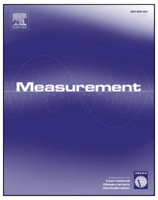
Contents lists available at ScienceDirect

Measurement

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



A robust instance segmentation framework for underground sewer defect detection

Yanfen Li [a], Hanxiang Wang [a], L.Minh Dang [b], Md Jalil Piran [a], Hyeonjoon Moon [a,*]

[a] *Department of Computer Science and Engineering, Sejong University, Seoul, Republic of Korea*
[b] *Department of Information Technology, FPT University at Ho Chi Minh city, Vietnam*

ARTICLE INFO

Keywords:
Deep learning
Defect inspection
Underground sewer
Instance segmentation

ABSTRACT

The inspection of underground sewer defects plays a considerable role in estimating the structural integrity and avoiding various unforeseen functional failures. However, the conventional sewer defect inspection approaches suffer from the blurry and vaporous environment inside the sewer pipes, which significantly lowers the performance. Besides, it is challenging to achieve efficient and accurate condition assessment by the common manual inspection. Therefore, this manuscript introduces an automatic instance segmentation-based defect analysis framework. The main contributions include 1) a novel defect segmentation model called *Pipe-SOLO* is firstly presented to segment six common types of defects at the instance level by proposing an efficient backbone structure *(Res2Net-Mish-BN-101)* and designing an enhanced BiFPN *(EBiFPN)*, 2) a GAN-based dehazing model is applied to effectively solve the image blurring problem, and 3) a publicly available sewer defect segmentation dataset. The experimental results show the proposed Pipe-SOLO achieved an improvement of 7.3% compared with the state-of-the-art method in terms of the mean Average Precision (mAP). Therefore, the proposed defect segmentation method is promising to be integrated with real-life applications that require defect localization and estimation.

1. Introduction

The underground sewage pipelines are a significant part of the urban infrastructure that is distributed throughout the city. More attention has been paid recently in order to treat the sewage and rainwater due to the city's development and the significant growth of its population in a short time. However, the sewer pipes were inevitably found in various stages of damage and leakage due to corrosion and poor management [1], which seriously affected the sewer treatment performance and affected the environment [2]. Therefore, it is necessary to take early intervention measures like defect inspection, evaluation, and rehabilitation for the deteriorating pipes. Previously, sewer inspectors identified defects and assessed the risk level based on the onsite manual observation. Beyond a doubt, this kind of subjective approach is improper and impractical for massive sewer pipes. In order to address this issue, an automatic method that can effectively inspect distinct defects in closed-circuit television (CCTV) videos or images is needed to manage underground sewer system.

In recent years, various technologies have been introduced to inspect and estimate the defects effectively. The common defect inspection techniques mainly include defect classification, detection, and segmentation [3], each technique has its notable characteristics. For example, deep learning (DL)-based methods have been increasingly used because they can obtain significantly better results when the data is sufficient [1,4,5]. Nevertheless, the image classification that only considers each defect's label provide less information than the object detection methods, because the main purpose of the object detection is to indicate the defect type and the precise location of the detected defect through the bounding box [6]. Currently, the image segmentation technique that can obtain the most comprehensive information has shown its superiority. The principal image segmentation methods are divided into morphological segmentation [7,8], semantic segmentation [9], and instance segmentation [10]. The semantic and instance segmentation methods based on fully-supervised learning usually achieve better performances than the morphological segmentation methods based on unsupervised learning [3]. Compared with semantic segmentation, the instance segmentation is specific to each instance instead of each class. This provides great convenience for the sewer inspectors to distinguish and analyze different defect samples in the same class. As a result, this study proposes a practical defect inspection framework that

* Corresponding author at: Department of Computer Science and Engineering, Sejong University, Seoul, Republic of Korea.
E-mail address: hmoon@sejong.ac.kr (H. Moon).

https://doi.org/10.1016/j.measurement.2022.110727
Received 24 September 2021; Received in revised form 7 January 2022; Accepted 9 January 2022
Available online 12 January 2022
0263-2241/

reports the objective boundary, area, and the corresponding class for the individual sample to evaluate the defect thoroughly.

Image blurring is an inevitable problem in vision-based inspection tasks that directly affect the performances of the machine learning (ML) models, so many researchers have shifted the attention to tackle it in recent studies [11,12]. For instance, a dark channel prior (DCP)-based dehazing process was introduced to preprocess the low visibility CCTV images [13]. In addition, an end-to-end dehazing system was proposed in [14], and it shows excellent performance on the benchmark dataset. Even though their dehazing processing modules can remove the underlying haze, the brightness of the processed image is changed. Motivated by those studies, an image preprocessing technology that can perform well on dehazing and maintaining the brightness is integrated into the defect segmentation system to improve the proposed framework's performance.

The main contributions of this study are as follows.

1. An instance-based sewer defect segmentation model (Pipe-SOLO) is firstly proposed by introducing an efficient backbone structure (Res2Net-Mish-BN-101) and designing an enhanced BiFPN (EBiFPN).
2. A dehazing algorithm is integrated into the proposed defect inspection framework in order to handle the image blurring problem and then increase the detection rate.
3. A manually validated and annotated dataset for the instance segmentation task is provided.

The rest of the paper is organized as follows. Section 2 summarizes and analyzes the related literature. The sewer defect analysis system is explained in Section 3. Section 4 explains the data collection process and evaluation metrics used in this study. After that, several experiments are discussed in Section 5 to evaluate the performance of the proposed approach. Section 6 concludes the research by showing current limitations and future research directions.

2. Related work

2.1. Image preprocessing

Image blurring is a common problem during the data acquisition process, so different dehazing technologies were proposed to solve it. For example, a conventional and effective dehazing algorithm called DCP was proposed to deblur a single image. However, the DCP algorithm showed poor performance on the images with low contrast [12]. Even though the Bayesian-based dehazing method suggested by Ju et al. was robust to all hazy images, some coefficients related to the atmospheric conditions needed to be predetermined [15]. In another work, an end-to-end dehazing system called DehazeNet was introduced to estimate the medium transmission map and perform dehazing [14]. Nevertheless, an error usually occurred in the intermediate process of the DehazeNet that affected the final dehazing performance. With a different approach than the mentioned methods, a dehazing network called the gated context aggregation network (GCANet) was offered recently to perform the dehazing processing. The main advantage of the approach was that it could perform dehazing well and retained the original brightness without any prior knowledge [11]. As a result, the GCANet dehazing model is integrated into the defect inspection framework proposed in this paper in order to efficiently remove noise and enhance the overall image quality.

2.2. Defect inspection

Recently, various methods have been proposed to analyze defects in different structural monitoring systems, such as pavements [16], tunnel surface [10], and steels [17,18]. The following subsections discuss the strengths and weaknesses of three kinds of common methods (defect classification, detection, and segmentation) and reveal the primary purpose of the study.

2.2.1. Defect classification

Image classification is an essential topic of the computer vision (CV) field, which has drawn a lot of attention from the research community. There is a growing number of studies that have been introduced recently in order to identify the defects in the public infrastructure. For example, an ML-based diagnosis system was suggested in [19] to recognize seven types of sewer defects. The experimental results showed that the support vector machine (SVM) classifier obtained an accuracy of 84.1%. The authors also revealed that the accuracy was highly related to the number of training images. Considering the conventional ML methods that require an additional feature engineering process, an ensemble of deep convolutional neural networks (CNNs) was presented to classify three classes of defects [4]. Even though the proposed model obtained an overall accuracy of 86.2%, the method failed to recognize some subclass defects that were commonly encountered. In another work, a deep learning-based sewer pipe condition evaluation system was presented to recognize six predefined defects on CCTV videos [1]. The technique achieved a promising classification result at 96.33%, but it ignored the multi-class defects in an image. A hierarchical classification method was recently introduced to detect and classify the defects from an extremely imbalanced CCTV dataset [5]. Although the overall accuracy of the binary classification in the high-level task was improved by 4.8% by using hierarchical softmax, the model performed poorly on individual defects in the low-level task.

2.2.2. Defect detection

Apart from the defect labels provided by the classification algorithms, the locations of different defects in sewer images are also important information. In recent years, there exist two object detection algorithms that are applied to detect distinct sewer pipe defects: one-stage detectors and two-stage detectors. For instance, the proposed framework in [20] was concerned with the issue of real-time automatic defect detection by streamlining the data and customizing a YOLOv3 model. Their customized one-stage method achieved a mean Average Precision (mAP) of 85.37%, and the detection speed reached about 33 frames per second (FPS). However, less training data collected in their study cannot provide the adequate features for each class of defect. As for a two-stage detector, Cheng, J. C. and Wang, M. developed a defect detection system based on the faster R-CNN model. Experiments demonstrate that the detection accuracy keeps increasing by extending the dataset, adding some convolutional layers, and modifying appropriate hyper-parameters. But this approach is limited to the detection of the defects with similar color or geometry [6].

2.2.3. Defect segmentation

Since the defect segmentation technique can provide the detailed information, such as the defect's class, location, and boundary, it is considered a crucial tool to assess any sewer pipe's condition. Most of the previous methods mainly focused on the morphological segmentation approach. For example, three distinct morphological methods were implemented to segment two typical defects (cracks and open joints) [7]. The experimental results showed that the morphological segmentation using the edge detection (MSED) method worked well on the crack class, whereas the opening top-hat operation (OTHO) method achieved high performance on the open joint class. Su et al. introduced a similar sewer segmentation system using the MSED method [8] and verified that the results generated by the MSED method were better than the OTHO method. All the conventional morphological segmentation methods that require many complicated processing steps are time-consuming and error-prone. Therefore, a modified version of the U-Net structure called PipeUNet was presented to perform the semantic segmentation [9]. The experimental results confirmed that the PipeUNet model achieved the Mean Intersection over Union (MIoU) of 76.37%

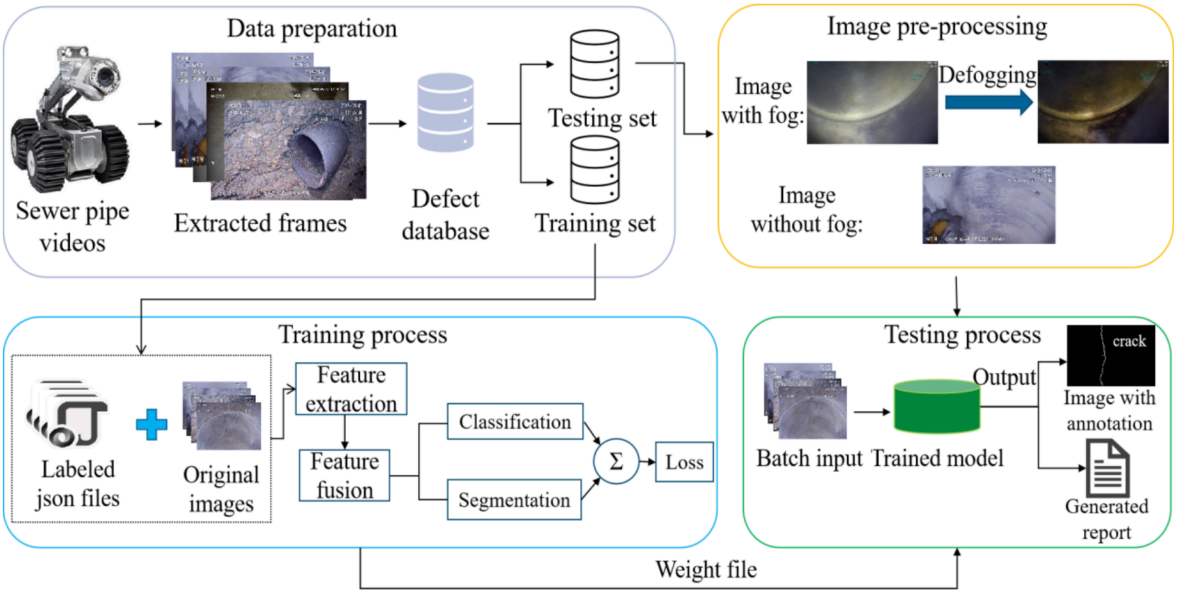


Fig. 1. Illustration for the overall process of the proposed sewer pipe inspection system.

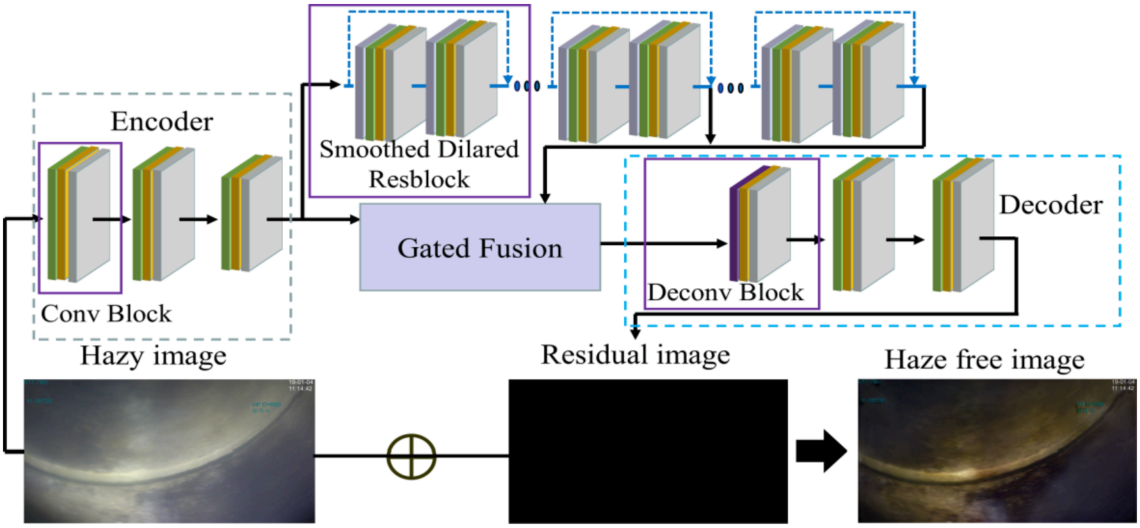


Fig. 2. Four main modules of the GCANet algorithm, which include encoder module, smoothed dilated resblock module, gated fusion sub-network, and decoder module.

and a high processing speed at 32 FPS. Nevertheless, the semantic segmentation method was unable to cover a comprehensive analysis for each defect sample. As a result, Xu et al. applied an instance segmentation approach to obtain the detailed information for only three types of defects on tunnel surfaces [10].

Given the drawbacks of previous studies, this manuscript proposed the first instance segmentation-based sewer defect inspection model that requires the pixel-wise semantic labeling and instance labeling simultaneously. Unlike the common segmentation, the segmentation model presented in this study can distinguish different classes and instances in the same image. The experimental results showed that the proposed model outperformed previous methods with the highest mAP of 59.3% on the collected six types of defects.

3. Proposed defect segmentation framework

The main processes of the proposed sewer pipe inspection system are described in Fig. 1. Firstly, all frames are extracted from the CCTV videos, and the frames with defects are manually validated and stored in the defect database. The polygonal annotations of defects were then manually labeled, which are required to train the proposed model. A preprocessing module is applied to remove the underlying haze before testing the foggy images (Section 3.1). After that, 80% of the collected defect dataset is fed into the instance segmentation-based model, whereas the remaining 20% of the dataset is used as the testing dataset to evaluate the model's performance. The proposed defect segmentation model produces the segmented images and the corresponding reports (Section 3.2).

3.1. Image preprocessing

The evaporation of water due to the various temperatures inside the sewer pipelines is an inevitable problem during the data acquisition process, which leads to blurry and noise in the collected CCTV videos that can directly affect the defect analysis's performance. As a result, an effective dehazing model called GCANet [11] was adopted to remove fog

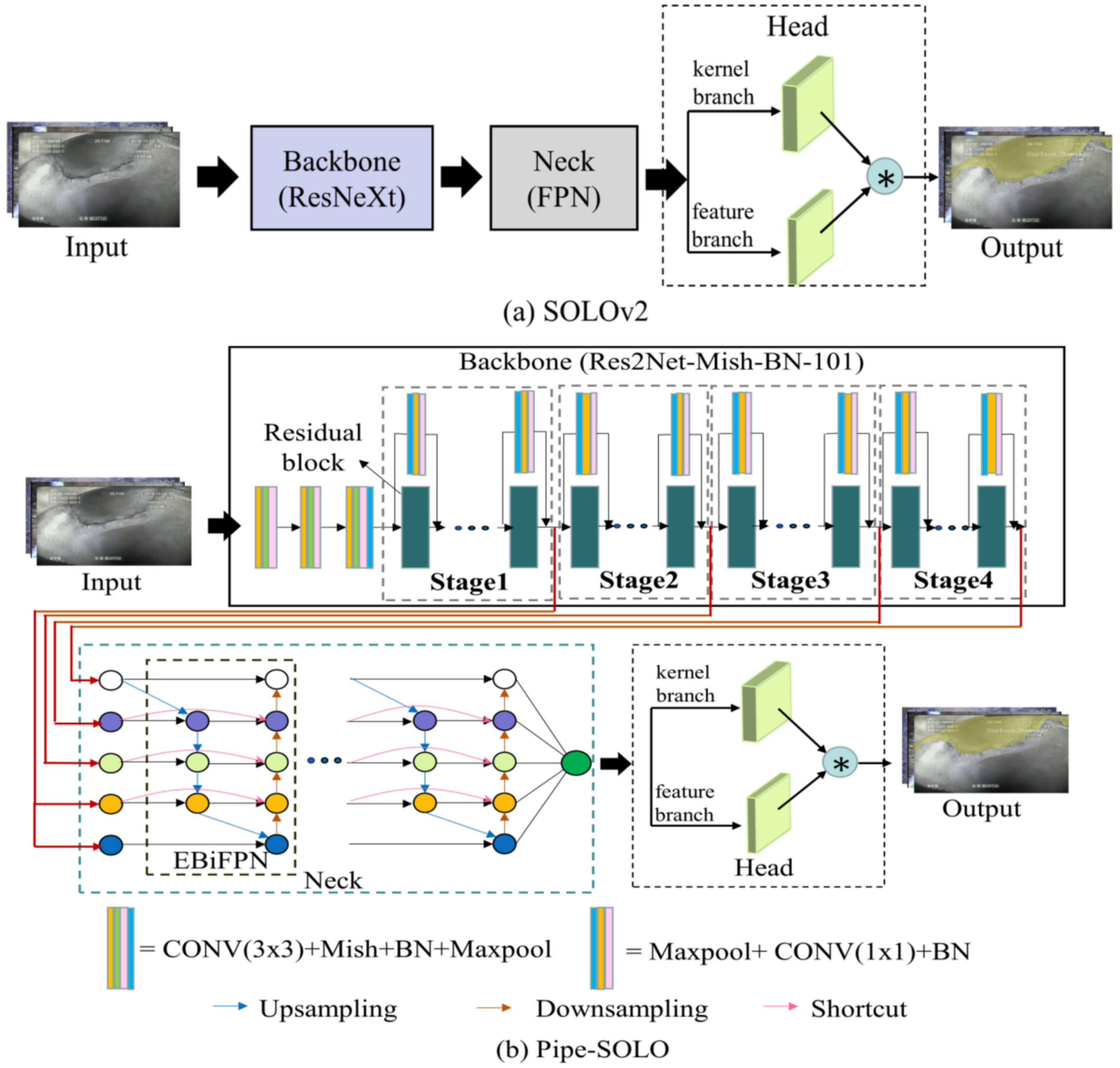


Fig. 3. Overall structures of (a) the original SOLOv2 model and (b) the proposed Pipe-SOLO model.

and improve the image quality in this study. The GCANet model directly figures out the relationship between the hazy image and the haze-free image without applying any prior knowledge. Fig. 2 shows that the GCANet model has four main modules, which include encoder module, smoothed dilated resblock module, gated fusion sub-network, and decoder module. The encoder module consists of three convolution blocks that are used to encode the blurred image as a feature map. Moreover, the smoothed dilated resblock and gated fusion sub-network are applied to replace the conventional down sampling block in order to aggregate more information and fuse the features with different levels. Finally, the fused feature map from the gated fusion sub-network is decoded back to the original image space and is added to the blurred image as a residual image.

The main innovations of the GCANet model are mainly reflected in two aspects. Firstly, the smoothed dilated convolutions are added between the encoder module and the decoder module. Common dehazing algorithms use down sampling operations, such as pooling layer and the convolution layer with large stride, to increase receptive fields. However, these down sampling operations reduce the spatial resolutions of the feature maps, resulting in gridding artifacts in the output images. Therefore, the smoothed dilated convolutions are applied to expand the receptive field without reducing the spatial resolution. Secondly, a new feature fusion approach is presented. GCANet adopts gated fusion sub-network f to calculate the weight coefficients (W_1, W_2, W_3) of the feature maps (F_1,F_2,F_3) with three different scales, as shown in Equation (1). The weight coefficients are weighted to the corresponding features before the future fusion process, which enhances the utilization of effective features. Equation (2) represents the feature fusion process.

$$(W_1, W_2, W_3) = f(F_1, F_2, F_3), \tag{1}$$

$$F = W_1{*}F_1 + W_2{*}F_2 + W_3{*}F_3, \tag{2}$$

Seven resblocks with dilation rates of 2, 2, 2, 4, 4, 4, and 1 are connected between the encoder module and the decoder module in order to improve the model's feature extraction ability. The number of the channels per convolution layer is set to 64. The instance normalization and ReLU function are used to further process the output features of convolution layers.

3.2. Defect segmentation

In order to provide a comprehensive representation, we attempted to explain the model's improvements from the aspects of the model architecture (Section 3.2.1) and the parameter optimization (Section 3.2.2). On the one hand, the detailed structures of the original SOLOV2 model and the proposed Pipe-SOLO model are described in Section 3.2.1. On the other hand, the principle of the loss function proposed in this work is introduced in Section 3.2.2 by exploring an important

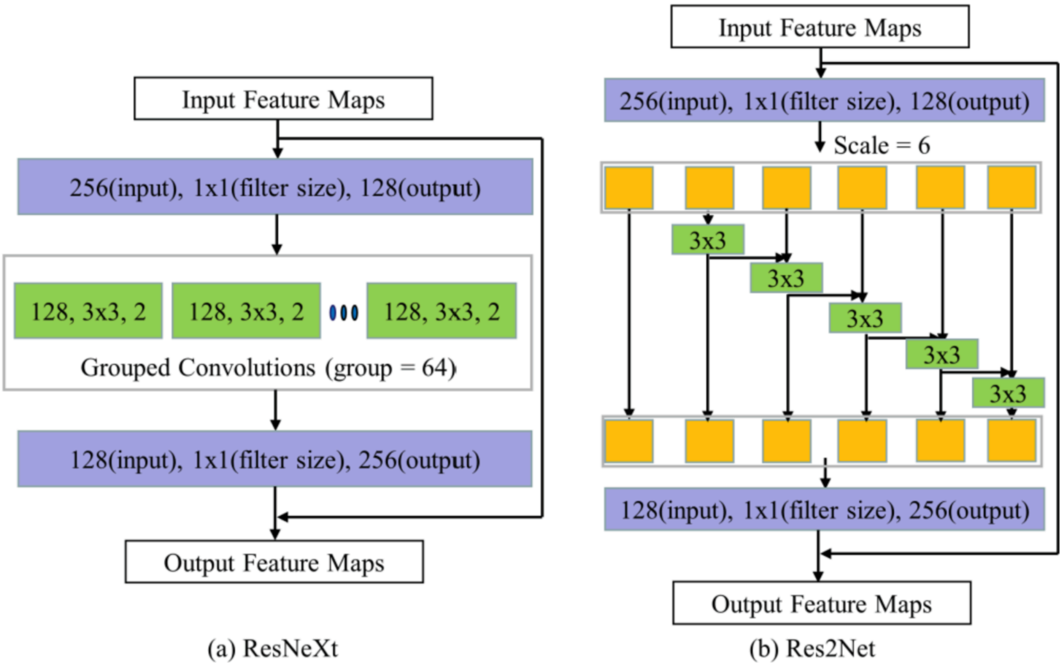

Input Feature Maps
256(input), 1x1(filter size), 128(output)
128, 3x3, 2
128, 3x3, 2
128, 3x3, 2
Grouped Convolutions (group = 64)
128(input), 1x1(filter size), 256(output)
Output Feature Maps
(a) ResNeXt
Input Feature Maps
256(input), 1x1(filter size), 128(output)
Scale = 6
3x3
3x3
3x3
3x3
3x3
128(input), 1x1(filter size), 256(output)
Output Feature Maps
(b) Res2Net

Fig. 4. Residual block of (a) ResNeXt and (b) Res2Net.

parameter.

3.2.1. *Pipe-SOLO architecture*

In this paper, an efficient instance segmentation network, which is called Pipe-SOLO, is proposed to segment defects using information, such as the location and size of the object. The proposed Pipe-SOLO model improved the overall structure and optimized the parameters of the segmenting objects by locations v2 model (SOLOv2) [21]. Fig. 3(a) describes the overall structure of the original SOLOv2, which includes the backbone that is responsible for extracting features, the neck that generates and fuses features, and the head that calculates the loss of the classification and the segmentation branches. Fig. 3(b) shows detailed descriptions of the proposed Pipe-SOLO, which improve the structure of the SOLOv2 model in Fig. 3(a). Firstly, a Res2Net-Mish-BN-101 module was proposed as the backbone of the Pipe-SOLO model (Section **Backbone**). An enhanced BiFPN (EBiFPN) was designed as the neck instead of the FPN network (Section **Neck**). Finally, the effectiveness of the mentioned Res2Net-Mish-BN-101 and EBiFPN structures will be discussed in Section 5.2.

Backbone: Three models (ResNet-50, ResNet-101 [22], and ResNeXt-101 [23]) were adopted as the backbone of the original SOLOv2 model, and the experiment results showed that the ResNeXt-101 outperformed other backbones. The grouped convolutions idea of ResNeXt combines the residual structure of ResNet and the split-transform-merge scheme of the inception net [24], as illustrated in Fig. 4 (a). In another paper, a novel residual network called Res2Net [25] showed excellent performance in various experiments. The main innovation of the Res2Net is that it added a small residual block in a residual unit, which enables the network to extract more fine-grained features and increase the receptive field of each layer. Res2Net uses the strategy of splitting first and then merges the output, as shown in Fig. 4 (b), in order to enable the convolution layers to obtain extra features efficiently. The input feature maps went through a 1x1 convolution layer, and the output was then evenly divided into six predefined blocks according to the number of channels and was passed through a following 3x3 convolution layer. Before the feature maps were fed into the convolution layer, the feature map of the current layer was fused with the output of the previous layer. Finally, the final six outputs were fused and fed into a 1x1 convolution layer.

This study proposes the Res2Net-Mish-BN-101 network by optimizing the structure of the Res2Net. Firstly, all the ReLU activation functions in the Res2Net are replaced by Mish because the Mish activation function does not have the gradient saturation phenomenon and enables better generalization ability due to the smooth activation function [26]. Secondly, the order of the activation layer and the normalization layer is changed to normalize the input of each layer,

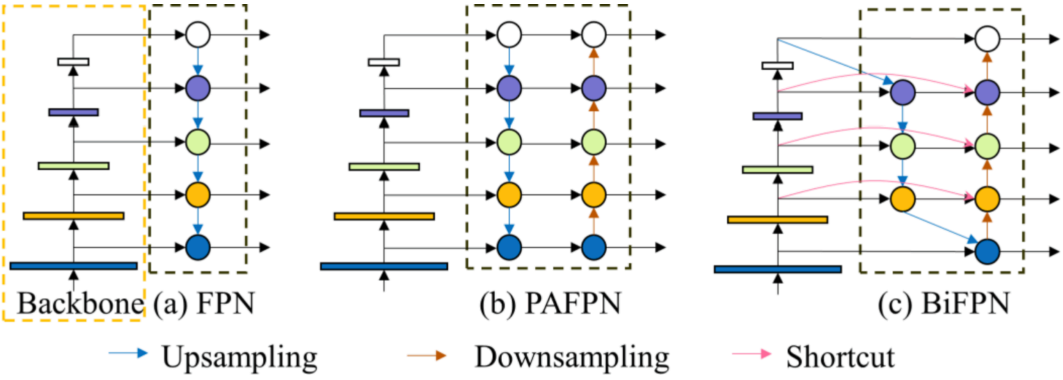


Fig. 5. The detailed structure of three different feature fusion networks, which include (a) FPN, (b) PAFPN, and (c) BiFPN.

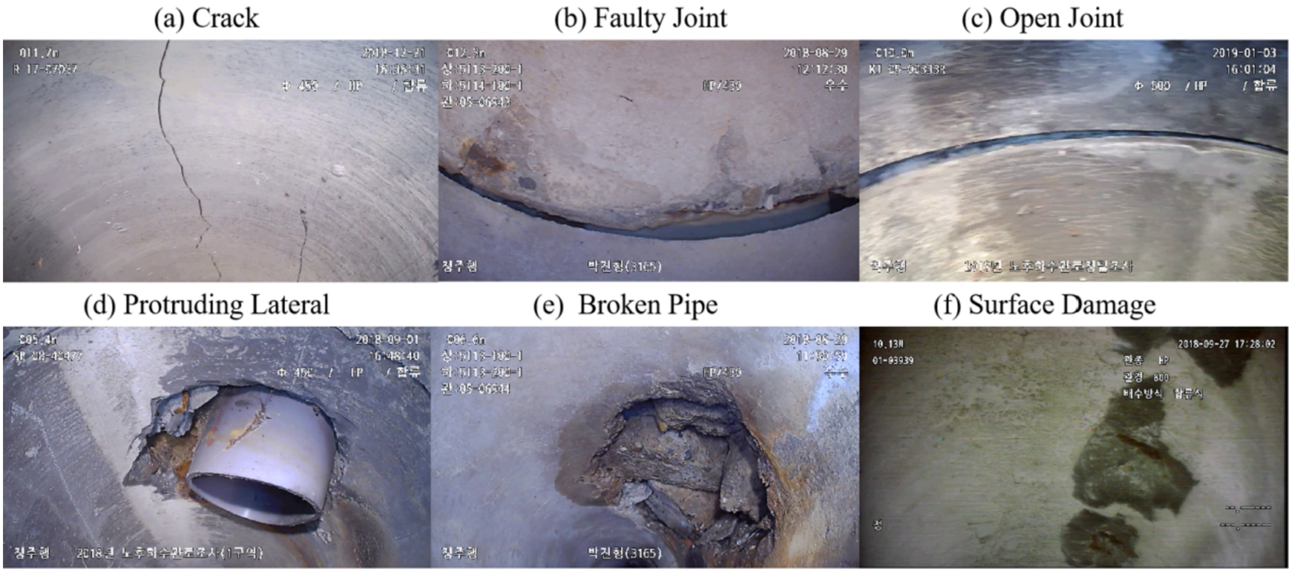


Fig. 6. Sample images for the six types of defects, which include (a) crack, (b) faulty joint, (c) open joint, (d) protruding lateral, (e) broken pipe, (f) surface damage.

which is more conducive to gradient descent [27]. Thirdly, the 7x7 convolutional layer in the Res2Net is replaced by three 3x3 convolutional layers, which can deepen the depth of the network and reduce a huge number of parameters but still retain the same scale of the receptive field. Finally, a Max-pooling layer is added before the 1x1 convolutional layer of the original projection shortcut structure of the Res2Net to reduce the parameters to be processed by the model.

Neck: The feature pyramid structure is widely used in the neck of the instance segmentation task to reuse and fuse multi-scale features. This study analyzes and compares standard feature pyramid structures, such as the feature pyramid network (FPN) [28], path aggregation feature pyramid network (PAFPN) [29], and bi-directional feature pyramid network (BiFPN) [30], as presented in Fig. 5. The FPN contains the top-down path that applies the up sampling operation to restore high-resolution features and preserve the high-level semantic information. On the other hand, a bottom-up path that uses the traditional convolution network to detect high-level semantic information is added in both the PAFPN and BiFPN to enhance the information flow in the entire neck network. Moreover, the shortcut structure is introduced to the BiFPN to improve the feature fusion process. To further improve the feature representation ability of neck, the output channel of BiFPN is increased from 256 to 384. In addition, the performance of batch normalization depends on the setting of batch size. A small batch size decreases the convergence speed and accuracy of the model, while a big batch size can increase the burden of machine memory [31]. To address this question, the batch normalization is replaced by the group normalization that assigns the same kind of features such as shape, color, and texture to a group for normalization operation.

Head: Some novel methods were proposed in the SOLOv2 to reduce the computational complexity of the head structure [21]. For example, the segmentation branch was divided into the mask kernel branch and the mask feature branch. The idea of a dynamic convolution kernel is used in the mask kernel branch to reduce the number of redundant prediction channels. In addition, a novel matrix non-maximum suppression (NMS) algorithm was applied to select the optimal mask and reduce the inference time of the model. The original head structure of the SOLOv2 is adopted in the proposed framework due to the mentioned advantages mentioned.

3.2.2. *Parameter optimization*

In this section, various methods that include a custom loss function and hyperparameter optimization are carried out to improve the model's performance and robustness. Based on the original loss function of SOLOv2, the optimal setting for the weight coefficient is explored. As defined in Equation (3), the custom loss function combines the focal loss of the classification task and the dice loss of the segmentation task according to a specific weight coefficient.

$$Loss = \frac{1}{N}\sum_{i} L_{cla}(p_i, p_i^{*}) + \lambda \frac{1}{N}\sum_{i} p_i^{*} L_{seg}(q_i, q_i^{*}), \tag{3}$$

where L_{cla} and L_{seg} are the loss values for object classification and segmentation, respectively. N represents the number of the predicted samples in the mini-batch, i is the index of the predicted sample, and λ is the weight coefficient. p_i is the specific probability value of the i^{th} predictions, and p_i^{*} means the negative prediction or positive prediction. q_i and q_i^{*} are the predicted mask and the ground truth mask, respectively. The hyperparameters of the proposed model are optimized by implementing and evaluating different optimizers, which include SGD and Adam [32,33], and learning rates that range from 0.01 to 0.0001.

4. Dataset and evaluation metric

4.1. *Dataset*

The sewer inspection CCTV videos of different locations of Seoul, South Korea, which were used in this study, were provided by The Seoul Digital Foundation. The videos were recorded by the inspection robots equipped with high-resolution RGB cameras with the video length ranges from 1 to 20 min. Some fundamental information, such as pipe ID, inspection distance, and inspection time, was printed on each frame of the videos. The sewer pipes investigated in this work are concrete pipes, and sample images are displayed in Fig. 6. This study divides the defects into six categories, which include open joint (OJ), faulty joint (FJ), protruding lateral (PL), crack (C), broken pipe (BP), and surface damage (SD). Some of the mentioned defect categories are at the subclass level, which can be used to evaluate the model's effectiveness because they are challenging for defect detection and segmentation. The specific description for each class of defect is explained as follow:

Table 1
Description of the six classes of the collected sewer defect dataset, which include Crack (C), Faulty Joint (FJ), Open Joint (OJ), Protruding Lateral (PL), Broken Pipe (BP), and Surface Damage (SD).

Index	Defects	Training set	Testing set	Total
1	OJ	417	102	519
2	FJ	498	126	624
3	PL	735	183	918
4	C	486	123	609
5	BP	459	120	579
6	SD	510	129	639
	Total	3,105	783	3,888

- Open joint: An open joint indicates the displacement in pipe joints.
- Faulty joint: A faulty joint represents the deterioration that happens around pipe joints.
- Protruding lateral: Protruding lateral refers to a connecting pipe section that protrudes from the internal diameter of the original pipe.
- Crack: A crack or fracture in the sewer pipeline is caused by the poor original installation or soil bedding.
- Broken pipe: A broken pipe refers to severe structural damage occurring in the pipe.
- Surface damage: Surface damage indicates the slight damage caused by erosion or tribological stresses on the surface.

All defect images are extracted from the original videos, with the resolution of the extracted images ranges from 640x480 to 1280x720. They are then manually labeled by an annotation tool called LabelMe [34] to make the ground truth files containing the detection and segmentation information (bounding box and polygon labelling) in the JSON format. The dataset introduced in this study will be publicly available with the acceptance of the academic research, which is significant for the future studies to make this research area more replicable and transparent. Table 1 describes the detailed information of the collected defect dataset that contains 3888 images. 80% of the data (3105 images) is used as the training set, and the remaining 20% of the data is considered the testing set. In addition, data augmentation technology is used to improve the model's generalization ability, such as Mixup, rotation, adding noise, and color jitter. After data augmentation, the number of training images (9,173) is almost three times more than the previous training amount (3,105).

4.2. Evaluation metric

This section describes the evaluation protocols adopted to examine the performance of the image dehazing and defect segmentation.

The peak signal-to-noise ratio (PSNR) and Structural SIMilarity (SSIM), which are standard evaluation methods used in previous works, are used to evaluate the image dehazing performance. PSNR is calculated based on the deviation between corresponding pixels and does not consider the visual characteristics of the human eye, whereas SSIM is a measure of image similarity from brightness, contrast, and structure, which can better reflect the subjective feeling of human eyes.

The formulas of PSNR and SSIM are described as follows.

$$PSNR(A,B) = 10\log_{10}\left(\frac{255^2}{(m\bullet n)^{-1}\sum_{x,y=1}^{m,n}[A(x,y)-B(x,y)]^2}\right), \tag{4}$$

$$SSIM(A,B) = \frac{(2\mu_A\mu_B + c_1)(2\sigma_{AB} + c_2)}{(\mu_A^2 + \mu_B^2 + c_1)(\sigma_A^2 + \sigma_B^2 + c_2)}, \tag{5}$$

where A and B represent the original haze image and haze free image with the size of m*n, respectively. In (3), μ, σ^2, and σ_{AB} represent the mean, variance, and covariance, respectively. c_1 and c_2 are two constants that are added to avoid the denominator zero.

The mean average precision (mAP), which is the standard evaluation protocol for the COCO dataset [35], was used to calculate the defect segmentation accuracy. The mean value of AP for n classes can be calculated by computing the average precision (AP) for each class using the precision-recall curves under a specific threshold of intersection-over-union (IOU):

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n}, \tag{6}$$

$$Precision = \frac{\text{TruePositive(TP)}}{\text{FalsePositive(FP)} + \text{TruePositive(TP)}}, \tag{7}$$

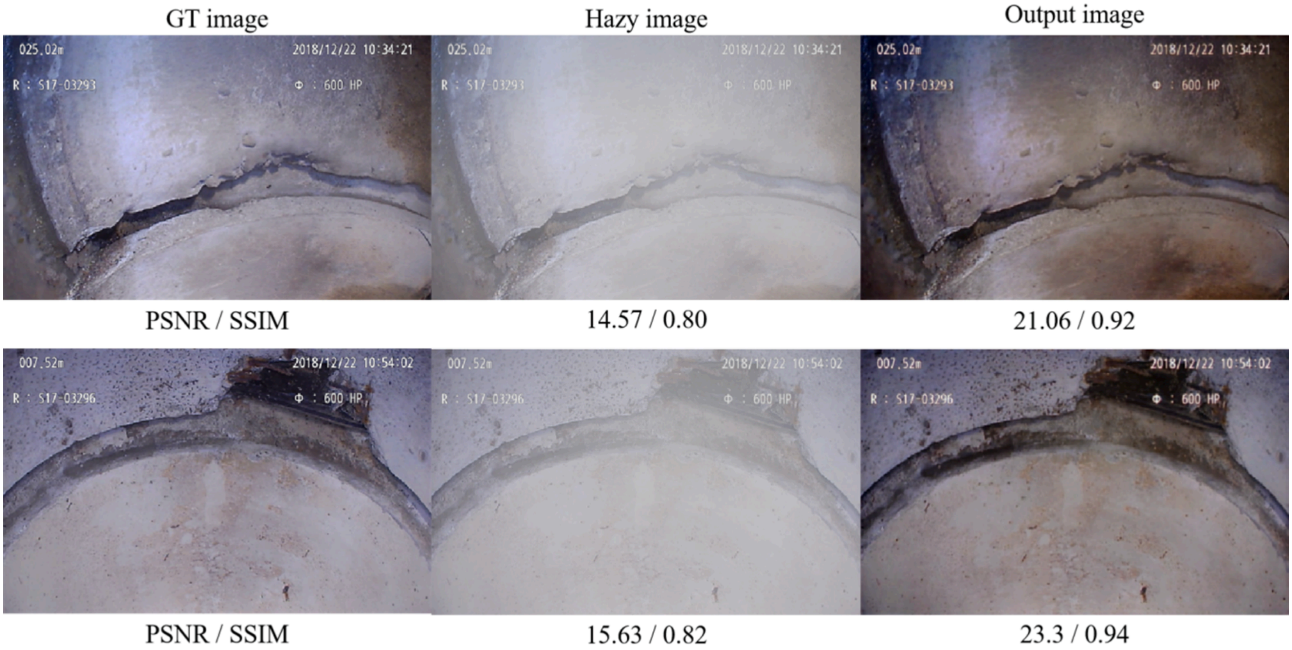


Fig. 7. Evaluation of the preprocessing module that uses the GCANet model by showing the PSNR and SSIM values.

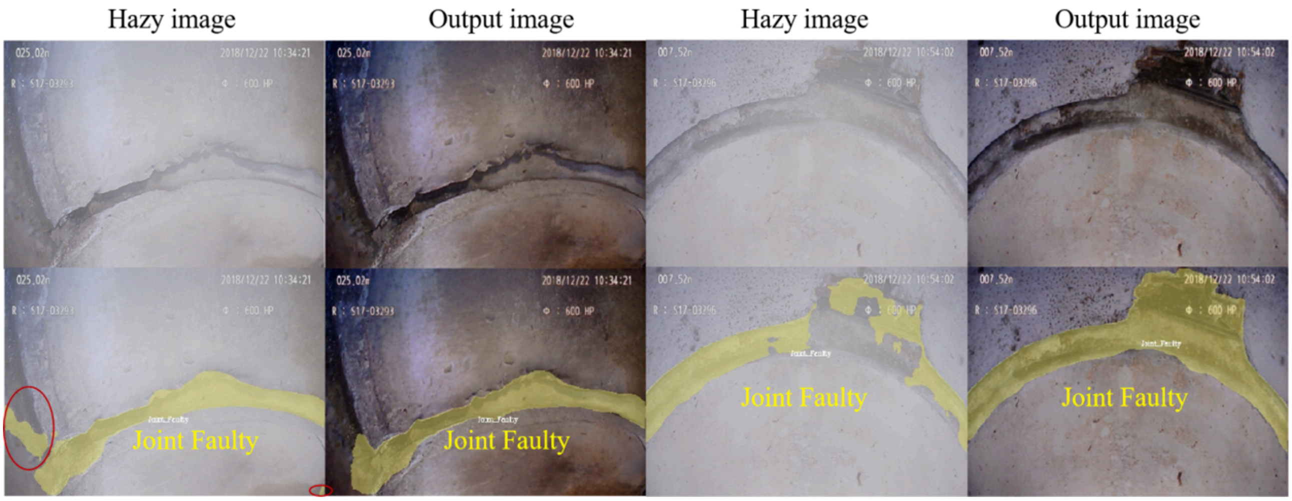


Fig. 8. Visualized segmentation results of hazy images and output images after dehazing.

$$Recall = \frac{\text{TruePositive(TP)}}{\text{FalseNegtive(FN)} + \text{TruePositive(TP)}}. \tag{8}$$

5. Experimental results

Various experiments were conducted to verify the proposed framework's robustness. All the experiments were performed on a pre-installed Linux machine with a Ubantu16.04 OS with four Titan X 12 GB GPUs and 64 GB of DDR4 RAM. The effectiveness of the dehazing process was confirmed in the first experiment (Section 5.1). After that, two separate experiments were designed to validate whether the proposed model (Section 5.2) and the parameter optimization process (Section 5.3) improved the overall defect segmentation's performance. The model's robustness was further analyzed by typical successful cases and rare failure cases (Section 5.4). Finally, the suggested defect segmentation model was compared to recent state-of-the-art studies to highlight its superiority and robustness (Section 5.5).

5.1. Image preprocessing

A collection of hazy images was manually selected and fed into the preprocessing algorithm to verify the GCANet's effectiveness in improving the image quality. The visual comparison between the ground truth image, hazy image, and the corresponding haze-free image is described in Fig. 7. The SSIM values between the Ground Truth (GT) images and the output images are 0.92 and 0.94, which indicates that the output images are consistent with human subjective feelings. Compared to the hazy images, the PSNR values of the two output images are improved significantly by 6.49 and 7.67, proving the dehazing algorithm's effectiveness. Besides, the method effectively removed the haze in the image while maintaining the original brightness of the image.

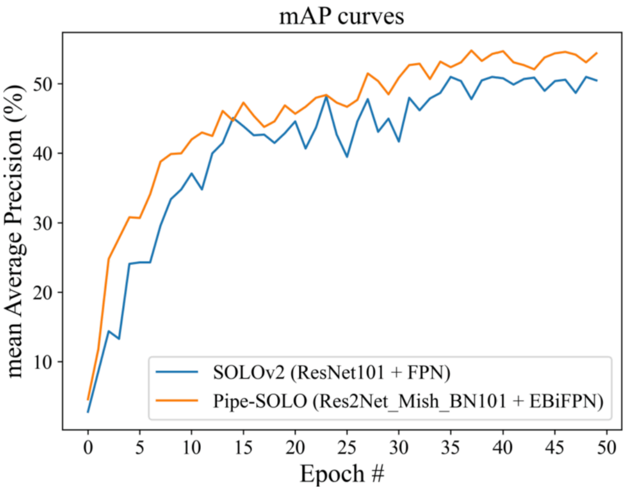


Fig. 9. mAP curves of the original SOLOv2 model and the proposed Pipe-SOLO model.

Accordingly, the segmentation effects of hazy images and output images after dehazing are visualized and compared to demonstrate the significance of the dehazing technique in the proposed defect segmentation framework. As shown in Fig. 8, both output images after dehazing obtained more complete segmentation results than the results of hazy input images. In addition, there are several mistakenly segmented regions that are marked by red circles in the hazy images due to the fog noises. As a result, defogging is necessary to improve the qualities of sewer images before defect segmentation.

Table 2
The experimental results of the 12 combinations of the Pipe-SOLO model using different networks. "√" indicates the selected networks for each combination.

	Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Backbone	ResNet-101	√	√	√	√												
	ResNeXt-101					√	√	√	√								
	Res2Net-101									√	√	√	√				
	Res2Net-Mish-BN-101													√	√	√	√
Neck	FPN	√				√				√				√			
	PAFPN		√				√				√				√		
	BiFPN			√				√				√				√	
	EBiFPN				√				√				√				√
Performance	mAP	50.9	50.8	51.7	52.4	52.0	52.3	52.6	53.3	52.8	52.5	53.9	54.7	53.5	53.8	54.5	**55.1**
	Loss	0.37	0.38	0.33	0.32	0.31	0.34	0.29	0.29	0.32	0.30	0.26	0.26	0.28	0.28	0.26	**0.24**

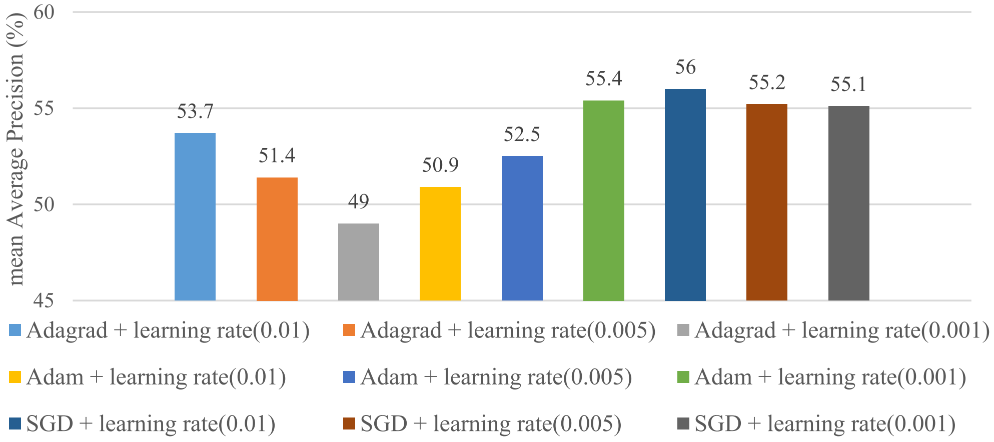


Fig. 10. Model's performance using different sets of hyperparameters.

5.2. Pipe-SOLO's structure analysis

This experiment analyzes the impact of various networks, including ResNet-101 [22], ResNeXt-101 [23], Res2Net-101 [25], Res2Net-Mish-BN-101 (proposed backbone), FPN [28], PAFPN [29], BiFPN [30], and EBiFPN (proposed neck), on the Pipe-SOLO's structure. We implemented 16 different combinations of different networks that are used in the backbone and neck structures of the proposed Pipe-SOLO model. Table 2 shows the segmentation performance of the 16 combinations on the proposed dataset using the same parameter settings. Res2Net-Mish-BN-101 outperformed other models that were used in the backbones because it inherited the Res2Net's ability to represent the multi-scale features at a fine-grained level, and the utilization of the Mish activation function helped it gain a better generalization ability. On the other hand, the proposed EBiFPN structure achieved the highest performance when it was implemented in the neck structure of the proposed Pipe-SOLO model. The 16th combination, which used the Res2Net-Mish-BN-101 in the Pipe-SOLO's backbone and the EBiFPN in the Pipe-SOLO's neck, achieved the best mAP value of 55.1% and minimum loss of 0.24.

Moreover, the mAP curves of the original SOLOv2 model and the proposed Pipe-SOLO model are plotted to highlight the improvement of the designed backbone and neck structures. As shown in Fig. 9, almost all the values of the purple curve are higher than the values on the blue curve. The best mAP of Pipe-SOLO is 55.1, which is 4.2 higher than the original SOLOv2. The SOLOv2 achieved stable performance after the 38th epoch, whereas our model stabilized at the 30th epoch. It suggests our Pipe-SOLO has a faster convergence speed during the training process.

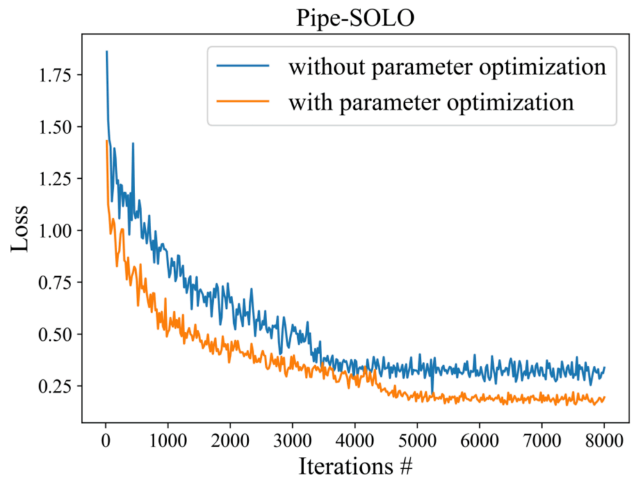
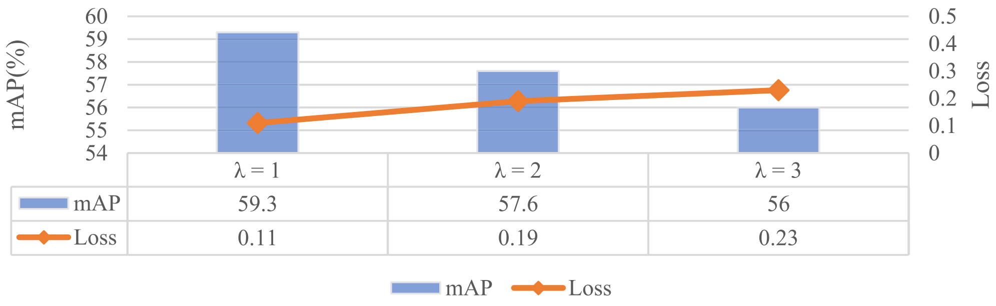


Fig. 12. Training performance of the proposed Pipe-SOLO model with and without the parameter optimization process.

5.3. Parameter analysis

The hyperparameters are crucial variables of a network, and there exists a set of optimal hyperparameters that helps a model to achieve the highest performance. For example, the learning rate determines how the network is trained. The loss fluctuates continuously when the learning rate is set too high, whereas the speed of convergence is slow when the learning rate is too low [32,33]. In order to find the optimal learning rate, the initial learning rate is set to 0.01, 0.005, and 0.001 in this experiment, and then it is changed gradually when the number of epochs increases [36]. In addition, this section also evaluates the mAP values of the defect segmentation model using different optimizers, including Adagrad, Adam, and SGD [32,33], to find the most appropriate optimizer.

Fig. 10 shows the model's performance using different learning rates and optimizers, which suggests the impact of the learning rate and



	λ = 1	λ = 2	λ = 3
mAP	59.3	57.6	56
Loss	0.11	0.19	0.23

Fig. 11. Performance of the model when different values of the weight coefficient λ are applied to calculate the loss function.

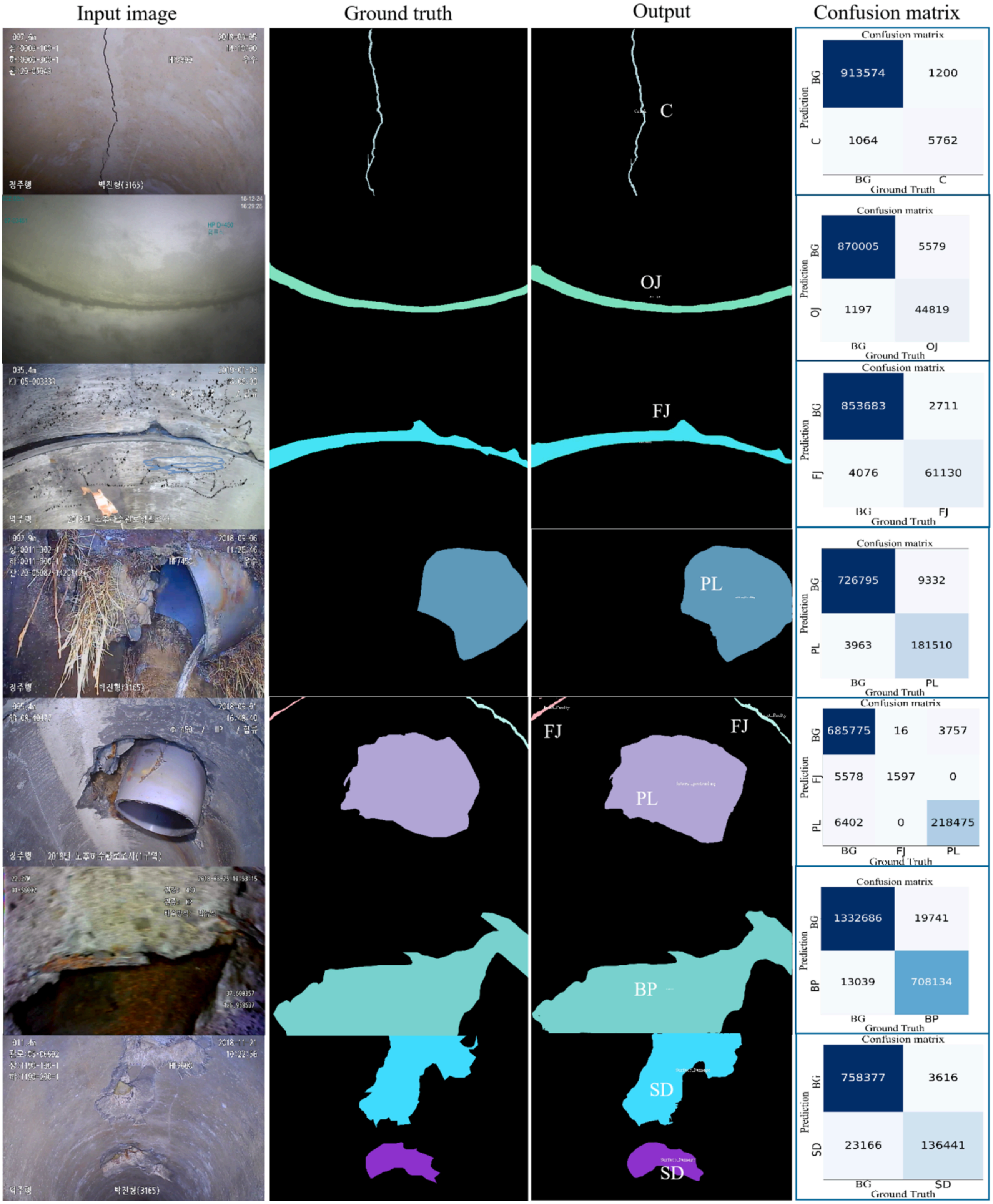


Fig. 13. Successful defect segmentation cases using the proposed dataset. From top to bottom (different defect classes): Crack (C), Open Joint (OJ), Faulty Joint (FJ), Protruding Lateral (PL), FJ & PL, Broken Pipe (BP), Surface Damage (SD). From left to right: Input image, Ground Truth, Output, and Confusion matrix. Note: ‘BG’ represents the background.

optimizer on the model. The highest mAP value of 56% was obtained when the optimizer was SGD, and the learning rate was 0.01 with a momentum of 0.9, which was 7% higher than the combination of the Adagrad optimizer and the learning rate of 0.001. Therefore, the SGD optimizer and the learning rate of 0.01 were selected as the optimal hyperparameters because they helped the model converged quickly during the training process.

In addition, the weight coefficient λ, which was used to compute the loss function (3), also had a considerable impact on the final performance. Fig. 11 displays the fluctuation of the model’s mAP and loss

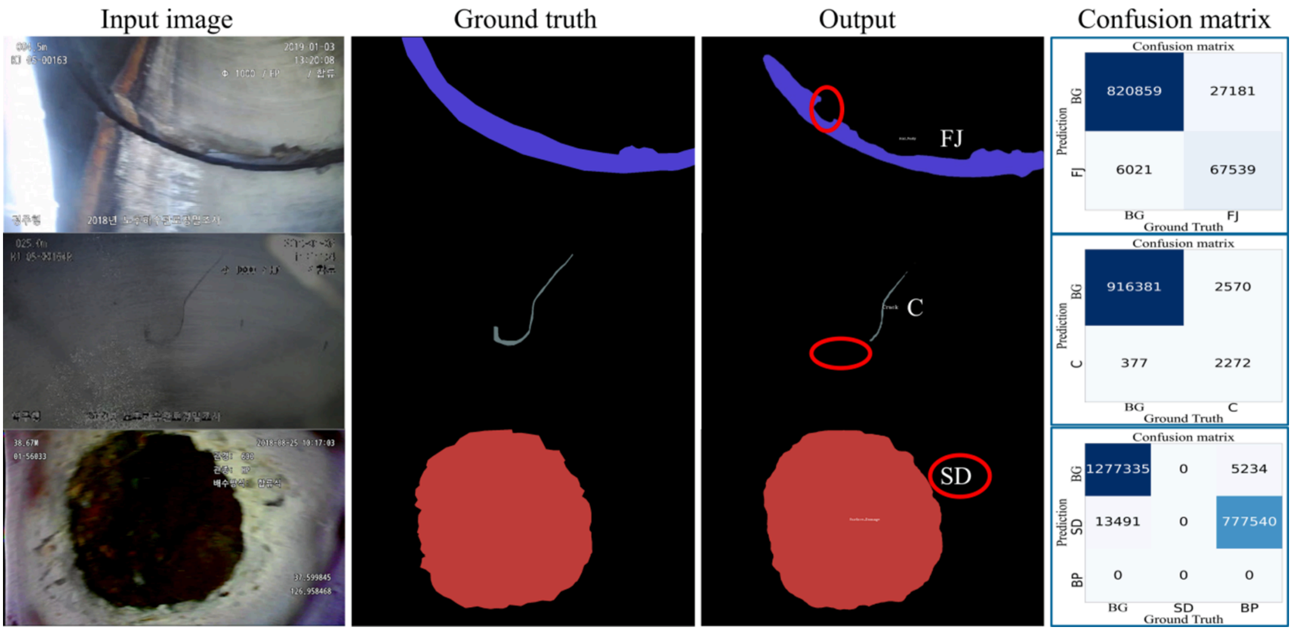


Fig. 14. Unsuccessful segmentation cases using the proposed dataset. From top to bottom (different defect class): Faulty Joint (FJ), Crack (C), Broken Pipe (BP). From left to right: Input image, Ground Truth, Output, and Confusion matrix. Note: 'BG' represents the background.

values when the weight coefficient parameter λ is changed. The model achieved the highest mAP value of 59.3% and the lowest loss of 0.11 when the λ was equal to 1.

The loss curves of the proposed model with and without the parameter optimization process are displayed in Fig. 12, which proves the effectiveness of the parameter optimization. The number of epochs was set to 50, and the iteration was configured to 8,000 to enable the model to learn discriminative features from the collected dataset. The loss of the model with the parameter optimization process decreases significantly to below 0.37 after 4,000 iterations, and then it decreases gradually and becomes stable at approximately 0.25. The loss of the proposed model is remarkably reduced by 0.13 compared to the original Pipe-SOLO.

5.4. Quantitative evaluation

This section quantitatively analyzes the segmentation result and the corresponding confusion matrix of the proposed instance segmentation model for each type of defect, as illustrated in Fig. 13 and Fig. 14. From the aspect of the segmentation result, Fig. 13 presents defects that were successfully segmented and classified, which confirmed that the Pipe-SOLO correctly predicted, localized, and segmented most of the defects. Fig. 13 (row 7 column 3) shows that the network correctly performed the instance segmentation, which treated two objects of the surface damage class as two separate objects. In particular, the faultless segmentation result of the challenging input image with lots of man-made noises at row 3, column 1 shows the robustness of the proposed model. In addition, our method performed well on the noisy image that exists extensive tree roots around the defective area from a real-life scenario at row 4, which demonstrates the proposed framework is practical and feasible in underground sewer defect inspection applications. Moreover, some other noises like haze (row 2 column 1) and blur (row 6 column 1) commonly appear in CCTV videos due to the evaporation of water and the motion of the crawler. The desirable results verify that our model is robust against different noises from real scenarios. From the aspect of confusion matrix, the values of TP and TN are always much bigger than the values of FP and FN. That results in a high precision and recall.

Nevertheless, the model failed to segment some images or gave a wrong label to the detected defect during the testing process, as displayed in Fig. 14. For the first case in Fig. 14 (row 1 column 3), the left section of the segmented Faulty Joint area in the red circle was mistakenly localized. That is caused by the overexposed image and the water stain around the defective region. Fig. 14 (row 2 column 3) illustrates that the whole crack in the image (row 2 column 3) was not detected completely due to the low-contrast background. Finally, the model predicted the Broken Pipe class as the Surface Damage class in Fig. 14 (row 3 column 3). The reason for this case may be low image resolution or subtle difference between Surface Damage class and Broken Pipe class in the training set.

5.5. Comparison with other work

The main purpose of this section is to demonstrate the superiority of

Table 3
Comparisons between the proposed model and the recent state-of-the-art defect segmentation approaches.

Image preprocessing	Segmentation approach	Defect type	Sample size	Speed	Accuracy	Reference
×	MSED, OTHO, and CBHO (morphological Segmentation)	Crack and open joint (2 classes)	100	1 ~ 10FPS	N/A	[7]
×	DilaSeg-CRF (semantic segmentation)	Crack, deposit, and root (3 classes)	1,885	9FPS	MIoU of 84.85%	[38]
×	PipeUNet (semantic segmentation)	Crack, infiltration, joint offset, and intruding lateral (4 classes)	3,654	32FPS	MIoU of 76.37%	[9]
√	Pipe-SOLO (instance segmentation)	Crack, Joint faulty, joint open, lateral protruding, pipe broken, and surface damage (6 classes)	3,888	15FPS	mAP of 59.3%	This study

Table 4
Performance comparison of different instance segmentation models using the proposed dataset.

Instance segmentation model	Backbone & Neck	mAP (%)	Loss
Mask R-CNN [39]	Res2Net-101 & FPN	51.4	0.28
HTC [40]	ResNeXt-101 & FPN	50.6	0.27
MS R-CNN [37]	ResNeXt-101 & FPN	48.3	0.30
SOLOv2 [21]	ResNeXt-DCN-101 & FPN	52.0	0.33
Pipe-SOLO (proposed)	Res2Net-Mish-BN-101 & EBiFPN	**59.3**	**0.11**

the proposed model compared to the previous approaches. Firstly, recent state-of-the-art defect segmentation approaches are summarized and analyzed thoroughly. Different instance segmentation networks are then implemented to validate the performance of the proposed network. Table 3 presents the performance in terms of accuracy and speed for different models. The number of defect types that were examined in this work was higher than the previous studies. Moreover, the preprocessing process was added in the proposed defect segmentation framework, which was ignored in previous methods. Table 4 shows the performance of different instance segmentation models based on the collected defect dataset. In order to conduct a fair comparison, all the experimental models are assessed with the same testing images that are preprocessed by a dehazing algorithm. Pipe-SOLO accurately predicted each class with the highest mAP of 59.3%, which was 11% higher than the MS R-CNN model [37]. The minimum loss value obtained by the proposed model was 0.22 lower than the SOLOv2 model [21]. The experimental results prove that the proposed model based on instance segmentation has the highest performance on the proposed dataset.

6. Conclusion

This study introduced a manually collected and annotated defect segmentation dataset for six main types of defects, including Crack, Faulty Joint, Open Joint, Protruding Lateral, Broken Pipe, and Surface Damage. After that, a dehazing model was applied to preprocess the hazy images before the defect segmentation process in order to improve the model's segmentation accuracy. Finally, a novel defect segmentation model called Pipe-SOLO was presented in this study. Pipe-SOLO optimized the original SOLOv2 network's structure using a new module (Res2Net-Mish-BN-101) in the backbone. Moreover, the EBiFPN module was adopted as the neck of the Pipe-SOLO structure because it helped the model obtain better performance. The experimental results show the proposed model obtained promising results on the collected dataset, which outperformed existing defect segmentation methods. As a result, the proposed dataset and defect detection framework are significant for the research on underground sewer pipelines assessment and maintenance.

In the future, the algorithm that can evaluate the defect risk or damage degree should be studied based on the proposed defect segmentation framework by measuring the areas and mean widths of defects. In addition, the current segmentation method only works with still images, so an automatic defect segmentation method for videos or a series of images can be further developed to deal with the defect analysis.

CRediT authorship contribution statement

Yanfen Li: Conceptualization, Methodology, Data curation, Writing – review & editing. **Hanxiang Wang:** Conceptualization, Methodology, Data curation, Writing – review & editing. **L.Minh Dang:** Visualization, Investigation. **Md Jalil Piran:** Visualization, Investigation. **Hyeonjoon Moon:** Supervision.

Declaration of Competing Interest

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

Acknowledgement

This work was supported by 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).

References

[1] S.I. Hassan, et al., Underground sewer pipe condition assessment based on convolutional neural networks, Autom. Constr. 106 (2019), 102849.
[2] S. Zamanian, J. Hur, A. Shafieezadeh, A high-fidelity computational investigation of buried concrete sewer pipes exposed to truckloads and corrosion deterioration, Eng. Struct. 221 (2020), 111043.
[3] J.B. Haurum, T.B. Moeslund, A Survey on Image-Based Automation of CCTV and SSET Sewer Inspections, Autom. Constr. 111 (2020), 103061.
[4] S.S. Kumar, D.M. Abraham, M.R. Jahanshahi, T. Iseley, J. Starr, Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks, Autom. Constr. 91 (2018) 273–283.
[5] D. Li, A. Cong, S. Guo, Sewer damage detection from imbalanced CCTV inspection data using deep convolutional neural networks with hierarchical classification, Autom. Constr. 101 (2019) 199–208.
[6] J.C.P. Cheng, M. Wang, Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques, Autom. Constr. 95 (2018) 155–171.
[7] T.-C. Su, M.-D. Yang, Application of morphological segmentation to leaking defect detection in sewer pipelines, Sensors 14 (5) (2014) 8686–8704.
[8] T.-C. Su, M.-D. Yang, T.-C. Wu, J.-Y. Lin, Morphological segmentation based on edge detection for sewer pipe defects on CCTV images, Expert Syst. Appl. 38 (10) (2011) 13094–13114.
[9] G. Pan, Y. Zheng, S. Guo, Y. Lv, Automatic sewer pipe defect semantic segmentation based on improved U-Net, Autom. Constr. 119 (2020), 103383.
[10] Y. Xu, D. Li, Q. Xie, Q. Wu, J. Wang, Automatic defect detection and segmentation of tunnel surface using modified Mask R-CNN, Measurement 178 (2021), 109316.
[11] D. Chen *et al.*, "Gated context aggregation network for image dehazing and deraining," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019: IEEE, pp. 1375–1383.
[12] K. He, J. Sun, X. Tang, Single image haze removal using dark channel prior, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2010) 2341–2353.
[13] X. Xin, Y. Cai, L. Changcheng, W. Xiaokang, Z. Jiandong, in: IEEE, 2018, pp. 270–273.
[14] B. Cai, X. Xu, K. Jia, C. Qing, D. Tao, Dehazenet: An end-to-end system for single image haze removal, IEEE Trans. Image Process. 25 (11) (2016) 5187–5198.
[15] M. Ju, C. Ding, D. Zhang, Y.J. Guo, BDPK: Bayesian dehazing using prior knowledge, IEEE Trans. Circuits Syst. Video Technol. 29 (8) (2018) 2349–2362.
[16] J. Gao, D. Yuan, Z. Tong, J. Yang, D. Yu, Autonomous pavement distress detection using ground penetrating radar and region-based deep learning, Measurement 164 (2020), 108077.
[17] X. Huang, Z. Liu, X. Zhang, J. Kang, M. Zhang, Y. Guo, Surface damage detection for steel wire ropes using deep learning and computer vision techniques, Measurement 161 (2020), 107843.
[18] Y. He, K. Song, Q. Meng, Y. Yan, An end-to-end steel surface defect detection approach via fusing multiple hierarchical features, IEEE Trans. Instrum. Meas. 69 (4) (2019) 1493–1504.
[19] X. Ye, J. Zuo, R. Li, Y. Wang, L. Gan, Z. Yu, X. Hu, Diagnosis of sewer pipe defects on image recognition of multi-features and support vector machine in a southern Chinese city, Front. Environm. Sci. Eng. 13 (2) (2019), https://doi.org/10.1007/s11783-019-1102-y.
[20] X. Yin, Y. Chen, A. Bouferguene, H. Zaman, M. Al-Hussein, L. Kurach, A deep learning-based framework for an automated defect detection system for sewer pipes, Autom. Constr. 109 (2020), 102967.
[21] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic, Faster and Stronger," *arXiv preprint arXiv:2003.10152,* 2020.
[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[23] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.
[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
[25] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P.H. Torr, Res2net: A new multi-scale backbone architecture, IEEE Trans. Pattern Anal. Mach. Intell. (2019).

[26] D. Misra, “Mish: A self regularized non-monotonic neural activation function,” *arXiv preprint arXiv:1908.08681,* 2019.

[27] G. Chen, P. Chen, Y. Shi, C.-Y. Hsieh, B. Liao, and S. Zhang, “Rethinking the Usage of Batch Normalization and Dropout in the Training of Deep Neural Networks,” *arXiv preprint arXiv:1905.05928,* 2019.

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

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

[30] M. Tan, R. Pang, Q.V. Le, Efficientdet: Scalable and efficient object detection, in: in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10781–10790.

[31] Y. Wu, K. He, Group normalization, in: in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[32] J. Yang, G. Yang, Modified convolutional neural network based on dropout and the stochastic gradient descent optimizer, Algorithms 11 (3) (2018) 28.

[33] H. Wang, Y. Li, L.M. Dang, J. Ko, D. Han, H. Moon, Smartphone-based bulky waste classification using convolutional neural networks, Multimedia Tools Appl. 79 (39) (2020) 29411–29431.

[34] A. Torralba, B.C. Russell, J. Yuen, Labelme: Online image annotation and applications, Proc. IEEE 98 (8) (2010) 1467–1484.

[35] D. Novotny, S. Albanie, D. Larlus, A. Vedaldi, Semi-convolutional operators for instance segmentation, in: in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 86–102.

[36] D. Han, Q. Liu, W. Fan, A new image classification method using CNN transfer learning and web data augmentation, Expert Syst. Appl. 95 (2018) 43–56.

[37] Z. Huang, L. Huang, Y. Gong, C. Huang, X. Wang, Mask scoring r-cnn, in: in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 6409–6418.

[38] M. Wang, J.C.P. Cheng, A unified convolutional neural network integrated with conditional random field for pipe defect segmentation, Comput.-Aided Civ. Infrastruct. Eng. 35 (2) (2020) 162–177.

[39] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[40] K. Chen, et al., Hybrid task cascade for instance segmentation, in: in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4974–4983.