use of yolov5 object detection algorithm in mask wearing recognition," World Scientific Research Journal, pp. 276–284, 2020.

[31] https://github.com/ultralytics/yolov5, "Yolov5," Ultralytics open-source, 2021.

[32] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 390–391.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.

[34] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

[35] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8759–8768.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[37] J. Zhao, X. Zhang, J. Yan, X. Qiu, X. Yao, Y. Tian, Y. Zhu, and W. Cao, "A wheat spike detection method in uav images based on improved yolov5," Remote Sensing, vol. 13, no. 16, p. 3095, 2021.

[38] M. Jaderberg, K. Simonyan, A. Zisserman et al., "Spatial transformer networks," Advances in neural information processing systems, vol. 28, pp. 2017–2025, 2015.

[39] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369–376.

[40] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4168–4176.

[41] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: Towards accurate text recognition in natural images," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5076–5084.

[42] G. Salton and M. J. McGill, Introduction to modern information retrieval. mcgraw-hill, 1983.

[43] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in European conference on computer vision. Springer, 2016, pp. 21–37.

[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[45] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[46] A. Singh, G. Pang, M. Toh, J. Huang, W. Galuba, and T. Hassner, "Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8802–8812.

[47] "MicrosoftDocs/azure-docs," https://github.com/MicrosoftDocs/azure-docs/blob/main/articles/cognitive-services/Computer-vision/overview-ocr.md, accessed 2020-03-03.

• • •